

Mensura et Numerus

MANUSCRIPT

William F. Barnes

1

May 25, 2026

Data and Reference

Unit Conversion

Dimensionless

360°	= 2π rad
1 dozen	= 12
12 dozen	= 1 gross
1 mol	= 6.022×10^{23}

Time

1 h	= 60 min = 3600 s
1 day	= 24 h
1 yr	= 365.25 days

Length

1 in	= 2.54 cm
1 ft	= 12 in = 0.3048 m
1 yd	= 3 ft = 0.9144 m
1 mi	= 5280 ft = 1.609 km
1 cm	= 0.393701 in
1 m	= 100 cm = 1.09361 yd
1 km	= 1000 m = 0.621371 mi

Volume

1 gal	= 128 fl oz
1 gal	= 231 in ³ = 3.78541 L
1 ft ³	= 1728 in ³ = 7.48052 gal = 0.0283168 m ³
1 cm ³	= 1 mL
1 mL	= $\frac{1}{1000}$ L

Mass & Weight

1 g	= 10^{-3} kg
1 kg	= 1000 g
1 lb	= 16 oz
1 lb	= 453.5929 g (sea level)
1 kg	= 2.20462 lb (sea level)
1 ton	= 2000 lb
1 amu	= $1.66053906660 \times 10^{-27}$ kg

Pressure

1 atm	= 1.013×10^5 Pa
1 bar	= 10^5 Pa
1 psi	= 6894.757 Pa
1 atm	= 14.7 psi = 760 mmHg
1 in w.c.	= 249.089 Pa

Temperature

$T(^{\circ}\text{C})$	= $T(K) - 273.15$
$T(^{\circ}\text{F})$	= $\frac{9}{5}T(^{\circ}\text{C}) + 32$
$T(^{\circ}\text{C})$	= $\frac{5}{9}(T(^{\circ}\text{F}) - 32)$

Energy & Power

1 J	= 1 kg m ² /s ²
1 cal	= 4.186 J
1 Btu	= 1054 J
1 eV	= 1.602×10^{-19} J
1 W	= 1 J/s = 1 kg m ² /s ³
1 hp	= 745.701 W

Water at STP

Weight per gallon	8.344 lb
Weight per cubic foot	62.417 lb
PSI per vertical foot	0.43527
Freezing point	0°C = 32°F
Boiling point	100°C = 212°F

Alphabetic Symbols

Greek

Uppercase	Lowercase	Name
A	α	Alpha
B	β	Beta
Γ	γ	Gamma
Δ	δ	Delta
E	ϵ, ε	Epsilon
Z	ζ	Zeta
H	η	Eta
Θ	θ, ϑ	Theta
I	ι	Iota
K	$\kappa, \varkappa$	Kappa
Λ	λ	Lambda
M	μ	Mu
N	ν	Nu
Ξ	ξ	Xi
O	o	Omicron
Π	π, ϖ	Pi
P	ρ, ϱ	Rho
Σ	σ, ς	Sigma
T	τ	Tau
Y	υ	Upsilon
Φ	ϕ, φ	Phi
X	χ	Chi
Ψ	ψ	Psi
Ω	ω	Omega

Hebrew

$\aleph$	Aleph
$\beth$	Beth
$\daleth$	Daleth
$\gimel$	Gimel

Astronomical

$\lrcorner$	Moon
$\mercury$	Mercury
$\venus$	Venus
$\odot$	Sun
$\mars$	Mars
$\jupiter$	Jupiter
$\saturn$	Saturn

Mathematical Symbols

Set Notation

$\supset$	Superset
$\subset$	Subset
$\in$	Element of
$\notin$	Not an element
$\cup$	Union
$\cap$	Intersection
$\emptyset$	Empty set

Logic

$\exists$	There exists
$\exists!$	Unique existence
$\nexists$	Does not exist
$\forall$	For all
$:$	Such that
$\neg$	Logical NOT
$\vee$	Logical OR
$\wedge$	Logical AND
$\implies$	Implies

Numbers

$\mathbb{R}, \Re$	Real numbers
$\mathbb{Q}$	Rational numbers
$\mathbb{Q}'$	Irrational numbers
$\mathbb{Z}$	Integers
$\mathbb{N}$	Natural numbers
$\mathbb{I}, \Im$	Imaginary numbers
$\mathbb{C}$	Complex numbers
∞	Infinity
$\cdot _a^b$	Evaluation

Bracketing

$()$	Parentheses
$[]$	Square brackets
$\langle \rangle$	Angle brackets
$\{ \}$	Curly brackets
$\overbrace{A}$	Overbrace
$\underbrace{A}$	Underbrace

Variables

x	Variable x
x'	x prime
x''	x double prime
x_y	x subscript y
x^y	x superscript y
$\bar{x}$	x overline
$x_\star$	x star (subscript)
$x^\star$	x star (superscript)
(x, y)	Ordered pair

Unary Operators

$!$	Factorial
$\sqrt{}$	Square root
$\sqrt[n]{}$	Radical
$ \cdot $	Absolute value
$\lfloor \cdot \rfloor$	Floor
$\lceil \cdot \rceil$	Ceiling

Binary Operators

a^b	Exponent / power
$\times$	Multiplication, Cross product
$\cdot$	Dot product, Inner product
$\div$ or $/$	Division
$+$	Addition
$-$	Subtraction, Negation
$\pm$	Plus/minus
$\mp$	Minus/plus

Equality

$=$	Equality
$\approx$	Approximately equal
$\propto$	Proportional
$\sim$	Similar
$\neq$	Not equal

Comparison

$>$	Greater
$\geq$	Greater or equal
$\gg$	Much greater
$<$	Lesser
$\leq$	Lesser or equal
$\ll$	Much less

Geometry

$\angle$	Angle
$^\circ$	Degree
$\overline{AB}$	Line AB
$\parallel$	Parallel
$\perp$	Perpendicular

Linear Algebra

A^{-1}	Inverse
A^T	Transpose
$\mathbf{A}$ or $\mathbf{A}$	Vector
$\hat{\mathbf{A}}$	Unit vector

Calculus

∂	Partial derivative
∂_x	Partial w.r.t. x
$\dot{x}$	Time derivative of x
∇	Nabla / Gradient
D_μ	Covariant derivative
$\int$	Integral
$\oint$	Closed contour integral

Vector Spaces

$\wedge$	Wedge / Exterior product
$\oplus$	Direct sum
$\otimes$	Tensor / Outer product
$\langle a $	Bra
$ a\rangle$	Ket
$\langle a b\rangle$	Braket
$\ \cdot\ $	Norm

Mathematical Constants

Name	Symbol	Value
Square root of 2	$\sqrt{2}$	1.4142135623730950488016887242097 ...
Golden ratio	ϕ	1.6180339887498948482045868343656 ...
Natural log of 10	$\ln(10)$	2.3025850929940456840179914546844 ...
Euler's constant	e	2.7182818284590452353602874713527 ...
Pi	π	3.1415926535897932384626433832795 ...

Physical Constants and Data

Name	Symbol	Value
Avogadro	N_A	$6.02214076 \times 10^{23}$
Boltzmann	k_B	$1.380649 \times 10^{-23} \text{ J/K}$
Electron charge	e	$1.602176634 \times 10^{-19} \text{ C}$
Electron mass	m_e	$9.1093837015 \times 10^{-31} \text{ kg}$
Proton mass	m_p	$1.67262192369 \times 10^{-27} \text{ kg}$
Grav. accel.	g	9.80665 m/s^2
Grav. constant	G	$6.67430 \times 10^{-11} \text{ N m}^2/\text{kg}^2$
Gas constant	R	$8.314462618 \text{ J mol}^{-1} \text{ K}^{-1}$
Planck const.	h	$6.62607015 \times 10^{-34} \text{ J s}$
Speed of light	c	299,792,458 m/s
Vacuum perm.	ϵ_0	$8.854187817 \times 10^{-12} \text{ C}^2/\text{N m}^2$
Vacuum permeab.	μ_0	$1.25663706212 \times 10^{-6} \text{ N/A}^2$

Contents

I	Algebra	10	7	Geometric Series	93
1	Numbers and Arithmetic	11	1	Introduction	93
1	Numbers	11	2	Alternate Derivations	95
2	Operators and Expressions	13	3	Applicataions	97
3	Algebraic Operations	16	4	Zeno's Paradox	100
4	Exponent and Logarithm	22			
5	Variables and Equations	26			
6	Rate, Interval, and Work	31			
2	Units and Dimension	35	II	Pre-Calculus	101
1	Dimensional Analysis	35			
2	Scientific Notation	41	8	Trigonometry	102
3	Fundamental Units	46	1	Angles and Triangles	102
3	Cartesian Plane	49	2	Circles	107
1	Graphing	49	3	Trigonometric Identities	111
2	Straight Lines	53	4	Inverse Trigonometry	113
3	Quadratic Curves	56	5	Trigonometry Tables	115
4	Factoring	59	6	Trigonometry and Geometry	122
1	Factoring Quadratics	59	7	Polar Coordinate System	132
2	Factoring Cubics	66	8	Circular Motion	135
5	Kinematics	70	9	Lissasjous Curves	136
1	One-Dimensional Motion	70	9	Conic Sections	137
2	Two-Dimensional Motion	76	1	Ellipse	137
6	Polynomials	80	2	Hyperbola	144
1	Binomial Theorem	80	3	Parabola	149
2	Long division	82	4	Slicing the Cone	153
3	Partial Fractions	85	5	General Conic Sections	155
4	Factoring by Division	87	10	Taylor Polynomial	158
5	Recursive Sequences	89	1	Time-Shifted Kinematics	158
6	Lucas Numbers	90	2	Taylor Polynomial	161
7	Fibonacci Numbers	91	3	Euler's Constant	164
			4	Periodic Curve	167
			5	Laws of Motion	169
			11	Limits, Functions, Sequences	171
			1	Limits	171
			2	Functions	176
			3	Sequences	185
			4	Series	188
			12	Probability and Statistics	191
			1	Events and Probability	191
			2	Applications	197
			3	Combinatorics	199
			4	Variables and Data	202
			13	Vectors	205
			1	Introduction to Vectors	205
			2	Vector Addition	207
			3	Scalar Multiplication	209
			4	Vector Products	210

5	Basis Vectors	214	18 Analytic Geometry	340
6	Polar Representation	216	1	Parametric Equations 340
7	Change of Basis	217	2	Parametric Derivatives 342
8	Rotations	220	3	Parametric Integrals 343
9	Vectors and Limits	221	4	Vector Derivatives 345
14 Complex Algebra		225	5	Plane Curve Analysis 348
1	History of Complex Numbers	225	6	Curves in Three Dimensions 351
2	Complex Numbers	227	7	Cartesian 3D Coordinates 353
3	Complex Plane	230	8	Non-Cartesian Coordinates 355
4	Euler's Formula	236	9	Intersections 358
5	Roots and Branches	239	19 Multivariate Calculus	364
6	Complex Functions	240	1	Surfaces and Solids 364
15 Linear Systems		243	2	Multiple Integration 366
1	Matrix Formalism	244	3	Partial Derivative 370
2	Matrix Operations	246	4	Vectors and Surfaces 375
3	Linear Systems	248	5	Surface Integral 378
4	Determinant	250	6	Fields and Derivatives 380
5	Inverse Matrix	252	7	Divergence 385
6	Special Matrices	255	8	Central Forces 387
7	Elimination	257	9	Gauss's Law 389
8	Eigenvectors and Eigenvalues	259	10	Line Integral 391
9	Degenerate Systems	262	20 Complex Analysis	392
III Calculus		264	1	Complex Algebra Review 392
16 Differential Calculus		265	2	Solving Classic Systems 395
1	Slope at a Point	265	3	Complex Differentiation 400
2	Techniques of Differentiation	273	4	Contour Integrals 403
3	Mixed Techniques	278	5	Residue Calculus 406
4	Applied Differentiation	280	21 Variational Calculus	414
5	Second Derivative	285	1	Introduction 414
6	Taylor's Theorem	289	2	Euler-Lagrange Equation 415
7	Antiderivative	299	3	Formalism 418
8	Simple Harmonic Oscillator	305	4	Unconstrained Systems 421
17 Integral Calculus		307	5	Constrained Systems 425
1	Area Under a Curve	307	22 Differential Equations	432
2	The Integral	310	1	Separable Equation 432
3	Techniques of Integration	313	2	Integrating Factor 435
4	Integrals and Geometry	329	23 Fourier Analysis	436
5	Series Analysis	332	IV Advanced Mathematics	437
6	Applied Integration	333	24 Special Functions	438
			1	Gamma Function 438

25 Systems and Distributions	440	31 Thermodynamics	555
1 Large-N Systems	440	1 Heat and Energy	555
2 Discrete Distributions	443	2 Kinematic Gas Theory	563
3 Binomial Distribution	445	3 First Law	569
4 Continuous Distributions	447	4 Second Law	576
26 Vector Spaces	451	5 Thermodynamic Variables	579
1 Foundations	451	6 Chemical Physics	584
2 Vector Space	453	32 Variational Mechanics	587
3 Inner Product	455	33 Classical Field Theory	588
4 Linear Combinations	457	VI Modern Physics	589
5 Orthonormal Basis	458	34 Special Relativity	590
6 Normed Vector Space	461	1 Minkowski Space	590
7 Countably Finite System	464	2 Rindler Space	613
8 Countably Infinite System	465	35 Quantum Mechanics	618
9 Operators	466	36 Statistical Mechanics	619
10 Eigen-Calculations	468	1 Introduction	619
11 Operator as Matrix	470	2 Microcanonical Ensemble	621
12 Hermitian Matrix	472	3 Canonical Ensemble	631
13 Matrix in Hilbert Subspace	474	4 Grand Canonical Ensemble	645
14 Functions of Operators	476	VII Computation	648
15 Unitary Operators	477	37 Iterative Methods	649
16 Differential Equations	480	1 Matrix Toolkit	649
17 Normal Operators	482	2 Curve Fitting	653
27 Differential Geometry	483	3 Numerical Integration	661
1 Tensors and Index Notation	483	4 Newton's Method	664
2 Manifold and the Metric	487	5 Euler's Method	669
3 Covariant Derivative	494	6 Runge-Kutta Techniques	675
4 Flat Manifolds	500	7 Symplectic Method	679
28 Riemannian Geometry	507	8 Driven Pendulum	682
1 Geodesic Equation	507	38 Functional Constructs	692
2 Curvature	513	1 Turing Machine	692
3 Curved Manifolds	518	2 The Lambda Calculus	693
4 Symmetry and Invariance	521	3 The SK Calculus	696
V Classical Physics	524	4 Functions	704
29 Newtonian Mechanics	525	5 Automata	707
1 Foundation	525		
30 Central Forces	531		
1 Planetary Motion	531		
2 Central Potential	546		
3 Power Law Potential	550		

39 Logic Circuits	708	42 Collatz Fractal	779
1 Introduction	708	1 Introduction	779
2 Simulating Logic Circuits	714	2 Collatz Set	781
3 Switching and Latching	722	3 Collatz Fractal	782
4 Binary Addition	724	4 Collatz Gems	784
5 Edge Detection	730	5 Futility	789
6 Counters	733		
7 And-Inv Completeness	737	43 Sxript Programming Language	800
40 Sequence Analysis	739	1 Hello World	800
1 Sequences	739	2 Numerical Operations	802
2 Random Number Analysis	742	3 Logical Operators	804
3 Alphanumeric Analysis	747	4 Quote Operations	805
4 Binary Sequence Prediction	750	5 Vectors	808
5 Binary Case Studies	755	6 Mixed-Type Operations	810
6 Creating a PRNG	761	7 Apply, Map, Reduce	813
41 Epicycles	762	8 Variables	814
1 Introduction	762	9 Code Blocks	815
2 Universe of Ptolemy	764	10 Functions	818
3 Drawing Continua	767	11 Flow and Procedure	822
4 Epicycle Clock	771	12 Large Numbers	827
5 Connecting Discrete Points	773	13 ASCII Plotting	832
6 Drawing Arbitrary Curves	776	14 Functional Programming	835
		15 Structured Programs	844
		16 Applications	846
		17 Problems and Solutions	849

Part I

Algebra

Chapter 1

Numbers and Arithmetic

1 Numbers

Numbers are abstract symbols used for counting, measuring, ordering, and describing quantities. Each number stands for a specific value, and no two different numbers represent the same value.

To proceed carefully, we should make a distinction: a number is not merely a digit or a string of numerals. It is a *concept* with structure and relationships. What qualifies as a number depends on what system of mathematics we are working in—and the most foundational system begins with a simple idea: comparison.

1.1 Number Line

Any two numbers can be compared. One of them is either less than, greater than, or equal to the other. This comparison implies an ordering. We can therefore imagine a straight, infinite line - called the *number line* - on which every number is assigned a unique position. By convention, smaller numbers lie to the left, and greater numbers to the right, as shown in Figure 3.1.

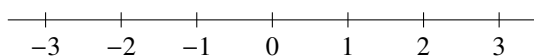


Figure 1.1. Number line (partial).

The number line is divided into two infinite halves: the *positive* numbers to the right of 0, and the *negative* numbers to the left. The special number 0 is neither positive nor negative. In both directions, the number line continues indefinitely. This idea of *unboundedness* leads naturally to the notion of infinity.

Infinity

Infinity is not a number in the ordinary sense. It does not have a fixed location on the number line, and it cannot be treated as a value that obeys the familiar rules of arithmetic. Expressions like

$$\infty + 1 = \infty$$

$$\infty + \infty = \infty$$

$$\infty - \infty = \text{undefined}$$

are not strictly false, but they are not equations in the algebraic sense. Infinity is more like a *concept* than a number: it describes a direction or an idea of size, not a quantity. Manipulating ∞ as though it were a number can lead to contradictions, such as falsely concluding that $1 = \infty - \infty = 0$.

Continuity

The number line is said to be *continuous*, meaning there are no gaps or holes. Between any two distinct numbers A and B , there are infinitely many other numbers. No matter how close two numbers are, there is always room for another one between them.

More formally, this property is called *density*: between any two real numbers, there exists another real number. But continuity means even more—it means that every point on the line corresponds to some real number, including those that require an infinite decimal expansion to describe.

1.2 Real Numbers

The full set of values that lie along the number line is called the set of *real numbers*, denoted $\mathbb{R}$. Every real number can be thought of as a decimal - finite, repeating, or infinite -

and every point on the number line corresponds to exactly one such number.

While we sometimes write

$$-\infty < \mathbb{R} < \infty$$

as symbolic shorthand, neither $-\infty$ nor $+\infty$ is actually a member of $\mathbb{R}$.

Rational and Irrational Numbers

The real numbers can be subdivided into two main types:

- **Rational numbers:** These are numbers that can be expressed as a ratio of two integers, that is, in the form a/b where $a, b \in \mathbb{Z}$ and $b \neq 0$. The set of rational numbers is denoted $\mathbb{Q}$. Rational numbers either terminate (like 0.75) or repeat (like 0.333...).
- **Irrational numbers:** These are real numbers that are *not* rational. They cannot be expressed as a ratio of integers, and their decimal expansions neither terminate nor repeat. Examples include $\sqrt{2}$, π , and e . The set of irrational numbers is sometimes denoted $\mathbb{Q}'$ or $\mathbb{R} \setminus \mathbb{Q}$.

In fact, there are *far more* irrational numbers than rational ones. This is a deep truth of set theory: while $\mathbb{Q}$ is *countable*, the set of all irrationals is *uncountable*. That means even though both are infinite, they differ in *magnitude* of infinity.

We can summarize the relationship using set theory notation:

$$\begin{aligned}\mathbb{Q} &\subset \mathbb{R} \\ \mathbb{Q}' &\subset \mathbb{R} \\ \mathbb{Q} \cup \mathbb{Q}' &= \mathbb{R}\end{aligned}$$

1.3 Integers

The integers, denoted $\mathbb{Z}$ (from the German *Zahlen*, meaning “numbers”), are the set of whole-valued numbers including positives, negatives, and zero:

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

All integers are rational numbers, since any integer n can be written as $n/1$. However, not all rational numbers are integers. So:

$$\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$$

Whole and Natural Numbers

Within the integers, we can identify still smaller subsets:

- The **whole numbers** are the non-negative integers:

$$\text{Whole Numbers} = \{0, 1, 2, 3, \dots\}$$

- The **natural numbers**, often denoted $\mathbb{N}$, are the counting numbers:

$$\mathbb{N} = \{1, 2, 3, \dots\}$$

Some authors include 0 in $\mathbb{N}$; others exclude it. For clarity, we will adopt the convention:

$$\mathbb{N} = \{1, 2, 3, \dots\} \quad \text{and} \quad \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$$

This distinction is especially important in number theory and formal logic, where the role of zero must be carefully specified.

Triangle Inequality

Consider any two real numbers a and b . There is an important relationship called the *triangle inequality* that tells us

$$|a + b| \leq |a| + |b|$$

is always true.

To prove this, first note for any variable x , it's always true that

$$\pm x \leq |x|.$$

This means the sum $a + b$ obeys

$$a + b \leq |a| + b \leq |a| + |b|,$$

and also

$$-a - b \leq |a| - b \leq |a| + |b|.$$

To be absolutely pedantic, we also need to examine the cases $a - b$, $-a + b$, and so on. You're welcome to chug through these, but the result is always going to be less than or equal to $|a| + |b|$. This is enough to call the proof complete.

2 Operators and Expressions

A *mathematical operation* is any process that turns existing information into new information. For dealing with numbers, we deal with a subset called *arithmetic operations*.

In no particular order, the simple *arithmetic operators* are:

Addition	+
Subtraction	-
Multiplication	×
Division	÷

When an arithmetic operator is situated between two numbers, we have an *arithmetic expression*. An expression is like a sentence fragment with numbers analogous to nouns and the operator analogous to the verb. For instance, the expression

$$4 + 6$$

embeds the intent to add four and six.

Carrying out a mathematical operation - doing the hard part - is called *evaluation* or *simplification*. For the example on hand, we know that adding four and six equals ten, so we write:

$$4 + 6 = 10$$

The presence of the equality sign (=) makes the above an *arithmetic equation*. The information on the left must always be the same as the information on the right. On the number line the location corresponding to $4 + 6$ is identical to that corresponding to 10.

Nested Expressions

Bracketing symbols called *parentheses* () are used to embed an expression within an expression. There must always be parenthetical balance, meaning the number of opening- and closing-parentheses must be equal.

When simplifying an expression, *the most-embedded parenthesized contents must be evaluated first*. To keep your work organized, it's good practice to keep all equality (=) symbols in a column. For instance:

$$\begin{aligned} 4 + (3 - (1 \times 2)) &= 4 + (3 - (1 \times 2)) \\ &= 4 + (3 - 2) \\ &= 4 + 1 \\ &= 5 \end{aligned}$$

Exponents

There are a slew of arithmetic operators in addition to the basic four. In the same way that multiplication is like generalized addition, i.e.

$$N + N + N = N \times 3,$$

the *exponent* operation, also called the *power* operation, is like generalized multiplication:

$$N \times N \times N = N^3$$

2.1 Order of Operations

By now we have six tools for basic arithmetic. Listing in a particular order, these are:

Operator	Symbol	Result
Parentheses	(N)	N
Exponent	^	product
Multiplication	× or ·	product
Division	÷ or /	ratio or quotient
Addition	+	sum
Subtraction	-	difference

Order of Operations = PEMDAS

The so-called *order of operations* is summarized by the letters **P E M D A S**, and can be recovered from the phrase **Please Excuse My Dear Aunt Sally**. This means to look for parentheses first, then exponents, multiplication, division, addition, and subtraction follow in order.

For some more terminology, operators have an order of *precedence*, where, for instance, exponents have higher precedence than sums. The subtraction operator has the lowest precedence. Another term one encounters is *binding*, where one could say the multiplication operator binds more tightly (to a number) than an addition operator.

Problem 1

Use the order of operations to verify

$$4 + 5 \times 9 = 49$$

and make sure the following gives a different answer:

$$(4 + 5) \times 9 = 180$$

Problem 2

Verify each of the following:

$$\begin{aligned}
 3 - (-4) + 5 &= 12 \\
 3^2 + 4^2 &= 5^2 \\
 (-3)^2 - (-4)^2 &= -7
 \end{aligned}$$

Problem 3

Insert any combination of operators to make the following statement true:

$$1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9 = 100$$

Suggestions:

$$\begin{aligned}
 1 \times 2 - 3 + 4 - 5 + 6 + 7 + 89 &= 100 \\
 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 \times 9 &= 100
 \end{aligned}$$

2.2 Special Operators**Inequality**

When two numbers or expressions are not necessarily equivalent, there are special symbols for *inequality*. These are summarized as follows:

Symbol	Meaning	Example
<	Less than	$2 < 3$
>	Greater than	$4 > 3$
$\leq$	Less or Equal	$2 \leq 2 + 2$
$\geq$	Greater or Equal	$4 \geq -4 $

Implied Multiplication

An abuse of notation that students of mathematics quickly accept is the omission of any symbol for multiplication. When a number ‘buts up’ against a set of parentheses, it is assumed we multiply the number into the parenthesized content:

$$2(3) = 2 \times 3 = 6$$

A classic example showing the trouble with this is the expression

$$6 \div 2(2 + 1) .$$

Some calculators simplify the above to 1, while others arrive at 9. Explicitly, one could turn the above into either correct statement:

$$\begin{aligned}
 6 \div (2 \times (2 + 1)) &= 1 \\
 (6 \div 2) \times (2 + 1) &= 9
 \end{aligned}$$

The precedence of implied multiplication can vary per computation regime, which leads to different answers.

One could perhaps argue that choosing the arithmetic division operator ($\div$) versus the forward slash ($/$) could break the tie on which way the expression is interpreted. With no solid rule, one could play games like this forever, indicating why numerical systems have more serious (or less serious) evaluation rules than implied by PEMDAS.

2.3 Unary Operators

There is a class of special operators that take one number instead of two, called *unary* operators, and these bind *more* tightly than the two-input operators in the ‘PEMDAS’ hierarchy.

Absolute Value

The *absolute value* of a number is an operation that converts any negative number to positive number while leaving positive numbers alone. A number enclosed by tall slashes ($| |$) tells us to take the absolute value. For example:

$$\begin{aligned}
 |-3| &= 3 \\
 |2 - 7| &= 5 \\
 |4/3| &= 4/3 \\
 |0| &= 0
 \end{aligned}$$

Factorial

The *factorial* operator occurs as an exclamation symbol (!) after a number, and (for our purposes) is only defined for whole numbers. The factorial operator tells us to take the base number and multiply it by every whole number less than the base number.

For example, the quantity 5! is pronounced ‘five factorial’, and is given by

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120 .$$

For shorthand, the same number can also be written:

$$5! = 5 \times 4! = 120$$

Since the factorial operator binds more tightly than multiplication, the above is quite different than:

$$20! = (5 \times 4)! = 2432902008176640000$$

Something you may not guess is what mathematicians have agreed to do with the factorial of zero. By definition, the factorial of zero is one:

$$0! = 1$$

To establish this, we note that

$$n! = n \times (n-1)!,$$

and then let $n = 1$ to have

$$1! = 1 \times 0!,$$

which can only be true if $0! = 1$.

Problem 4

Insert any combination of parentheses and operators to make the following statement true:

$$1 \ 1 \ 1 \ 1 = 5$$

Suggestion:

$$(1 + 1 + 1)! - 1 = 5$$

Problem 5

Insert any combination of parentheses and operators to make the following statement true:

$$2 \ 0 \ 1 \ 6 = 1$$

Suggestion:

$$\frac{(2 + 0 + 1)!}{6} = 1$$

Floor and Ceiling

Two operations useful in computer science are *floor* and *ceiling*. The ‘floor’ operation ($\lfloor \]$) encloses any real base number and returns the greatest integer that is less than the base number. The ‘ceil’ operation ($\lceil \]$) encloses any real base number and returns the smallest integer that is greater than the base number.

The following table demonstrates a few use cases of the floor and ceiling operations:

N	floor $\lfloor N \rfloor$	ceil $\lceil N \rceil$
3	3	3
3.1	3	4
3.4	3	4
3.9	3	4
-3.1	-4	-3
-3.9	-4	-3
-4	-4	-4

2.4 Special Numbers

Zero and one are two numbers that behave unlike the rest in many ways. In the following suppose N is any nonzero real number.

Properties of Zero

- Adding or subtracting zero to any number N leaves the number unchanged:

$$N \pm 0 = N$$

- Multiplying any number by zero results in zero:

$$N \times 0 = 0$$

- Dividing zero by any number results in zero:

$$\frac{0}{N} = 0$$

- Division by zero produces no useful information:

$$\frac{N}{0} = \text{Undefined}$$

- The only number equal to the negative of itself is zero:

$$0 = -0$$

Properties of One

- Multiplying or dividing a number by one leaves the number unchanged:

$$N \times 1 = \frac{N}{1} = N$$

- Fractions with the same numerator and denominator are equivalent to one:

$$\frac{N}{N} = 1$$

- Raising one to any real power results in one:

$$1 \times 1 \times 1 \times 1 \times \dots = 1^0 = 1^{-1.75} = 1$$

- The infinite repeating decimal $0.999\bar{9}$ is equivalent to one:

$$0.999\bar{9} = 1$$

(Proven later with geometric series.)

3 Algebraic Operations

Let the symbols A, B, C, D, N, M represent any real numbers. To be ‘safe’, which means to avoid subtle errors like division by zero, you can imagine each number being nonzero.

3.1 Commutative Property

Commutative Property of Addition

Consider the sum

$$N = A + B .$$

The *commutative property of addition* tells us that the order in which the terms A, B occur in the operation does not change the resulting number N . That is

$$\begin{aligned} N &= A + B \\ N &= B + A , \end{aligned}$$

or equivalently:

$$A + B = B + A$$

Note that the subtraction operator doesn’t yield an analogously true statement. The difference $A - B$ is not the same as $B - A$.

Commutative Property of Multiplication

Consider the product

$$N = A \times B .$$

The *commutative property of multiplication* tells us that the order in which the terms A, B occur in the operation does not change the resulting number N . That is

$$\begin{aligned} N &= A \times B \\ N &= B \times A , \end{aligned}$$

or equivalently:

$$A \times B = B \times A$$

Note that the division operator doesn’t yield an analogously true statement. The ratio A/B is not the same as B/A .

3.2 Associative Property

Associative Property of Addition

Consider the sum

$$N = A + B + C .$$

The *associative property of addition* tells us that the order in which the terms A, B, C are added does not change the resulting number N . That is

$$\begin{aligned} N &= (A + B) + C \\ N &= A + (B + C) , \end{aligned}$$

or equivalently:

$$(A + B) + C = A + (B + C)$$

Associative Property of Multiplication

Consider the product

$$N = A \times B \times C .$$

The *associative property of multiplication* tells us that the order in which the terms A, B, C are multiplied does not change the resulting number N . This means the two products

$$\begin{aligned} N &= (A \times B) \times C \\ N &= A \times (B \times C) \end{aligned}$$

are equivalent, equivalently:

$$(A \times B) \times C = A \times (B \times C)$$

3.3 Distributive Property

The *distributive property of multiplication* tells what to do with expressions of the form

$$(A + B) \times (C + D) ,$$

where $(A + B), (C + D)$ are called *binomials*. Note that the multiplication operator $(\times)$ is optional, and the above is the same as

$$(A + B) (C + D) .$$

To multiply two binomials, multiply each term in the first binomial into each term in the second:

$$\begin{aligned} (A + B) (C + D) &= A (C + D) + B (C + D) \\ &= AC + AD + BC + BD \end{aligned}$$

By the commutative property, the same result is attained when swapping the binomials:

$$\begin{aligned} (C + D) (A + B) &= C (A + B) + D (A + B) \\ &= AC + BC + AD + BD \end{aligned}$$

FOIL Method for Binomials

A trick for remembering the steps to multiply two binomials is contained in the word 'FOIL', where each respective letter stands for **F**irst, **O**uter, **I**nnner, **L**ast. For instance, consider the product:

$$(4 + 2)(5 - 3)$$

1. Multiply the **F**irst respective terms to get

$$4 \times 5 = 20 .$$

2. Multiply the **O**uter terms, resulting in

$$4 \times -3 = -12 .$$

3. Multiply the **I**nnner terms, namely

$$2 \times 5 = 10 .$$

4. Finally, the **L**ast terms multiply to give

$$2 \times -3 = -6 .$$

The final answer is the sum of the four respective products. In our example, we have

$$(4 + 2)(5 - 3) = 20 - 12 + 10 - 6 = 12 .$$

Multiplication with Trinomial

The FOIL method works on *all* binomials, but *only* works on binomials. This can cause confusion when confronted with a problem like

$$(A + B)(C + D + E) ,$$

where the combination $C + D + E$ is a *trinomial*.

Simplifying the above requires a direct application of the distributive property. Particularly, one must have:

$$\begin{aligned} (A + B)(C + D + E) \\ &= A(C + D + E) + B(C + D + E) \\ &= AC + AD + AE + BC + BD + BE \end{aligned}$$

Problem 6

Let x and y be two real numbers. Simplify

$$(y - 4)(y^2 - 3x - 2)$$

in two different ways. First distribute each term in the first polynomial into each term of the second polynomial. Then do it again in the other order. Either way, find the result

$$y^3 - 4y^2 - 3xy + 12x - 2y + 8 .$$

3.4 Fractions

Any real number that is not an integer is some kind of fraction or decimal. These are numbers such as $1/3$, 9.5 , -0.6666666 , and so on. The conventional symbol for division is the forward-slash ($/$), equivalent to the numerator-over-denominator notation (top number over the bottom number). That is, the following are equivalent:

$$2/3 \leftrightarrow \frac{2}{3}$$

Multiplying Fractions

When fractions are multiplied, first multiply all terms in each numerator, and then separately multiply all terms in each denominator. Explicitly, Consider two fractions

$$\begin{aligned} N &= A/B \\ M &= C/D . \end{aligned}$$

The product NM is

$$NM = \frac{A}{B} \times \frac{C}{D} = \frac{A \times C}{B \times D} .$$

For example:

$$\frac{2}{3} \times \frac{5}{4} = \frac{2 \times 5}{3 \times 4} = \frac{5}{6}$$

Notion of Reciprocal

The *reciprocal* or *inverse* of a fraction arises when exchanging the numerator and the denominator. For instance, the reciprocal of $2/7$ is $7/2$. In the most general sense, the reciprocal of any value equals one divided by that value:

$$\text{Reciprocal} = \frac{1}{\text{Fraction}}$$

Dividing Fractions

Dividing by a fraction is equivalent to multiplying by its reciprocal. Consider two fractions

$$N = A/B$$
$$M = C/D .$$

The ratio N/M is written:

$$\frac{N}{M} = \left(\frac{A}{B}\right) \div \left(\frac{C}{D}\right) = \frac{A \times D}{B \times C}$$

Problem 7

Convert to a single fraction and simplify

$$(3/4) / (2/5) .$$

Answer:

$$(3/4) / (2/5) = \frac{3}{4} \times \frac{5}{2} = \frac{15}{8}$$

Problem 8

Convert to a single fraction and simplify

$$\frac{3/5}{7/8} .$$

Answer:

$$\frac{(3/5)}{(7/8)} = \frac{3}{5} \times \frac{8}{7} = \frac{24}{35}$$

Change of Denominator

Fractions can be modified to have a different numerator and denominator, but only if the overall value of the fraction remains the same. For example, we know that

$$\frac{1}{2} \leftrightarrow \frac{3}{6}$$

are *equivalent*, and evaluate to 0.5, yet the actual numbers used in each fraction are different.

The conversion between equivalent fractions is done by multiplying by a carefully chosen factor of one. For example, we convert $1/2$ to $3/6$ by:

$$\frac{1}{2} = \frac{1}{2} \times \left(\frac{3}{3}\right) = \frac{1 \times 3}{2 \times 3} = \frac{3}{6}$$

Choosing a different factor of one, we can convert to any fraction that is ultimately equivalent to the original ratio. For instance:

$$\frac{1}{2} = \frac{1}{2} \times \left(\frac{13}{13}\right) = \frac{1 \times 13}{2 \times 13} = \frac{13}{26}$$

Problem 9

Consider the ratio $9/3$. Change the denominator to 17. Answer:

$$\frac{9}{3} = \frac{9/3}{1} \times \left(\frac{17}{17}\right) = \frac{3 \times 17}{17} = \frac{51}{17}$$

Problem 10

Consider the pair of fractions $2/3$ and $3/4$. Rewrite each fraction such that the denominators are the same.

Answer:

$$\frac{2}{3} = \frac{2}{3} \times \left(\frac{4}{4}\right) = \frac{8}{12}$$
$$\frac{3}{4} = \frac{3}{4} \times \left(\frac{3}{3}\right) = \frac{9}{12}$$

Adding and Subtracting Fractions

There is one single rule for adding and subtracting fractions: only fractions having the *same* denominator can be added or subtracted.

To demonstrate, consider the true statement:

$$\frac{3}{4} - \frac{1}{4} = \frac{3-1}{4} = \frac{2}{4}$$

The reason the above statement executes without error is the denominator of each fraction being added (subtracted in this case) is the same. On the left side, the two terms $3/4$ and $1/4$ share a denominator (namely 4), thus the two terms combine easily.

Common Denominator

Two fractions can be added or subtracted if they share a *common denominator*.

Problem 11

Evaluate the sum of $2/3$ and $3/4$.

Answer:

$$\frac{2}{3} + \frac{3}{4} = \frac{8}{12} + \frac{9}{12} = \frac{17}{12}$$

Problem 12

Evaluate the expression

$$1/2 + 1/3 - 1/7 .$$

Answer:

$$\frac{1}{2} + \frac{1}{3} - \frac{1}{7} = \frac{21}{42} + \frac{14}{42} - \frac{6}{42} = \frac{29}{42}$$

Terminating Decimal Conversion

Consider any real decimal number N where the digits are non-repeating, a case called *terminating*. Given N , there must be some power of ten that, when multiplied by N , results in an integer:

$$M = N \times 10^B = \text{Integer}$$

Then, the original number N can be written in fraction form

$$N = \frac{M}{10^B}.$$

For example, if $N = 1.2345$, then we must have $B = 4$ to have integer M , and thus

$$1.2345 = \frac{12345}{10000}.$$

Percentages

When the denominator of a fraction is exactly 100, this translates to ‘per one hundred’, or ‘per cent’. The ‘percent’ symbol (%) works as:

$$\frac{A}{100} = A\%$$

Problem 13

Source: *PHCC of MA*. Slide 14.

While ordering plumbing materials, the bill was \$4,860.00. You also received discounts of 20%, 10% and 2% from the \$4,860.00 amount. How much was the final bill?

Answer:

$$\begin{aligned} N &= \$4860.00 \times (1 - .20) \times (1 - .10) \times (1 - .02) \\ &= \$3429.22 \end{aligned}$$

Problem 14

Source: *PHCC of MA*. Slide 17.

While ordering plumbing materials, the bill was \$6,229.00. You also received discounts of 25%, 5% and 2% from the \$6,229.00 amount. How much was the final bill?

Answer:

$$\begin{aligned} N &= \$6229.00 \times (1 - .25) \times (1 - .05) \times (1 - .02) \\ &= \$4349.40 \end{aligned}$$

3.5 Factorization

Any real number N can be considered as the product of *factors* A , B , C , etc., such that:

$$N = A \times B \times C \times \cdots$$

Greatest Common Factor

Given a list of real numbers, the *common factors* are those that occur in the factorization of each member on the list. The greatest of these is called the *greatest common factor*, or GCF.

Problem 15

List all factors of 12, 30 and identify the GCF. Answer:

$$12 \rightarrow \mathbf{1, 2, 3, 4, 6, 12}$$

$$30 \rightarrow \mathbf{1, 2, 3, 5, 6, 10, 15, 30}$$

$$\text{GCF} = 6$$

Problem 16

List all factors of 15, 30, 105 and identify the GCF. Answer:

$$15 \rightarrow \mathbf{1, 3, 5, 15}$$

$$30 \rightarrow \mathbf{1, 2, 3, 5, 6, 10, 15, 30}$$

$$105 \rightarrow \mathbf{1, 3, 5, 7, 15, 21, 35, 105}$$

$$\text{GCF} = 15$$

Problem 17

Find the GCF of 24 and 108. Answer:

$$24 = 2 \times (2 \times 2 \times 3)$$

$$108 = (2 \times 2 \times 3) \times 3 \times 3$$

$$\text{GCF} = (2 \times 2 \times 3) = 12$$

Problem 18

Bobby has 160 red marbles and 144 blue marbles. He wants to split the marbles into identical groups (not necessarily the same number of each color in a group). What is the greatest number of groups he can make?

Answer: Solve this by finding the GCF of 160 and 144, or:

$$160 \rightarrow 10, \mathbf{16}$$

$$144 \rightarrow 9, \mathbf{16}$$

$$\text{GCF} = 16$$

Thus 16 groups can be made. Each group has $160/16 = 10$ red marbles and $144/16 = 9$ blue marbles.

Problem 19

Alice has 9 forks and 6 spoons. She intends to lay the utensils out in groups around the dinner table with none left

over. How many groups can she make, and how many of each utensil per group?

Problem 20

Betty is preparing Halloween treats for her classmates. She has 72 orange candies and 24 red candies in total, and wants to divide them into the greatest number of bags. How many bags can she prepare, and how much of each type of candy should be in each bag?

3.6 Prime Numbers

Prime numbers are natural numbers that cannot be divided into smaller natural numbers (besides 1). Apart from 2, no even numbers are prime. There is otherwise no simple pattern to the prime numbers.

The following table shows the first fifteen prime numbers in boldface situated in the lattice of integers.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50

Prime Factorization

For a whole number N , the *prime factorization* of N is defined as a list of prime numbers whose product is N .

To find the prime factorization of a whole number, try dividing by the first prime number M , namely $M = 2$. The result

$$Q = \frac{N}{M}$$

is handled as follows:

1. If N/M is a prime number, stop. The prime factorization includes M , $Q = N/M$.
2. If Q is a non-prime whole number, then M is a factor. Replace N with Q and go to Step 1.
3. If Q is fraction or decimal, replace M with the next prime number and go to Step 1.

The above steps can be repeated until the prime being tested exceeds $N/2$. The original number N is the product of factors.

Problem 21

Use the above as a guide to verify the prime factorization for each:

$$12 = 2 \times 2 \times 3$$

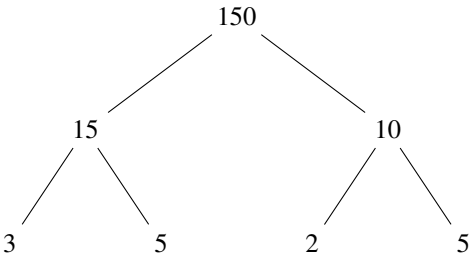
$$231 = 3 \times 7 \times 11$$

$$150 = 2 \times 3 \times 5 \times 5$$

$$225 = 3 \times 3 \times 5 \times 5$$

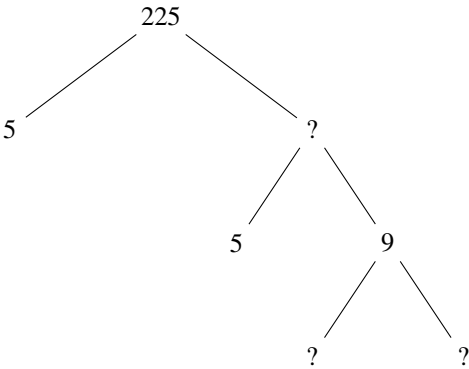
Tree Method

A pictorial technique for prime decomposition is the so-called *tree method*, which represents the factors of a number as 'branches'. The final extremity of each branch is the set of prime factors as demonstrated for the number 150.



Problem 22

Fill in the missing numbers in the following tree method factorization of 225: Answer: 45, 3, 3



Least Common Multiple

Given a list of real numbers, the *least common multiple* (LCM) is the product of common prime factors raised to the highest power in which they occur. In other words, the LCM is the smallest number that can be evenly divided by two or more numbers.

Problem 23

Find the LCM of 84, 147. Answer:

$$84 = 2^3 \times 3^1 \times 7$$

$$147 = 3 \times 7^2$$

$$\text{LCM} = 2^3 \times 3 \times 7^2 = 588$$

Problem 24

What is the smallest number that can be divided by 3, 9, 21? Answer:

$$3 = 3$$

$$9 = 3^2$$

$$21 = 3 \times 7^1$$

$$\text{LCM} = 3^2 \times 7 = 63$$

Problem 25

Cindy works in the orchard picking peaches, and fits 8 peaches per bag. Bob has the same job, but fits 9 peaches per bag. At the end of the day, they have picked the same number of peaches. What is the smallest number of peaches they each could have picked? (Hint: Find the LCM of 8, 9.)

Problem 26

For the numbers 2940, 3150, find the GCF and LCM. Answer: GCF = 210, LCM = 210^2

4 Exponent and Logarithm

Exponent Notation

When a number is multiplied by itself more than twice, or many times, we don't want to be burdened by writing things like $6 \times 6 \times 6 \times 6$. This is remedied by exponent notation, demonstrated as follows:

$$\begin{aligned} 4 &= 4^1 \\ 3 \times 3 &= 3^2 \\ 5 \times 5 \times 5 &= 5^3 \\ 2 \times 2 \times 9 \times 9 &= 2^2 \times 9^2 \end{aligned}$$

Exponents obey certain conventions that we shall establish here.

- The *base* number is the 'lower' number to which the exponent applies.
- The *exponent* is written smaller and to the upper-right of the base number.

Special Exponents

- When the exponent is zero, the result equals one:

$$N^0 = 1$$

- When the exponent is one, the result equals the base number:

$$N^1 = N$$

- When the exponent is 2, the result is the *square* of the base number.
- When the exponent is 3, the result is the *cube* of the base number.

Negative Base Number

Negative numbers with exponents must be treated carefully. For instance, if asked to square the number -3 , one may write

$$(-3)^2 = -3 \times -3 = (-1)^2 \times 3^2 = 9,$$

which has the same result as (positive) $3^2 = 9$. Indeed, raising to any base number *even* power results in a positive number.

On the other hand, if a negative base number is raised to an *odd* number, factors of -1 occur an odd number of times, as in

$$(-2)^3 = (-1)^3 \times 2^3 = -8.$$

Beware the placement of parentheses when handling exponents. For instance the number -2^4 translates to

$$-2 \times 2 \times 2 \times 2 = -16.$$

However, if we were handed $(-2)^4$, the full statement would be

$$(-2) \times (-2) \times (-2) \times (-2) = 16.$$

Multiplying Numbers with Exponents

For multiplying two numbers having equal base and unequal exponents, the product is the base number raised to the sum of the exponents:

$$A^B \times A^C = A^{B+C}$$

For example, the product $3^2 \times 3^4$ can be immediately translated to 3^6 :

$$(3 \times 3) \times (3 \times 3 \times 3 \times 3) = 3^{2+4} = 3^6$$

Negative Exponent

A number raised to a negative exponent is equivalent to the reciprocal of that number raised to the positive exponent:

$$A^{-B} = \left(\frac{1}{A}\right)^B = \frac{1}{A^B}$$

For an example, one may have

$$6^2 \times 3^{-2} = 6^2 \times \left(\frac{1}{3}\right)^2 = \left(\frac{6}{3}\right)^2 = 4,$$

or similarly

$$6^2 \times 3^{-2} = \frac{6^2}{3^2} = \left(\frac{6}{3}\right)^2 = 4.$$

Compound Exponent

A number that is raised to an exponent, and then raised to an exponent *again*, as in

$$(2^3)^2 = 8^2 = 64$$

is equivalent to *multiplying* the two exponents:

$$(2^3)^2 = 2^{3 \times 2} = 2^6 = 64$$

In general the rule for compound exponents reads:

$$(A^B)^C = A^{B \times C}$$

Zero Exponent

By the established properties of exponents, it can only be that any number raised to the zero power equals one. In detail, consider the trivial ratio A^B/A^B , which simplifies according to:

$$1 = \frac{A^B}{A^B} = A^B \times \frac{1}{A^B} = A^{B-B} = A^0$$

Radicals

Numbers raised to non-integer exponents are sometimes called *radicals*. Valid radical numbers may look like:

$$\begin{aligned} 3^{\cdot 5} &= \sqrt{3} \\ 4^{3/2} &= \left(\sqrt{4}\right)^3 = \sqrt[3]{4} \\ 7^{-1.5} &= 7^{-3/2} = \frac{1}{7^{3/2}} = \frac{1}{\sqrt[3]{7}} \end{aligned}$$

The base number is also denoted the *radicand*, and the exponent is called the *degree* of the radical. In the general case, we write:

$$A^{B/2} = \sqrt[B]{A}$$

Roots

When the exponent is precisely $1/2$, the result of the operation is the *square root* of the base number. Similarly, if the exponent is precisely $1/3$, the result of the operation is the *cube root* of the base number, and so on.

4.1 Logarithm Operator

Consider the equation

$$N = A^B,$$

where A is the base number and B is the exponent. The *logarithm operator* can be used to solve for B in terms of A and N :

$$\log_A(N) = B$$

In words, the above reads ‘log base A of N equals B ’. To have it all in one statement, one may write:

$$\log_A(A^B) = B \quad (1.1)$$

Inverse Logarithm

The inverse to the logarithm is the exponent operator. To invert the above, start with the base number A and use each side as the exponent in the equation:

$$A^{\log_A(N)} = N = A^B$$

Isolating the Base

To solve for A , raise each side of the above to the $1/B$ power:

$$N^{1/B} = A^{B/B} = A$$

Log of One

Right away, we can establish an important identity for when the exponent is zero:

$$\log_A(1) = 0$$

That is, for any base number, the log of one equals zero.

Log Equals One

For another identity we start with $B = 1$ to have $N = A^1$, and thus

$$\log_A(A) = 1,$$

true for any base number.

Log of Zero

One should note that the logarithm of zero is undefined. In particular, suppose we have

$$0 = A^B.$$

Apart from the trivial solution of $A = 0$, $B \neq 0$, the only way to satisfy the above is to have B equal to an infinitely large negative number. For this reason, we establish

$$\log_A(0) = -\infty,$$

which is technically undefined.

Base-Ten Logarithm

The base-ten logarithm operator quickly isolates the magnitude of any real number. For instance:

$$\begin{aligned} \log_{10}(1000) &= 3 \\ \log_{10}(10) &= 1 \\ \log_{10}\left(\frac{1}{1000}\right) &= -3 \\ \log_{10}(10^{-8}) &= -8 \end{aligned}$$

Change of Base

From the equation $N = A^B$, we can attain an equivalent form $N = C^D$ by a *change of base*, which implies the exponent changes as well. Start with

$$\log_A (N) = B$$

to eliminate B from the original equation:

$$N = A^{\log_A (N)}$$

In this form, we see A can be replaced with any base number C without changing anything, i.e.

$$N = C^{\log_C (N)} .$$

From here we see

$$D = \log_C (N) ,$$

and solving for the base C yields

$$C = N^{1/D} .$$

Addition Identity

Consider the sum

$$N = \log_A (B) + \log_A (C) .$$

To simplify, use each side as an exponent with A as the base number:

$$A^N = A^{\log_A (B) + \log_A (C)} ,$$

simplifying to

$$A^N = A^{\log_A (B)} \times A^{\log_A (C)} = BC .$$

Apply the $\log_A$ operator to each side to find

$$N = \log_A (BC) ,$$

and in conclusion:

$$\log_A (B) + \log_A (C) = \log_A (BC) \quad (1.2)$$

Not surprisingly, multiplying the product BC by a factor $D = 1$ leaves the result unchanged:

$$\log_A (B) + \log_A (C) + \log_A (\cancel{1}) = \log_A (BC \times 1)$$

Subtraction Identity

Consider the difference

$$N = \log_A (B) - \log_A (C) .$$

To simplify, use each side as an exponent with A as the base number:

$$A^N = A^{\log_A (B) - \log_A (C)} ,$$

simplifying to

$$A^N = \frac{A^{\log_A (B)}}{A^{\log_A (C)}} = \frac{B}{C} .$$

Apply the $\log_A$ operator to each side to find

$$N = \log_A \left(\frac{B}{C} \right) ,$$

and in conclusion:

$$\log_A (B) - \log_A (C) = \log_A \left(\frac{B}{C} \right) \quad (1.3)$$

Product Identity

Start with the equation

$$N = C \times \log_A (B) ,$$

and use each side as an exponent with A as the base number:

$$A^N = A^{C \times \log_A (B)}$$

Recognizing the compound exponent, the above is also

$$A^N = \left(A^{\log_A (B)} \right)^C = B^C .$$

Apply the $\log_A$ operator to each side and simplify:

$$N = \log A \left(B^C \right)$$

Eliminating N gives the product identity:

$$C \times \log_A (B) = \log_A \left(B^C \right) \quad (1.4)$$

4.2 Natural Logarithm

There is a special base number called *Euler's constant*, namely

$$e \approx 2.7182818284590 \dots$$

Precisely *why* this number is important is beyond the scope of this discussion. Nonetheless, when the base number is equal to e , the corresponding operation is called the *natural logarithm*:

$$\log_e (B) = \ln (B)$$

Correspondingly, we also have

$$e^{\ln(B)} = B$$

$$\ln(e^B) = B.$$

It's sometimes useful to convert a logarithm to base e . For a number N , take any base number A and use the $\log_A$ operator to establish

$$\log_A (N) = \log_A (e^{\ln(N)}) = \ln(N) \times \log_A (e),$$

or

$$\ln(N) = \frac{\log_A (N)}{\log_A (e)}.$$

Similarly, one may write

$$\ln(N) = \ln(A^{\log_A(N)})$$

to find

$$\ln(N) = \log_A(N) \times \ln(A),$$

or equivalently

$$\log_A(N) = \frac{\ln(N)}{\ln(A)}$$

Summary

To summarize the properties of the logarithm, Equations (1.1)-(1.4) tell us:

$$\log_A(A^B) = B$$

$$\log_A(B) + \log_A(C) = \log_A(BC)$$

$$\log_A(B) - \log_A(C) = \log_A\left(\frac{B}{C}\right)$$

$$C \times \log_A(B) = \log_A(B^C)$$

$$\log_A(B) \times \log_B(A) = 1$$

Problems

Problem 27

Derive the identity:

$$\ln(A) \times \log_A(e) = 1$$

Problem 28

For two base numbers A, B , derive

$$\log_A(N) = \log_B(N) \times \log_A(B)$$

$$\log_B(N) = \log_A(N) \times \log_B(A),$$

and also:

$$\log_A(B) \times \log_B(A) = 1 \quad (1.5)$$

Problem 29

Given

$$5^x = 17,$$

solve for x . Answer:

$$x = \frac{\log_A(17)}{\log_A(5)}$$

Problem 30

Spot the error in the following 'proof' that: $1 = 2 = 3 = \dots$

$$1 = 1^2 = 1^3 = 1^4 = \dots$$

$$\ln(1) = \ln(1^2) = \ln(1^3) = \ln(1^4) = \dots$$

$$\ln(1) = 2 \ln(1) = 3 \ln(1) = 4 \ln(1) = \dots$$

$$\ln(1) = 2 \ln(1) = 3 \ln(1) = 4 \ln(1) = \dots$$

$$1 = 2 = 3 = 4 = \dots$$

5 Variables and Equations

A *variable* is any symbol used to represent a number (or anything). Common symbols used are familiar letters x , y , z , a , b , c - and some unfamiliar (Greek) letters α , β , γ , etc.

Variables behave exactly as numbers do: any place where a number is needed, a variable will work just as well.

Variable Notation

When a variable is multiplied by a number or another variable, the multiplication symbol is usually ignored. This means $4 \times x$ or $4 \cdot x$ should be written $4x$. When a variable appears alone, there is always an invisible 1 next to it. That is, x is equivalent to $1x$ or $x/1$.

Coefficients

When a variable is multiplied by a number, i.e. $3x$, $-5y$, etc., the leading number is called a *coefficient*. When a variable is 'by itself', there is an implied coefficient of one.

Problem 31

Show that the GCF and LCM of $4a^2b$, $6ab$, $8ab^2$ are:

$$\text{GCF} = 2 \times a \times b = 2ab$$

$$\text{LCM} = 2^3 \times 3 \times a^2 \times b^2 = 24a^2b^2$$

Variable Assignment

To *set a variable* is to associate a number with a symbol. For example, the statement

$$x = 3$$

will 'tie' to the value of 3 to the symbol x . Any time x occurs, it really means 3. The value of a variable is the number represented by the symbol. That is, we say 3 is the 'value' of x .

Variable names such as x , y , z , etc. are heavily recycled. That is, the same letters are prone to show up in different problems. Needless to mention, variables in one problem are not the same in a different problem. (They wouldn't be variables otherwise.)

Combining Like Terms

When an expression contains two or more instances of the same variable raised to the same exponent, the coefficients may be added (or subtracted) in a process called *combining like terms*.

To demonstrate, consider the expression:

$$2x - 5x + 7x^2$$

Notice that $2x$ and $-5x$ each contain a single instance of x . It follows that these terms may be added together to give $-3x$. The $7x^2$ term can't be combined with anything, because there are no other x^2 terms in play. The simplified expression is then:

$$-3x + 7x^2$$

Problem 32

Add the two trinomials:

$$(2x^2 + 2y^2 - 3) + (y^2 - 9x^2 - 2)$$

to get

$$-7x^2 + 4y^2 - 5.$$

Problem 33

Add the two polynomials

$$(2x^2 + 2y^2 - 3) - (2y^2 - 4x^2)$$

to get

$$6x^2 - 3.$$

Problem 34

Multiply

$$(y - 4)(y^2 - 3x - 2)$$

to get

$$y^3 - 4y^2 - 3xy + 12x - 2y + 8.$$

5.1 Manipulating an Equation

Formally, an *equation* is made of two expressions joined by an equal sign. However 'ugly' an equation may appear, the left side and right side are equal. Examples of equations are:

$$5 + 5 = 10$$

$$3 - x = 2$$

$$x^2 = 24 + 25$$

Whatever manipulations are performed on an equation, the information it contains must not change: the left side must always equal the right side. The number of allowed manipulations boil down to just *two*: you may add zero to an equation, or multiply one into an equation, that's it. In the following we unpack this notion.

Adding or Subtracting a Number

Adding or subtracting the same number on both sides of an equation is the same as adding zero. For example, consider the equation

$$x - 5 = 3 .$$

If we simply add 5 to each side, the equation becomes

$$x - \cancel{5} + \cancel{5} = 3 + 5 ,$$

which simplifies very nicely:

$$x = 8$$

By isolating x , we have successfully ‘solved for’ its value.

Multiplying or Dividing a Number

Multiplying or dividing the same number on both sides of an equation is the same as multiplying by one. For example, let us isolate x in the equation

$$2x = 24 .$$

To do so, multiply both sides by a factor of $1/2$ to get

$$\left(\frac{1}{2} \times \cancel{2}\right) x = \frac{1}{2} \times 24 ,$$

and the coefficients cancel out on the left:

$$x = 12$$

Of course, one exception to the above occurs when attempting to divide by zero. If you ever do this, turn back, it’s an error.

Raising to a Power

Raising each side of an equation to a power is another allowed operation. For example, consider the equation:

$$\sqrt{x} = 3$$

If we raise each side to the power 2, the equation becomes

$$(\sqrt{x})^2 = 3^2 ,$$

or

$$(x^{1/2})^2 = x = 3^2 ,$$

and thus $x = 9$.

Utilizing Logarithm

The logarithm operators $\log_A$, $\ln$, can also be applied to each side of an equation. Recycling the previous example, we could solve $\sqrt{x} = 3$ by writing

$$\log_x (\sqrt{x}) = \frac{1}{2} \log_x (x) = \log_x (3) ,$$

or

$$\log_x (x) = \log_x (3^2) ,$$

reducing (again) to $x = 9$.

Taking a Root

Taking the square root (or cube root, etc.) of each side of an equation is an allowed operation for positive numbers. For example, consider the equation:

$$x^2 = 49$$

Putting the square root symbol around the entire left side and the entire right side, the equation becomes

$$\sqrt{x^2} = \sqrt{49} ,$$

where the square root and the 2-power cancel, giving

$$\sqrt{x^2} = x = 7 .$$

Paying closer attention, note that $x = -7$ also satisfies the original equation $x^2 = 49$. This is in fact true for all square roots: whenever we take the square root of a number, there are *two* results, one positive and one negative. In the most general case, we will always have

$$\sqrt{a^2} = \begin{cases} +a \\ -a \end{cases} , \quad (1.6)$$

usually condensed using the ‘plus-or-minus’ symbol ($\pm$):

$$\sqrt{a^2} = \pm a$$

Strategy

To summarize the allowed manipulations on an equation, we say **you may multiply by one or add zero**. Anything else is error. Do not be sloppy when applying operators across an equation. To illustrate, consider the equation

$$3x + 6 = 12y + 18.$$

There are many ways to manipulate the equation above, but for the sake of demonstration, let us try the square root operation. It would be **WRONG** to try:

$$\sqrt{3x} + \sqrt{6} = \sqrt{12y} + \sqrt{18},$$

as this approach forgets to treat the left side and right side equally. To proceed correctly, we must write

$$\sqrt{3x + 6} = \sqrt{12y + 18}.$$

Conversely, if we wanted to square each side of the equation, it would be **WRONG** to try

$$3x^2 + 6^2 = 12y^2 + 18^2,$$

which again forgets to treat each side equally. Instead, we must have

$$(3x + 6)^2 = (12y + 18)^2.$$

Problem 35

Given $a = b$, spot the error in the following ‘proof’ that $2 = 1$:

$$\begin{aligned} a &= b \\ a^2 &= ab \\ a^2 - b^2 &= ab - b^2 \\ (a - b)(a + b) &= b(a - b) \\ \cancel{(a - b)}(a + b) &= b\cancel{(a - b)} \\ a + b &= b \\ b + b &= b \\ 2b &= b \\ 2 &= 1 \end{aligned}$$

5.2 Solving for a Variable

When a variable is embedded in an equation, we often need to determine its exact value to solve a problem. This is done by manipulating the equation so as to ‘get the variable by itself’ on the left side or the right side, a process called *solving for a variable*.

Problem 36

Solve for x :

$$\frac{4x}{5} + \frac{7x}{9} = 4$$

Step 1: Rewrite fractions with a common denominator:

$$\frac{36x}{45} + \frac{35x}{45} = 4$$

Step 2: Combine like terms:

$$\frac{71x}{45} = 4$$

Step 3: Multiply both sides by 45:

$$\frac{71x}{\cancel{45}} \times \cancel{45} = 4 \times 45$$

Step 4: Divide both sides by 71 and simplify:

$$x = \frac{4 \times 45}{71} = \frac{180}{71}$$

Problem 37

Solve for x :

$$6 = 2\sqrt{x + 3}$$

Step 1: Divide by 2 on both sides:

$$3 = \sqrt{x + 3}$$

Step 2: Raise each side to the power 2:

$$3^2 = (\sqrt{x + 3})^2$$

Step 3: Simplify:

$$9 = x + 3$$

Step 4: Subtract 3 from both sides:

$$6 = x$$

Problem 38

Find two solutions for x :

$$x^2 + 3 = 12$$

Step 1: Subtract 3 from both sides:

$$x^2 = 9$$

Step 2: Take the square root of each side:

$$\sqrt{x^2} = \sqrt{9}$$

Step 3: Simplify to write the result:

$$x = \pm 3$$

Problem 39

If $y = 2$, determine all solutions for x :

$$x^2 - 10y - 5 = 0$$

Step 1: Add $10y + 5$ to both sides:

$$x^2 = 10y + 5$$

Step 2: Take the square root of both sides:

$$x = \pm\sqrt{10y + 5}$$

Step 3: Substitute $y = 2$ and simplify:

$$x = \pm\sqrt{25}$$

Step 4: Use the a table, calculator, or memory to write the final answer:

$$x = \pm 5$$

Problem 40

What three digits are represented by X , Y , and Z in the addition problem shown?

$$\begin{array}{r} \text{XZY} \\ + \text{XYZ} \\ \hline \text{YZX} \end{array}$$

Answer: The first column indicates $Y \neq 0$. The second column suggests $Z + Y \geq 10$, and furthermore $10 + 10Z + 10Y = 10Z + 100$, giving $Y = 9$. Thus, the first column is only satisfied by $X = 4$. Finally, the third column tells us $Y + Z = 10 + X$, or $Z = 5$.

Problem 41

Shown is a long division with some digits hidden. Which of (5, 6, 7, 8) could represent one of the empty boxes?

$$\begin{array}{r} \boxed{4} \\ \boxed{} \overline{) \boxed{} \boxed{3}} \\ \underline{ \boxed{} \boxed{2}} \\ \boxed{1} \end{array}$$

Answer: Label the divisor as x , where y and z are natural numbers. Observe $10x + 3 - (10z + 2) = 1$ to conclude $y = z$. Observe next that

$$\frac{10y + 3}{x} = 4 + \frac{1}{x},$$

which only delivers a natural number if $y = 3$ and $x = 8$.

Problem 42

At temperature T in Celsius, the speed of sound v in air (in meters per second) is given by

$$v \approx 331.1 \times \sqrt{1 + \frac{T}{273.15}}.$$

If the speed of sound in a certain place is measure to be exactly 350 m/s, calculate the required temperature.

Problem 43

The area of a circle is πR^2 , where $\pi \approx 3.14$ and R is the radius (center-to-edge distance). If the area of a certain pizza is 100 units, calculate the length of a slice (same as R).

5.3 Systems of Equations

When handed multiple equations containing multiple variables, this is called a system of equations. The number of equations/variables is called the *order* of the system. For example, below we have an order-two system, having two equations and two variables:

$$4x + 2y = 20$$

$$x + 3y = 15$$

There is a reliable technique for solving a system of two equations and two variables. Using the above system as an example, the following procedure generally works:

1. Solve either equation for x or y , whichever is easier.

$$y = 10 - 2x$$

$$x + 3y = 15$$

2. Substitute the isolated variable into the unused equation.

$$x + 3(10 - 2x) = 15$$

3. The resulting equation should contain only one variable:

$$x = 3$$

4. Substitute the known variable value into either of the original equations.

$$3 + 3y = 15$$

5. Solve for the last unknown:

$$y = 4$$

If the number of equations/variables does not match, the system is either *underdetermined* (not enough information), or *overdetermined* (conflicting information). In either case, there is no clear solution to the system.

Problem 44

For the linear system of equations

$$2x + 3y = 19$$

$$4x - y = 3,$$

show that the solution to the system is $x = 2, y = 5$.

Celsius and Fahrenheit

In standard conditions, water freezes at $0^\circ\text{C} = 32^\circ\text{F}$, and it boils at $100^\circ\text{C} = 212^\circ\text{F}$. The two temperature scales (Fahrenheit and Centigrade) obey a linear relation

$$T_F = m \cdot T_C + b.$$

Use a system of equation to determine m and b (as in $y = mx + b$).

Generate two equations and two unknowns:

$$32^\circ\text{F} = m \cdot 0 + b$$

$$212^\circ\text{F} = m \cdot 100^\circ\text{C} + b$$

First eliminate b to solve for m :

$$m = \frac{180^\circ\text{F}}{100^\circ\text{C}}$$

Next solve for b to get

$$b = 32^\circ\text{F},$$

and write the final answer:

$$T_F = \frac{9}{5} \cdot T_C + 32^\circ\text{F}$$

Problem 45

Solve for T_C in terms of T_F .

5.4 Linearity of Equations

Linear System

A(n) (system of) equation(s) is considered *linear* if all variables x, y, z , etc. have an exponent of exactly *one*. A linear equation should never have terms such $\sqrt{x}$, or $x \cdot y$, or z^2 , as these would be considered *nonlinear* terms. For example, the set of equations

$$4x + 2y = 20$$

$$x + 3y = 15$$

qualifies as an order-two linear system, and can always be solved by the procedure outlined above.

Nonlinear System

On the other hand, the system

$$x^2 + y = 21$$

$$xy = 20$$

contains two nonlinear equations due to the x^2 - and xy -terms. Nonetheless, the solution can be attained by reducing the system to one equation with one variable. In this case, substitute $y = 20/x$ into the first equation and simplify to get

$$x^3 - 21x + 20 = 0,$$

which is still nonlinear, but *can* be solved. (There are two solutions, one with $x = 1$ and the other with $x = 4$.)

Transcendental Equation

An equation is *transcendental* if the variable(s) being solved for cannot be isolated by the usual tools of algebra. For instance, the equation

$$2^x = x^2$$

qualifies as transcendental, as no ordinary process can be used to isolate x . Many transcendental equations can still be solved by analytic means, or when that fails, approximated by computer.

6 Rate, Interval, and Work

Units and Dimension

A number that is measured in a countable quantity such as dollars, seconds, or pounds are said to carry *units*, also called *dimension*. A plain number such as 3 (ant not 3 dollars or seconds) is called *dimensionless*.

Unit quantities can be manipulated as one would do with variables, and are subject to the same rules. For instance, we may add numbers carrying units by combining like terms as in

$$2 \text{ meter} + 3 \text{ meter} = 5 \text{ meter} ,$$

or multiply as in

$$2 \text{ second} \times 3 \text{ second} = 6 \text{ second}^2 .$$

As long as we obey the rules for combining fractions, rates can also be combined, as in

$$14 \frac{\text{dollar}}{\text{hour}} - 3 \frac{\text{dollar}}{\text{hour}} = 11 \frac{\text{dollar}}{\text{hour}} .$$

Using fewer letters, the above calculations are identical to

$$2x + 3x = 5x$$

$$2y \times 3y = 6y^2$$

$$14z - 3z = 11z ,$$

where x replaces meter, y replaces second, and z replaces dollar/hour.

6.1 Rate

A *rate* is defined as some quantity, such as a cost, weight, or area, measured against another quantity, such as time or distance. Rates are *always* ratios, containing two pieces of information, hence the similarity in name. Common examples of rates may occur as:

$$\text{Hourly Wage} = 14 \text{ dollar/hour}$$

$$\text{Speed} = 15 \text{ meter/second}$$

$$\text{Pressure} = 55 \text{ pound/inch}^2$$

$$\text{Density} = 55 \text{ gram/cm}^3$$

In general, we shall denote a rate with the variable letter R . For instance, for a car traveling at 40 miles per hour (abbreviated mi/hr or mph), we may write

$$R = 40 \frac{\text{mi}}{\text{hr}} = 40 \text{ mph} .$$

6.2 Interval

Looking again at the definition of a rate, there is a numerator of a certain dimension (dollars, meters, etc.), along with a denominator of a different dimension (usually *time*, but not always.)

For any given rate R , we can conceive a number called an *interval*, denoted I , having units matching those in the denominator of R . For instance, consider an example rate

$$R = 11 \frac{\text{dollar}}{\text{hour}} .$$

To construct an interval, choose *some* quantity I measured specifically in hours, perhaps

$$I = 3 \text{ hour} .$$

By computing the product $R \times I$, we find

$$R \times I = 11 \frac{\text{dollar}}{\text{hour}} \times 3 \text{ hour} = 33 \text{ dollar} ,$$

which cancels the ‘hours’ unit altogether, leaving the result in only dollars.

6.3 Work

The product of the rate and the interval is called *work*, denoted W :

$$W = R \times I \quad (1.7)$$

Reiterating the previous example, we say that a rate of $R = 11$ dollar/hour multiplied by an interval of $I = 3$ hour amounts to the work value of $W = 33$ dollar.

The so-called ‘work equation’ can be solved for R , namely

$$R = \frac{W}{I} ,$$

telling us that a rate can be interpreted as the work value divided by the interval. Similarly, the interval

$$I = \frac{W}{R}$$

is the ratio of the work value to the rate.

Problem 46

Joseph works as a journeyman electrician for a pay rate of 30 dollars per hour. After a 9-hour day, what are his daily earnings?

Step 1: Identify the rate, the interval, and the work value:

$$\begin{aligned}R &= 30 \frac{\text{dollar}}{\text{hour}} \\I &= 9 \text{ hour} \\W &= ?\end{aligned}$$

Step 2: Apply the work equation with W as the unknown:

$$W = R \times I = 30 \frac{\text{dollar}}{\text{hour}} \times 9 \text{ hour}$$

Step 3: Cancel the hour units, and simplify the fraction:

$$\begin{aligned}W &= R \times I = 30 \frac{\text{dollar}}{\text{hour}} \times 9 \text{ hour} \\&= 30 \times 9 \text{ dollar} = 270 \text{ dollar}\end{aligned}$$

Problem 47

Joseph keeps working as a journeyman electrician for a pay rate of 30 dollars per hour. After a long workday, his earnings were 330 dollars. How many hours did he work?

Step 1: Identify the rate, the interval, and the work value:

$$\begin{aligned}R &= 30 \frac{\text{dollar}}{\text{hour}} \\I &= ? \\W &= 330 \text{ dollar}\end{aligned}$$

Step 2: Apply the work equation with I as the unknown:

$$I = \frac{W}{R} = \frac{330 \text{ dollar}}{30 \text{ dollar/hour}}$$

Step 3: Cancel the dollar units, and simplify the fraction:

$$I = \frac{W}{R} = \frac{330 \cancel{\text{dollar}}}{30 \cancel{\text{dollar}}/\text{hour}} = \frac{330}{30} \text{ hour} = 11 \text{ hour}$$

Problem 48

Joseph agrees to work on a special project for 10 hours. His earnings for the project were 350 dollars. Calculate the pay rate for the special project.

Step 1: Identify the rate, the interval, and the work value:

$$\begin{aligned}R &= ? \\I &= 10 \text{ hour} \\W &= 350 \text{ dollar}\end{aligned}$$

Step 2: Apply the work equation with R as the unknown:

$$R = \frac{W}{I} = \frac{350 \text{ dollar}}{10 \text{ hour}}$$

Step 3: Simplify to write the answer:

$$R = 35 \frac{\text{dollar}}{\text{hour}}$$

Problem 49

A bee can visit 644 flowers in 7 hours. How many flowers can the bee visit in 9 hours?

Step 1: Set up a work equation containing known information

$$\begin{aligned}W &= R \times I \\644 \text{ flower} &= R \times 7 \text{ hour}\end{aligned}$$

Step 2: Solve for the rate R :

$$R = \frac{644 \text{ flower}}{7 \text{ hour}} = 92 \frac{\text{flower}}{\text{hour}}$$

Step 3: Step up a work equation with R known and $I = 9$ hour to get the result:

$$\begin{aligned}W &= R \times I \\W &= 92 \frac{\text{flower}}{\text{hour}} \times 9 \text{ hour} \\W &= 92 \times 9 \text{ flower} = 828 \text{ flower}\end{aligned}$$

Mixed Units

Certain problems call for a rate-like quantity to have mixed units.

Suppose a certain brand of concentrated iced tea advises that 2 oz of product mixed with 12 oz of water makes the perfect serving. How much of each ingredient is needed to fill a 64 oz container at the proper ratio?

Proceed by writing an equation containing known information:

$$14 \text{ oz tea} = 2 \text{ oz product} + 12 \text{ oz water},$$

and then divide each side by the number 14 to write a 'unit work' equation:

$$1 \text{ oz tea} = \frac{1}{7} \text{ oz product} + \frac{6}{7} \text{ oz water}$$

We can interpret the work value as 1 oz tea, but the quantity $I \times R$ is tangled up in the sum on the right. The natural

choice for the I -variable is the number of ounces of each ingredient, so we write the above in equivalent form:

$$(1 \text{ oz}) \text{ tea} = \left(\frac{1}{7} \text{ product} + \frac{6}{7} \text{ water} \right) \times (1 \text{ oz})$$

Now, it's clear that 1 oz can be replaced with any 'work' quantity, not restricted to ounces. Choosing 64 oz, we find

$$\begin{aligned} (64 \text{ oz}) \text{ tea} &= \left(\frac{1}{7} \text{ product} + \frac{6}{7} \text{ water} \right) \times (64 \text{ oz}) \\ &= 9.14 \text{ oz product} + 54.86 \text{ oz water} \end{aligned}$$

6.4 Adding Rates

Certain situations call for individual rates to be combined into an *effective rate*. That is, if we have a handful of compatible rates R_1 , R_2 , etc., the effective rate is

$$R_{\text{eff}} = R_1 + R_2 + \cdots,$$

obeying the same work equation

$$W = R_{\text{eff}} \times I.$$

The obvious assumption is that all rates being combined share the same units. It makes no sense (in any context) to add rates with dissimilar units.

Problem 50

An inlet pipe can fill a swimming pool in 5 hours, while an outlet pipe can drain the pool in 8 hours. By mistake, a maintenance worker left both pipes open. Will the pool overflow?

Step 1: Identify the rate, the interval, and the work value:

$$R = R_{\text{eff}} = R_{\text{fill}} + R_{\text{drain}}$$

$$I = (\text{not needed})$$

$$W = 1 \text{ pool}$$

Step 2: Determine the effective rate: (Hint: Find the LCM of 5, 8.)

$$\begin{aligned} R_{\text{eff}} &= \frac{1 \text{ pool}}{5 \text{ hour}} - \frac{1 \text{ pool}}{8 \text{ hour}} \\ R_{\text{eff}} &= \frac{8 \text{ pool}}{40 \text{ hour}} - \frac{5 \text{ pool}}{40 \text{ hour}} \\ R_{\text{eff}} &= \frac{3 \text{ pool}}{40 \text{ hour}} = 0.075 \frac{\text{pool}}{\text{hour}} \end{aligned}$$

Step 3: Observe that R_{eff} came out positive, meaning the pool will overflow.

Problem 51

A cold water faucet can fill a bathtub in 12 minutes, and a hot water faucet can fill the bathtub in 18 minutes. The drain can empty the bathtub in 24 minutes. If both faucets are on and the drain is open, how long would it take to fill the bathtub?

Step 1: Identify the rate, the interval, and the work value:

$$R = R_{\text{eff}} = R_{\text{hot}} + R_{\text{cold}} + R_{\text{drain}}$$

$$I = ?$$

$$W = 1 \text{ tub}$$

Step 2: Determine the effective rate: (Hint: Find the LCM of 12, 18, and 24.)

$$\begin{aligned} R_{\text{eff}} &= \frac{1 \text{ tub}}{12 \text{ minute}} + \frac{1 \text{ tub}}{18 \text{ minute}} - \frac{1 \text{ tub}}{24 \text{ minute}} \\ R_{\text{eff}} &= \frac{6 \text{ tub}}{72 \text{ minute}} + \frac{4 \text{ tub}}{72 \text{ minute}} - \frac{3 \text{ tub}}{72 \text{ minute}} \\ R_{\text{eff}} &= \frac{7 \text{ tub}}{72 \text{ minute}} \end{aligned}$$

Step 3: Apply the work equation with I as the unknown:

$$I = \frac{W}{R_{\text{eff}}} = W \times \frac{1}{R_{\text{eff}}} = 1 \text{ tub} \times \frac{72 \text{ minute}}{7 \text{ tub}}$$

Step 4: Cancel the tub units, and simplify the fraction:

$$\begin{aligned} I &= 1 \cancel{\text{tub}} \times \frac{72 \text{ minute}}{7 \cancel{\text{tub}}} \\ &= \frac{72}{7} \text{ minute} = 10.29 \text{ minute} \end{aligned}$$

Problem 52

It takes Tom 4 hours to build a fence. If he hires Jack to help him, together they can do the job in just 3 hours. If Jack built the same fence alone, how long would it take him?

Step 1: Discern the following equations from the problem statement:

$$1 \text{ fence} = R_{\text{Tom}} \times 4 \text{ hour}$$

$$1 \text{ fence} = (R_{\text{Tom}} + R_{\text{Jack}}) \times 3 \text{ hour}$$

$$1 \text{ fence} = R_{\text{Jack}} \times I_{\text{Jack}}$$

Step 2: Solve the middle equation for R_{Jack} :

$$\begin{aligned} R_{\text{Jack}} &= \frac{1 \text{ fence}}{3 \text{ hour}} - R_{\text{Tom}} \\ &= \frac{1 \text{ fence}}{3 \text{ hour}} - \frac{1 \text{ fence}}{4 \text{ hour}} = \frac{1 \text{ fence}}{12 \text{ hour}} \end{aligned}$$

Step 3: Solve the third equation for I_{Jack} :

$$\begin{aligned} I_{\text{Jack}} &= \frac{1 \text{ fence}}{R_{\text{Jack}}} \\ &= 1 \text{ fence} \times \frac{12 \text{ hour}}{1 \text{ fence}} = 12 \text{ hour} \end{aligned}$$

Problem 53

A woodworking shop is open eight hours per day. Tony can build three birdhouses in the eight-hour day. Joe is new, and accidentally destroys one birdhouse every four hours. If the shop hires a third builder that can make x birdhouses every six hours, solve for x such that the shop builds five total birdhouses per day.

Step 1: Let H denote ‘house’ and ‘hr’ stand for ‘hour’. Identify the rate, the interval, and the work value:

$$R = R_{\text{eff}} = R_{\text{Tony}} + R_{\text{Joe}} + \frac{x H}{6 \text{ hr}}$$

$$I = 8 \text{ hr}$$

$$W = 5 H$$

Step 2: Write the work equation using the information provided and simplify:

$$\begin{aligned} W &= R_{\text{eff}} \times I \\ 5 H &= \left(\frac{3 H}{8 \text{ hr}} - \frac{1 H}{4 \text{ hr}} + \frac{x H}{6 \text{ hr}} \right) \times 8 \text{ hr} \\ 5 \cancel{H} &= \left(\frac{3 \cancel{H}}{8 \cancel{\text{hr}}} - \frac{1 \cancel{H}}{4 \cancel{\text{hr}}} + \frac{x \cancel{H}}{6 \cancel{\text{hr}}} \right) \times 8 \cancel{\text{hr}} \\ 5 &= \left(\frac{3}{8} - \frac{1}{4} + \frac{x}{6} \right) \times 8 \end{aligned}$$

Step 3: Solve the above for x :

$$x = 6 \cdot \left(\frac{5}{8} - \frac{3}{8} + \frac{1}{4} \right) = 3$$

Problem 54

Source: *PHCC of MA*. Slide 26.

A water storage tank contains 6,500 gallons of water. The tank is 1/3 full and supplied water at a rate of 75 gpm. The tank empties at a rate of 32 gpm. How long will it take to fill the tank?

Answer:

$$\begin{aligned} 6500 \text{ gal} &= (1/3) 6500 \text{ gal} + (75 - 32) \frac{\text{gal}}{\text{min}} \times T \\ T &= \frac{(1 - 1/3) \times 6500 \text{ gal}}{(75 - 32) \text{ gal/min}} \\ &= 100.8 \text{ min} \end{aligned}$$

Canoe Problem

Problem 55

On a canoe trip, you paddle upstream for one mile, at which point your hat falls into the river. Ten minutes later, you turn around and catch up to the hat precisely where the trip began. Calculate the speed of the river.

Answer: Assume you can paddle the canoe at speed v on still water. If the speed of the river is $\tilde{v}$, then the canoe's net speed is $v - \tilde{v}$ upstream, and $v + \tilde{v}$ downstream.

Let D equal the distance traveled before the canoe turns around, thus we write

$$1 \text{ mile} + (v - \tilde{v}) 10 \text{ min} = D$$

for upstream. For downstream we have

$$(v + \tilde{v}) T = D,$$

where T is the time required to paddle back to where the canoe began.

Note that the time required for the hat to float downstream to its starting point is T plus ten minutes, so we may also write

$$1 \text{ mile} = \tilde{v} (T + 10 \text{ min}).$$

Eliminate D between the first two equations and eliminate factors of $\tilde{v}$ using the third equation to find $T = 10 \text{ min}$. It quickly follows that

$$\tilde{v} = \frac{1 \text{ mile}}{20 \text{ min}}.$$

Chapter 2

Units and Dimension

1 Dimensional Analysis

Quantities carrying dimension can be converted from one set of units to another using *dimensional analysis*, also known as *unit conversion*. One may convert inches to centimeters, or seconds to hours, or liters to gallons because these conversions preserve the fundamental dimension of the quantity. On the other hand, trying to convert microseconds to pounds is a nonsensical attempt to equate a time to a weight.

Unit Rate

Starting with a typical entry from a unit conversion table, for instance, $1 \text{ in} = 2.54 \text{ cm}$, this can be converted into a *unit rate* by dividing through by either side, giving

$$1 = \frac{2.54 \text{ cm}}{1 \text{ in}}$$

or

$$1 = \frac{1 \text{ in}}{2.54 \text{ cm}}.$$

In each case, we have a dimensionless number 1 on the left side of the equation. On the right, we have a rate with definite units.

From algebra, we know that multiplying by one leaves the quantity unchanged. It follows that multiplying, say, twelve inches by a carefully chosen factor of one should yield the equivalent in new units:

$$\begin{aligned} 12 \text{ in} &= 12 \text{ in} \times (1) = 12 \cancel{\text{in}} \times \left(\frac{2.54 \text{ cm}}{1 \cancel{\text{in}}} \right) \\ &= 12 \times 2.54 \text{ cm} = 30.48 \text{ cm} \end{aligned}$$

Dimensional analysis must also work in reverse. Driving the previous example backwards, start with with 30.48 cm,

and multiply by another carefully-chosen factor of one so that centimeters cancel:

$$\begin{aligned} 30.48 \text{ cm} &= 30.48 \text{ cm} \times (1) \\ &= 30.48 \cancel{\text{cm}} \times \left(\frac{1 \text{ in}}{2.54 \cancel{\text{cm}}} \right) = 12 \text{ in} \end{aligned}$$

Often, quantities of length in inches, centimeters, etc., occur as their square or cube, i.e. in^2 , cm^3 , etc. For these we take the corresponding conversion factor to the appropriate power so that units cancel properly.

For instance, converting from square inches to square centimeters looks like:

$$\begin{aligned} 25 \text{ in}^2 &= 25 \text{ in}^2 \times (1^2) = 25 \cancel{\text{in}^2} \left(\frac{2.54^2 \text{ cm}^2}{1^2 \cancel{\text{in}^2}} \right) \\ &= 25 \cdot 2.54^2 \text{ cm}^2 \approx 161 \text{ cm}^2 \end{aligned}$$

1.1 Distance, Time, Speed

Problem 1

Convert 12.69 feet to inches.

Answer:

$$12.69 \cancel{\text{ft}} \times \left(\frac{12 \text{ in}}{1 \cancel{\text{ft}}} \right) = 152.28 \text{ in}$$

Problem 2

Joe finds a wrench on the shop floor. Comparing to his own wrench set, he learns the unknown wrench is slightly larger than $9/16 \text{ in}$, and slightly smaller than $5/8 \text{ in}$. Estimate the size of the wrench in millimeters.

Answer $\approx 15 \text{ mm}$

Problem 3

How many seconds are there in 30 years?

Problem 4

Explain why six weeks is 10! seconds.

Problem 5

The speed limit on a Canadian highway is 120 kilometers per hour. Convert this speed to miles per hour.

Answer: 74.6 mph

Problem 6

Convert 70 meters per second to miles per hour. Answer: 156.6 mph

1.2 Area and Volume

Squares, Rectangles, Boxes

Let the width W and height H be perpendicular dimensions of a square or rectangle.

$$\text{Area} = W \times H$$

$$\text{Diagonal} = \sqrt{W^2 + H^2}$$

$$\text{Perimeter} = 2(W + H)$$

Let the length L represent a third dimension perpendicular to W and L . We then have:

$$\text{Volume} = W \times H \times L$$

$$\text{Diagonal} = \sqrt{W^2 + H^2 + L^2}$$

$$\text{Surface} = 2(WH + HL + LW)$$

For the special case of a perfectly cubic box, we have $W = H = L$, and then

$$\text{Cube Surface} = 6L^2.$$

Triangular Area and Volume

Let the width W and height H be perpendicular dimensions of a triangle.

$$\text{Area} = \frac{1}{2} \times W \times H$$

$$\text{Diagonal} = \sqrt{W^2 + H^2}$$

$$\text{Perimeter} = W + H + \sqrt{W^2 + H^2}$$

Let the length L represent a third dimension perpendicular to W and L . For the the volume of such a 'tent' we have

$$\text{Volume} = \frac{1}{2} \times W \times H \times L.$$

The surface area, including both end caps, takes some more work. In particular:

$$\text{Surface} = WH + HL + LW + L\sqrt{W^2 + H^2}$$

Circles and Cylinders

A circle is characterized by its diameter D , which is the distance across the circle through the center. The radius is defined as half the diameter:

$$R = \frac{D}{2}$$

The area of any circle is given by

$$\text{Area} = \pi R^2 = \frac{\pi}{4} D^2,$$

and the circumference C is given by

$$C = 2\pi R = \pi D.$$

Let the length L represent a third dimension perpendicular to the diameter to stretch the circle to a cylinder. The volume of is given by

$$\text{Volume} = \pi R^2 L = \frac{\pi}{4} D^2 L.$$

The surface area including both end caps is

$$\begin{aligned} \text{Surface} &= 2\pi R^2 + 2\pi RL \\ &= \frac{\pi}{2} D^2 + \pi DL. \end{aligned}$$

Problem 7

Calculate the area of a four-inch diameter circle. Convert the result to square centimeters.

Problem 8

Calculate the volume of one gallon of (any) liquid in cubic feet.

Answer:

$$\begin{aligned} 1 \text{ gal} &= 3.785 \cancel{\text{L}} \frac{1000 \cancel{\text{mL}}}{\cancel{\text{L}}} \frac{1 \text{ cm}^3}{\cancel{\text{mL}}} \frac{1 \cancel{\text{in}}^3}{2.54^3 \cancel{\text{cm}}^3} \frac{1 \text{ ft}^3}{12^3 \cancel{\text{in}}^3} \\ &= \frac{3.785 \times 1000}{2.54^3 \times 12^3} \text{ ft}^3 = 0.13367 \text{ ft}^3 \end{aligned}$$

Problem 9

One gallon of liquid is equal to how many cubic inches?

Answer:

$$1 \text{ gal} = 0.13367 \text{ ft}^3 \left(\frac{12 \text{ in}}{\text{ft}} \right)^3 = 231 \text{ in}^3$$

Problem 10

One cubic foot of liquid is equal to how many gallons?

Answer:

$$1 \text{ ft}^3 = \frac{1 \text{ gal}}{0.13367} = 7.4813 \text{ gal}$$

Problem 11

One cubic inch of liquid is equal to how many gallons?

Answer:

$$1 \text{ ft}^3 \left(\frac{12^3 \text{ in}^3}{\text{ft}^3} \right) = 7.4813 \text{ gal}$$

$$1 \text{ in}^3 = \frac{7.4813}{12^3} \text{ gal} = 0.004329456 \text{ gal}$$

Problem 12

Source: *PHCC of MA*. Slide 20.

If 3 inches of rain was trapped on a 62 ft long $\times$ 40 ft wide roof, how many gallons would that equal?

Answer:

$$V = (62 \times 40 \times 0.25) \text{ ft}^3 \times 7.48052 \frac{\text{gal}}{\text{ft}^3}$$

$$= 4637.92 \text{ gal}$$

Problem 13

One cubic foot is equal to how many cubic meters?

Answer:

$$1 \text{ ft}^3 = \left(\frac{12 \times 2.54}{100} \right)^3 \text{ m}^3 = 0.028317 \text{ m}^3$$

Problem 14

Convert 4500 gallons to cubic meters. Answer: 17.03m³

Problem 15

Three gallons of paint are uniformly spread over a wall having area 600 ft². Calculate the thickness of the paint in millimeters. Answer: 0.204 mm

Problem 16

A plumbing license study guide *PHCC of MA* (Slide 25) poses the question: *What is the capacity in gallons of a 3-inch diameter pipe which is 27 inches long?* The solution is claimed to be

$$\text{Capacity} = \text{Diameter}^2 \times \text{Length} \times 0.0034 .$$

Derive the formula for the factor of 0.0034.

Answer:

$$\frac{\pi}{4 \times 231} \approx 0.003399992$$

Problem 17

The cylindrical gallon capacity formula

$$\text{Capacity} = \text{Diameter}^2 \times \text{Length} \times X$$

with $X \approx 0.0034$ only applies when the diameter and length are in inches. Derive similar factors Y , Z that apply for (i) diameter in inches and length in feet, (ii) diameter in feet and length in feet.

Answer:

$$Y = \frac{12\pi}{4 \times 231} \approx 0.04079990$$

$$Z = \frac{12^3\pi}{4 \times 231} \approx 5.875186$$

Problem 18

Source: *PHCC of MA*. Slide 27.

How many square feet of insulation is required to wrap the outside of a storage tank which is 18 inches in diameter and 5 feet long?

Answer:

$$A = 2 \times \text{Endcaps} + \text{Body}$$

$$= 2 \times (\pi 9^2 \text{ in}^2) + (\pi \times 18 \text{ in} \times 60 \text{ in})$$

$$= \pi \times (162 + 1080) \text{ in}^2 \left(\frac{1 \text{ ft}}{12 \text{ in}} \right)^2$$

$$= \frac{\pi \times 1242}{144} \text{ ft}^2 = 27.0962 \text{ ft}^2$$

Problem 19

The mass of Earth is approximately $M = 5.972 \times 10^{24} \text{ kg}$. Using radius $R = 6371 \text{ km}$, calculate the average density (total mass over total volume) of Earth in grams per cubic centimeter.

Answer: 5.513 g/cm³

Problem 20

Suppose one cubic foot of liquid passes through a hole of diameter 1.25 inches in one minute. Calculate the speed of the liquid through the hole.

Answer: Take a cylinder of liquid having volume

$$1 \text{ ft}^3 = \pi \left(\frac{1.25}{2} \text{ in} \right)^2 L ,$$

so

$$L = \frac{1 \text{ ft}^3}{\pi (0.625 \text{ in})^2} = 1408 \text{ in} = 117.3 \text{ ft}.$$

It takes one minute for the whole volume of liquid to move through length L , so the speed is:

$$|v| = \frac{1408 \text{ in}}{60 \text{ sec}} = 23.47 \frac{\text{in}}{\text{sec}} = 117.3 \frac{\text{ft}}{\text{min}}$$

1.3 Mass vs. Weight

It is (perhaps) common knowledge that ≈ 2.2 pounds is equivalent to one kilogram. Paying attention to units though, we know this *can't* be exactly correct, as mass (just plain old mass) is not the same as weight (a force). The proper equivalency is written in the format

$$W = mg,$$

or

$$2.20462 \text{ lb} = 1 \text{ kg} \times g,$$

where g is the local gravity constant given as

$$g = 9.80665 \frac{\text{m}}{\text{s}^2}.$$

Problem 21

At sea level, a 2.205 pound rock weighs the same as a 1 kilogram object. On Mt. Everest, a 2.199 pound rock weighs the same as a 1 kilogram object. What is the weight in pounds of an 80.0 kg person at each location?

Answer: 176.4 lb, 175.9 lb

1.4 Density

The *density* of a given substance, denoted ρ (Greek *rho*), is defined as the mass per unit volume of material present:

$$\rho = \frac{m}{V}$$

At atmospheric pressure, the density of pure liquid water

varies per temperature as suggested:

$$\rho_{0^\circ\text{C}} = 999.84 \text{ kg/m}^3$$

$$\rho_{4^\circ\text{C}} = 998.97 \text{ kg/m}^3$$

$$\rho_{20^\circ\text{C}} = 998.20 \text{ kg/m}^3$$

$$\rho_{22^\circ\text{C}} = 997.77 \text{ kg/m}^3$$

$$\rho_{25^\circ\text{C}} = 997.05 \text{ kg/m}^3$$

$$\rho_{30^\circ\text{C}} = 995.65 \text{ kg/m}^3$$

$$\rho_{40^\circ\text{C}} = 992.22 \text{ kg/m}^3$$

$$\rho_{60^\circ\text{C}} = 983.20 \text{ kg/m}^3$$

$$\rho_{80^\circ\text{C}} = 971.79 \text{ kg/m}^3$$

$$\rho_{100^\circ\text{C}} = 958.40 \text{ kg/m}^3$$

Problem 22

Using $\rho_{\text{water}} = 1000 \text{ kg/m}^3$, calculate the weight in pounds of one gallon of water at sea level. Express the final answer as a weight density.

Answer:

$$\begin{aligned} m &= \rho V \\ &= \frac{1000 \text{ kg}}{\text{m}^3} (1 \text{ gal}) \frac{3.785 \cancel{\text{L}}}{1 \cancel{\text{gal}}} \left(\frac{1000 \cancel{\text{cm}^3}}{1 \cancel{\text{L}}} \right) \frac{1 \cancel{\text{m}^3}}{100^3 \cancel{\text{cm}^3}} \\ &= \frac{1000 \times 3785}{10^6} \text{ kg} = 3.785 \text{ kg} \end{aligned}$$

$$W = 3.785 \cancel{\text{kg}} \left(\frac{2.20462 \text{ lb}}{1 \cancel{\text{kg}}} \right) = 8.344 \text{ lb}$$

$$\tilde{\rho} = 8.344 \text{ lb/gal}$$

Problem 23

What is the weight in pounds of fifteen gallons of water at sea level?

Answer:

$$W = 15 \text{ gal} \left(\frac{8.344 \text{ lb}}{\text{gal}} \right) = 125.16 \text{ lb}$$

Problem 24

What is the weight in pounds of one cubic foot of water at sea level?

Answer:

$$\begin{aligned} W &= \tilde{\rho} \times 1 \text{ ft}^3 \\ &= 8.344 \frac{\text{lb}}{\text{gal}} \frac{\text{gal}}{231 \text{ in}^3} \left(\frac{12 \text{ in}}{\text{ft}} \right)^3 \times 1 \text{ ft}^3 \\ &= \frac{8.344 \times 12^3}{231} \text{ lb} = 62.417 \text{ lb} \end{aligned}$$

Problem 25

What is the weight in pounds of fifteen cubic feet of water at sea level?

Answer:

$$W = 15 \text{ ft}^3 \left(\frac{62.417 \text{ lb}}{\text{ft}^3} \right) = 936.255 \text{ lb}$$

Problem 26

What is the weight in pounds (at sea level) of one cubic inch of water?

Answer:

$$W = \frac{8.344}{231} \text{ lb} = 0.03612121 \text{ lb}$$

Problem 27

The volume of a cylindrical pipe is $V = \pi R^2 L$, where R is the radius and L is the length. Calculate the volume of water in a length $L = 1 \text{ ft}$ of pipe having diameters (i) 8 in, (ii) 6 in, (iii) 4 in, (iv) 3 in, (v) 2 in, (vi) 1.5 in, (vii) 1 in, (viii) 0.5 in. Also calculate the weight in pounds (at sea level) of each volume using ρ calculated previously. Answer:

Diam. (in)	Volume (gal)	Weight (lb)
8	2.6115	21.791
6	1.4689	12.258
4	0.65287	5.4479
3	0.3672	3.0644
2	0.16322	1.3620
1.5	0.091809	0.76610
1	0.040804	0.34049
0.5	0.010201	0.085122

1.5 Dimensional Black Magic

Dimensional analysis can occasionally substitute for the lack of proper theory. To demonstrate, it's well-known in the American trades that gas pressures are measured in units of 'inches of water column', so we need something to justify pressure being measured in inches. In particular, the pressure of a gas is attained by attaching a water column to the system and then measuring the upward displacement of water.

Assuming we don't know how to calculate pressure vs. distance relationship in water, let us list the relevant variables on hand:

$$P = 1 \frac{\text{lb}}{\text{inch}^2}$$

$$\rho_{\text{water}} = 1000 \frac{\text{kg}}{\text{m}^3}$$

$$g = 9.80665 \frac{\text{m}}{\text{s}^2}$$

Before proceeding, it's best practice to convert all units to grams and centimeters, in which case the above becomes (dropping the subscript on ρ):

$$P = \frac{453.5929 \times 980.665}{2.54^2} \frac{\text{gram cm}}{\text{cm}^2 \text{ s}^2}$$

$$\rho = 1.000 \frac{\text{gram}}{\text{cm}^3}$$

$$g = 980.665 \frac{\text{cm}}{\text{s}^2}$$

Next, let us arrange the variables P , ρ , g such that the final dimension is length. Right away, we see the unit gram needing to cancel, so this means the answer involves either P/ρ or ρ/P . Right away we rule out the latter case, as the final result (inches) should increase with increasing P . The P -variable carries an extra factor of cm/s^2 , which happens to be the dimensionality of g . Thus we deduce

$$\text{Height} = \frac{P}{\rho g} = \frac{453.5929 \times 1.000 \times \cancel{980.665}}{2.54^3 \times \cancel{980.665}} \text{ in},$$

simplifying to

$$\text{Height} = 27.67993 \text{ in} = 2.306661 \text{ ft}.$$

Problem 28

Use the above to derive the formula for P in pounds per square inch as a function of height H in feet.

Answer:

$$P = \left(\frac{2.54^3 \times 2.20462 \times 12}{1000} \right) \left(\frac{H}{\text{ft}} \right) \text{ Psi}$$

$$= 0.433527 \left(\frac{H}{\text{ft}} \right) \text{ Psi}$$

1.6 Pressure

By definition, the pressure on a surface is equal to the force on the surface divided by the surface area:

$$P = \frac{F}{A}$$

Problem 29

Calculate the force in pounds, also known as weight, on the bottom surface of a four inch diameter pipe filled with water standing vertically.

Answer:

$$\begin{aligned}F &= P \times A \\&= 0.433527 \left(\frac{10 \cancel{\text{ft}}}{\cancel{\text{ft}}} \right) \text{Psi} \times (\pi 2^2) \text{in}^2 \\&= 4.33527 \times 4\pi \frac{\text{lb}}{\text{in}^2} \text{in}^2 \\&\approx 54.47861 \text{ lb}\end{aligned}$$

1.7 Energy

In thermodynamics, it's easily established that the amount of heat energy Q transferred to an object of mass m with specific heat c scales with the change in temperature ΔT :

$$Q = mc\Delta T$$

Problem 30

Derive a formula for the energy increase in Joules if one gallon of water is raised by one degree Fahrenheit. The specific heat of water is 4187 Joules per kilogram per degree centigrade. Convert the result to British thermal units.

$$\begin{aligned}Q &= \frac{1 \text{ gal}}{g} \left(\frac{8.344 \text{ lb}}{\text{gal}} \frac{1 \text{ kg} \times g}{2.20462 \text{ lb}} \right) \\&\quad \times \left(4187 \frac{\text{J}}{\text{kg} \cdot 1^\circ\text{C}} \right) \left(\frac{5}{9} \times 1^\circ\text{C} \right) \\&= 8803.817 \text{ J} \\&= \frac{8803.817}{1054} \text{ Btu} = 8.352768 \text{ Btu}\end{aligned}$$

2 Scientific Notation

Consider the list of real numbers:

$$\begin{aligned} I &= 1 \\ J &= 10003 \\ K &= 0.0001357 \\ L &= 31.4159 \\ M &= 28318530718 \end{aligned}$$

While each is a valid real number, it would be neater to use a unified notation for each of these. For this we employ *scientific notation*, in where any number N is represented by:

$$N = A \times 10^B$$

Precision and Magnitude

Any number represented in scientific notation consists of a ‘precision number’ $|A| < 10$, and an order-of-ten ‘magnitude’ number B . Both A and B are themselves real numbers.

In this convention, the above list of numbers reads:

$$\begin{aligned} I &= 1 \times 10^0 \\ J &= 1.0003 \times 10^4 \\ K &= 1.357 \times 10^{-4} \\ L &= 3.14159 \times 10^1 \\ M &= 2.8318530718 \times 10^{10} \end{aligned}$$

In this form, one can quickly read off the magnitude of the number from the exponent attached to 10. The rest of the information carries the precision.

Large Numbers and Scientific Notation

Converting a standard fraction or decimal to scientific notation is a straightforward task. For exotically-large numbers (or small) numbers, some care must be taken.

To work an example, let us convert the number the number

$$N = 55^{99}$$

to scientific notation, which means to find A and B such that

$$55^{99} = A \times 10^B.$$

To proceed, recast the variable A via $A = 10^C$, where because $|A| < 10$, it follows that C obeys $0 < C < 1$:

$$55^{99} = 10^{B+C}$$

Apply the $\log_{10}$ operator to each side and simplify to write

$$99 \times \log_{10}(55) = B + C.$$

Supposing B and C are positive, it turns out that the ‘floor’ operator can be used to tease the values apart. In particular, we must have

$$B = \text{floor}(B + C) = \text{floor}(99 \times \log_{10}(55)),$$

and subsequently

$$C = 99 \times \log_{10}(55) - B.$$

For the example on hand, we use a machine or lookup table to find, approximately,

$$99 \times \log_{10}(55) \approx 172.296,$$

and thus:

$$\begin{aligned} B &= 172 \\ C &\approx 0.296 \end{aligned}$$

From C we calculate $A \approx 10^{0.296} \approx 1.977$, and the work is done:

$$N \approx 1.977 \times 10^{172}$$

Despite the approximate nature of the result, the order of magnitude B is exact and the precision C can be made finer by stressing the calculator that produced it.

Problem 31

Convert the number

$$X = 5^{256}$$

to scientific notation. Answer:

$$\log_{10}(X) = \log_{10}(5^{256}) = 256 \log_{10}(5) \approx 178.936$$

$$X \approx 10^{0.936} \times 10^{178} \approx 8.63 \times 10^{178}$$

Factorials and Scientific Notation

The factorial operator, which to remind, occurs as

$$N! = N(N-1)(N-2) \cdots (2)(1),$$

gives very large results for increasing N , and often such numbers can outsize the ability of a standard calculator.

To work with very large factorials, it helps to attain $N!$ in scientific notation. To proceed, apply the $\log_{10}$ operator to both sides of the above to write

$$\begin{aligned} \log_{10}(N!) &= \log_{10}(N) + \log_{10}(N-1) \\ &\quad + \log_{10}(N-2) + \cdots + \log_{10}(2) + 0 \end{aligned}$$

Sigma Notation

To contain the large sum on the right, we employ sigma notation as follows:

$$\log_{10}(N!) = \sum_{j=0}^{N-1} \log_{10}(N-j) = M$$

Rather than deal with $N!$ as a whole, the work turns into adding up the base-ten logarithm values for all natural numbers up to N . The sum of this effort is stored in the variable M . With a value for M on hand, use the exponent operator to solve for $N!$, particularly

$$N! = 10^M.$$

Converting to scientific notation, we write, as previously,

$$N! = 10^M = A \times 10^B = 10^{B+C},$$

where $|A| < 10$ and thus $|C| < 1$. Moreover, we have

$$M = B + C,$$

which means we simply look at M and assign the whole number part to B and the decimal to C .

Problem 32

Convert the number $X = 1000!$ to scientific notation.

Answer:

Step 1: Using a computer or calculator, compute:

$$M = \sum_{j=0}^{1000-1} \log_{10}(1000-j) = 2567.604644$$

Step 2: Identify

$$B = 2567$$

$$C = 0.604644$$

Step 3: Assemble the result:

$$1000! \approx 10^{0.604644} \times 10^{2567} \approx 4.02387 \times 10^{2567}$$

2.1 Order of Magnitude

The advantage to scientific notation is we don't have to waste time writing zeros. Many of the powers of ten have a special name, listed below from small to large. Common entries are emphasized in bold.

Power	Prefix	Symbol
10^{-24}	yocto	y
10^{-21}	zepto	z
10^{-18}	atto	a
10^{-15}	femto	f
10^{-12}	pico	p
10^{-9}	nano	n
10^{-6}	micro	u
10^{-3}	milli	μ
10^{-2}	centi	c
10^{-1}	deci	m
Power	Prefix	Symbol
10^3	kilo	k
10^6	mega	M
10^9	giga	G
10^{12}	tera	T
10^{15}	peta	P
10^{18}	exa	E
10^{21}	zetta	Z
10^{24}	yotta	Y

One can easily see the utility of scientific notation in practice. It would be a great burden to write, for instance, thirty extra zeros whenever we want to calculate with the solar mass. Following is a brief tour through the magnitude of typical quantities used in the sciences.

Order	Example
10^{-30}	Mass of an electron in kilograms.
10^{-27}	Mass of a hydrogen atom in kilograms.
10^{-15}	Diameter of a hydrogen atom in meters.
10^{-11}	Radius of a hydrogen atom in meters.
10^{-9}	Volume of a flea in cubic meters.
10^{-6}	Lifetime of a muon particle in seconds.
10^{-3}	Oscillation period of a guitar string in seconds.
10^0	Height of a human child in meters.
10^3	Mass of a car in kilograms.
10^6	Radius of Earth in meters.
10^8	Earth-Moon distance in meters.
10^9	Duration of a century in seconds.
10^{12}	Number of cells in human body.
10^{16}	Radius of the solar system in meters.
10^{19}	Number of Rubik's Cube positions.
10^{20}	Number of sand grains on Earth.
10^{24}	Number of atoms in twelve grams of carbon-12.
10^{27}	Mass of Jupiter in kilograms.
10^{30}	Mass of the Sun in kilograms.
10^{80}	Number of particles in the universe.

Problem 33

The thickness of paper is closest to:

- 10^{-4} m
- 10^{-1} m
- 10^1 cm
- 10^{-7} m

Suppose a very wide sheet of paper has thickness 0.1 mm. If the paper is folded once, the thickness doubles to 0.2 mm. If the resulting sheet is folded again, the total thickness doubles again to 0.4 mm. A third fold makes 0.8 mm, a fourth fold makes 1.6 mm, and so on. How thick is the paper after 20 folds? Answer:

$$T = 2^{20} \times 0.1 \text{ mm} \approx 104.9 \text{ m}$$

Problem 34

The length of an American football field (100 yards) is closest to:

- 10^4 m
- 10^{-1} km
- 10^5 mm
- 10^3 foot

Problem 37

If a very wide sheet of paper with thickness 0.1 mm is folded N times, determine N such that the resultant thickness is the radius of the solar system, approximately 10^{16} m. Answer:

$$N = \log_2 \left(\frac{10^{16} \text{ m}}{0.1 \text{ mm}} \right) \approx 63$$

Problem 35

The radius of Earth is about 10^6 meters. If the volume of a sphere is given by $V = (4/3)\pi R^3$, roughly estimate Earth's volume in cubic meters. Answer: $\approx 10^{18} \text{ m}^3$

Lightyears

The speed of light as defined (or measured) in meters per second is equal to

$$c = 2.99792498 \times 10^8 \text{ meter/sec ,}$$

which is close enough to

$$c \approx 3.00 \times 10^8 \text{ m/s ,}$$

Folding Paper**Problem 36**

much easier to remember. A lightyear is defined as the *distance* that light will travel in one year. With this, it follows that one lightyear is

$$1 \text{ lightyear} = c \times 1 \text{ year}.$$

We can calculate the length of a lightyear in meters by unit conversion. Start with

$$1 \text{ lightyear} = c \times 1 \text{ year} \left(\frac{365.25 \text{ day}}{1 \text{ year}} \right) \left(\frac{24 \text{ hr}}{1 \text{ day}} \right),$$

yielding the intermediate step

$$1 \text{ lightyear} = c \times 365.25 \times 24 \text{ hr}.$$

The units on the right need to be seconds to cleanly multiply by c , so keep going:

$$\begin{aligned} 1 \text{ lightyear} &= c \times 365.25 \times 24 \text{ hr} \left(\frac{3600 \text{ sec}}{1 \text{ hr}} \right) \\ &= c \times 365.25 \times 24 \times 3600 \text{ sec} \end{aligned}$$

Finally, we have

$$\begin{aligned} 1 \text{ lightyear} &= 2.9979 \times 10^8 \times 31557600 \frac{\text{sec}}{\text{sec}} \text{ m} \\ &= 9.4607 \times 10^{15} \text{ m} \end{aligned}$$

Problem 38

Calculate the lightyear distance in miles.

Problem 39

Express the speed of light in miles per second and miles per hour.

Answer: $1.86 \times 10^5 \text{ mi/sec}$, $6.71 \times 10^8 \text{ mi/hr}$

Problem 40

Convert the Massachusetts highway speed limit of 65 mph into lightyears per second.

Answer: $3.071 \times 10^{-15} \text{ lightyear/sec}$

2.2 Numerical Precision

In any laboratory, engineering, or other real-world setting, a number's precision is subject to the methods or apparatus used to attain that number. For instance, a length measured by a ruler might be precise to one millimeter, which means two similar things: (i) Any object smaller than one millimeter won't be 'visible' to the ruler. (ii)

Any length measured by the millimeter-precision ruler cannot be known to greater precision than 1 mm. For instance, suppose your lab assistant measures his shoelace with the above-mentioned ruler and reports the result $L = 26.5912 \text{ cm}$. You should immediately round this to $L = 26.6 \text{ cm}$, because the 'junk' digits 912 occupy decimal places that are smaller than the precision of the ruler.

A way to justify throwing away the 'junk' digits is to consider what happens when the measurement is repeated. While it's likely that the length L will be close to 26.6 cm again, it's unlikely that the digits 912 would follow a second time. In fact, we aren't confident that the shoelace really is 26.6 cm until the measurement is repeated.

Significant Figures

The precision in a number is represented by how many decimal places are *confidently* known, where each decimal place is called a *significant figure* or *significant digit*. Supposing a quantity is handed to us *without* any junk digits, there are several rules for counting the number of significant figures. These are:

1. Zeros to the left of the number are not significant.
2. Zeros to the right of the number are significant.
3. All nonzero digits in the number are significant.
4. Any zeros between significant digits are significant.

The following table lists a few numbers with varying significant figures as expressed in scientific notation:

Number	Sig. Figs.	Sci. Not.
7	1	7×10^0
10^{-3}	1	1×10^{-3}
3.0×10^8	2	3.0×10^8
000.0000010	2	1.0×10^{-6}
2.998×10^8	4	2.998×10^8
0623.53	5	6.2353×10^2
12.340	5	1.2340×10^1
90000	5	9.0000×10^4
90000.00	7	9.000000×10^4

Strategy

Sometimes the precision of a number is not obvious as given. If someone says they earned \$4500 last month, does this precisely mean \$4500.00 dollars and zero cents, or was \$4500 rounded from some nearby number? Should we trust only the 4 and the 5? This issue is resolved by the

following convention: *When the precision is ambiguous, do intermediate calculations to four significant figures, and report answers to three significant figures.* For the example on hand, we can agree that an honest representation of \$4500 is $\$4.50 \times 10^3$.

Notice that purely mathematical factors like 2 or $\pi = 3.14159265358979323846264 \dots$ or $2/3 = 0.66\bar{6}$ are more-or-less infinitely precise numbers, causing no penalty in the overall precision of a calculation.

2.3 Propagation of Precision

When two numbers of finite precision are combined, the natural question arises: what's the precision of the result? The answer depends on the operation being performed, but there are two main ways to proceed.

Multiplication, Division, Roots, Powers

Be pessimistic: In your calculation, seek the quantity with the lowest precision, and adopt *that* precision for the final answer, as illustrated in the following examples:

$$\begin{aligned} 2.345 \times 5.4 &\approx 13 \\ 55555 \times 0.0100 &\approx 556 \\ \sqrt{\pi \times 2.56/5.567} &\approx 1.20 \\ 70 \times 23.846251 &\approx 1.7 \times 10^3 \\ \left((2/3) \times (6.984 \times 10^{-8}) \right)^2 &\approx 2.170 \times 10^{-15} \end{aligned}$$

Addition and Subtraction

If the quantities being added or subtracted do not involve any decimals, you're off the hook, and may proceed as usual. When there *are* decimals though, be pessimistic again: the number of decimal places carried by the answer is determined by the contributing factor with the *least* number of decimal places. See the examples that follow.

$$\begin{aligned} 12551 + 3114 &= 15665 \\ 1000 - 3 &= 997 \\ 2.345 + 5.4 &\approx 7.7 \\ 100.2 + 15.438 &\approx 115.6 \\ 1000 - 0.3 &\approx 1000 \end{aligned}$$

3 Fundamental Units

3.1 Physical Units

Any physical quantity carries some kind of dimensionality, i.e. physical units. Whether the quantity represents duration in years, force in pounds, distance in lightyears, etc., the associated physical units are always reducible to some combination of (i) length, (ii) mass, and (iii) time. These are called the *fundamental dimensions*.

Following are several combinations of the fundamental dimensions linked with their colloquial names:

Combination	Label
L^2	Area
L^3	Volume
L/T	Speed
L/T^2	Acceleration
$M \cdot L/T$	Momentum
$M \cdot L/T^2$	Force
M/L^3	Mass Density
$M/(LT^2)$	Pressure
$M \cdot L^2/T^2$	Energy

SI System

The abbreviation *SI* stands for *Système International*, more commonly known as the *metric system*. There are two factions within the metric system, one called *MKS*, and the other called *CGS*. The letters *M-K-S* tell us to use meters, kilograms, seconds as fundamental units. Similarly, *C-G-S* uses centimeters, grams, seconds.

Imperial System

As of 2024, three countries have resisted the metric system's campaign for domination: Liberia, Myanmar, and the United States. Due to this, we still accommodate the *English system*, whose fundamental dimensions are (i) the foot, (ii) the slug, and (iii) seconds.

Expensive Mistake by NASA

Forgetting to properly handle units can be an expensive mistake, namely \$328 million when NASA lost the Mars Climate Orbiter. In September 1999, NASA attempted to put a satellite close enough to Mars to study its atmosphere and surface features. However, a ground-based computer in communication with the Orbiter was sending information using the *English* system of units! Specifically, the Orbiter expected a packet of information expressed in Newtons,

but it was interpreted as Pounds. The Orbiter didn't steer correctly and slammed into Mars.

3.2 Standard Units

Standard of Length

Length standards trace back to ancient history, with one popular unit of measurement being the the elbow-to-fingertip distance on a human arm in an L-shape, known as the 'cubit'. Obviously, the cubit varies from person to person, and more drastically between separate populations.

From Europe we gained another measure of length, the familiar 'foot', which was once defined by the anatomy of the prevailing royalty. It wasn't until the days of Louis XIV that the foot was standardized in France. The name was upgraded to the 'Royal Foot'.

Around 1795, the standard length unit was decidedly the *meter*, and was defined as 1/20000000 the length of Earth's prime meridian (this was before orthodox geology was developed).

By the late 1800's, the standard length unit was determined by the length of one carefully protected platinum-iridium bar, whose length was simply declared to equal precisely one meter. Precise copies of this bar were produced and shipped off to spread the length standard around the world.

Eventually, experimental physics would demand a more precise definition for the length standard that was robust with respect to special relativity, which inevitably involves the speed of light. *The meter is precisely defined as the distance light travels in 1/299792458 sec in vacuum.*

Standard of Mass

The modern definition of the Kilogram was settled in 2019. Before the decision to uproot tradition, the *kilogramme* was a another carefully-protected chunk of metal housed in France made of 90% platinum, 10% iridium, whose mass was decidedly one kilogram by popular vote. It's likely that the calibration of all science-grade mass scales and standards of the time could be traced back to the *kilogramme*.

There was curious problem with the *kilogramme* though: it's mass has slowly *increased* over history, and much effort has been devoted to finding out why this appears so. With or without a perfect explanation for this, the conclusion we must draw is that a chunk of metal is ultimately *not* a good mass standard.

The General Conference on Weights and Measures (CGPM) tried repeatedly to settle a definition of the kilo-

gram that arises from quantum mechanics, specifically involving the Planck mass. They couldn't decide in 2011, and procrastinated again in 2014. Eventually, laboratory measurements of Planck's constant became sufficiently refined (about 13 ppb uncertainty) such that the definition of one kilogram is simply back-read out of the result.

Standard of Time

One reasonable definition for the *second* is to slice a 24-hour interval into $60 \cdot 60 \cdot 24$ parts. This was considered reliable until 1967, until it was decided that the fundamental unit of time had better take on a less 'Earthly' definition. It was then decided that one second would be defined as precisely 9192631770 periods of the light waves emitted from cesium decay.

There was nothing special about cesium except it was an available atom for study. There's nothing terribly special about the number 9192631770, except it seemed good enough for reproducing the 'old' value of one second. Going through the trouble to define the second this way allows physicists to deal more easily with many practical issues, such as adding leap-seconds to the calendar, and calibrating clocks far from Earth. (The latter was on their minds in the 1960's more so than it is today.)

3.3 Elemental Mass Units

Avogadro's Number

Certain problems in thermodynamics, chemistry, and quantum mechanics require calculating with a large number of particles, atoms, or molecules. To this end, we employ a quantity like a 'chemists dozen', called a *mole*. One mole of any material is equal to 6.022×10^{23} particles, known as *Avogadro's Number*, denoted N_A :

$$N_A = 6.022 \times 10^{23}$$

Mass Units

By definition, one mole of carbon-12 contains N_A atoms and weighs precisely 12 g. From this we define the *unified atomic mass unit*, a single number used to approximate both the proton mass and the neutron mass having value 1.660539×10^{-24} g. In *physical units*, the unified amu is roughly 0.9928 proton masses or 0.9914 neutron masses. The *molecular mass* is defined as the number of *atomic mass units* (masses of protons and neutrons) in an element or molecule, ignoring electron masses.

Elemental masses are not reported as whole numbers in the periodic table. This arises from averaging over various *isotopes* of a given element occurring in nature, whose number of neutrons vary from the number in the pure element.

For example, 99.2% of hydrogen atoms observed in nature contain one proton and zero neutrons, denoted ${}^1_1\text{H}$. Meanwhile, 0.8% of hydrogen occurs with one proton and one neutron, denoted ${}^2_1\text{H}$. Statistically 0% of hydrogen contains two neutrons per atom, meaning ${}^3_1\text{H}$ is exceedingly rare. The weighted average of the respective masses per isotope is therefore

$$(0.992) 1 \text{ amu} + (0.008) 2 \text{ amu} + 0 = 1.008 \text{ amu} ,$$

matching the reported molecular mass of hydrogen, and the same goes for all other elements in the periodic table.

Problem 41

An H_2O molecule consists of 10 protons, 8 neutrons, and 10 electrons. Use the periodic table values $m_H = 1.008$ and $m_O = 15.999$ to compute the mass of a single water molecule in grams. Answer: 2.99×10^{-23} g

3.4 Units in Quantum Mechanics

Fundamental constants in quantum theory can be expressed in terms of Planck's length (L), Planck mass (M), and Planck time (T). Particularly, one can state:

- The speed of light c in units LT^{-1} .
- The electron mass m in units of M .
- Planck's reduced constant $\hbar = h/2\pi$ in units ML^2T^{-1} .
- The squared electron charge $e^2/4\pi\epsilon_0$ in units ML^3T^{-2} .

Reduced Compton Wavelength

Consider three unknown integers x, y, z such that the quantity

$$\lambda_c = \hbar^x c^y m^z$$

has units of length. From dimensional analysis, we generate three equations

$$x + z = 0$$

$$2x + y = 1$$

$$x + y = 0 ,$$

solved by $x = 1$ with $y = z = -1$. Such a length is called the *reduced Compton wavelength*:

$$\lambda_C = \frac{\hbar}{mc}$$

Fine Structure Constant

For the proper choice of x, y, z , the combination

$$\alpha = \hbar^x c^y \left(e^2 / 4\pi\epsilon_0 \right)^z$$

can be made dimensionless. From dimensional analysis, we generate three equations

$$\begin{aligned} x + z &= 0 \\ -x - y - 2z &= 0 \\ 2x + y + 3z &= 0, \end{aligned}$$

solved by $x = y = -1, z = 1$. The result is called the *fine structure constant*:

$$\alpha = \frac{e^2 / 4\pi\epsilon_0}{\hbar c} \approx \frac{1}{137}$$

Bohr Radius

The scale of a hydrogen atom is approximated by the so-called *Bohr radius*, denoted a_0 , which is the ratio of the reduced Compton wavelength to the fine structure constant:

$$a_0 = \frac{\lambda_C}{\alpha} = \frac{\hbar^2}{m (e^2 / 4\pi\epsilon_0)} \approx 137\lambda_C \approx 5.29 \times 10^{-11} \text{ m}$$

Chapter 3

Cartesian Plane

1 Graphing

In mathematics, a *graph* is a visual or otherwise sensory representation of data. Data can occur in the forum of numbers, relationships, flowchart components, etc. It follows that ‘graphing’ mathematical data can mean a variety of things.

1.1 One-Dimensional Graph

The simplest non-trivial graph is facilitated by the standard number line as shown in Figure 3.1. For brevity there is no attempt to draw the entire number line. Any portion of the number line can be used as the basis for a graph.

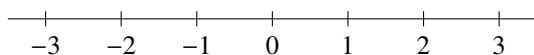


Figure 3.1. Number line.

To make use of the number line as a graph, we can consider the solutions to any equation of one variable. For simplicity, suppose we have

$$x^2 = 4.$$

Knowing the above is satisfied by $x = 2$ and also $x = -2$, these solutions can be marked, or *plotted* at their respective locations on the number line as shown in Figure 3.2.

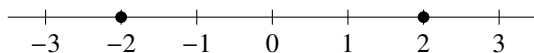


Figure 3.2. Number line with solutions to $x^2 = 4$.

A more informative name for ‘one dimensional number line’ is the *x-axis*, particularly when a variable named x is to be represented on the line.

Analog Clock Face

A one-dimensional number line that repeats itself is called *periodic*, as exhibited by a standard clock face in Figure 3.3.

Despite being arranged on a two-dimensional surface, the analog clock is a one-dimensional graph of the time being displayed.

When the clock hand sweeps through increasing time, the motion is called *clockwise*. The clock running backwards is going in the *counterclockwise*, or *anticlockwise* direction.

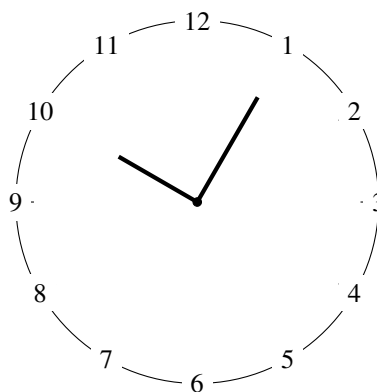


Figure 3.3. Analog clock is a one-dimensional graph.

1.2 Two-Dimensional Graph

We’ve witnessed that solutions to single-variable systems can be represented on a one-dimensional graph. It should follow that a visual representation of a systems with two independent variables need an extra ‘degree of freedom’, meaning a two-dimensional graph must be considered.

For example, suppose you're given the information

$$\begin{aligned}x &= 2 \\ y &= -2.\end{aligned}$$

One way to visualize this involves two number lines, one for the x -axis and another for the y -axis, marked accordingly as in Figures 3.4, 3.5.

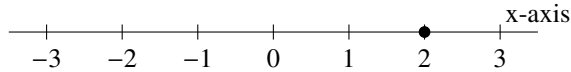


Figure 3.4. $x = 2$.

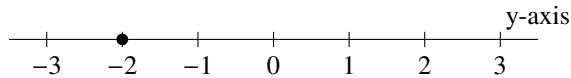


Figure 3.5. $y = -2$.

It would be more efficient, though, if the same solution can be represented in one picture instead of two. Since the x -axis and y -axis are independent, the two can be arranged to form a plane as shown in Figure 3.6.

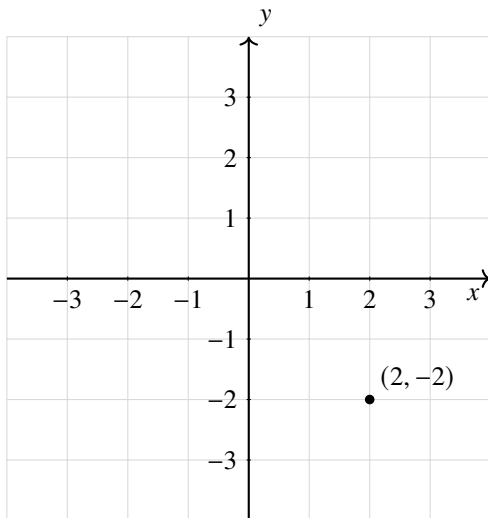


Figure 3.6. A single point representing $x = 2$, $y = -2$.

In constructing a two-dimensional graph, it is by popular convention that the point $x = 0$ overlaps with the point $y = 0$. One could easily imagine laying the y -axis across the x -axis at other locations, but it's not often necessary.

What *is* necessary, however, is that the x -axis and y -axis meet at a right angle. This way, contours of constant x span vertically, and contours of constant y span horizontally.

1.3 Ordered Pairs

If two quantities x , y represent a location on a two-dimensional graph, then the item

$$(x, y)$$

is called an *ordered pair*. The order in which x and y occur *does* matter, meaning (y, x) is different than (x, y) .

To demonstrate, suppose there is a need to visualize the equation

$$x^2 + y = 3$$

in two dimensions. Proceed by choosing a sequence of x -values labeled x_1, x_2 , etc., such that

$$\begin{aligned}x_1 &= -2 \\ x_2 &= -1 \\ x_3 &= 0 \\ x_4 &= 1 \\ x_5 &= 2.\end{aligned}$$

Of course, the choices $-2, -1$, etc. are arbitrary. From these, generate the corresponding y -values that satisfy the equation given. Particularly:

$$\begin{aligned}y_1 &= 3 - (-2)^2 = -1 \\ y_2 &= 3 - (-1)^2 = 2 \\ y_3 &= 3 - 0^2 = 3 \\ y_4 &= 3 - (1)^2 = 2 \\ y_5 &= 3 - (2)^2 = -1\end{aligned}$$

Matching x_1 to y_1 and so on, we generate five ordered pairs:

$$(-2, -1), (-1, 2), (0, 3), (1, 2), (2, -1)$$

Plotting each of these on the same two-dimensional graph gives rise to Figure 3.7.

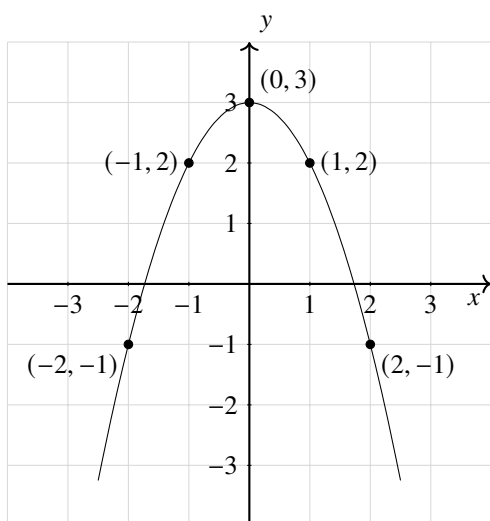


Figure 3.7. Ordered pairs representing $x^2 + y = 3$.

Supposing we sample many more x -values to generate more ordered pairs, it should be possible to fill in enough points on the graph so as to form a convincing 'curve'. Superimposed onto Figure 3.7 is a curve that represents doing such an effort.

Problem 1

Consider the equation

$$4y - 3x + 2 = 0.$$

On a two-dimensional graph, find two ordered-pair solutions (x_1, y_1) , (x_2, y_2) , draw a straight line through each point on the graph. Find a third solution (x_3, y_3) and check that it occurs along the same line.

1.4 Graphical Solutions

Two-dimensional graphs can be used to hunt for solutions to otherwise very hard problems.

Consider the notorious transcendental equation

$$x^2 = 2^x,$$

and the job is to find solutions for x . To begin we let each side of the equation equal a variable y such that:

$$y_1 = x^2$$

$$y_2 = 2^x$$

Then, choosing a flurry of x -values to the left and right of $x = 0$, we can generate the following table:

x	$y_1 = x^2$	$y_2 = 2^x$
-2	4	0.25
-1.5	2.25	0.3536
-1	1	0.5
-0.5	0.25	0.7071
0	0	1
0.5	0.25	1.4142
1	1	2
1.5	2.25	2.8284
2	4	4
2.5	6.25	5.6569

From the table, we construct the ordered pairs $(-2, 4)$, $(-2, 0.25)$, $(-1.5, 2.25)$, etc. Letting a machine calculate and plot many such ordered pairs gives rise to Figure 3.8.

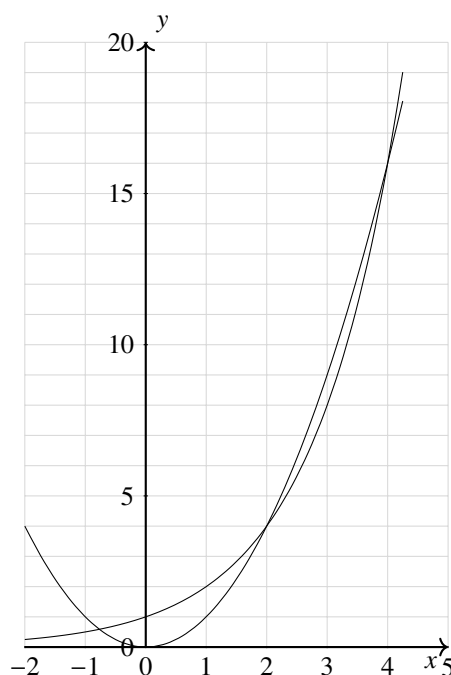


Figure 3.8. Graphical solution to $x^2 = 2^x$.

Right away, we see from the table or the Figure that the solution $x = 2$ readily satisfies $x^2 = 2^x$. Slightly less obvious is the solution at $x = 4$, because $4^2 = 2^4$ is a valid statement.

The least conspicuous solution is implicated somewhere to the left of the y -axis where the two curves intersect. In Figure 3.9 we get a closer look at the window

$$-1 \leq x \leq 0.5$$

$$-0.5 \leq y \leq 1$$

as shown.

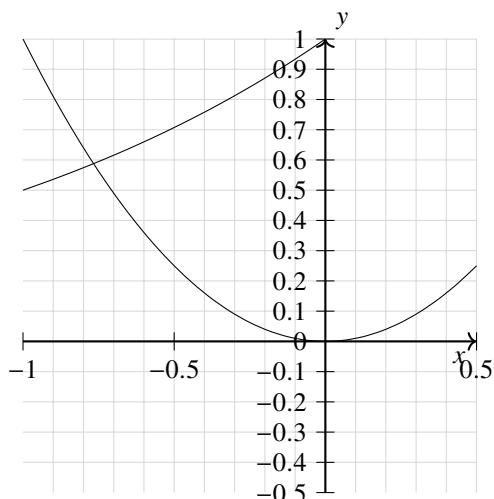


Figure 3.9. Graphical solution to $x^2 = 2^x$, reduced window.

By re-centering and re-zooming the graph, we can ‘hone in’ on any point we wish, and thus another solution at

$$x \approx -0.766,$$

give or take, becomes evident. For a test, we can use a calculator to establish

$$(-0.766)^2 - 2^{(-0.766)} = -0.00129$$

in hopes that the right side comes out as close to zero as possible.

Problem 2

Consider the system of equations

$$\begin{aligned} x^2 + y &= 21 \\ xy &= 20. \end{aligned}$$

By plotting ordered-pair solutions to each equation, show that the two solutions to the system are $(1, 20)$, $(4, 5)$.

Problem 3

For the system

$$\begin{aligned} x + y - 5 &= 0 \\ y - 2x + 10 &= 0, \end{aligned}$$

use graphical methods to determine the only ordered-pair solution. Answer: $(5, 0)$

Problem 4

Consider the equation

$$x^2 = x + 6.$$

Convert this to a system of equations

$$\begin{aligned} y_1 &= x^2 \\ y_2 &= x + 6, \end{aligned}$$

and use graphical methods to find two values for x that satisfy the system. Answer: $x = -2$ and $x = 3$

1.5 Cartesian Plane

Formally, the two-dimensional graphing space in which the x -axis is perpendicular to the y -axis is called the *Cartesian plane*.

Origin

Like the x - and y -number lines, the Cartesian plane spans infinitely in the horizontal and vertical directions so that every possible ordered pair (x, y) of real numbers has a unique ‘home’ in the plane. The point $(0, 0)$ is called the *origin*.

Quadrants

The rest of the Cartesian plane is divided into four *quadrants*, also known as *quarter-planes*, depending on whether x , y are positive or negative numbers as shown in Figure 3.10.

On a technical note, each quadrant has infinite size. The entire Cartesian plane has infinite size, but it’s a ‘larger’ infinity.

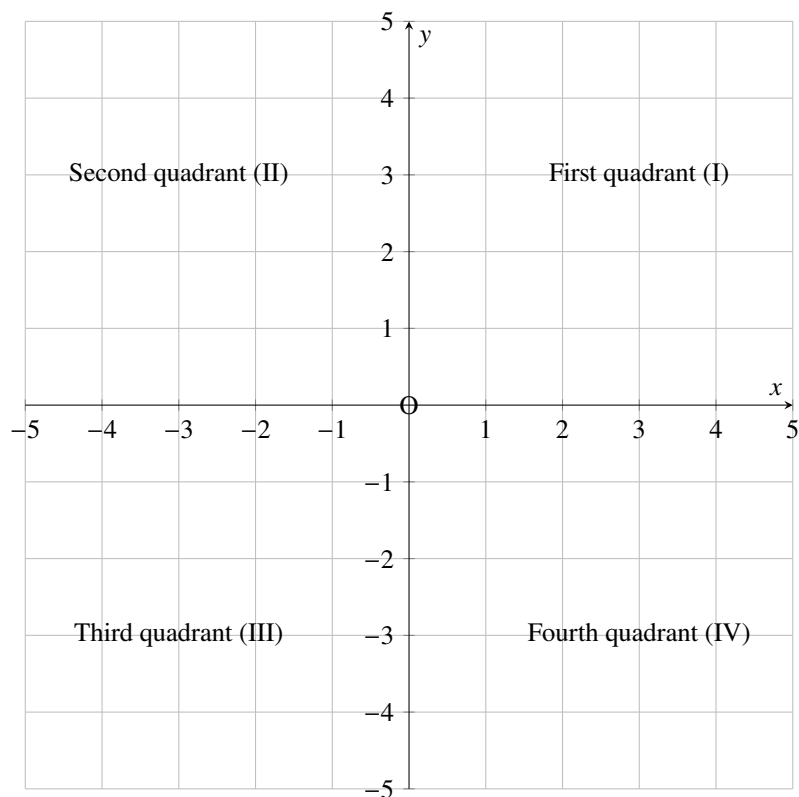


Figure 3.10. Cartesian Plane.

2 Straight Lines

Linear Equation

Any equation of the form

$$Ax + By + C = 0 \quad (3.1)$$

is a *linear equation*. By ‘linear’, we mean that x and y only occur to the first power. The presence of terms like x^2 or $\sqrt{y}$ would make the equation nonlinear.

In the Cartesian plane, a linear equation corresponds to a straight line, easily verified by plotting ordered pairs (x, y) that satisfy the equation. Each coefficient A , B , C , each being any real number, dictates the position and orientation of the line.

Intercepts

The special point $(x_{\text{int}}, 0)$ is called the x -intercept, which corresponds to where a straight line crosses the x -axis. An exception occurs when the line is purely horizontal.

Correspondingly, the special point $(0, y_{\text{int}})$ is called the y -intercept, which is where a straight line crosses the y -axis. An exception occurs when the line is purely vertical.

From Equation (3.1), one readily solves for the x - and y -intercepts:

$$x_{\text{int}} = \frac{-C}{A}$$

$$y_{\text{int}} = \frac{-C}{B}$$

Horizontal and Vertical Lines

Looking at each denominator in the above, we see that x_0 only exists if $A \neq 0$. Similarly, y_0 exists only if $B \neq 0$. This makes sense in light of Equation (3.1), for if $A = 0$ then y is always equal to y_0 to make a horizontal line. Likewise, if $B = 0$ then x is always equal to x_0 to make a vertical line.

2.1 Slope

From a linear equation, one may generate two ordered-pair solutions (x_1, y_1) , (x_2, y_2) so long as $x_1 \neq x_2$ and $y_1 \neq y_2$. That is, we want a pair of *unique* points in the Cartesian plane as that satisfy Equation (3.1). Doing so gives

$$\begin{aligned} Ax_1 + By_1 + C &= 0 \\ Ax_2 + By_2 + C &= 0. \end{aligned}$$

Given the points (x_1, y_1) , (x_2, y_2) , the difference $x_2 - x_1$ is called the *run*, and is written:

$$\Delta x = x_2 - x_1 = \text{run} \quad (3.2)$$

The triangular Greek symbol Δ is called ‘delta’. Similarly, the difference $y_2 - y_1$ is called the *rise*:

$$\Delta y = y_2 - y_1 = \text{rise} \quad (3.3)$$

This construction is illustrated in Figure 3.11.

The ratio $\Delta y / \Delta x$ is called the *slope* of the line, denoted m :

$$m = \frac{\Delta y}{\Delta x} = \text{slope} \quad (3.4)$$

In terms of the coefficients in Equation (3.1), the slope comes out to

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{-A}{B}.$$

The special case $A = 0$ corresponds to a horizontal line, i.e. a line with zero rise, and thus zero slope. On the other hand, the case $B = 0$ causes division by zero, making the slope undefined. This corresponds to a vertical line, which has ‘infinite’ slope - all rise and no run.

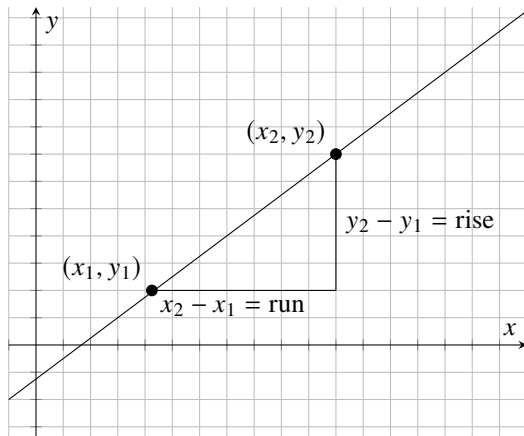


Figure 3.11. Rise and run.

Line through the Origin

In terms of the x - and y -intercepts, it's straightforward to show that the slope can be written

$$m = \frac{-y_{\text{int}}}{x_{\text{int}}}$$

for a straight line.

Any line that passes through the origin at $(0, 0)$ always has

$$x_{\text{int}} = y_{\text{int}} = 0,$$

which causes the slope m to be unreadable from the above. Such a line through the origin can have *any* slope as dictated by the ratio $-A/B$.

Point-Slope Form

To summarize, we've seen that any two points on a line given by Equation (3.1) can be used to calculate the slope.

The same line can be determined a slightly different configuration. Namely, if we start with one given point (x_0, y_0) on the line, along with the slope m , we can write an equation that accounts for the entire line. The run and the rise become

$$\begin{aligned} \Delta x &= x - x_0 \\ \Delta y &= y - y_0, \end{aligned}$$

and similar to Equation (3.4), we can write

$$m = \frac{y - y_0}{x - x_0}.$$

Rearranging a little, the *point-slope form* of the line emerges:

$$y - y_0 = m(x - x_0) \quad (3.5)$$

Slope-Intercept Form

Another useful equation for a straight line is the so-called *slope-intercept form*. Starting from Equation (3.1), solve for the y -variable to find

$$y = \frac{-A}{B}x - \frac{C}{B},$$

or equivalently

$$y = mx + y_{\text{int}}.$$

By convention, the y -intercept is given the shorthand label b such that

$$b = \frac{-C}{B},$$

and the finished equation takes the form:

$$y = mx + b \quad (3.6)$$

The slope-intercept form requires two pieces of information, namely m and b , to determine the entire straight line.

Perpendicular Lines

Suppose we take Figure 3.11 and rotate the page by ninety degrees counterclockwise so that the y -axis now points along $-x$, and similarly the x -axis now points vertically. Any line $y = mx + b$ drawn on the unrotated page snaps to a new orientation with different x - and y -intercepts. Similar comments apply if we rotate the page clockwise instead.

Whether we rotate the page clockwise or counterclockwise, the slope m of a line takes on a new value $m_{\perp}$, particularly

$$m_{\perp} = \frac{-\Delta x}{\Delta y} = \frac{-1}{m} .$$

Problems

Problem 5

Starting from Equation (3.5), derive the form

$$Ax + By = Ax_0 + By_0$$

and show that

$$C = -(Ax_0 + By_0) .$$

Problem 6

Determine the slope of a line connecting the points $(2, 3)$ and $(7, 4)$.

Problem 7

Write the point-slope equation of a line connecting the points $(-1, 10)$ and $(5, 2)$.

Problem 8

Write the slope-intercept equation of a line connecting the points $(-8, -8)$ and $(-7, 9)$.

Problem 9

Find the y -intercept of a line having $m = 2$ that passes through $(2.5, 0)$.

Problem 10

Write the equation of a line with an x -intercept of 3, and a y -intercept of 6.

3 Quadratic Curves

Quadratic Equation

Starting with the point-slope form a straight line in Equation (3.5), one may wonder what happens if $(x - x_0)$ is replaced by $(x - x_0)^2$. Right away we replace the slope term m with a different arbitrary coefficient a to write a *quadratic equation*:

$$y - y_0 = a(x - x_0)^2 \quad (3.7)$$

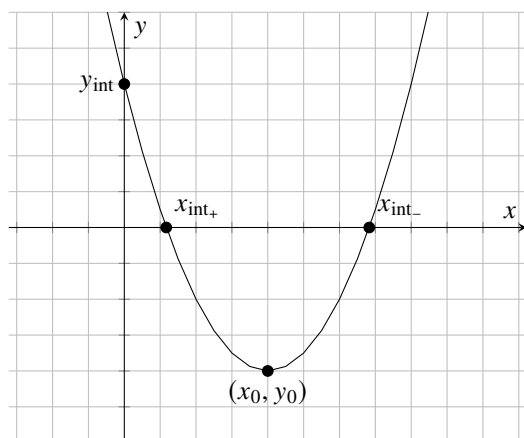


Figure 3.12. Quadratic curve.

Vertex and Symmetry

Figure 3.12 depicts a sketch of Equation (3.7) in the Cartesian plane. The overall U-shape of the curve is called a *parabola*. The point (x_0, y_0) is called the *vertex*, and the vertical line defined by $y = y_0$ is the axis of symmetry.

Concavity

When the parabola is ‘facing up’ as depicted, the curve is *concave up*. The entire curve flips upside-down (keeping the same vertex) if a is a negative number, and becomes *concave down*.

The magnitude of a determines the wideness or sharpness of the parabola. For smaller a , the distance between the x -intercepts widens and the whole parabola becomes flatter. The limit case $a = 0$ means y is always y_0 for a horizontal line. Conversely, the parabola becomes sharper for increasing a .

Intercepts

As written in Equation (3.7), a parabola always has a y -intercept given by

$$y_{\text{int}} = y_0 + ax_0^2.$$

As for x -intercepts, let $y = 0$ to find

$$x_{\text{int}} = x_0 \pm \sqrt{\frac{-y_0}{a}}.$$

Note there are up to two solutions for x_{int} , but no x -intercept exists when y_0/a is a negative number.

General Form

Equation (3.7) can be configured in general form. Multiply out $(x - x_0)^2$ and combine like terms to write

$$y = ax^2 - 2axx_0 + (y_0 + ax_0^2),$$

and the parenthesized term is familiar as y_{int} .

Seeking a general form for the quadratic curve, define three coefficients such that

$$A = a$$

$$B = -2ax_0$$

$$C = y_0 + ax_0^2 = y_{\text{int}},$$

so the quadratic curve is written

$$y = Ax^2 + Bx + C. \quad (3.8)$$

Starting from Equation (3.8), it’s handy to be able to get back to a, x_0, y_0 in terms of A, B, C . Solving this system is straightforwardly done via:

$$a = A$$

$$x_0 = \frac{-B}{2A}$$

$$y_0 = C - \frac{B^2}{4A}$$

3.1 Quadratic Formula

Writing the general form of the parabola in terms of a, x_0, y_0 , one has

$$y = \left(C - \frac{B^2}{4A}\right) + A\left(x + \frac{B}{2A}\right)^2.$$

With this, let us repeat the x -intercept calculation by setting $y = 0$. Doing so gives the famed *quadratic formula*:

$$x_{\text{int}_{\pm}} = \frac{-B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A} \quad (3.9)$$

From the above, we also see that the sum of intercepts obeys

$$x_{\text{int}_{+}} + x_{\text{int}_{-}} = \frac{-B}{A}$$

Product of Quadratic Intercepts

A seldom-utilized identity involves the product of the two x -intercepts $x_{\text{int}_{+}}$, $x_{\text{int}_{-}}$. Carrying this out, one finds:

$$x_{\text{int}_{+}} \cdot x_{\text{int}_{-}} = \left(\frac{-B}{2A} + \frac{\sqrt{B^2 - 4AC}}{2A} \right) \left(\frac{-B}{2A} - \frac{\sqrt{B^2 - 4AC}}{2A} \right)$$

This readily simplifies to:

$$x_{\text{int}_{+}} \cdot x_{\text{int}_{-}} = \frac{C}{A}$$

Alternate Quadratic Equation

Since each x -intercept is a point where the parabola has $y = 0$, we can also write

$$y = A(x - x_{\text{int}_{+}})(x - x_{\text{int}_{-}}) \quad (3.10)$$

Simplify the right side to get

$$\begin{aligned} y &= Ax^2 - A(x_{\text{int}_{+}} + x_{\text{int}_{-}})x + Ax_{\text{int}_{+}}x_{\text{int}_{-}} \\ &= Ax^2 - A\left(\frac{-B}{A}\right)x + A\frac{C}{A} \\ &= Ax^2 + Bx + C, \end{aligned}$$

recovering the general form.

Discriminant

The quantity $B^2 - 4AC$ is called the *discriminant*, denoted $\mathcal{D}$:

$$\mathcal{D} = B^2 - 4AC$$

For the x -intercepts of a quadratic equation be real-valued, the discriminant must be positive. The special case $B^2 = 4AC$ corresponds to there being one intercept, i.e. $x_{\text{int}_{+}} = x_{\text{int}_{-}}$. When the discriminant is negative, there are no x -intercepts.

3.2 Line Intersecting Parabola

Consider the following nonlinear system of two equations and two unknowns:

$$\begin{aligned} 4x - y &= 3 \\ 2x^2 + 3y &= 19 \end{aligned}$$

In the Cartesian plane, this system represents the intersection of a line and a parabola.

To solve for x or y analytically, we first combine the two equations to eliminate one of the variables. Solving for y in the first equations and substituting into the second, we find

$$x^2 + 6x = 14,$$

which is clearly a quadratic equation, having two solutions:

$$x_1 = -3 + \sqrt{23} \approx 1.796$$

$$x_2 = -3 - \sqrt{23} \approx -7.796$$

For a similar exercise, re-combine the equations in the system to eliminate x instead, yielding a quadratic equation

$$y^2 + 30y = 143,$$

which itself has two solutions:

$$y_1 = -15 + \sqrt{368} \approx 4.183$$

$$y_2 = -15 - \sqrt{368} \approx -34.183$$

From the information gained, we have four possible solutions to the system

$$(x_1, y_1)$$

$$(x_1, y_2)$$

$$(x_2, y_1)$$

$$(x_2, y_2),$$

however not all are valid. Each must be checked against the original equations, which is equivalent to checking with a graphical method. Doing so leads to Figure 3.13.

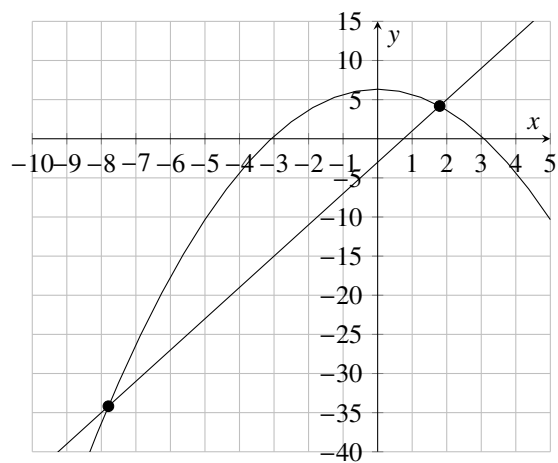


Figure 3.13. Line intersecting Parabola.

Evidently, intersections occur at $(1.7953, 4.1833)$ and $(-7.7958, -34.183)$, thus the valid solutions to the system are (x_1, y_1) and (x_2, y_2) .

Chapter 4

Factoring

1 Factoring Quadratics

1.1 Box Method

The most grotesque means of factoring is surely the *box method*, usually confined to pre-college pedagogy.

To factor the quadratic expression

$$2x^2 - x - 6,$$

first draw a box, and then put the first and third terms along the diagonal as shown:

$2x^2$	
	-6

Off to the side, multiply the first and third coefficients to get -12 . List every factorization of this number:

$$-12 = \begin{cases} \pm 1 \times \mp 12 \\ \pm 2 \times \mp 6 \\ \pm 3 \times \mp 4 \end{cases}$$

The pair of factors that sums to the expression's *middle* coefficient, namely -1 , is the pair we need. The box is then updated:

$2x^2$	$-4x$
$3x$	-6

Finally, look across each row and each column of the box. In the margin of each, write the greatest common divisor occurring in that row or column:

x	-2	
$2x^2$	$-4x$	$2x$
$3x$	-6	3

The final result is read from the margins:

$$2x^2 - x - 6 = (x - 2)(2x + 3)$$

The box method is popular for numerous reasons, one being that it's easy for an instructor to prepare factoring worksheets. Correctness can be judged based on whether the student writes the correct term in the correct box. Easy to fill in, easy to grade.

On the other hand, one immediately sees the departure from mathematical thinking inherent to the box method. Once the box is drawn, the student is asked to set aside the tools of algebraic procedure, giving way to something akin to spreadsheet heuristics.

Examples

For another complete example, the equation

$$6x^2 + 11x + 4 = (3x + 4)(2x + 1)$$

is encoded by:

$3x$	4	
$6x^2$	$8x$	$2x$
$3x$	4	1

Similarly, the equation

$$3x^2 - 2x - 8 = (x - 2)(3x + 4)$$

is encoded by:

x	-2	
$3x^2$	$-6x$	$3x$
$4x$	-8	4

1.2 Split-Term Method

A relatively quick means of factoring a quadratic expression involves splitting the middle term.

For the trinomial

$$2x^2 + 7x + 3,$$

first record the product of the two outer coefficients, namely $(2)(3) = 6$. The task becomes splitting the middle coefficient (7) into two numbers whose product is 6, which is straightforwardly done:

$$2x^2 + 7x + 3 = 2x^2 + \underbrace{6x + x}_{7x} + 3$$

We then proceed to write

$$\begin{aligned} 2x^2 + 7x + 3 &= 2x(x + 3) + (x + 3) \\ &= (2x + 1)(x + 3), \end{aligned}$$

and the problem is solved after some regrouping of terms.

The split-term method works well when it's easy to divine *how* the middle term should be split. Like most common methods, this can depend on a bit of luck, and works best when the coefficients involved are tame integers.

1.3 Completing the Square

Completing the square is a robust tool for factoring quadratic expressions.

Supposing we're given

$$x^2 + 6x + 2,$$

the first move is to look at the middle coefficient, namely 6, and divide that by two. This number becomes the term inside a 'square' factor as follows:

$$x^2 + 6x + 2 = (x + 3)^2 - 3^2 + 2$$

Note that a constant -3^2 is introduced to keep algebraic balance.

To proceed, seek solutions to

$$(x + 3)^2 - 3^2 + 2 = 0,$$

resulting in

$$\begin{aligned} x_1 &= -3 + \sqrt{7} \\ x_2 &= -3 - \sqrt{7}, \end{aligned}$$

so the final answer is written

$$x^2 + 6x + 2 = (x + 3 - \sqrt{7})(x + 3 + \sqrt{7}).$$

1.4 Algebraic Identities

The limit case of a lucky factoring problem is one that maps perfectly onto a known algebraic identity. Committing identities to memory such as

$$\begin{aligned} a^2 + 2ab + b^2 &= (a + b)^2 \\ a^2 - 2ab + b^2 &= (a - b)^2 \\ a^2 - b^2 &= (a + b)(a - b), \end{aligned}$$

quick work can be made of problems that fit them, i.e.:

$$\begin{aligned} x^2 + 10x + 25 &= (x + 5)^2 \\ x^2 - 6x + 9 &= (x - 3)^2 \\ 9x^2 - 16 &= (3x + 4)(3x - 4) \end{aligned}$$

1.5 Normalizing a Quadratic

Factoring a quadratic expression is undoubtedly simpler when the leading coefficient A doesn't complicate things, epitomized by $A = 1$. When this isn't the case, one can always factor A from the whole expression

$$Ax^2 + Bx + C = A \left(x^2 + \frac{B}{A}x + \frac{C}{A} \right),$$

and focus on the parenthesized quantity, proceeding as if there is no leading coefficient. Of course, the coefficients B and C are then modified by A , which can make the problem much less penetrable.

Alternatively, one can define a variable

$$w = x\sqrt{A}$$

such that

$$Ax^2 + Bx + C = w^2 + \frac{B}{\sqrt{A}}w + C.$$

Then, factor in the w -variable and restore x when finished.

Unit Leading Coefficient

When $A = 1$, the transformed quadratic is

$$w^2 - Bw + C = 0,$$

and the kernel-root link simplifies to

$$\{w_1, w_2\} = \{-x_1, -x_2\}$$

so $w = -x$ at the level of solutions.

Unit First and Third

In the special case that the first and third coefficients are equal to one, we have

$$x^2 + Bx + 1 = (x - x_1)(x - x_2) ,$$

implying

$$x_1 + x_2 = -B$$

$$x_1 \cdot x_2 = 1 .$$

With $x_1 \cdot x_2 = 1$ established, we can define a variable

$$z = \frac{1}{x}$$

and seek solutions to a modified expression

$$1 + \frac{B}{z} + z^2 = 0 ,$$

equivalent to

$$z^2 + Bz + 1 = 0 ,$$

exactly what we started with, up to a change of variables $x \rightarrow z$.

where $x_{1,2}$ is given by:

$$x_{1,2} = \frac{-B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A}$$

Furthermore, it's straightforward to show that

$$x_1 + x_2 = \frac{-B}{A}$$

and

$$x_1 \cdot x_2 = \frac{C}{A} .$$

The quadratic formula can be derived by completing the square on

$$Ax^2 + Bx + C = 0 .$$

The key intermediate step involves writing

$$\left(\sqrt{A}x + \frac{B}{2\sqrt{A}} \right)^2 = \frac{B^2}{4A} - C ,$$

which is straightforwardly solved for x .

1.6 Quadratic Formula

One of the surest ways to factor a quadratic equation is using the quadratic formula. For this, we have

$$Ax^2 + Bx + C = A(x - x_1)(x - x_2) ,$$

width=!,height=!, pages=-, addtotoc=1,section,1,Method
of Transform,sec:anon

Problems

Use any factoring method to show the following:

Problem 1

$$6x^2 + 11x + 4 = (2x + 1)(3x + 4)$$

Hint: $w_1 = 8, \quad w_2 = 3, \quad \alpha = 2, \quad \beta = 1$
 $w_1 = 3, \quad w_2 = 8, \quad \alpha = 3, \quad \beta = 4$

Problem 2

$$3x^2 - 2x - 8 = (3x + 4)(x - 2)$$

Hint: $w_1 = -6, \quad w_2 = 4, \quad \alpha = 3, \quad \beta = 4$
 $w_1 = 4, \quad w_2 = -6, \quad \alpha = 1, \quad \beta = -2$

Problem 3

$$x^2 + 3x - 28 = (x - 4)(x + 7)$$

Hint: $w = -x$

Problem 4

$$4x^2 + 2x - 12 = 2(2x - 3)(x + 2)$$

Hint: $d = 2$
 $w_1 = 4, \quad w_2 = -3, \quad \alpha = 2, \quad \beta = -3$
 $w_1 = -3, \quad w_2 = 4, \quad \alpha = 1, \quad \beta = 2$

Problem 5

$$36x^2 - 121y^2 = (6x + 11y)(6x - 11y)$$

Hint: $B = 0$

Problem 6

$$12x^2 + 11x - 15 = (4x - 3)(3x + 5)$$

Hint: $w_1 = 20, \quad w_2 = -9, \quad \alpha = 4, \quad \beta = -3$
 $w_1 = -9, \quad w_2 = 20, \quad \alpha = 3, \quad \beta = 5$

Problem 7

$$15x^2 + 14x - 8 = (5x - 2)(3x + 4)$$

Hint: $w_1 = 20, \quad w_2 = -6, \quad \alpha = 5, \quad \beta = -2$
 $w_1 = -6, \quad w_2 = 20, \quad \alpha = 3, \quad \beta = 4$

Problem 8

$$15x^2 + 44x + 21 = (5x + 3)(3x + 7)$$

Hint: $w_1 = 35, \quad w_2 = 9, \quad \alpha = 5, \quad \beta = 3$
 $w_1 = 9, \quad w_2 = 35, \quad \alpha = 3, \quad \beta = 7$

Problem 9

$$4x^2 + 15x + 9 = (4x + 3)(x + 3)$$

Hint: $w_1 = 12, \quad w_2 = 3, \quad \alpha = 4, \quad \beta = 3$
 $w_1 = 3, \quad w_2 = 12, \quad \alpha = 1, \quad \beta = 3$

Problem 10

$$6x^2 + 23x - 4 = (6x - 1)(x + 4)$$

Hint: $w_1 = 24, \quad w_2 = -1, \quad \alpha = 6, \quad \beta = -1$
 $w_1 = -1, \quad w_2 = 24, \quad \alpha = 1, \quad \beta = 4$

Problem 11

$$10x^2 + x - 3 = (2x - 1)(5x + 3)$$

Hint: $w_1 = 6, \quad w_2 = -5, \quad \alpha = 2, \quad \beta = -1$
 $w_1 = -5, \quad w_2 = 6, \quad \alpha = 5, \quad \beta = 3$

Problem 12

$$15x^2 - 7x - 4 = (3x + 1)(5x - 4)$$

Hint: $w_1 = -12, \quad w_2 = 5, \quad \alpha = 3, \quad \beta = 1$
 $w_1 = 5, \quad w_2 = -12, \quad \alpha = 5, \quad \beta = -4$

Problem 13

$$8x^2 - 2xy - 3y^2 = (2x + y)(4x - 3y)$$

Hint: $w_1 = -6y, \quad w_2 = 4y, \quad \alpha = 2, \quad \beta = 1y$
 $w_1 = 4y, \quad w_2 = -6y, \quad \alpha = 4, \quad \beta = -3y$

Problem 14

$$(x + 4)^3 - 9x - 36 = (x + 4)(x + 1)(x + 7)$$

Problem 15

$$x^2 + 2x + 1 - y^2 = (x + 1 + y)(x + 1 - y)$$

Problem 16

$$x^2 - y^2 + 12y - 36 = (x + y - 6)(x - y + 6)$$

Problem 17

$$3x^3 + 5x^2y + xy^2 - y^3 = (3x - y)(x + y)^2$$

Problem 18

Find the GCF and LCM of:

$$x^2y - xy^2$$

$$3x - 3y$$

$$x^2 - 2xy + y^2$$

Answer:

$$x^2 - xy^2 = xy(x - y)$$

$$3x - 2y = 3(x - y)$$

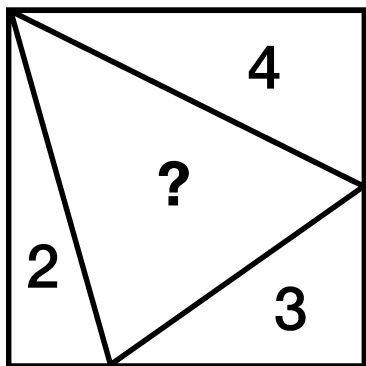
$$x^2 - 2xy + y^2 = (x - y)^2$$

$$\text{GCF} = (x - y)$$

$$\text{LCM} = 3xy(x - y)^2$$

Problem 19

Determine the unknown area contained in the square as shown.



Answer: Starting on the left side and reading counter-clockwise, the four perimeter segments are (A) , $(B + C)$, $(D + E)$, (A) . Note also that

$$AB = 4$$

$$CD = 6$$

$$AE = 8.$$

Use $(A - B)(A - E) = CD$ to arrive at

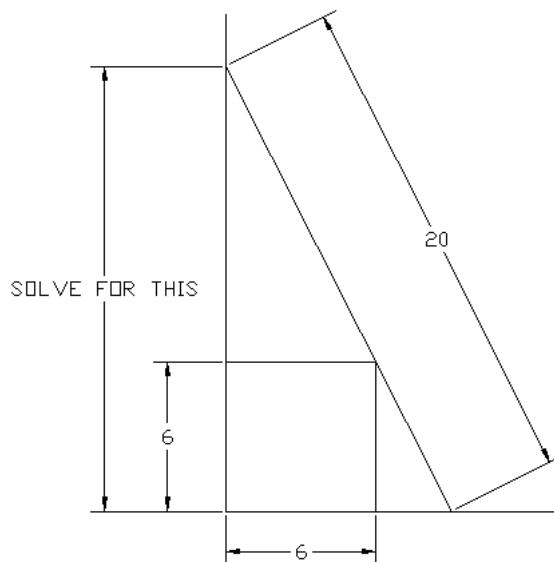
$$A^2 + \frac{32}{A^2} = 18,$$

which resolves to $A = 4$, making the unknown area equal to 7.

Problem 20

Source: stirlingsouth.com

A 20-foot ladder leans on a perpendicular wall such that it touches the edge of a $6 \times 6 \times 6$ -foot cube flatly pushed against the wall, as seen in the figure. Find the vertical height of the ladder above the cube.



Answer: Denote x as the unknown vertical component of the ladder's projection, and y as the unknown horizontal component. By area arguments, or by considering similar triangles, observe that $xy = 36$. Further, the Pythagorean theorem dictates

$$(x + 6)^2 + (y + 6)^2 = 20^2,$$

where completing the square in the variable $x + y$ eventually gives

$$x + y = -6 + 2\sqrt{109},$$

resolving to $x \approx 11.840$, $y \approx 3.0405$.

Problem 21

Consider the system of equations

$$2x^2 + 3y = 19$$

$$4x - y = 3$$

Eliminate y , x separately to derive:

$$x^2 + 6x = 14$$

$$y^2 + 30y = 143$$

Solve each of these for x , y , respectively. Find two ordered-pair solutions to the system. Answer: $(1.7953, 4.1833)$, $(-7.7958, -34.183)$

2 Factoring Cubics

A *cubic expression* is a *third-order* polynomial:

$$Ax^3 + Bx^2 + Cx + D$$

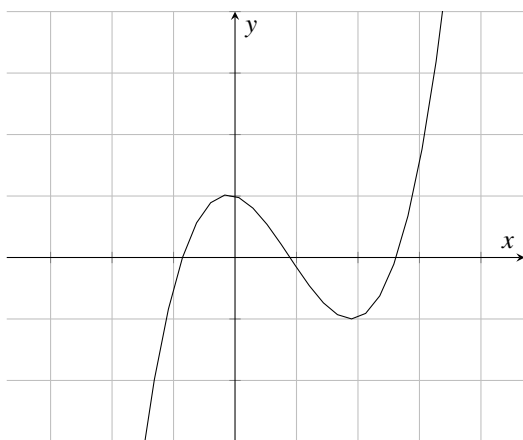
As a generalized quadratic expression, one might wonder how a cubic expression could be factored, and whether there exists an order-three analog to the quadratic formula. As it turns out, this answer isn't so trivial to come by. Unlike quadratic expressions, whose mysteries were shooed away millennia ago, it took until the sixteenth century to crack the problem of cubics.

2.1 Cubic Equations

To approach this problem we first construct a *cubic equation*

$$y = Ax^3 + Bx^2 + Cx + D,$$

qualitatively appearing as follows:



In the above, the coefficients are chosen such that

$$A = 1/2$$

$$B = -4/3$$

$$C = -1/3$$

$$D = 1$$

The leading coefficient A determines the overall steepness of the curve. For $A > 0$, the curve grows upward for increasing x , while dipping (very) negative for decreasing x . Far from the origin, this is true regardless of B, C, D , as the x^3 -term grows much faster than the lower-order terms. For $A < 0$, similar comments apply.

The coefficient D plays the role of the y -intercept, controlling the vertical placement of the plot on the Cartesian plane. The coefficients B and C are responsible for the overall structure near $x = 0$. In the general case, a cubic equation flaunts two vertex points and up to three x -intercepts. The number of x -intercepts can vary depending on the values of B, C, D , but there is always at least one.

Factoring Cubic Equations

We now develop a method to factor any cubic expression, which amounts to looking for x -intercepts in the cubic equation

$$y = Ax^3 + Bx^2 + Cx + D.$$

2.2 Depressed Cubic

To begin, we employ a trick to do away with the x^2 -term by making the curious substitution

$$x = z - \frac{B}{3A}.$$

Letting the algebra carry forth, we find

$$\begin{aligned} 0 = z^3 + z \left(-\frac{B^2}{3A^2} + \frac{C}{A} \right) \\ + 2 \left(\frac{B}{3A} \right)^3 - \frac{C}{A} \left(\frac{B}{3A} \right) + \frac{D}{A}, \end{aligned}$$

which is a bit ugly, but conveniently omits a z^2 -term. Proceed by setting

$$\begin{aligned} b = -\frac{B^2}{3A^2} + \frac{C}{A} \\ -c = 2 \left(\frac{B}{3A} \right)^3 - \frac{C}{A} \left(\frac{B}{3A} \right) + \frac{D}{A} \end{aligned}$$

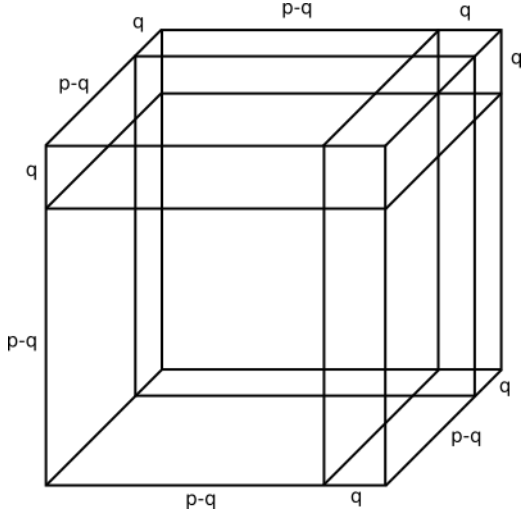
to arrive at the equation of the *depressed cubic*:

$$z^3 + bz = c$$

Henceforth we'll work in the variable z , keeping in mind that we get back to x using the initial substitution.

Geometric Form of the Cubic

Using geometry as an aid to solving cubic equations is a trick attributed to Gerolamo Cardano in 1545. Tracing Cardano's steps, begin with a cube of side p , and then introduce three planes inside the cube, parallel to the top, right, and back faces. Set each plane length q from the cube's respective sides as shown.



The total volume inside the total cube consists of the 'primary' cube of side $p - q$, three slabs of volume $q(p - q)^2$, three bars of volume $q^2(p - q)$, and a small cube of side q . Meanwhile, the total volume is simply p^3 . Equating the total volumes, we write

$$p^3 = (p - q)^3 + 3q^2(p - q) + 3(p - q)^2q + q^3,$$

readily simplifying to

$$(p - q)^3 + 3pq(p - q) = p^3 - q^3.$$

Looking closely, note the above is a depressed cubic equation. Identifying

$$p - q = z$$

$$3pq = b$$

$$p^3 - q^3 = c,$$

we recover the form $z^3 + bz = c$.

Ferro-Tartaglia Formula

Picking up from Cardano's geometric formulation of the depressed cubic problem

$$z^3 + bz = c,$$

eliminate q between the latter two equations to get

$$p^6 - cp^3 - \left(\frac{b}{3}\right)^3 = 0,$$

which is in fact a quadratic equation in the variable p^3 , easily isolated by the quadratic formula:

$$p^3 = \frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}$$

This automatically gives us q^3 , specifically

$$q^3 = p^3 - c = -\frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3},$$

allowing a solution for z to be written:

$$z_0 = p - q$$

$$z_0 = \left(\frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}\right)^{1/3} + \left(\frac{c}{2} - \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}\right)^{1/3}$$

The above is associated with Italian mathematician N. F. Tartaglia (1500-1557), although its discovery is credited to another Italian mathematician S. del Ferro (1465-1526).

In conclusion, we have the Ferro-Tartaglia formula as

$$z_0 = \left(\frac{c}{2} + q\right)^{1/3} + \left(\frac{c}{2} - q\right)^{1/3},$$

where

$$q = \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}.$$

Problem 22

Find one solution to:

$$z^3 - \frac{z}{3} - \frac{2}{27} = 0$$

Step 1: Identify the above as a depressed cubic equation and pick out coefficients:

$$3pq = b = -\frac{1}{3}$$

$$p^3 - q^3 = c = \frac{2}{27}$$

Step 2: Solve for p^3 , and write p and q :

$$p^3 = \frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}$$

$$p^3 = \frac{1}{27} + \sqrt{\left(\frac{1}{27}\right)^2 - \left(\frac{1}{9}\right)^3} = \frac{1}{27}$$

$$p = \frac{1}{3}$$

$$q = -\frac{1}{3}$$

Step 3: Write the solution for z in terms of p and q :

$$z = p - q = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

2.3 Completing the Cubic Solution

In the general case, the geometric approach to solving a depressed cubic equation

$$z^3 + bz = c$$

produces one solution, however there should exist (up to) three total solutions to the equation. Labeling the known solution as z_0 , it follows that $(z - z_0)$ can be factored out of the depressed cubic equation to get

$$z^3 + bz - c = (z - z_0) \left(z^2 + z_0 z + \frac{c}{z_0} \right).$$

Remaining solutions to the depressed cubic equation are given by

$$0 = z^2 + z_0 z + \frac{c}{z_0},$$

an easy application of the quadratic formula.

Problem 23

Use the known solution $z_0 = 2/3$ to continue factoring the expression:

$$z^3 - \frac{z}{3} - \frac{2}{27}$$

Step 1: Substitute $x_0 = 2/3$ and $c = -2/27$ into the quadratic equation for z and simplify:

$$z = \frac{2/3}{2} \left(-1 \pm \sqrt{1 - \frac{4 \cdot 2/27}{(2/3)^3}} \right) = -\frac{1}{3}$$

Step 2: Pick out any new solution(s) gained. In this case, we have two copies of the same number:

$$z_1 = z_2 = -\frac{1}{3}$$

Step 3: Write the final form:

$$z^3 - \frac{z}{3} - \frac{2}{27} = \left(z - \frac{2}{3} \right) \left(z + \frac{1}{3} \right)^2$$

Depressed Cubic in Disguise

Problem 24

Evaluate:

$$\left(7 + \sqrt{50} \right)^{1/3} + \left(7 - \sqrt{50} \right)^{1/3}$$

Answer: Proceeding with caution, let us store the whole quantity in a variable x , and then calculate x^3 :

$$\begin{aligned} x^3 &= \left(\left(7 + \sqrt{50} \right)^{1/3} + \left(7 - \sqrt{50} \right)^{1/3} \right)^3 \\ &= 14 + 3 \left(\left(7 + \sqrt{50} \right)^{1/3} \left(7 - \sqrt{50} \right)^{1/3} \right) \\ &\quad \cdot \left(\left(7 + \sqrt{50} \right)^{1/3} + \left(7 - \sqrt{50} \right)^{1/3} \right) \\ &= 14 + 3(-1)^{1/3} x \end{aligned}$$

Evidently then, we find

$$x^3 = -3x + 14,$$

the equation of a depressed cubic. By standard means we find the solution to be $x = 2$, which finally means:

$$\left(7 + \sqrt{50} \right)^{1/3} + \left(7 - \sqrt{50} \right)^{1/3} = 2$$

2.4 Negative Radicals

It will turn out that the Ferro-Tartaglia formula isn't so straightforwardly applied in all cases, especially from the point of view of a 1500s mathematician. To demonstrate, let us factor the expression:

$$x^3 + 5x^2 - 2x - 24$$

The first target is the depressed cubic form. Start with the transformation $x = z - B/3A$ and simplify to find

$$\begin{aligned} b &= -\frac{31}{3} \\ c &= \frac{308}{27}, \end{aligned}$$

and the corresponding depressed cubic equation reads

$$z^3 + z \left(\frac{-31}{3} \right) = \frac{308}{27}.$$

With b and c , a solution z_0 to the depressed cubic is generated from the Ferro-Tartaglia formula. Applying this, we inevitably hit

$$q = \sqrt{\frac{c^2}{4} + \frac{b^3}{27}} = \sqrt{-\frac{25}{3}}.$$

Alarming, there is a minus sign inside the square root term.

Bombelli's Wild Thought

The courage to work with an equation that contains a negative term inside a square root came first to Rafael Bombelli, now known as his *wild thought*. To replicate Bombelli's wild thought, define a pair of numbers U, V where

$$\begin{aligned}\left(\frac{c}{2} + q\right)^{1/3} &= U + \sqrt{-1}V \\ \left(\frac{c}{2} - q\right)^{1/3} &= U - \sqrt{-1}V,\end{aligned}$$

such that the sum

$$z_0 = 2U$$

becomes an identity.

To proceed, raise the top equation to the third power to write

$$\frac{c}{2} + q = U \left(U^2 - 3V^2 \right) + \sqrt{-1}V \left(3U^2 - V^2 \right),$$

which was enough for Bombelli to see the solution. By separating the 'real' term without the factor of $\sqrt{-1}$ from its 'imaginary' counterpart, we end up with a pair of equations where *neither* makes reference to $\sqrt{-1}$. Moreover, Bombelli went on to seek integer solutions for U, V , which isn't always easy, or always possible.

Continuing the example on hand, we have

$$\frac{c}{2} + q = \frac{154}{27} + \sqrt{-\frac{25}{3}},$$

and comparing this to the expanded form in terms of U, V , gives

$$\begin{aligned}\frac{154}{27} &= U \left(U^2 - 3V^2 \right) \\ \sqrt{\frac{25}{3}} &= V \left(3U^2 - V^2 \right).\end{aligned}$$

To make headway on the problem, some fiddling leads one to the substitution

$$\begin{aligned}2\sqrt{3}V &= \tilde{V} \\ 6U &= \tilde{U},\end{aligned}$$

and the above becomes

$$\begin{aligned}11 \cdot 14 \cdot 8 &= \tilde{U} (\tilde{U} + 3\tilde{V}) (\tilde{U} - 3\tilde{V}) \\ 1 \cdot 12 \cdot 10 &= \tilde{V} (\tilde{U} + \tilde{V}) (\tilde{U} - \tilde{V}),\end{aligned}$$

solved by two integers

$$\begin{aligned}\tilde{U} &= 11 \\ \tilde{V} &= 1,\end{aligned}$$

meaning

$$\begin{aligned}U &= \frac{11}{6} \\ V &= \frac{1}{2\sqrt{3}},\end{aligned}$$

from which we get

$$z_0 = 2U = \frac{11}{3}.$$

Restoring the x - variable, we finally have

$$x_0 = z_0 - \frac{B}{3A} = \frac{11}{3} - \frac{5}{3} = 2$$

as a solution to the original cubic expression.

With the hard part finished, we can find two more solutions to the example on hand by solving

$$0 = z^2 + \left(\frac{11}{3}\right)z + \frac{308 \cdot 3}{27 \cdot 11},$$

having solutions

$$\begin{aligned}z_1 &= -\frac{7}{3} \\ z_2 &= -\frac{4}{3}.\end{aligned}$$

Restoring the x variable, these read

$$\begin{aligned}x_1 &= -4 \\ x_2 &= -3.\end{aligned}$$

Finally then, we have

$$x^3 + 5x^2 - 2x - 24 = (x - 2)(x + 4)(x + 3),$$

finishing the example.

Chapter 5

Kinematics

1 One-Dimensional Motion

Any system that has one *degree of freedom*, such as a the position of a train on a straight track, is said to be a *one-dimensional* system. To represent location in a one-dimensional system, naturally, we need just a one-dimensional graph. By choosing a point of origin, any position can be represented by some variable x that is a real-number displacement from the origin.

Position Plot

When the position of a one-dimensional system changes with respect to some variable, usually time, it makes sense to represent the states of the system on a two-dimensional graph. For this, it is customary to let displacement be represented vertically, and time represented horizontally as demonstrated in Figure 5.1.

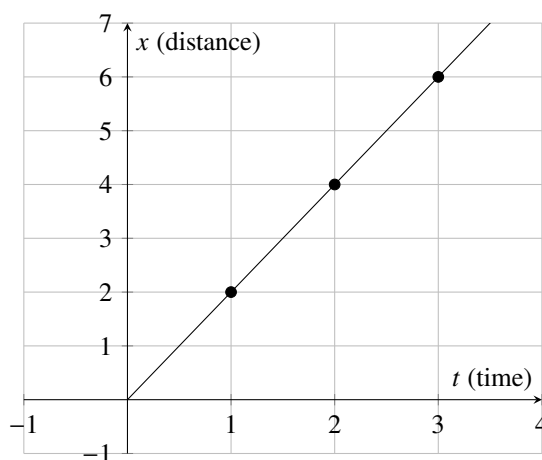


Figure 5.1. Position plot showing uniform one-dimensional motion.

The above Figure depicts the motion of an object that started at the origin at $t = 0$, and then moves two distance units for every time unit. The distance could be centimeters, miles, kilometers, etc. The time unit is most often seconds, hours, or years.

Note that a stationary object with $x(t)$ constant traces a horizontal line in the position plot. Correspondingly, a vertical line would imply that the object's position is undefined except for a small instant of time, during which all possible positions are occupied before becoming undefined again. Such a thing is non-physical at least, and makes very little sense at most.

Slope = Velocity

When one dimensional displacement is represented vertically with time horizontally, then the slope of the line

representing the motion is the *velocity*:

$$v(t) = \frac{\Delta x}{\Delta t} \quad (5.1)$$

In one dimension, the velocity can be positive, negative, or zero. At a given moment in time, the quantity $v(t)$ is also known as the *instantaneous velocity*.

Speed

The magnitude $|v|$ of the velocity is the speed *speed*. As an absolute value, the speed is never negative.

Problem 1

If the average speed of an object is zero, what is the average velocity? If the average velocity is zero, why isn't the speed necessarily zero?

1.1 Uniform Velocity

When the velocity of an object throughout its motion does not change, the object is said to have *constant velocity* or *uniform velocity*, denoted v_0 . In this case, all of the apparatus for straight-line equations can be used. On the basis of Equation (5.1), we can also write

$$v_0 = \frac{x - x_0}{t - t_0}.$$

Position in Time

One-dimensional motion need not begin at the origin at $t = 0$, as offsets in initial position or initial time don't inherently affect the slope of the line. To this end, use the above to write something like the point-slope form for constant velocity:

$$x - x_0 = v_0(t - t_0) \quad (5.2)$$

As it turns out, Equation (5.2) contains more freedom than we really need, and it is customary to set $t_0 = 0$ so that x_0 handles the initial (or otherwise given) condition of the motion. Solving for x delivers the first equation of *kinematics* for uniform velocity:

$$x = x_0 + v_0 t$$

Problem 2

A car travels 30 miles at 50 mph and then 90 miles at 30 mph. Show that the average speed is 33.33 mph.

Problem 3

Mr. Smith drives to the store at 55 mph. Mrs. Smith drives half way to the same store at 45 mph, and then drives the other half way at V mph. Solve for V such that Mr. and Mrs. Smith spend the same time driving to the store.

1.2 Uniform Acceleration

Upping the ante, now consider the regime there the velocity $v(t)$ is not a constant, but instead increases or decreases linearly via *acceleration*. The acceleration term, denoted a , is a constant having units

$$[a] = [\text{distance}/\text{time}^2].$$

Instantaneous Velocity Equation

In terms of the uniform acceleration, the velocity at time t is given by

$$v = v_0 + at, \quad (5.3)$$

where v_0 is the velocity at $t = 0$. As t increases, v 's behavior is governed by at .

Velocity Plot

As a linear equation in time, a plot with v on the vertical axis and t on the horizontal axis will trace a straight line with a v -intercept at v_0 as shown in Figure 5.2.

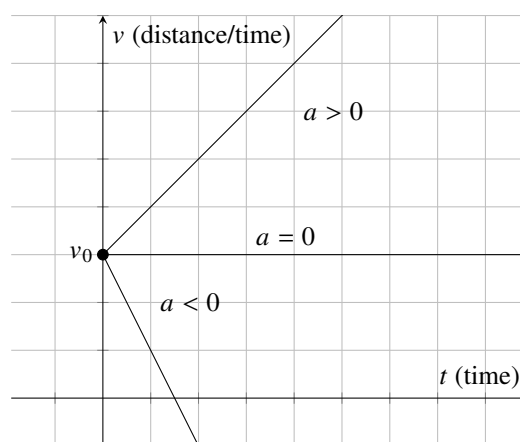


Figure 5.2. Velocity plot showing uniform acceleration in one dimension.

In the Figure, one sees that the slope of the velocity with respect to time is the constant a . The case $a = 0$ corresponds to constant velocity. The 'negative acceleration' case, also known as *deceleration*, causes v to reduce in time and eventually go negative.

Problem 4

Convert $g = 9.80665 \text{ m/s}^2$ to ft/s^2 and also to miles per hour per second.

Problem 5

An arrow traveling at 175 ft/sec strikes the ground, coming to rest in 0.1 sec. Assuming the acceleration of the arrow inconstant as it moves through the ground, determine its magnitude in ft/sec^2 .

Problem 6

A vehicle traveling at 45 mph skids to a stop over a 70 ft distance. Calculate the acceleration of the car. How long did it take to stop?

Average Velocity

The *average velocity* $\bar{v}$ is the average of v_0 and v at time t . Calculating this out, we find

$$\bar{v} = \frac{1}{2} (v_0) + \frac{1}{2} (v_0 + at) ,$$

or

$$\bar{v} = v_0 + \frac{1}{2} at . \quad (5.4)$$

Net Displacement

In terms of the average velocity, the net displacement can be calculated via

$$x = x_0 + v_0 t + \frac{1}{2} at^2 . \quad (5.5)$$

in agreement with the previous calculation.

Problem 7

A car stopped at a red light begins accelerating when the light turns green. At the same moment, a truck coasting at a constant 20 m/s passes the car. If the car's acceleration is 7.5 m/s^2 , calculate when (t) and where (x) the car catches up to the truck.

Displacement as Area

At a given time t , the total area under the line $v = v_0 + at$ equals the net displacement $x - x_0$ of the motion. (Read that twice!)

From Figure 5.2, one sees the area under the line is the sum of a rectangle and a triangle such that:

$$\text{rectangle area} = v_0 t$$

$$\text{triangle area} = \frac{1}{2} (v - v_0) t = \frac{1}{2} at^2$$

If our incantation linking displacement to area is true, then we must have

$$x - x_0 = v_0 t + \frac{1}{2} at^2 .$$

Time-Shift Analysis

One more way to derive the displacement equation for uniform acceleration is by time-shift analysis. For this, introduce an unknown variable A and write the generic form

$$x = x_0 + v_0 t + At^2 ,$$

where the job is to solve for A .

To proceed, introduce the time shift $t \rightarrow t + h$, making the position transform via $x \rightarrow w$. For this, we write

$$w = x_0 + v_0 (t + h) + A (t + h)^2 .$$

Distribute all products and rearrange in powers of h to get

$$w = (x_0 + v_0 t + At^2) + h (v_0 + 2At) + Ah^2 .$$

Now, suppose t is held constant while h does the job of being the time variable. For this, let $x = x_t$ and $v = v_t$ to remind that these values are 'stuck' at time t . The above becomes

$$w = x_t + h (v_0 + 2At) + Ah^2 ,$$

however the parenthesized term multiplying h must have the form of the velocity at time t , i.e.

$$v_0 + 2At = v_0 + at ,$$

revealing finally that

$$A = \frac{1}{2} a .$$

The final form for w reads

$$w = x_t + v_t h + \frac{1}{2} ah^2 .$$

Comparing to

$$x = x_0 + v_0 t + \frac{1}{2} at^2 ,$$

we see that the w -equation is the same as the x -equation up to choice of initial condition.

1.3 Position Plot

For uniform acceleration, we established three ways that the position of an object obeys the Equation (5.5). Due to the presence of the t^2 -term, the position is quadratic in time as sketched in Figure 5.3.

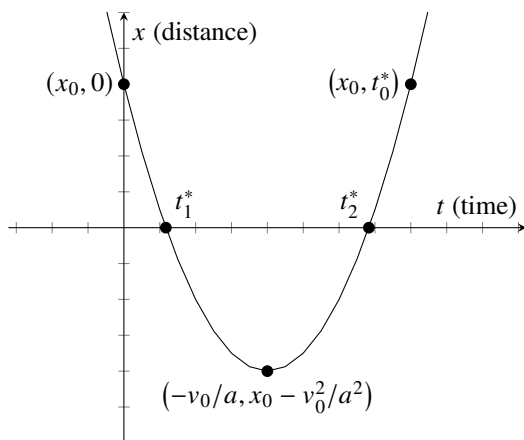


Figure 5.3. Quadratic position curve in one dimension.

Problem 8

Starting with Equation (5.5), complete the square in t to derive

$$x = \left(x_0 - \frac{v_0^2}{2a} \right) + \frac{a}{2} \left(t + \frac{v_0}{a} \right)^2 \quad (5.6)$$

Problem 9

Starting with Equation (5.5), use the quadratic formula to derive

$$t = \frac{-v_0}{a} \pm \frac{v_0}{a} \sqrt{1 + \frac{2a(x - x_0)}{v_0^2}} \quad (5.7)$$

Problem 10

Show that the vertex of the parabola is located at

$$(t_{\text{vert}}, x_{\text{vert}}) = \left(\frac{-v_0}{a}, x_0 - \frac{v_0^2}{2a} \right).$$

Problem 11

Find the time t_0^* at which the position returns to x_0 and discuss its validity. Answer:

$$t_0^* = \frac{-2v_0}{a}$$

Problem 12

Find the times $t_{1,2}^*$ at which the position touches $x = 0$. Answer:

$$t_{1,2}^* = \frac{-v_0}{a} \left(1 \pm \sqrt{1 - \frac{2x_0 a}{v_0^2}} \right)$$

1.4 Kinematic Identities

In addition to Equation (5.3), namely

$$v = v_0 + at,$$

and Equation (5.5) given by

$$x = x_0 + v_0 t + \frac{1}{2} at^2,$$

we can derive a flurry of kinematic identities to quantify one-dimensional motion with constant acceleration.

Recall that in terms of the average velocity via (5.4)

$$\bar{v} = \frac{v_0 + v}{2} = v_0 + \frac{1}{2} at,$$

the position x can also be written

$$x = x_0 + \bar{v} t. \quad (5.8)$$

Eliminating the v_0 term between the above, one has

$$x = x_0 + vt - \frac{1}{2} at^2. \quad (5.9)$$

An important kinematic identity is established by eliminating t between Equations (5.3), (5.8). From these, one writes:

$$\begin{aligned} x - x_0 &= \bar{v} t \\ x - x_0 &= \frac{v_0 + v}{2} \left(\frac{v - v_0}{a} \right) \\ 2a(x - x_0) &= v^2 - v_0^2 \end{aligned}$$

Solve for v^2 to land at:

$$v^2 = v_0^2 + 2a(x - x_0) \quad (5.10)$$

Summary

To summarize, we now have a set of equations describing one-dimensional motion under constant acceleration, each purposefully containing all but one of the variables in play:

Identity:	Missing:
$v = v_0 + at$	$x - x_0$
$x = x_0 + v_0 t + at^2/2$	v
$x = x_0 + (v_0 + v)t/2$	a
$x = x_0 + vt - at^2/2$	v_0
$v^2 = v_0^2 + 2a(x - x_0)$	t

Problem 13

Pick out the equation that can't possibly be true:

$$\Delta x = (v_i + v_f) (v_f - v_i) / 2a$$

$$t = \sqrt{\frac{2\Delta x}{a}}$$

$$x_f = x_i + v^2 t + at$$

1.5 Freefall Acceleration

Near Earth's surface, we know from experience that objects tend to fall downward due to gravity. It turns out that the gravitational effect on a freefalling object resolves to a *constant* acceleration, denoted g , having magnitude

$$g = 9.80665 \frac{\text{m}}{\text{s}^2}. \quad (5.11)$$

If we orient $+x$ or $+y$ as the 'upward' direction, the acceleration due to gravity is:

$$a_{\text{freefall}} = -g$$

Problem 14

Suppose a rod of length $L = 1408$ in begins at rest and freefalls downward. How long does it take for the rod to fall a distance equal to its own length?

Answer:

$$t = \sqrt{\frac{2L}{g}} = \sqrt{\frac{(2)(1408)(2.54)}{(100)(9.80665)}} \text{ s} = 2.70 \text{ s}$$

Problem 15

On the moon, the acceleration due to gravity is approximately one sixth of that on Earth's surface. Starting from rest, how much longer does it take an object to fall a distance H on the moon than it does on Earth? After falling for 75 m, how fast is the object falling in each case?

Wishing Well Problem

A stone is dropped from rest into a well. The splash is heard 2.059 seconds after release. The speed of sound in air is approximately 330 m/s. Calculate the depth of the well. Refrain from using a machine.

From the information given, we have

$$2.059 \text{ s} = \frac{D}{330 \text{ m/s}} + \sqrt{\frac{2D}{g}},$$

which ends up being quadratic in D :

$$(2.059 \text{ s})^2 - 2D \frac{2.059 \text{ s}}{330 \text{ m/s}} + \frac{D^2}{(330 \text{ m/s})^2} = \frac{2D}{9.80665 \text{ m/s}^2}$$

The denominator of the D^2 -term overwhelms all others, thus the whole term is close enough to zero to be omitted. Then, the distance D in meters is given by

$$(2.059)^2 \approx D \left(\frac{1}{4.9} + \frac{4.12}{330} \right)$$

$$20.78 \approx D \left(1 + \frac{20.19}{330} \right),$$

or finally:

$$D \approx \frac{20.78}{1.06} \approx 19.6$$

Two Balls and a Building

A ball is thrown straight up at speed v_0 from the edge of the roof of a building of height H . A second ball is dropped from the roof t_0 seconds later. Ignoring air drag, determine the condition for the two balls hitting the ground simultaneously.

Let y_1 represent the trajectory of the first ball, and y_2 represent the second ball. According to the information given, we may write

$$y_1 = H + v_0 t - \frac{g}{2} t^2$$

$$y_2 = H - \frac{g}{2} (t - t_0)^2,$$

where H is the height of the building, and v_0 is the initial velocity of the first ball.

Now impose the condition that each ball reaches the ground at via $y_1 = y_2 = 0$ at the same time $t = t^*$ to write

$$0 = H + v_0 t^* - \frac{g}{2} (t^*)^2$$

$$0 = H - \frac{g}{2} (t^* - t_0)^2.$$

The difference of these two results eliminates H to avail a formula for t^* , namely

$$t^* = \frac{gt_0^2/2}{gt_0 - v_0}.$$

From this, we see the special case $gt_0 = v_0$ causes t^* to be undefined, and $gt_0 < v_0$ causes t^* to be negative, neither of which is physically possible. Evidently, we must have

$$v_0 < gt_0 = v_{\text{max}}$$

for a valid solution.

Meanwhile, recall there is a time $\tilde{t} = 2v_0/g$ at which the first ball returns to its original height, and this time must be greater than t_0 , which means

$$v_0 > \frac{gt_0}{2} = v_{\min}.$$

We may solve for t^* in each equation to get

$$t^* = \frac{v_0}{g} \pm \sqrt{\frac{v_0^2}{g^2} + \frac{2H}{g}}$$
$$t^* = t_0 \pm \sqrt{\frac{2H}{g}},$$

combining to give

$$\frac{v_0}{g} \pm \sqrt{\frac{v_0^2}{g^2} + \frac{2H}{g}} = t_0 \pm \sqrt{\frac{2H}{g}},$$

which avails an equation for H , after some simplifying:

$$H = \frac{gt_0^2}{2} \left(\frac{gt_0/2 - v_0}{gt_0 - v_0} \right)^2$$

2 Two-Dimensional Motion

For motion not limited to one dimension, an object free to move in two dimensions occupies some position in the Cartesian plane, thus the description of any position is an ordered pair (x, y) . If the x - and y -coordinates are independent (which we may always assume), then the whole kinematics apparatus for one dimension applies to x and y separately.

In particular, Equation (5.5) manifests twice according to

$$\begin{aligned}x &= x_0 + v_{x0}t + \frac{1}{2}a_x t^2 \\ y &= y_0 + v_{y0}t + \frac{1}{2}a_y t^2,\end{aligned}$$

where subscripts on initial velocity v_0 and acceleration a have been added accordingly. Similar applies to Equation (5.3), namely

$$\begin{aligned}v_x &= v_{x0} + a_x t \\ v_y &= v_{y0} + a_y t.\end{aligned}$$

The speed v is constructed using the Pythagorean theorem:

$$v = \sqrt{v_x^2 + v_y^2} \quad (5.12)$$

Freefall Motion

For freefall motion near Earth's surface, the effect of gravity links a_y to the local acceleration constant via

$$a_y = -g$$

in analog to Equation (5.11). By a similar token, $a_x = 0$ unless a gravity-like force acts horizontally (which doesn't exist). For this reason, the motion in the x -direction is given by

$$\begin{aligned}x &= x_0 + v_{x0}t \\ v_x &= v_{x0},\end{aligned}$$

and the motion in the y direction is given by

$$\begin{aligned}y &= y_0 + v_{y0}t - \frac{1}{2}gt^2 \\ v_y &= v_{y0} - gt.\end{aligned}$$

2.1 Deparameterized Equations

For two-dimensional motion, having separate equations for x , v_x , y , v_y all in terms of time t constitutes a parameterized description of the motion. It's often useful to have a set of equations that relates y directly to x , thereby eliminating the parameter t .

Eliminating t between the pair of velocity equations, one finds

$$v_y = v_{y0} - \frac{g}{2} \left(\frac{x - x_0}{v_{x0}} \right)^2.$$

More useful is the equation that eliminated t between x and y . Solve for t in the x -equation and replace all instances of t in the y -equation:

$$y = y_0 + v_{y0} \left(\frac{x - x_0}{v_{x0}} \right) - \frac{g}{2} \left(\frac{x - x_0}{v_{x0}} \right)^2 \quad (5.13)$$

Completing the square in x , the above becomes

$$y = y_0 + \frac{v_{y0}^2}{2g} - \frac{g}{2} \left(\frac{x - x_0}{v_{x0}} - \frac{v_{y0}}{g} \right)^2. \quad (5.14)$$

Perhaps not surprisingly, the y -equation is quadratic in the variable x as it was in the variable t . This means the motion of a projectile in uniform gravity free of air drag always follows a parabolic path as shown in Figure 5.4.

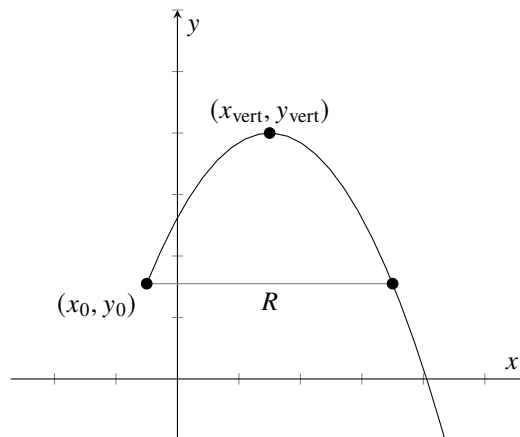


Figure 5.4. Two-dimensional motion in uniform gravity.

Vertex

The above Figure is sketched with $v_{x0} > 0$ and $v_{y0} > 0$. The vertex of the parabolic motion is located at $(x_{\text{vert}}, y_{\text{vert}})$,

which is easily determined by setting the second term in Equation (5.14) to zero. From this we quickly see

$$y_{\text{vert}} = y_0 + \frac{v_{y0}^2}{2g}$$

and

$$x_{\text{vert}} = x_0 + \frac{v_{x0} v_{y0}}{g}.$$

In terms of the vertex location, Equation (5.14) is written more simply:

$$y = y_{\text{vert}} - \frac{g}{2} \left(\frac{x - x_{\text{vert}}}{v_{x0}} \right)^2$$

Not forgetting time is still relevant, we can write down the time t_{vert} required for the object to reach the vertex. From the x -equation we find

$$t_{\text{vert}} = \frac{x_{\text{vert}} - x_0}{v_{x0}},$$

and from the y -equation,

$$t_{\text{vert}} = \frac{v_{y0}}{g}.$$

Horizontal Range

An object moving upward against uniform gravity will eventually reach a maximum height at y_{vert} and fall to its original height y_0 . The horizontal distance traversed during this part of the motion is the *horizontal range*, depicted as R in Figure 5.4.

By construction, the horizontal range is two times the distance from x_0 to x_{vert} , or:

$$R = 2(x_{\text{vert}} - x_0) = \frac{2v_{x0}v_{y0}}{g} \quad (5.15)$$

Interestingly, notice that the horizontal range remains the same if mutually swapping $v_{x0} \leftrightarrow v_{y0}$. More strongly, notice that the range is the same so long as the product $v_{x0}v_{y0}$ remains the same as sketched in Figure 5.5.

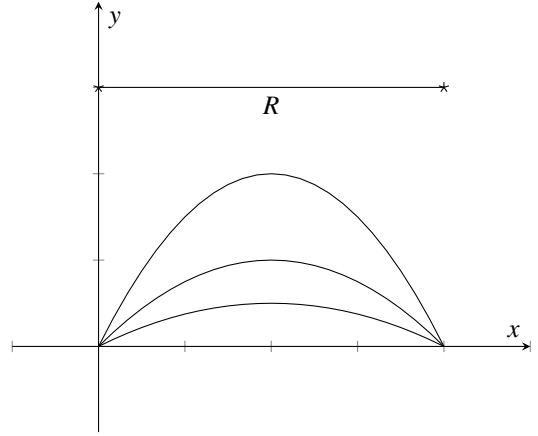


Figure 5.5. Range equivalence of projectile motion.

Flight Time

The time t_R required to traverse the horizontal range is twice the time t_{vert} :

$$t_R = 2 \left(\frac{x_{\text{vert}} - x_0}{v_{x0}} \right) = \frac{2v_{y0}}{g}$$

Multiplying through by v_{x0} gives a tidy identity relating the above to R :

$$v_{x0}t_R = R$$

It's also worth having another equation that relates t directly to y . Adapting Equation (5.7), write

$$t = \frac{v_{y0}}{g} \mp \frac{v_{y0}}{g} \sqrt{1 - \frac{2g(y - y_0)}{v_{y0}^2}},$$

which simplifies to:

$$t = t_{\text{vert}} \mp \sqrt{\frac{2}{g} (y_{\text{vert}} - y)}$$

2.2 Envelope of Trajectories

For a given initial speed

$$v_0 = \sqrt{v_{x0}^2 + v_{y0}^2},$$

the *envelope of trajectories* is a curve in the Cartesian plane that characterizes all possible paths of motion. For simplicity we'll assume each velocity is greater than zero.

To find the envelope of trajectories for freefall motion, note from the above that

$$\frac{1}{v_{x0}^2} = \frac{1}{v_0^2} \left(1 + \frac{v_{y0}^2}{v_{x0}^2} \right),$$

and substitute into Equation (5.13) for deparameterized kinematics, yielding

$$y = y_0 + \frac{v_{y0}}{v_{x0}} (x - x_0) - \frac{g}{2} (x - x_0)^2 \frac{1}{v_0^2} \left(1 + \frac{v_{y0}^2}{v_{x0}^2} \right).$$

The above is in fact quadratic in the variable v_{y0}/v_{x0} . To keep the algebra tame, write

$$0 = A \left(\frac{v_{y0}}{v_{x0}} \right)^2 + B \left(\frac{v_{y0}}{v_{x0}} \right) + C$$

such that

$$A = - \left(\frac{g}{2} (x - x_0)^2 \frac{1}{v_0^2} \right)$$

$$B = (x - x_0)$$

$$C = \left(y_0 - \frac{g}{2} (x - x_0)^2 \frac{1}{v_0^2} \right) - y = -y + y_0 + A.$$

As a quadratic equation, thereby solved by the quadratic formula, the discriminant $B^2 - 4AC$ is the most telling quantity. For the trajectory to exist, the discriminant must be non-negative. The limit case $B^2 = 4AC$ will provide the outline of all allowed trajectories for the given v_{x0}, v_{y0} . Setting $B^2 = 4AC$ and simplifying, we find

$$\begin{aligned} (x - x_0)^2 &= -4A (y - y_0) + 4A^2 \\ \frac{-2Av_0^2}{g} &= -4A(y - y_0) + 4A \cdot A \\ y - y_0 &= \frac{v_0^2}{2g} + A, \end{aligned}$$

and finally:

$$y - y_0 = \frac{v_0^2}{2g} - \frac{g}{2v_0^2} (x - x_0)^2 \quad (5.16)$$

Evidently, the envelope of trajectories has parabolic character with intercepts occurring at

$$\begin{aligned} x_{\text{int}} &= x_0 + \frac{v_0^2}{g} \\ y_{\text{int}} &= y_0 + \frac{v_0^2}{2g}. \end{aligned}$$

For simplicity, let $x_0 = y_0 = 0$ to visualize the envelope as shown in Figure 5.6.

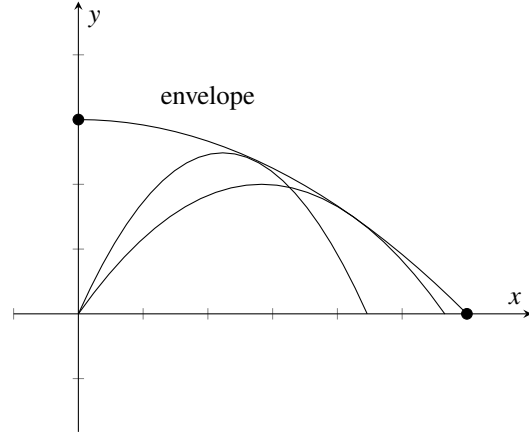


Figure 5.6. Envelope of trajectories.

Maximal Range

Knowing the envelope of trajectories, particularly the x -intercept, we see that the maximum horizontal range for any projectile (assuming $x_0 = 0$) is given by

$$R = \frac{v_0^2}{g}.$$

Meanwhile, we already have a formula for R via Equation (5.15), namely

$$R = \frac{2v_{x0}v_{y0}}{g}.$$

Equating these, we find a restriction on v_{x0}, v_{y0} that correspond to maximal range:

$$\begin{aligned} 0 &= v_0^2 - 2v_{x0}v_{y0} \\ 0 &= v_{x0}^2 + v_{y0}^2 - 2v_{x0}v_{y0} \\ 0 &= (v_{x0} + v_{y0})(v_{x0} - v_{y0}) \end{aligned}$$

For the condition of maximal horizontal range to be satisfied, it must be that:

$$v_{x0} = v_{y0} = \frac{v_0}{\sqrt{2}}$$

2.3 Rifle Problems

Identical Rifles

Suppose a pair of identical rifles mounted at the origin are tilted with unequal angles with respect to the horizon. Let us find the condition for which the projectiles strike the same target located at (x_*, y_*) .

To proceed, use Equation (5.13) for each rifle to write

$$y_* = \frac{v_{y0}}{v_{x0}} x_* - \frac{gx_*^2}{2} \left(\frac{1 + (v_{y0}/v_{x0})^2}{v_0^2} \right)$$

$$y_* = \frac{v_{y'_0}}{v_{x'_0}} x_* - \frac{gx_*^2}{2} \left(\frac{1 + (v_{y'_0}/v_{x'_0})^2}{v_0^2} \right).$$

Eliminate y_* and simplify to end up with

$$\frac{2v_0^2}{gx_*} = \frac{v_{y0}}{v_{x0}} + \frac{v_{y'_0}}{v_{x'_0}}.$$

If the target height y_* is the same height as each rifle, then the horizontal distance x_* is equivalent to the range R , leading to

$$v_{x0} v_{y0} = v_{x'_0} v_{y'_0},$$

and furthermore:

$$v_{y'_0} = v_{x0}$$

$$v_{x'_0} = v_{y0}$$

Hunter and the Monkey

A hunter aims his rifle at a monkey at height H in a tree that is horizontal distance L downrange. The instant the hunter

shoots his rifle, the monkey lets go of the tree. Does the shot miss the monkey?

The position of the projectile starting from $(0, 0)$ is given by

$$x(t) = v_{x0} t$$

$$y(t) = v_{y0} t - \frac{1}{2} g t^2.$$

Meanwhile, the height z of the monkey is given by

$$z(t) = H - \frac{1}{2} g t^2.$$

Eliminate $gt^2/2$ between the y - and z -equations to write the height difference between the monkey and the projectile

$$z(t) - y(t) = H - v_{y0} t,$$

and evaluate at $t_* = L/v_{x0}$ to write

$$z(t_*) - y(t_*) = H - \frac{v_{y0}}{v_{x0}} L.$$

The left side of the above must resolve to zero if the shot hits the monkey. The hunter aims directly at the monkey initially, so we use similar triangles to state

$$\frac{L}{H} = \frac{v_{x0}}{v_{y0}}.$$

This means the right side of the above cancels out, and

$$z(t_*) = y(t_*).$$

Poor monkey.

Chapter 6

Polynomials

1 Binomial Theorem

For two values x and y , consider the quantity

$$P(n) = (x + y)^n,$$

where n is any of

$$n = 0, 1, 2, 3, \dots$$

The quantity $P(n)$ means *polynomial of degree n* .

Explicitly, we find, using the distributive property and the properties of exponents:

$$P(0) = (x + y)^0 = 1$$

$$P(1) = (x + y)^1 = x + y$$

$$P(2) = (x + y)^2 = x^2 + 2xy + y^2$$

$$P(3) = (x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

There is no upper limit to the allowed n . Jotting down a few more for the sake of establishing a pattern, one finds:

$$P(4) = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

$$P(5) = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$$

1.1 Pascal's Triangle

Ignoring any x - and y -symbols and listing *only* the coefficients associated with a given $P(n)$, one discovers *Pascal's Triangle*:

$$\begin{array}{cccccccc} & & & & 1 & & & \\ & & & 1 & & 1 & & \\ & & 1 & & 2 & & 1 & \\ & 1 & & 3 & & 3 & & 1 \\ 1 & & 1 & & 4 & & 6 & & 4 & & 1 \\ & 1 & & 5 & & 10 & & 10 & & 5 & & 1 \\ 1 & & 1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1 \end{array}$$

At a glance, one can see that a given number in the body of Pascal's triangle is the sum of the pair of adjacent numbers one row above.

1.2 Binomial Coefficients

A given number in Pascal's Triangle can be indexed by the row n and the column k , where $0 \leq k \leq n$. To denote the coefficient at position n, k , one writes

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad (6.1)$$

or equivalently:

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-(k-1))}{k!} \quad (6.2)$$

Reflection Identity

Due to the symmetry in Pascal's triangle, one may swap k with $n - k$ to produce the same binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{n-k} \quad (6.3)$$

Coefficient Sum Identity

The binomial coefficients obey a summation identity

$$\binom{n}{m} + \binom{n}{m+1} = \binom{n+1}{m+1}, \quad (6.4)$$

81

1.4 Binomial Theorem

Each expansion $(x + y)^n$, $(x - y)^n$ is wrapped up nicely in the *binomial theorem*:

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k \quad (6.7)$$

For completeness, a given binomial coefficient is equal to

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Abusing the Theorem

While the binomial theorem is clearly defined for integer n , it turns out that Equation (6.7) is robust enough to handle non-integer n . Of course, the sum can't terminate at a non-integer n , so a reasonable thing to do is let the sum run to *infinity*, meaning:

$$(x + y)^n = \sum_{k=0}^{\infty} \binom{n}{k} x^{n-k} y^k$$

For the above to remain valid, it must be that $(y/x)^k$ tends to zero for large k .

For a detailed example, let us attempt to calculate the numerical value of $\sqrt{3}$ by setting

$$\begin{aligned} x &= 4 \\ y &= -1 \\ n &= 1/2 \end{aligned}$$

on the left side of the above equation. Then, we should have

$$(4 - 1)^{1/2} = \sum_{k=0}^{\infty} \binom{1/2}{k} 4^{1/2} \left(\frac{-1}{4}\right)^k,$$

which leaves the question of what to do with a non-integer binomial coefficient.

Proceeding boldly as Newton did, use the definition of the binomial coefficient to write

$$\binom{1/2}{k} = \frac{\frac{1}{2} \left(\frac{1}{2} - 1\right) \left(\frac{1}{2} - 2\right) \cdots \left(\frac{1}{2} - (k-1)\right)}{k!},$$

or explicitly:

$$\begin{aligned} \binom{1/2}{0} &= 1 \\ \binom{1/2}{1} &= \frac{1}{2} \\ \binom{1/2}{2} &= \frac{\frac{1}{2} \left(\frac{1}{2} - 1\right)}{2!} = \frac{-1}{8} \\ \binom{1/2}{3} &= \frac{\frac{1}{2} \left(\frac{1}{2} - 1\right) \left(\frac{1}{2} - 2\right)}{3!} = \frac{1}{16} \\ \binom{1/2}{4} &= \frac{\frac{1}{2} \left(\frac{1}{2} - 1\right) \left(\frac{1}{2} - 2\right) \left(\frac{1}{2} - 3\right)}{4!} = \frac{-5}{128} \\ \binom{1/2}{5} &= \frac{\frac{1}{2} \left(\frac{1}{2} - 1\right) \left(\frac{1}{2} - 2\right) \left(\frac{1}{2} - 3\right) \left(\frac{1}{2} - 4\right)}{5!} = \frac{7}{256} \end{aligned}$$

Putting everything together, find

$$\begin{aligned} \frac{\sqrt{3}}{2} &= \frac{(4-1)^{1/2}}{2} = 1 + \frac{1}{2} \left(\frac{-1}{4}\right)^1 - \frac{1}{8} \left(\frac{-1}{4}\right)^2 \\ &\quad + \frac{1}{16} \left(\frac{-1}{4}\right)^3 - \frac{5}{128} \left(\frac{-1}{4}\right)^4 \\ &\quad + \frac{7}{256} \left(\frac{-1}{4}\right)^5 - \cdots, \end{aligned}$$

leading quite correctly to:

$$\sqrt{3} = 1.7320508075 \dots$$

2 Long division

A versatile and brutal method for manipulating a mathematical expression is the method of polynomial division. The setup and procedure for polynomial division is identical to elementary methods for arithmetic. To illustrate, consider a ratio such as

$$\frac{x^2 - 9x - 10}{x + 1},$$

and set up the corresponding division structure:

$$\begin{array}{r} x+1 \overline{) x^2 - 9x - 10} \end{array}$$

For the process to yield useful results, the numerator should always contain a higher-degree polynomial than the denominator.

2.2 Dividing Infinite Sums

Polynomial division also works well with infinite sums. A case worth exploring starts with the cosine and sine given by

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

The definition of $\tan(x)$ is the ratio of the sine to the

cosine, which can be calculated by brute force using polynomial division. Setting up the problem carefully, one finds

$$\tan(x) \approx x + \frac{x^3}{3} + \frac{2x^5}{15} + \cdots$$

This result comes with a warning, though. Unlike the sine and cosine, the above representation of the tangent is not periodic and *only* works near $x = 0$.

3 Partial Fractions

When confronted with a ratio of polynomials where the denominator is of higher degree than the numerator, a technique called *partial fractions* can be used to break apart the ratio.

Starting with a simple case, consider the scenario where the denominator has a degree-two polynomial in factored form. With this, observe that such a ratio can be split into the sum of two terms, each containing a degree-one polynomial:

$$\frac{cx + d}{(x - a)(x - b)} = \frac{A}{x - a} + \frac{B}{x - b}$$

The unknowns A, B are easily determined in terms of a, b, c, d . By setting $x = 0$, and then $x = 1/c$, respectively, we gain two equations

$$\begin{aligned} -d &= bA + aB \\ c &= A + B, \end{aligned}$$

solved by

$$\begin{aligned} A &= \frac{ac + d}{a - b} \\ B &= \frac{bc + d}{b - a}, \end{aligned}$$

which could have also been inferred by choosing values $x = a, x = b$.

This method generalizes to higher-degree polynomial denominators, as shown for the degree-three case:

$$\frac{1}{(x - a)(x - b)(x - c)} = \frac{A}{x - a} + \frac{B}{x - b} + \frac{C}{x - c}$$

Corollary

In general, if a polynomial $p(x)$ occurs in the denominator and is already factored into linear and quadratic terms, then for each factor $x - a$, there exists a term

$$\frac{A}{x - a},$$

where A must be determined in context.

Problem 1

Find the equivalent ratio as a sum of partial fractions:

$$\frac{2x + 1}{(x - 3)(x - 4)}$$

Step 1: Rewrite the ratio as a sum:

$$\frac{2x + 1}{(x - 3)(x - 4)} = \frac{A}{x - 3} + \frac{B}{x - 4}$$

Step 2: Solve for A and B to get:

$$\begin{aligned} A &= -7 \\ B &= 9 \end{aligned}$$

Step 3: Assemble the result:

$$\frac{2x + 1}{(x - 3)(x - 4)} = \frac{-7}{x - 3} + \frac{9}{x - 4}$$

Problem 2

Find the equivalent ratio as a sum of partial fractions:

$$\frac{1}{a^2 - x^2}$$

Step 1: Factor the denominator:

$$\frac{1}{a^2 - x^2} = \frac{1}{(a - x)(a + x)}$$

Step 2: Rewrite the ratio as a sum:

$$\frac{1}{(a - x)(a + x)} = \frac{A}{a - x} + \frac{B}{a + x}$$

Step 3: Solve for A and B to get:

$$\begin{aligned} A &= 1/2a \\ B &= 1/2a \end{aligned}$$

Step 4: Assemble the result:

$$\frac{1}{a^2 - x^2} = \frac{1}{2a} \left(\frac{1}{a - x} + \frac{1}{a + x} \right)$$

3.1 Repeated Roots

Of course, the partial fraction expansion is prone to error if we run into division by zero, i.e. the case $a = b$. To handle a ratio having two repeated roots in the denominator, we use a partial fraction expansion

$$\frac{1}{(x - a)^2(x - b)} = \frac{A_1}{x - a} + \frac{A_2}{(x - a)^2} + \frac{B}{x - b},$$

admitting a separate term for each instance of $(x - a)$. This pattern generalizes to three repeated roots, and so on:

$$\begin{aligned} \frac{1}{(x - a)^3(x - b)} = & \\ & \frac{A_1}{x - a} + \frac{A_2}{(x - a)^2} + \frac{A_3}{(x - a)^3} + \frac{B}{x - b} \end{aligned}$$

3.2 Quadratic Factors

Factors of the form $x^2 + ax + b$ occurring in the denominator can be balanced by an $Ax + B$ -term according to

$$\frac{1}{(x^2 + ax + b)(x - c)} = \frac{Ax + B}{x^2 + bx + c} + \frac{C}{x - c}.$$

If a factor like $(x^2 + ax + b)^2$ occurs, extra terms are needed:

$$\begin{aligned} \frac{1}{(x^2 + ax + b)^2 (x - c)} = \\ \frac{A_1x + B_1}{x^2 + bx + c} + \frac{A_2x + B_2}{(x^2 + bx + c)^2} + \frac{C}{x - c} \end{aligned}$$

Problem 3

Find the equivalent ratio as a sum of partial fractions:

$$\frac{1}{x^4 - 1}$$

Step 1: Factor the denominator:

$$\frac{1}{x^4 - 1} = \frac{1}{(x - 1)(x + 1)(x^2 + 1)}$$

Step 2: Rewrite the ratio as a sum:

$$\begin{aligned} \frac{1}{(x - 1)(x + 1)(x^2 + 1)} = \\ \frac{A}{x - 1} + \frac{B}{x + 1} + \frac{Cx + D}{x^2 + 1} \end{aligned}$$

Step 3: Multiply through by the left-hand denominator:

$$\begin{aligned} 1 &= A(x + 1)(x^2 + 1) + B(x - 1)(x^2 + 1) \\ &\quad + (Cx + D)(x - 1)(x + 1) \end{aligned}$$

Step 4: Let $x = 1$, $x = -1$, $x = 0$, and $x = 2$ to isolate each coefficient:

$$\begin{aligned} A &= 1/4 \\ B &= -1/4 \\ D &= -1/2 \\ C &= 0 \end{aligned}$$

Step 5: Assemble the result:

$$\frac{1}{x^4 - 1} = \frac{1}{4} \left(\frac{1}{x - 1} - \frac{1}{x + 1} \right) - \frac{1}{2} \left(\frac{1}{x^2 + 1} \right)$$

3.3 Mixed Division Cases

Certain situations call for polynomial division *and* partial fractions. For instance, in the ratio

$$\frac{x^3 + 4}{x^2 + x},$$

the numerator contains a higher-degree polynomial than the denominator. Carrying out the division problem

$$x^2 + x \overline{) x^3 + 4},$$

we end up with a quotient and a remainder as follows:

$$\frac{x^3 + 4}{x^2 + x} = (x - 1) + \frac{x + 4}{x^2 + x}$$

Next, take the remainder term in isolation and use partial fraction analysis to write

$$\frac{x + 4}{x^2 + x} = \frac{x + 4}{x(x + 1)} = \frac{A}{x} + \frac{B}{x + 1},$$

where we find $A = 4$, $B = -3$. In summary then, we find:

$$\frac{x^3 + 4}{x^2 + x} = x - 1 + \frac{4}{x} - \frac{3}{x + 1}$$

4 Factoring by Division

Consider the curious equation with a special n -degree polynomial

$$x^n - a^n = 0,$$

where a is an arbitrary real constant. In the most general case, there are n complex solutions to the above, which may or may not be difficult to come by. Regardless of n though, we *can* be sure that $x_0 = a$ is a valid real solution.

With a solution on hand, it's instructive to factor x_0 from the left-hand expression to check if anything interesting happens, i.e.

$$\frac{x^n - a^n}{x - a} = ?$$

Setting this up, we have:

$$\begin{array}{r} x - a \overline{) x^n - a^n} \end{array}$$

Without specifying n , it's not clear where the division process ought to terminate. Carrying out the division process *anyway*, we find, after four steps:

$$\frac{x^n - a^n}{x - a} = x^{n-1} + a^1 x^{n-2} + a^2 x^{n-3} + a^3 x^{n-4} + \frac{a^4 x^{n-4} - a^n}{x - a}$$

To be prudent, the maximum number of division steps should not exceed the degree number n , otherwise the exponent on x becomes negative.

Tidy up the equation by multiplying $x - a$ into each side:

$$x^n - a^n = (x - a) \left(x^{n-1} + a^1 x^{n-2} + a^2 x^{n-3} + a^3 x^{n-4} \right) + \left(a^4 x^{n-4} - a^n \right)$$

Of course, there was no real reason to stop the division process at four steps, so the above must also be true for j steps:

$$x^n - a^n = (x - a) \left(x^{n-1} + a^1 x^{n-2} + a^2 x^{n-3} + \dots + a^{j-1} x^{n-j} \right) + \left(a^j x^{n-j} - a^n \right)$$

By choosing $j = n$, the remainder term vanishes entirely, leaving

$$x^n - a^n = (x - a) \left(x^{n-1} + a^1 x^{n-2} + a^2 x^{n-3} + \dots + a^{n-1} \right). \quad (6.8)$$

This we'll take as the answer to the curious factoring problem. Starting with the polynomial $x^n - a^n$, solutions other than $x = a$ are contained in another polynomial of order $n - 1$ given by the above.

Sigma Notation

that replaces a^n with a :

In order to avoid always writing Equation (6.8) as a long polynomial basted with exponents, we switch to sigma notation as follows:

$$x^n - a = \left(x - a^{1/n} \right) \left(\sum_{k=1}^n a^{(k-1)/n} x^{n-k} \right) \quad (6.10)$$

$$x^n - a^n = (x - a) \left(\sum_{k=1}^n a^{k-1} x^{n-k} \right) \quad (6.9)$$

Problems

Sometimes it's convenient to work the modified version Problem 4

Factor:

$$x^3 - 8$$

Step 1: Identify variables:

$$n = 3$$

$$a = 2$$

Step 2: Write the factored expression in summation notation:

$$x^3 - 8 = (x - 2) \left(\sum_{k=1}^3 2^{k-1} x^{3-k} \right)$$

Step 3: Simplify:

$$x^3 - 8 = (x - 2) (x^2 + 2x + 4)$$

Problem 5

Factor:

$$x^4 - 9$$

Step 1: Identify variables:

$$n = 4$$

$$a = 3$$

Step 2: Write the factored expression in summation notation:

$$x^4 - 9 = (x - \sqrt{3}) \left(\sum_{k=1}^4 9^{(k-1)/4} x^{4-k} \right)$$

Step 3: Simplify:

$$x^4 - 9$$

$$= (x - \sqrt{3}) (x^3 + \sqrt{3}x^2 + 3x + 3\sqrt{3})$$

$$= (x - \sqrt{3}) (x^2 (x + \sqrt{3}) + 3 (x + \sqrt{3}))$$

$$= (x - \sqrt{3}) (x + \sqrt{3}) (x^2 + 3)$$

5 Recursive Sequences

Equation (6.8) representing the ‘curious identity’ lends to a variety of uses beyond factoring. Here we develop the notion of a *recursive sequence*, which in essence, is a sequence of numbers that uses itself to extend itself.

5.1 Applied Polynomial Division

By making the substitution $a = (-1/x)^n$, we use Equation (6.10) to write

$$\frac{x^n - (-1/x)^n}{x + 1/x} = \sum_{k=1}^n (-1)^{k-1} x^{n+1-2k},$$

where the right-hand sum is a sequence depending solely on x and n .

Choosing $n = 1, n = 2, n = 3$, and so on, a simple-enough pattern emerges:

$$\frac{x^n - (-1/x)^n}{x + 1/x} = \begin{cases} n = 1 : x^0 \\ n = 2 : x^1 - x^{-1} \\ n = 3 : x^2 - x^0 + x^{-2} \\ n = 4 : x^3 - x^1 + x^{-1} - x^{-3} \\ n = 5 : x^4 - x^2 + x^0 - x^{-2} + x^{-4} \\ n = 6 : x^5 - x^3 + x^1 - x^{-1} + x^{-3} - x^{-5} \end{cases}$$

Labeling the n th result as C_n , we equivalently write

$$C_n = \frac{x^n - (-1/x)^n}{x + 1/x} \quad (6.11)$$

$$= \begin{cases} C_1 = 1 \\ C_2 = x^1 - x^{-1} \\ C_3 = -C_1 + x^2 + x^{-2} \\ C_4 = -C_2 + x^3 - x^{-3} \\ C_5 = -C_3 + x^4 + x^{-4} \\ C_6 = -C_4 + x^5 - x^{-5} \end{cases}$$

Recursion Relations

By inspection of the above, the coefficients C_n are subject to *recursion relations*:

$$C_n = -C_{n-2} + x^{n-1} - x^{-(n-1)} \quad (6.12)$$

$$C_{n+1} = -C_{n-1} + x^n + x^{-n} \quad (6.13)$$

5.2 Large-n Recursion

Supposing we choose any even-valued n , the coefficient C_n and its next neighbor relate by the recursion relations (6.12), (6.13). The pair of these begs the ratio

$$R = \frac{C_{n+1}}{C_n} = \frac{-C_{n-1} + x^n + x^{-n}}{-C_{n-2} + x^{n-1} - x^{-(n-1)}}.$$

Within this ratio, let us examine the quantities x^n, x^{-n} with n growing very large. Regardless of whether x is less than one or greater than one (but not equal to one), either x^n or x^{-n} will grow very large, whereas the other will grow very small.

Taking the case with $x > 1$, then x^{-n} and $x^{-(n-1)}$ become negligible, and we find

$$R \approx x \left(\frac{-C_{n-1} + x^n}{-x \cdot C_{n-2} + x^n} \right) \approx x,$$

suggesting that, for large x^n :

$$C_{n+1} \approx x \cdot C_n$$

Taking $x < 1$ instead, the same reasoning boils down to, for small x^n :

$$C_{n+1} \approx \frac{-C_n}{x}$$

6 Lucas Numbers

In deriving Equation (6.11), we managed to avoid specifying the variable x . While we're free to mess with x directly, it's more interesting to direct this freedom into the C_2 coefficient.

Choosing the most nontrivial case of $C_2 = 1$, we have

$$1 = x - \frac{1}{x},$$

which forces x to be found by the quadratic equation. There are two solutions $x_1 = \phi$, $x_2 = \psi$ to the above, obeying

$$\begin{aligned}\phi \cdot \psi &= -1 \\ \phi + \psi &= 1.\end{aligned}$$

Then, making the association

$$\begin{aligned}x &= \phi \\ 1/x &= -\psi,\end{aligned}$$

the coefficients C_n represented in Equation (6.11) specialize to

$$F_n = \frac{\phi^n - \psi^n}{\phi - \psi}. \quad (6.14)$$

The terms F_n are labeled to foreshadow their formal name.

In terms of ϕ, ψ , the above occurs in list form as

$$F_n = \begin{cases} F_1 = 1 \\ F_2 = \phi + \psi \\ F_3 = -F_1 + \phi^2 + \psi^2 \\ F_4 = -F_2 + \phi^3 + \psi^3 \\ F_5 = -F_3 + \phi^4 + \psi^4 \\ F_6 = -F_4 + \phi^5 + \psi^5 \end{cases},$$

or as a recursion relation,

$$F_n = -F_{n-2} + L_{n-1}. \quad (6.15)$$

6.1 Lucas Generating Formula

The terms

$$L_n = \phi^n + \psi^n \quad (6.16)$$

are called *Lucas numbers*. These are a bit tricky to evaluate, but nonetheless with some grit one can produce:

$$\begin{aligned}L_1 &= \phi^1 + \psi^1 = 1 \\ L_2 &= \phi^2 + \psi^2 = (\phi + \psi)^2 - 2\phi\psi = L_1^2 + 2 \\ L_3 &= \phi^3 + \psi^3 = \phi^2 + 1 + \psi^2 = L_1 + L_2 \\ L_4 &= \phi^4 + \psi^4 = (\phi^2 + \psi^2)^2 - 2\phi^2\psi^2 = L_2^2 - 2 \\ L_5 &= \phi^5 + \psi^5 = \phi^4 + \phi^2 + 1 + \psi^2 + \psi^4 = L_3 + L_4 \\ L_6 &= \phi^6 + \psi^6 = (\phi^3 + \psi^3)^2 - 2\phi^3\psi^3 = L_3^2 + 2\end{aligned}$$

Evidently, the pattern in L_n splits between odd and even channels

$$\begin{aligned}L_{n \text{ odd}} &= L_{n-1} + L_{n-2} \\ L_{n \text{ even}} &= L_{n/2}^2 - 2 \cdot (-1)^{n/2},\end{aligned}$$

where explicitly:

$$\begin{aligned}L_0 &= 2 \\ L_1 &= 1 \\ L_2 &= 1^2 + 2 = 3 \\ L_3 &= L_2 + L_1 = 4 \\ L_4 &= L_2^2 - 2 = 3^2 - 2 = 7 \\ L_5 &= L_4 + L_3 = 7 + 4 = 11 \\ L_6 &= L_3^2 + 2 = 4^2 + 2 = 18 \\ L_7 &= L_6 + L_5 = 18 + 11 = 29\end{aligned}$$

Recursion Relations

Since the equation for $L_{n \text{ odd}}$ makes reference to the *two* previous terms, it suffices to write

$$L_n = L_{n-1} + L_{n-2} \quad (6.17)$$

as a single formula applying to both odd and even Lucas numbers.

6.2 Lucas Sequence

(6.16) Listing the Lucas numbers in order gives the *Lucas sequence*:

$$\{L\} = \{2, 1, 3, 4, 7, 11, 18, 29, \dots\}$$

7 Fibonacci Numbers

In discovering the Lucas numbers, the intermediate relation

$$F_n = -F_{n-2} + L_{n-1}$$

emerged before focusing on L_n . In the above, F_n is given by

$$F_n = \frac{\phi^n - \psi^n}{\phi - \psi},$$

and ϕ, ψ are solutions to $1 = x - 1/x$. With the result for L_n in hand, we can use the above to generate the *Fibonacci numbers*:

$$\begin{aligned} F_1 &= 1 \\ F_2 &= 1 \\ F_3 &= -1 + L_2 = 2 \\ F_4 &= -1 + L_3 = 3 \\ F_5 &= -2 + L_4 = 5 \\ F_6 &= -3 + L_5 = 8 \\ F_7 &= -5 + L_6 = 13 \\ F_8 &= -8 + L_7 = 21 \end{aligned}$$

Note that the Fibonacci numbers follow a recursion relation analogous to equation (6.17), namely

$$F_n = F_{n-1} + F_{n-2}. \quad (6.18)$$

Despite its obvious truth, the proof of the above is reserved for the more general treatment of the Lucas-Fibonacci system (see below).

7.1 Fibonacci Sequence

Listing the Fibonacci numbers in order gives the *Fibonacci sequence*:

$$\{F\} = \{1, 1, 2, 3, 5, 8, 13, 21, \dots\}$$

7.2 Negative Fibonacci Numbers

The calculation that gives rise to the Fibonacci sequence can be repeated with a slight tweak that involves swapping x for $1/x$. This gives rise to a modified generating formula

$$\tilde{F} = \frac{(1/x)^n - (-x)^n}{1/x + x} = \frac{(-\psi)^n - (-\phi)^n}{\phi - \psi},$$

inviting a similar algebraic puzzle to the one solved already.

Working this formula carefully, we find

$$\begin{aligned} \tilde{F}_1 &= 1 = F_1 \\ \tilde{F}_2 &= -\frac{(\phi + \psi)(\phi - \psi)}{(\phi - \psi)} = -1 = -F_2 \\ \tilde{F}_3 &= \frac{(\phi - \psi)(\phi^2 + \phi\psi + \psi^2)}{(\phi - \psi)} = -1 + L_2 = F_3 \\ \tilde{F}_4 &= \frac{((- \psi)^2 + (-\phi)^2)(\phi - \psi)((-\psi) + (-\phi))^{-1}}{(\phi - \psi)} \\ &= -(L_1^2 + 2) = -F_4, \end{aligned}$$

which is enough to see the pattern. Evidently, the even-indexed terms flip sign, where the odd-indexed terms remain the same. The extended Fibonacci sequence thus reads:

$$\{F\} = \{\dots, -8, 5, -3, 2, -1, 1, 0, 1, 1, 2, 3, 5, 8, \dots\}$$

7.3 Solving for x

Interestingly, we've made it this far without explicitly needing the numerical values of $x_1 = \phi, x_2 = \psi$. Recalling these are defined as solutions to

$$1 = x - \frac{1}{x},$$

it's straightforward to find

$$\begin{aligned} x_1 = \phi &= \frac{1 + \sqrt{5}}{2} \approx 1.618034\dots \\ x_2 = \psi &= -\frac{1}{\phi} = \frac{1 - \sqrt{5}}{2} \approx -0.618034\dots \end{aligned}$$

Golden Ratio

The constant

$$\phi = \frac{1 + \sqrt{5}}{2} \approx 1.618034\dots$$

is the famed *golden ratio*. Little ado is made of this number in the fundamental sciences, but plenty of attention is given to this number in areas pertaining to graphics, architecture, and biology.

Without using symbols, the n th Lucas or Fibonacci number can be straightforwardly expressed:

$$\begin{aligned} L_n &= \frac{(1 + \sqrt{5})^n + (1 - \sqrt{5})^n}{2^n} \\ F_n &= \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n \sqrt{5}} \end{aligned}$$

7.4 General L-F Numbers

The Lucas-Fibonacci rabbit-hole was entered by setting $C_2 = 1$ as it occurs as Equation (6.11). Of course, this can all be generalized by setting C_2 to an arbitrary constant p , leading to a generalized Lucas-Fibonacci regime characterized by

$$C_2 = p = x - \frac{1}{x},$$

having solutions x_1, x_2 obeying

$$\begin{aligned} x_1 \cdot x_2 &= -1 \\ x_1 + x_2 &= p. \end{aligned}$$

Pursuing this, we stumble upon a new recursion statement analogous to Equation (6.15), namely

$$C_n = -C_{n-2} + \tilde{L}_{n-1}, \quad (6.19)$$

where $\tilde{L}$ is a generalized Lucas number

$$\tilde{L}_n = x_1^n + x_2^n,$$

obeying the recursion relation

$$\tilde{L}_n = p\tilde{L}_{n-1} + \tilde{L}_{n-2}. \quad (6.20)$$

Recursion Relations

To derive a robust recursion relation, let us write three instances of Equation (6.15) based on respective indices n , $n+1$, and $n+2$:

$$\begin{aligned} C_n &= -C_{n-2} + \tilde{L}_{n-1} \\ C_{n+1} &= -C_{n-1} + \tilde{L}_n \\ C_{n+2} &= -C_n + \tilde{L}_{n+1} \end{aligned}$$

Multiply the first equation by a factor of -1 , and the second by a factor of $-p$

$$\begin{aligned} -C_n &= C_{n-2} - \tilde{L}_{n-1} \\ -pC_{n+1} &= pC_{n-1} - p\tilde{L}_n \\ C_{n+2} &= -C_n + \tilde{L}_{n+1} \end{aligned}$$

Next, take the sum of all three equations and regroup terms:

$$\begin{aligned} (C_{n+2} - pC_{n+1} - C_n) + (C_n - pC_{n-1} - C_{n-2}) \\ = (\tilde{L}_{n+1} - p\tilde{L}_n - \tilde{L}_{n-1}) \end{aligned}$$

The right side is identically zero by Equation (6.20). The rest can only be true if

$$C_n = pC_{n-1} + C_{n-2}, \quad (6.21)$$

which we take as the generalized recursion relation. In the special case $p = 1$, the above reduces to the Fibonacci case represented by Equation (6.18).

Modified Seed

The *seed* value $C_2 = p$ has direct bearing on the set of Lucas-Fibonacci-like numbers that emerge from the analysis. In general, the solutions x depend on p such that

$$x = \frac{p}{2} \pm \frac{1}{2}\sqrt{p^2 + 4},$$

where the special case $p = 1$ was studied in detail above. From this perspective, we see there is a continuous family of Lucas-Fibonacci numbers.

The first number in any Fibonacci-like sequence is always $C_1 = 1$, and the second number is always $C_2 = p$. Using the generalized recursion relation (6.21), we easily find the rest:

$$\begin{aligned} \{F_{p=1}\} &= \{1, 1, 2, 3, 5, 8, 13, 21, \dots\} \\ \{F_{p=2}\} &= \{1, 2, 5, 12, 29, 70, 169, 408, \dots\} \\ \{F_{p=3}\} &= \{1, 3, 10, 33, 109, 360, 1189, 3927, \dots\} \\ \{F_{p=4}\} &= \{1, 4, 17, 72, 305, 1292, 5473, 23184, \dots\} \end{aligned}$$

Chapter 7

Geometric Series

1 Introduction

An important identity that arises from playing with polynomial division is

$$x^n - a^n = (x - a) \left(\sum_{k=1}^n a^{k-1} x^{n-k} \right),$$

which can generate many handy results by choosing the proper a and proper n that fit a given situation.

1.1 Geometric Series

The derivation of an equally-important identity can begin by considering the ratio

$$\frac{x^n - 1}{x - 1}$$

for various values of n , which can be studied by taking the $a = 1$ -case of the above. Expanding out the cases $n = 1$, $n = 2$, $n = 3$, etc., we find

$$\begin{aligned}\frac{x^2 - 1}{x - 1} &= 1 + x \\ \frac{x^3 - 1}{x - 1} &= 1 + x + x^2 \\ \frac{x^4 - 1}{x - 1} &= 1 + x + x^2 + x^3,\end{aligned}$$

which suggests for arbitrary n :

$$\frac{x^n - 1}{x - 1} = 1 + x + x^2 + \cdots + x^{n-1} = \sum_{k=1}^n x^{k-1}$$

Reshuffling to put all n -dependence on the right, we evidently find:

$$\frac{1}{1 - x} = \sum_{k=1}^n x^{k-1} + \frac{x^n}{1 - x} \quad (7.1)$$

Convergence

Note that the right side contains x raised to steadily increasing exponents up to x^n . By letting n become arbitrarily large, the sum blows up to infinity unless we restrict the absolute value of x to be less than one. In such a case, the above converges to the *geometric series*,

$$\frac{1}{1 - x} = 1 + x + x^2 + x^3 + \cdots = \sum_{k=0}^{\infty} x^k \quad (7.2)$$

provided $|x| < 1$.

Basketball Bouncing Forever

A basketball is dropped from 10 feet and bounces up 6 feet. On each bounce, the ball recovers $3/5$ of its previous height. Bouncing forever, what is the total distance traveled by the ball?

Step 1: Add up the total distance accumulated during each movement downward:

$$D_1 = 10 \cdot \left(1 + \frac{3}{5} + \left(\frac{3}{5} \right)^2 + \cdots \right)$$

Step 2: Add up the total distance accumulated during each movement upward:

$$D_2 = 6 \cdot \left(1 + \frac{3}{5} + \left(\frac{3}{5} \right)^2 + \cdots \right)$$

Step 3: Compare each infinite sequence to the geometric series, and find:

$$1 + \frac{3}{5} + \left(\frac{3}{5}\right)^2 + \cdots = \frac{1}{1 - 3/5} = \frac{5}{2}$$

Step 4: Assemble the total distance moved in feet:

$$D_1 + D_2 = 10 \cdot \frac{5}{2} + 6 \cdot \frac{5}{2} = 40$$

2 Alternate Derivations

Polynomial Division Method

The most brutal way to derive the geometric series is to perform polynomial division on the quantity $1/(1-x)$:

$$\begin{array}{r}
 1-x \overline{) \begin{array}{r} 1 \\ 1-x \\ x \\ x-x^2 \\ x^2 \\ x^2-x^3 \\ x^3 \\ x^3-x^4 \\ x^4 \end{array}}
 \end{array}$$

After a few terms in, it's clear that such an exercise leads to Equation (7.1) again.

The G-Shortcut

If all you remember is the infinite version of the geometric series but not the finite one, let

$$G = 1 + x + x^2 + x^3 + \cdots + x^n,$$

and multiply through by x :

$$xG = x + x^2 + x^3 + \cdots + x^{n+1}$$

Next, take the difference $G - xG$ and divide out $(1-x)$ from the left to get

$$G = \frac{1 - \cancel{x} + \cancel{x} - \cancel{x^2} + \cancel{x^2} + \cdots - x^{n+1}}{1-x},$$

and simplify to recover Equation (7.1):

$$1 + x + x^2 + x^3 + \cdots + x^n = \frac{1 - x^{n+1}}{1-x}$$

Number Line Method

Consider the real numbers within the domain $[1 : 2]$, and divide the interval between 1 and 2 into n equally-sized bins. From the left, we can locate the upper boundary of each bin:

$$\begin{aligned}
 \text{first bin: } & 1 + 1/n \\
 \text{second bin: } & 1 + 2/n \\
 n-2\text{th bin: } & 1 + (n-2)/n \\
 n-1\text{th bin: } & 1 + (n-1)/n
 \end{aligned}$$

Note that the same locations can be listed from the right:

$$\begin{aligned}
 \text{first bin: } & 2 - (n-1)/n \\
 \text{second bin: } & 2 - (n-2)/n \\
 n-2\text{th bin: } & 2 - 2/n \\
 n-1\text{th bin: } & 2 - 1/n
 \end{aligned}$$

From these, you can check that the bin representations are equivalent.

To go further, consider the real numbers within the domain $[1 + 1/n : 1 + 2/n]$, which are the two boundaries of the second bin. Divide this interval into n equal 'new' bins. From the left, we can locate the upper boundary of the first new bin as

$$\text{first bin: } 1 + \frac{1 + 1/n}{n} = 1 + \frac{1}{n} + \frac{1}{n^2}.$$

From the right, locate the upper boundary of the second new bin:

$$\text{second bin: } 2 - \frac{n-2}{n} - \frac{n-2}{n^2}$$

Supposing some value z lies within the second new bin, it follows that

$$1 + \frac{1}{n} + \frac{1}{n^2} < z < 2 - \frac{n-2}{n} - \frac{n-2}{n^2}.$$

While the above is sufficient to continue, it's worthwhile to simplify the right side to get

$$1 + \frac{1}{n} + \frac{1}{n^2} < z < 1 + \frac{1}{n} + \frac{2}{n^2}.$$

There may be enough to spot a pattern. To be sure, consider the real numbers within the domain

$$\left[1 + \frac{1}{n} + \frac{1}{n^2} : 1 + \frac{1}{n} + \frac{2}{n^2} \right].$$

Divide this into n bins and find the boundaries of the second 'new new' bin, which results in

$$L < z < R,$$

where:

$$\begin{aligned}
 L &= 1 + \frac{1}{n} + \frac{1}{n^2} + \frac{1}{n^3} \\
 R &= 2 - \frac{n-2}{n} - \frac{n-2}{n^2} - \frac{n-2}{n^3} \\
 &= 1 + \frac{1}{n} + \frac{1}{n^2} + \frac{2}{n^3}
 \end{aligned}$$

Looking at the L and R terms above, it's clear what will happen if we execute q iterations of this game, namely, each sum accumulates a new term with increasing powers of n in the denominator up to $1/n^q$. The variable z gets squeezed into a smaller interval.

Furthermore, note that R can be rewritten in terms of L via

$$R = 2 - (n - 2) (L - 1) ,$$

which also means

$$L < z < 2 - (n - 2) (L - 1) .$$

Since L is less than the entire right side, we can skip over z and go with

$$L < 2 - (n - 2) (L - 1) ,$$

readily simplifying to

$$L < \frac{1}{1 - 1/n} .$$

Letting $1/n = x$, we finally get something very much like the geometric series:

$$1 + x + x^2 + x^3 + \cdots + x^q < \frac{1}{1 - x}$$

In the limit $q \rightarrow \infty$, this result is indistinguishable from Equation (7.2).

3 Applicataions

Squaring the Geometric Series

It's possible to multiply converging infinite sums together and get a new infinite sum. The simplest of these simply multiplies the geometric series into itself. Doing this carefully, we find

$$\begin{aligned} \left(\frac{1}{1-x}\right)^2 &= 1 + x + x^2 + x^3 + \dots \\ &\quad + x + x^2 + x^3 + x^4 + \dots \\ &\quad + x^2 + x^3 + x^4 + x^5 + \dots, \end{aligned}$$

simplifying to:

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \dots \quad (7.3)$$

As an exercise in brute force algebra, higher powers can be handled as well:

$$\frac{1}{(1-x)^3} = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + \dots \quad (7.4)$$

$$\frac{1}{(1-x)^4} = 1 + 4x + 10x^2 + 20x^3 + 35x^4 + \dots \quad (7.5)$$

Relation to Pascal's Triangle

Pausing a moment on the sequence of coefficients going with the above results, namely

$$\begin{aligned} &\{1, 2, 3, 4, 5, \dots\} \\ &\{1, 3, 6, 10, 15, \dots\} \\ &\{1, 4, 10, 20, 35, \dots\}, \end{aligned}$$

notice these sequences are already present in Pascal's Triangle:

$$\begin{array}{ccccccc} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 2 & 1 & & & & \\ 1 & 3 & 3 & 1 & & & \\ 1 & 4 & 6 & 4 & 1 & & \\ 1 & 5 & 10 & 10 & 5 & 1 & \\ 1 & 6 & 15 & 20 & 15 & 6 & 1 \end{array}$$

Choosing the n th row and reading down-and-right (diagonally) predicts the expansion of $1/(1-x)^n$.

Negative Argument

In the geometric series, setting $x \rightarrow -x$ has the effect of reversing the sign on all odd-powered terms while leaving even-powered terms the same. With this, we get a slew of results for free:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \dots \quad (7.6)$$

$$\frac{1}{(1+x)^2} = 1 - 2x + 3x^2 - 4x^3 + 5x^4 - \dots \quad (7.7)$$

$$\frac{1}{(1+x)^3} = 1 - 3x + 6x^2 - 10x^3 + 15x^4 - \dots \quad (7.8)$$

$$\frac{1}{(1+x)^4} = 1 - 4x + 10x^2 - 20x^3 + 35x^4 - \dots \quad (7.9)$$

These can be predicted by the alternating Pascal Triangle:

$$\begin{array}{ccccccc} 1 & & & & & & \\ 1 & -1 & & & & & \\ 1 & -2 & 1 & & & & \\ 1 & -3 & 3 & -1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ 1 & -5 & 10 & -10 & 5 & -1 & \\ 1 & -6 & 15 & -20 & 15 & -6 & 1 \end{array}$$

Choosing the n th row and reading down-and-right (diagonally) predicts the expansion of $1/(1+x)^n$.

Squared Argument

Consider the sum

$$\frac{1}{1-x^2} = \frac{1}{2} \left(\frac{1}{1-x} + \frac{1}{1+x} \right)$$

for $|x| < 1$. Using Equations (7.3), (7.6) to expand the right side, we find, after simplifying:

$$\frac{1}{1-x^2} = 1 + x^2 + x^4 + x^6 + \dots \quad (7.10)$$

Evidently, replacing $x \rightarrow x^2$ is as straightforward as it looks. All powers are doubled. In a similar way, we can try $x \rightarrow -x^2$ to establish

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \dots \quad (7.11)$$

Note also that the sum of the above results essentially repeats the problem with all powers doubled again:

$$\frac{1}{1-x^4} = 1 + x^4 + x^8 + x^{12} + \dots \quad (7.12)$$

$$\frac{1}{1+x^4} = 1 - x^4 + x^8 - x^{12} + \dots \quad (7.13)$$

Pascal Triangle Checksum

Introducing the shift $z = 1 - x$, the geometric series becomes

$$\frac{1}{z} = 1 + (1 - z) + (1 - z)^2 + (1 - z)^3 + \cdots,$$

which converges for $0 < z < 2$.

For a seemingly roundabout exercise, note that each term on the right is a polynomial $(1 - z)^n$ with all integer powers of n present. If each of these is expanded out, the right side of the equation ends up containing the entirety of Pascal's Triangle based on the quantity $(1 - z)$.

However, Equations (7.6) - (7.9) already claim the columns of the alternating Pascal's Triangle. Using this information to replace the right side entirely gives a curious representation of $1/z$:

$$\frac{1}{z} = \frac{1}{1+z} + \frac{1}{(1+z)^2} + \frac{1}{(1+z)^3} + \cdots \quad (7.14)$$

By this shifty move, the above no longer converges for $0 < z < 2$. Proceed by letting $y = 1/(1+z)$, and the above becomes

$$\frac{1}{1/y - 1} = y + y^2 + y^3 + \cdots,$$

and add 1 to both sides:

$$\frac{1/y}{1/y - 1} = \frac{1}{1 - y} = 1 + y + y^2 + y^3 + \cdots$$

Evidently, Equation (7.14) still embeds the geometric series in the y -variable, which converges for all $|y| < 1$. Since z is dependent on y , the restriction on z is therefore:

$$\left| \frac{1}{1+z} \right| < 1$$

In other words, all $z > 0$ lead to convergence, and so do all $z < -2$.

This allows for some interesting relationships between the rational numbers, particularly neighboring fractions, for instance:

$$\frac{1}{3} = \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \cdots$$

One attains the same result with the geometric series at the point $x = 1/4$.

Reciprocal Binomial Thm.

From polynomial multiplication, recall the binomial theorem

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k,$$

which, as advertised, provides the expansion of $(x + y)^n$ for positive integer n .

Any binomial coefficient is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

which is also the entry in the n th row, k th column of Pascal's Triangle.

We can write a similar theorem for negative exponents by extending Equations (7.3) - (7.9). The first thing to notice is each is an infinite series, having an infinite number of terms, which corresponds to reading diagonally along Pascal's Triangle.

Going from the patterns established, one discovers:

$$\begin{aligned} \frac{1}{(1+x)^n} &= 1 - \binom{n}{1}x + \binom{n+1}{2}x^2 \\ &\quad - \binom{n+2}{3}x^3 + \binom{n+3}{4}x^4 - \cdots \end{aligned}$$

In sigma notation, this reads

$$\frac{1}{(1+x)^n} = \sum_{k=0}^{\infty} (-1)^k \binom{n+k}{k} x^k.$$

The version for $x \rightarrow -x$ does not alternate signs:

$$\frac{1}{(1-x)^n} = \sum_{k=0}^{\infty} \binom{n+k}{k} x^k$$

Multiplying through by another variable y^{-n} and relabeling variables gives a more general result:

$$\frac{1}{(y+x)^n} = \sum_{k=0}^{\infty} (-1)^k \binom{n+k}{k} \frac{x^k}{y^{n+k}} \quad (7.15)$$

The above is most useful when the terms x^k aren't growing to infinity, which is to say x/y is a number less than one. For instance, the $n = 4$ case in this criteria yields the approximation:

$$\frac{1}{(y+x)^4} \approx \frac{1}{y^4} - \frac{4x}{y^5} + \frac{10x^2}{y^6} - \frac{20x^3}{y^7} + \cdots$$

Correspondingly, one has, for $x \rightarrow -x$:

$$\frac{1}{(y-x)^4} \approx \frac{1}{y^4} + \frac{4x}{y^5} + \frac{10x^2}{y^6} + \frac{20x^3}{y^7} + \cdots$$

Repeating Decimals

The geometric series helps make sense of decimal numbers whose digits eventually repeat. Consider a number of the format

$$N = 0.abcd \dots qabcd \dots q ,$$

where the sequence $abcd \dots q$ is Q digits in length. As a sum, N can be written

$$N = \left(\frac{a}{10} + \frac{b}{100} + \frac{c}{1000} + \dots + \frac{q}{10^Q} \right) \times \left(1 + \frac{1}{10^Q} + \frac{1}{10^{2Q}} + \frac{1}{10^{3Q}} + \dots \right) ,$$

which has been factored into a product of two terms.

For the first term we can define the shorthand

$$N' = \frac{a}{10} + \frac{b}{100} + \frac{c}{1000} + \dots + \frac{q}{10^Q}$$

as the truncation of N before the sequence repeats. The second term is precisely a geometric series:

$$1 + \frac{1}{10^Q} + \frac{1}{10^{2Q}} + \frac{1}{10^{3Q}} + \dots = \frac{1}{1 - 10^{-Q}}$$

Reconstituting N from these items, we have something that allows repeating decimals to be written in closed form:

$$N = \frac{N'}{1 - 10^{-Q}}$$

To have an example, the decimal $0.123123123 \dots$ has $N' = 0.123$ with $Q = 3$:

$$0.123123123 \dots = \frac{0.123}{1 - 10^{-3}} = \frac{123}{1000 - 1} = \frac{123}{999}$$

For another example, the special case $N' = 0.9$ with $Q = 1$ tells us

$$0.999 \dots = \frac{0.9}{1 - 10^{-1}} = \frac{9}{10 - 1} = \frac{9}{9} = 1 ,$$

which means $0.999 \dots$ repeating forever is indistinguishable from 1:

$$0.999\bar{9} = 1$$

Infinite Sum Analysis

The geometric series can help make sense of infinite sums that, at face value, don't appear to be penetrable. To demonstrate, consider the infinite sum

$$A = \sum_{k=0}^{\infty} \frac{k}{2^k}$$

The first term in the sum is identically zero, so it's harmless to start the index at one instead of zero:

$$A = \sum_{k=1}^{\infty} \frac{k}{2^k}$$

Next let $n = k - 1$ and the sum becomes

$$A = \sum_{n=0}^{\infty} \frac{n+1}{2^{n+1}} = \frac{1}{2} \sum_{n=0}^{\infty} \frac{n}{2^n} + \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{2^n} .$$

On the right, the first sum is simply A again. The second sum is a geometric series that resolves to one. Thus

$$A = \frac{1}{2} A + 1 ,$$

which can only mean

$$A = 2 .$$

The same trick works on harder sums. For instance, suppose

$$B = \sum_{k=0}^{\infty} \frac{k^2}{2^k} .$$

Observing that the first term in the sum is zero, and using the same substitution $n = k - 1$ leads to

$$B = \frac{1}{2} \sum_{n=0}^{\infty} \frac{n^2 + 2n + 1}{2^n} = \frac{B}{2} + A + 1 ,$$

which is only solved by

$$B = 6 .$$

4 Zeno's Paradox

An ancient 'paradox' originating in Greece began with Zeno of Elia, as recalled by Aristotle:

That which is in locomotion must arrive at the half-way stage before it arrives at the goal.

This sounds fine, but then the ancient Greeks take the argument off the rails:

In a race, the quickest runner can never overtake the slowest, since the pursuer must first reach the point whence the pursued started, so that the slower must always hold a lead.

According to Zeno, to reach a destination, an object must go half-way first, but to reach the half-way point, it has to reach the quarter-way point, and so on. The object in turn may never reach its destination, and even worse, it's not clear where the object gets stuck, or if the motion ever starts at all.

Linear Motion

Had the Greeks known about the geometric series, particularly the notion of convergence, then maybe there would have been no paradox.

Consider the one-dimensional motion of any object with constant velocity V that takes time T to move distance X , or

$$X = VT .$$

Spatial Sum

To pose the problem as the Greeks may have, suppose that the interval X were divided into sections of decreasing size starting with $X/2$, and then $X/4$, $X/8$, $X/16$, etc. Zeno claims that their sum can't tally to X , however we boldly

write:

$$\begin{aligned} \text{Spatial sum} &= \frac{X}{2} + \frac{X}{4} + \frac{X}{8} + \frac{X}{16} + \cdots \\ &= \frac{X}{2} \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \right) \end{aligned}$$

The parenthesized sum is a geometric series equivalent to $1/(1 - 1/2) = 2$, leaving

$$\text{Spatial sum} = X ,$$

simple as that. That is, the total distance X to somewhere really *is* the infinite sum of ever-halving lengths as described.

Temporal Sum

The paradox can be resolved in the time variable as well. For this, relate the distances $X/2$, $X/4$, $X/8$, etc. to the time required to traverse each:

$$X/2 = Vt_1$$

$$X/4 = Vt_2$$

$$X/8 = Vt_3$$

$$X/2^j = Vt_j$$

Then, the total time is

$$\text{Temporal sum} = \sum_{j=1}^{\infty} t_j = \frac{X}{V} \sum_{j=1}^{\infty} \left(\frac{1}{2^j} \right) ,$$

and the remaining sum is the same as above and resolves to one. In conclusion we find

$$\text{Temporal sum} = \frac{X}{V} = T$$

and no burden of a paradox.

Part II

Pre-Calculus

Chapter 8

Trigonometry

1 Angles and Triangles

Trigonometry is the mathematical study of triangles. A triangle is defined as the intersection of three non-parallel straight lines as shown in Figure 8.1. The place where two lines intersect is called a *vertex*, and a triangle has three vertices. At each vertex, the triangle has an *interior* angle A , B , C . The vertex-to-vertex distance is a *side* of the triangle, AB , BC or CA , respectively.

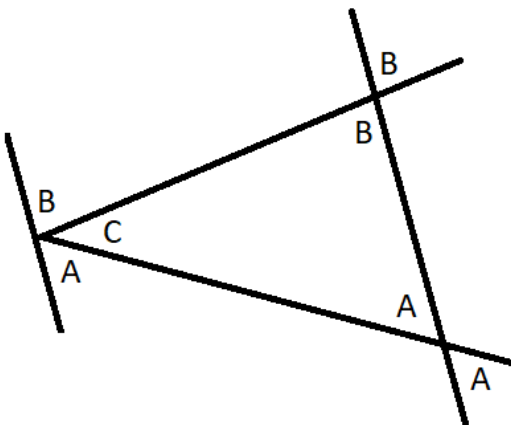


Figure 8.1. Triangle made from three lines.

Using the rules of Euclidean geometry, the angles A , B are portrayed with their *opposite* angles outside the triangle. Also from Euclidean geometry, we can imagine translating the segment AB to the left (as shown) until touching vertex C . From here, we see that the sum of angles $B + C + A$ is equivalent to the the angle represented by a straight line.

1.1 Angles

Any angle A , B , C , etc. is generally represented by the symbol θ (Greek *theta*), a parameter that must be a *dimensionless quantity*. That is, θ must be a pure number such as 3 or -17.5 , but never some measure of meters, seconds, or pounds.

Degrees and Radians

There are two standard units for representing angle, namely *degrees* and *radians*. By convention, a triangle encloses 180 degrees, also written 180° , which is equivalent to π radians:

$$A + B + C = 180^\circ$$

$$A + B + C = \pi$$

From these, we have a pair of unit conversion factors

$$1^\circ = \frac{\pi}{180} \text{ rad} \quad 1 \text{ rad} = \frac{180^\circ}{\pi},$$

which extrapolates to the following:

$$0^\circ = 0 \text{ rad}$$

$$45^\circ = \pi/4 \text{ rad}$$

$$90^\circ = \pi/2 \text{ rad}$$

$$135^\circ = 3\pi/4 \text{ rad}$$

$$360^\circ = 2\pi \text{ rad}$$

The primary domain for angles is represented by

$$0 \leq \theta < 360^\circ$$

$$0 \leq \theta < 2\pi,$$

which is to say that any quantity depending on angle regards 0° and 360° to be synonymous. Or, an angle of 375° is effectively the same as 15° .

1.2 Taxonomy of Triangles

Area of a Triangle

From geometry, we know that the area of any triangle is given by

$$\text{Area} = \frac{\text{Base} \times \text{Height}}{2}.$$

Equilateral Triangle

An *equilateral* triangle has all three sides of the same length. It follows too that all three angles must be the same, particularly 60° or $\pi/3$ rad, regardless of the size of the triangle. An equilateral triangle exhibits three-fold symmetry about its center, which is to say, there are three orientations of an equilateral triangle that appear identical.

Isoceles Triangle

An *isosceles* triangle has two equal sides and two equal angles. The third side and third angle are allowed to be larger or smaller than the other sides and angles. An isosceles triangle exhibits mirror symmetry about a line through the vertex of the two equal sides and the center of the triangle.

Scalene Triangle

A *scalene* triangle has no equal sides, no equal angles, and no symmetry. It's a typical 'unplanned' triangle.

Acute Triangle

An *acute* triangle has all internal angles less than 90° .

Obtuse Triangle

An *obtuse* triangle has any one internal angle greater than 90° .

1.3 Right Triangles

A *right triangle* is any triangle that has two sides meeting at $90^\circ = \pi/2$ rad. Labeling either of the 'unused' angles as θ , the sides of the right triangle take on unique names as shown in Figures 8.2, 8.3.

- The side across from the ninety-degree angle is the *Hypotenuse*.
- The side across from θ is the *Opposite*.
- The side touching θ is the *Adjacent*.

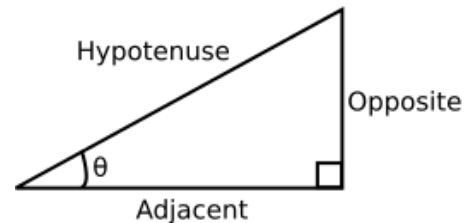


Figure 8.2. Right triangle.

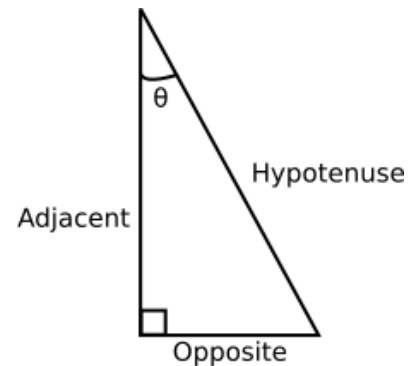


Figure 8.3. Right triangle.

Area of a Right Triangle

For the case of right triangles, it's convenient to associate Base, Height with Opposite, Adjacent (or vice versa):

$$A_{\text{Right}} = \frac{\text{Opposite} \times \text{Adjacent}}{2}$$

1.4 Pythagorean Theorem

The *Pythagorean theorem* is an equation that relates the sides of a right triangle to one another, and happens to also be the backbone equation of trigonometry. Figure 8.4 shows a typical right triangle with hypotenuse c , opposite a , and adjacent b .

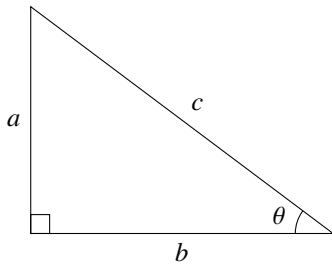


Figure 8.4. Right triangle.

To derive the Pythagorean theorem, imagine a line that extends from the ninety-degree vertex and intersects the hypotenuse at a right angle as shown in Figure 8.5.

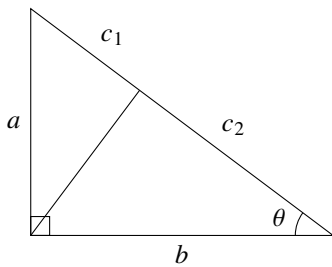


Figure 8.5. Similar right triangles.

With this, the hypotenuse is broken into two segments obeying

$$c_1 + c_2 = c .$$

Then, using similar triangles, we can write two observations:

$$\frac{c_1}{a} = \frac{a}{c}$$

$$\frac{c_2}{b} = \frac{b}{c}$$

Rearrange these and rewrite to get

$$c_1 c = a^2$$

$$c_2 c = b^2 ,$$

and then sum the two equations

$$c (c_1 + c_2) = a^2 + b^2 ,$$

and replace $c_1 + c_2$ with c to finish the job:

$$a^2 + b^2 = c^2 \quad (8.1)$$

Using the triangle side names in place of a, b, c yields the equivalent statement:

$$\text{Opposite}^2 + \text{Adjacent}^2 = \text{Hypotenuse}^2$$

Inverted Pythagorean Theorem

In the above construction, one can find straightforwardly that

$$\frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{c_1 c_2} .$$

As it turns out, the product $c_1 c_2$ has a significance.

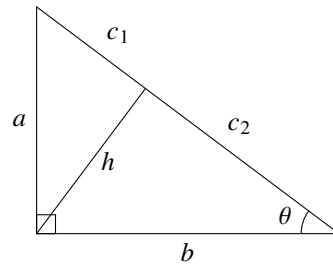


Figure 8.6. Similar right triangles.

Let the line connecting the ninety-degree corner to the hypotenuse be denoted h as shown in Figure 8.6. Using the Pythagorean theorem twice, we write

$$c_1^2 + h^2 = a^2$$

$$c_2^2 + h^2 = b^2 .$$

Add the pair of equations and simplify to find

$$h^2 = c_1 c_2 ,$$

leading to an ‘inverted’ Pythagorean theorem:

$$\frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{h^2} \quad (8.2)$$

1.5 Sine, Cosine, Tangent

On a right triangle, the opposite, adjacent, and hypotenuse can be stacked into ratios. These ratios have designated names:

$$\text{Sine} = \frac{\text{Opposite}}{\text{Hypotenuse}} \quad (8.3)$$

$$\text{Cosine} = \frac{\text{Adjacent}}{\text{Hypotenuse}} \quad (8.4)$$

$$\text{Tangent} = \frac{\text{Opposite}}{\text{Adjacent}} = \frac{\text{Sine}}{\text{Cosine}} \quad (8.5)$$

The ratio of sides of a triangle, i.e. the sine, cosine, or tangent, is equivalent governed by the angle θ formed between the hypotenuse and the adjacent. It's customary to

include the θ -dependence into the notation and create the abbreviations:

$$\text{Sine} = \sin(\theta)$$

$$\text{Cosine} = \cos(\theta)$$

$$\text{Tangent} = \tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)}$$

SohCahToa

A useful mnemonic for recovering Equations (8.3)-(8.5) on the fly is the fictitious name:

SohCahToa

In this, the letter a stands for adjacent, o for opposite, and h for hypotenuse. Meanwhile S is for the sine, C for cosine, and T for tangent. Then, the SocCahToa shorthand expands to:

$$S = o/h$$

$$C = a/h$$

$$T = o/a$$

Fundamental Trigonometric Identity

Immediately from the Pythagorean theorem, we can write the most important equation in trigonometry, known as the *fundamental identity*. Starting with Equations (8.3), (8.4), square each and take the sum:

$$(\sin(\theta))^2 + (\cos(\theta))^2 = \frac{\text{Opposite}^2 + \text{Adjacent}^2}{\text{Hypotenuse}^2}$$

The right side is identically one due to the Pythagorean theorem. On the left is the sum of sine squared and cosine squared, which is conventionally written with the exponent before the parentheses:

$$(\sin(\theta))^2 + (\cos(\theta))^2 = \sin^2(\theta) + \cos^2(\theta)$$

In concise form, the fundamental trigonometric identity reads:

$$\sin^2(\theta) + \cos^2(\theta) = 1 \quad (8.6)$$

1.6 Special Right Triangles

Right Isosceles Triangle

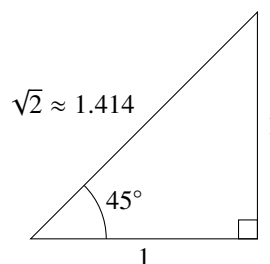


Figure 8.7. Right triangle: isosceles (unit sides)

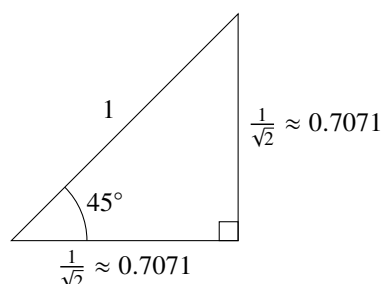


Figure 8.8. Right triangle: isosceles (unit hypotenuse)

Right Pi/8 Triangle

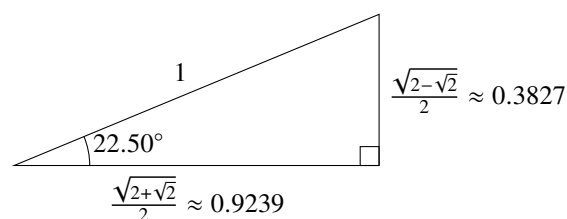


Figure 8.9. Right triangle: 22.5° (opposing angle 67.5° , unit hypotenuse)

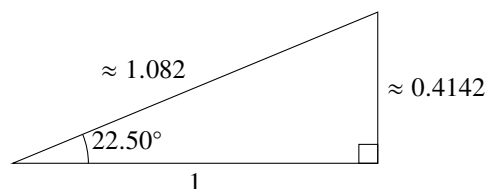


Figure 8.10. Right triangle: 22.5° (opposing angle 67.5° , unit adjacent)

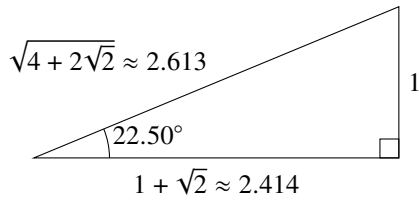


Figure 8.11. Right triangle: 22.5° (opposing angle 67.5° , unit opposite)

Right Pi/16 Triangle

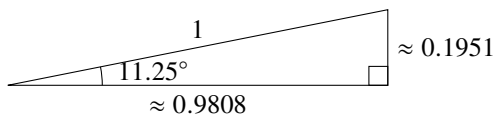


Figure 8.12. Right triangle: 11.25° (opposing angle 78.75° , unit hypotenuse)

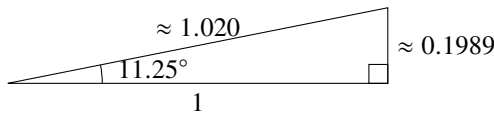


Figure 8.13. Right triangle: 11.25° (opposing angle 78.75° , unit adjacent)

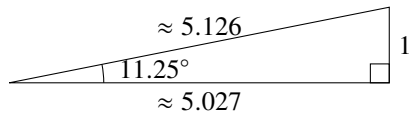


Figure 8.14. Right triangle: 11.25° (opposing angle 78.75° , unit opposite)

Right 30-60-90 Triangle

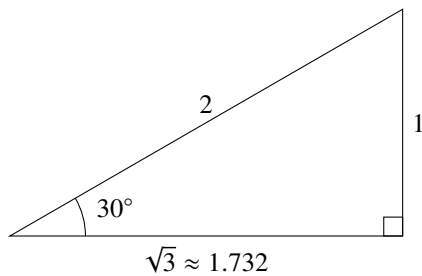


Figure 8.15. Right triangle: 30-60-90

Right 3-4-5 Triangle

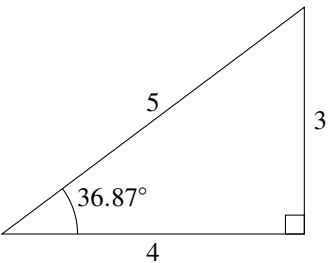


Figure 8.16. Right triangle: 3-4-5

Right 5-12-13 Triangle

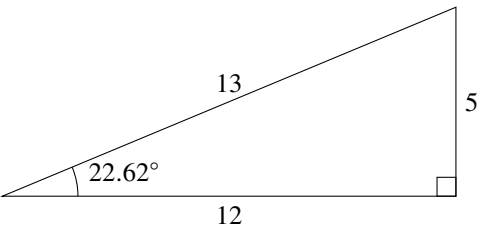


Figure 8.17. Right triangle: 5-12-13

Pythagorean Triplets

A *Pythagorean triplet* is any set of three integers that obey the Pythagorean theorem. In addition to the triplet (3, 4, 5), (5, 12, 13) detailed above, one can also have (but not limited to):

7, 24, 25
8, 5, 17
9, 40, 41

2 Circles

In the Cartesian plane, a circle is most generally described by

$$(x - h)^2 + (y - k)^2 = R^2, \quad (8.7)$$

where the center of the circle is located at (h, k) and the radius is R as depicted in Figure 8.18.

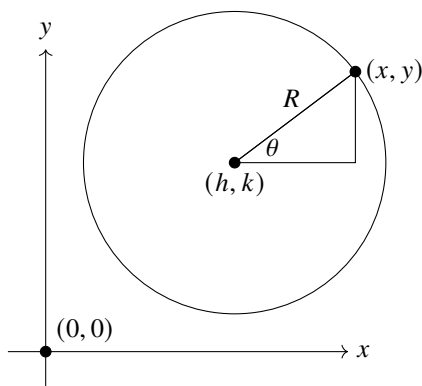


Figure 8.18. Offset circle.

Taking a second look at this construction, notice that the line joining the center of the circle to a point (x, y) on its perimeter is the hypotenuse of a right triangle where θ is defined to ‘rise off’ a line parallel to the x -axis.

Theta Convention

By tradition, the angle θ always ‘opens up’ in the counter-clockwise direction, starting from $\theta = 0$, measured from a ray parallel to the positive x -axis.

2.1 Taxonomy of Circles

As we’ve seen a circle is entirely characterized by its center (h, k) and its radius R .

Diameter

The *diameter* of any shape is distance between its maximally-separated points on its perimeter. On a circle, any point one chooses has a ‘twin’ across the circle precisely distance $2R$ away. It follows that the diameter of the circle is $2r$:

$$\text{Diameter} = 2R$$

Circumference

The *circumference* of a circle is the total length of its perimeter. This is in fact where the definition of π originates:

$$\text{Circumference} = 2\pi R$$

Area

It’s straightforward to show, although not using trigonometry alone, that the area of a circle is

$$A = \pi R^2.$$

Circular Arc Length

In terms of θ , the distance along a circular perimeter is given by

$$S = R\theta,$$

where S is called *arc length*. Let $\theta = 2\pi$ for the arc length to recover the circumference.

Circular Inscription Problem

Problem 1

Let ABC be a triangle with right angle A and hypotenuse $|BC|$ as shown in Figure 8.19. If the inscribed circle of radius R touches the hypotenuse at D , show that:

$$|CD| = \frac{|AC| + |BC| - |AB|}{2}$$

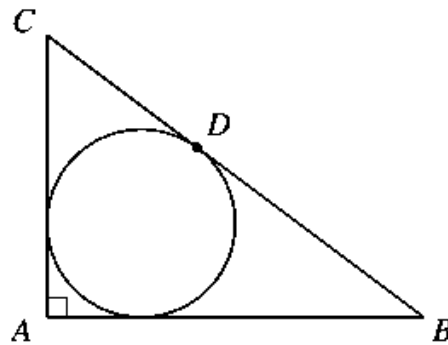


Figure 8.19. Inscribed Circle in Right Triangle

Answer: By inspecting the Figure, there are two ways to write the radius of the circle:

$$R = |AB| - |BD|$$

$$R = |AC| - |CD|$$

Eliminating R tells us

$$|AB| - |BD| = |AC| - |CD| \text{ .}$$

Replace $|BD|$ using

$$|BD| = |BC| - |CD| ,$$

and solve for $|CD|$ to get the answer.

2.2 Parameterized Circle

A triangle having a fixed hypotenuse with continuously adjustable opposite and adjacent sides is all one needs to trace out a circle. By letting θ sweep from 0 to 2π , the endpoint of the hypotenuse, having location (x, y) in the plane, is described by:

$$x(\theta) = h + R \cos(\theta) \quad (8.8)$$

$$y(\theta) = k + R \sin(\theta) \quad (8.9)$$

The above represents the *parameterized* equation of a circle. To quickly recover Equation (8.7), solve for $\cos(\theta)$, $\sin(\theta)$, respectively, and exploit Equation (8.6).

2.3 Unit Circle

A circle centered at the origin with unit radius, i.e. $(h, k) = (0, 0)$ and $R = 1$, is called the *unit circle*. The unit circle is a special case where the adjacent side is equal to $\cos(\theta)$ and the opposite side is equal to $\sin(\theta)$. In the Cartesian plane, the unit circle is:

$$x^2 + y^2 = 1$$

The unit circle is most useful as a data structure to help remember the sine and cosine values of key angles, as shown in Figure 8.20.

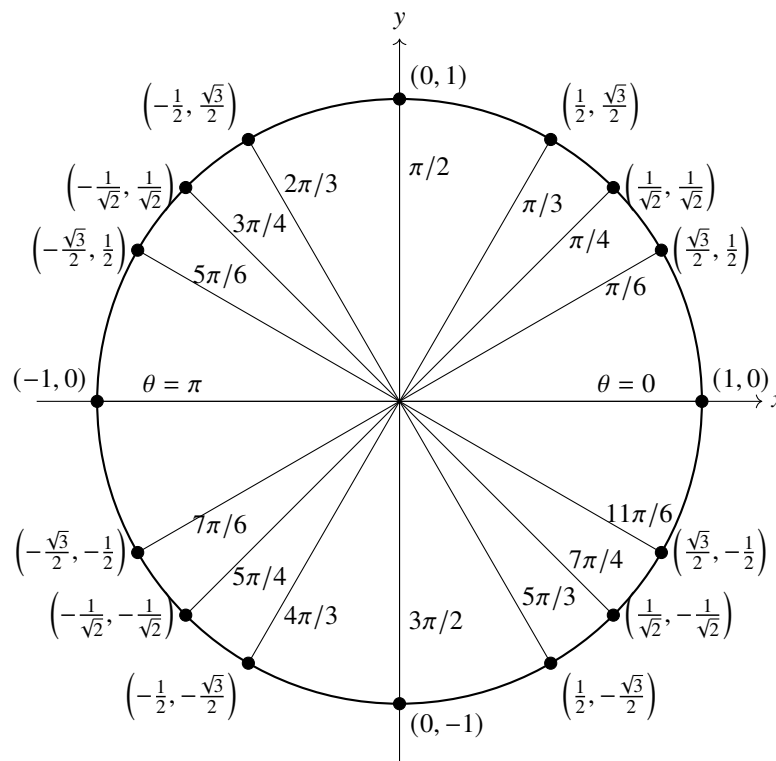


Figure 8.20. Unit circle.

2.4 Tangent Line

In the general sense, a *tangent* line is a straight line that touches a curve (locally) in one place, and the slope of the

curve at the point of contact equals the slope of the line. When it comes to circles, the line is relatively straightforward to analyze.

Consider the unit circle (radius one, centered at the origin) with a point (x_0, y_0) selected somewhere on the perimeter. The slope of the ‘position line’ from the origin to (x_0, y_0) is naturally y_0/x_0 , which is identically $\tan(\theta)$:

$$\tan(\theta) = \frac{y_0}{x_0}$$

The tangent line to the unit circle at (x_0, y_0) has slope $-x_0/y_0$, and is sketched in Figure 8.21. By standard straight line analysis, the equation of the tangent line obeys

$$\frac{y - y_0}{x - x_0} = -\frac{x_0}{y_0}.$$

More concisely, the same equation can be written

$$xx_0 + yy_0 = 1. \quad (8.10)$$

It just happens that the length of the tangent line segment from (x_0, y_0) to its intersection with the x -axis is equal to $\tan(\theta)$. To prove this, note that the line’s intersection with the x -axis occurs at $(1/x_0, 0)$, and then construct the distance

$$\sqrt{(x_0 - 1/x_0)^2 + y_0^2},$$

which simplifies to x_0/y_0 , the definition of $\tan(\theta)$ on the unit circle.

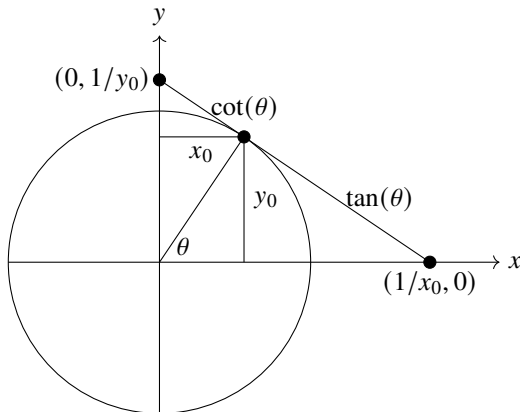


Figure 8.21. Unit circle with tangent line.

2.5 Cotangent

The *cotangent* (of angle θ) is defined as

$$\cot(\theta) = \frac{1}{\tan(\theta)} = \frac{\cos(\theta)}{\sin(\theta)}. \quad (8.11)$$

On the unit circle, $\cot(\theta)$ is the segment of the tangent line extending from $(0, 1/y_0)$ to (x_0, y_0) . To establish this, much like the tangent case, simplify the quantity

$$\sqrt{(1/y_0 - y_0)^2 + x^2},$$

which comes out to the ratio x_0/y_0 , the definition of $\cot(\theta)$ on the unit circle.

2.6 Secant

Continuing with Figure 8.21, the distance from the origin to the point $(1/x_0, 0)$ is called the *secant*. This is not to be confused with a ‘secant line’, which is the extension of a chord through the circle.

To relate the secant to the existing items of trigonometry, observe from the Figure that

$$(\cot(\theta) + \tan(\theta)) \sin(\theta) = \frac{1}{x_0},$$

which, using Equation (8.5) and Equation (8.6), simplifies to

$$\frac{1}{\cos(\theta)} = \frac{1}{x_0}.$$

Since x_0 is already claimed as the cosine of theta, we have:

$$\sec(\theta) = \frac{1}{\cos(\theta)} \quad (8.12)$$

2.7 Cosecant

The distance from the origin to the point $(0, 1/y_0)$ is called the *cosecant*.

Much like the secant case, observe from the Figure that

$$(\cot(\theta) + \tan(\theta)) \cos(\theta) = \frac{1}{y_0},$$

which simplifies to

$$\frac{1}{\sin(\theta)} = \frac{1}{y_0}.$$

Since y_0 is already claimed as the sine of theta, we have:

$$\csc(\theta) = \frac{1}{\sin(\theta)} \quad (8.13)$$

Inverted Pythagorean Theorem

For a proverbial a checksum on Figure 8.21, recall the inverted Pythagorean theorem

$$\frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{h^2} .$$

In this case, identify

$$a = 1/x_0$$

$$b = 1/y_0$$

$$h = \sqrt{x_0^2 + y_0^2} = 1$$

2.8 Periodicity

Due to the confined domain $[0 : 2\pi)$ of the θ -variable, it follows that quantities like $\sin(\theta)$, $\cos(\theta)$, $\tan(\theta)$ are only unique in this interval. It's just fine, however, to feed θ -values outside the standard domain. Before $\theta = 0$ or after $\theta = 2\pi$, everything repeats via

$$\sin(\theta \pm 2n\pi) = \sin(\theta) \quad (8.14)$$

$$\cos(\theta \pm 2n\pi) = \cos(\theta) \quad (8.15)$$

$$\tan(\theta \pm 2n\pi) = \tan(\theta) , \quad (8.16)$$

where n is any integer. This property is called *periodicity*.

2.9 Phase

A *phase shift* occurs when any quantity is added to θ .

Negative Angle

If we replace θ by $-\theta$, the symmetry of the unit circle demands:

$$\sin(-\theta) = -\sin(\theta) \quad (8.17)$$

$$\cos(-\theta) = \cos(\theta) \quad (8.18)$$

$$\tan(-\theta) = -\tan(\theta) \quad (8.19)$$

Phase Shift Pi

A phase shift of π radians jumps exactly across the unit circle. Accordingly, we have:

$$\sin(\theta \pm \pi) = -\sin(\theta) \quad (8.20)$$

$$\cos(\theta \pm \pi) = -\cos(\theta) \quad (8.21)$$

Phase Shift Pi/2

As they're defined, it turns out that $\sin(\theta)$ and $\cos(\theta)$ are related by the phase $\pi/2$:

$$\sin\left(\theta + \frac{\pi}{2}\right) = \cos(\theta) \quad (8.22)$$

$$\cos\left(\theta + \frac{\pi}{2}\right) = -\sin(\theta) \quad (8.23)$$

$$\sin\left(\theta - \frac{\pi}{2}\right) = -\cos(\theta) \quad (8.24)$$

$$\cos\left(\theta - \frac{\pi}{2}\right) = \sin(\theta) \quad (8.25)$$

Similar equations apply when θ is replaced with $-\theta$:

$$\sin\left(\frac{\pi}{2} - \theta\right) = \cos(\theta) \quad (8.26)$$

$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin(\theta) \quad (8.27)$$

3 Trigonometric Identities

It turns out that $\sin(\theta)$, $\cos(\theta)$, and $\tan(\theta)$, along with their reciprocated counterparts, fit into a slew of equations called *trigonometric identities*. In practice, the so-called ‘trig identities’ are bits of algebra that can be used to elaborate on or simplify a given situation.

3.1 Fundamental Trig Identities

The fundamental trigonometric identity first documented as Equation (8.6), namely

$$\sin^2(\theta) + \cos^2(\theta) = 1,$$

can be exploited to yield several more. Divide by $\sin^2(\theta)$ or by $\cos^2(\theta)$ to yield the following:

$$1 + \cot^2(\theta) = \csc^2(\theta) \quad (8.28)$$

$$\tan^2(\theta) + 1 = \sec^2(\theta) \quad (8.29)$$

For an interesting sanity check, take the sum of the two above equations to come up with

$$(\tan(\theta) + \cot(\theta))^2 = \csc^2(\theta) + \sec^2(\theta), \quad (8.30)$$

which is the summary of Figure 8.21.

3.2 Angle-Sum Formulas

Consider the sum of two angles α , β , as they embed in the unit circle as shown in Figure 8.22. The triangle swept out by β has adjacent side $\cos(\beta)$ and opposite side $\sin(\beta)$. Each of these sides is the hypotenuse of a pair of right triangles whose sides are the products denoted in the Figure.

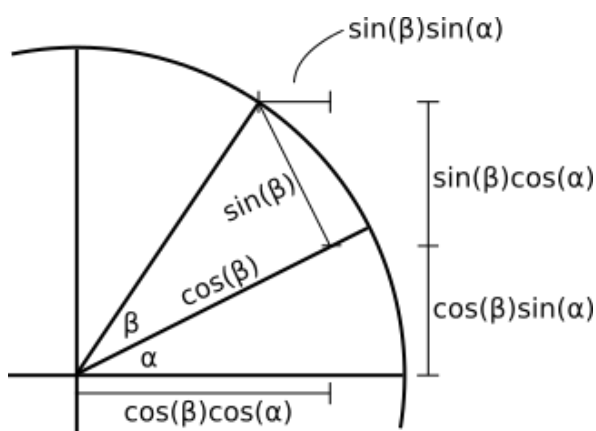


Figure 8.22. Angle-sum analysis.

Expressing $\sin(\alpha + \beta)$, $\cos(\alpha + \beta)$ in terms of the products of individual terms, we find by inspection the *angle-sum formulas*:

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \quad (8.31)$$

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \quad (8.32)$$

Using these two results, we can easily calculate the tangent of $\alpha + \beta$:

$$\tan(\alpha \pm \beta) = \frac{\tan(\alpha) \pm \tan(\beta)}{1 \mp \tan(\alpha)\tan(\beta)} \quad (8.33)$$

3.3 Product Formulas

Starting with the angle-sum formulas (8.31), (8.32), it's straightforward to derive the *product formulas*:

$$2\sin(\alpha)\cos(\beta) = \sin(\alpha + \beta) + \sin(\alpha - \beta) \quad (8.34)$$

$$2\cos(\alpha)\sin(\beta) = \sin(\alpha + \beta) - \sin(\alpha - \beta) \quad (8.35)$$

$$2\cos(\alpha)\cos(\beta) = \cos(\alpha + \beta) + \cos(\alpha - \beta) \quad (8.36)$$

$$2\sin(\alpha)\sin(\beta) = \cos(\alpha - \beta) - \cos(\alpha + \beta) \quad (8.37)$$

3.4 Double-Angle Formulas

Starting with the product formulas, let $\alpha = \beta = \theta$ to derive the *double-angle formulas*:

$$\sin(2\theta) = 2\sin(\theta)\cos(\theta) \quad (8.38)$$

$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta) \quad (8.39)$$

$$\tan(2\theta) = \frac{2\tan(\theta)}{1 - \tan^2(\theta)} \quad (8.40)$$

3.5 Half-Angle Formulas

Starting with equation (8.39), replace the $\sin^2(\theta)$ term and also replace $\theta \rightarrow \theta/2$ to write

$$\cos(\theta) = 2 \cos^2\left(\frac{\theta}{2}\right) - 1.$$

From here, it's a small matter of algebra to generate the *half-angle formulas*:

$$\sin\left(\frac{\theta}{2}\right) = \pm \sqrt{\frac{1}{2} - \frac{\cos(\theta)}{2}} \quad (8.41)$$

$$\cos\left(\frac{\theta}{2}\right) = \pm \sqrt{\frac{1}{2} + \frac{\cos(\theta)}{2}} \quad (8.42)$$

$$\tan\left(\frac{\theta}{2}\right) = \frac{\sin(\theta)}{1 + \cos(\theta)} \quad (8.43)$$

$$\cot\left(\frac{\theta}{2}\right) = \frac{\sin(\theta)}{1 - \cos(\theta)} \quad (8.44)$$

$$\sec\left(\frac{\theta}{2}\right) = \frac{2 \cos(\theta/2)}{1 + \cos(\theta)} \quad (8.45)$$

3.6 Superposition Relationships

Superposition of Sines

Consider the sum $\alpha + \beta$ and difference $\alpha - \beta$ of two angles. Take the sine and cosine, respectively, of each quantity and take their product

$$\begin{aligned} \sin(\alpha + \beta) \cos(\alpha - \beta) = \\ (\sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta)) \times \\ (\cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta)) , \end{aligned}$$

simplifying to, after a bit of work,

$$\sin(\alpha + \beta) \cos(\alpha - \beta) = \frac{\sin(2\alpha)}{2} + \frac{\sin(2\beta)}{2}.$$

Refactor the α, β variables to get the first result:

$$\sin(\alpha) + \sin(\beta) = 2 \sin\left(\frac{\alpha + \beta}{2}\right) \cos\left(\frac{\alpha - \beta}{2}\right) \quad (8.46)$$

Replace β with $-\beta$ to get the second superposition relationship for free:

$$\sin(\alpha) - \sin(\beta) = 2 \sin\left(\frac{\alpha - \beta}{2}\right) \cos\left(\frac{\alpha + \beta}{2}\right) \quad (8.47)$$

These are both called *superposition relationships*.

Superposition of Cosines

Starting with the superposition relationships above, introduce the phase shifts:

$$\alpha \rightarrow \alpha + \pi/2$$

$$\beta \rightarrow \beta - \pi/2$$

Inserting these and simplifying gives two more superposition relationships for the cosine:

$$\cos(\alpha) - \cos(\beta) = -2 \sin\left(\frac{\alpha + \beta}{2}\right) \sin\left(\frac{\alpha - \beta}{2}\right) \quad (8.48)$$

$$\cos(\alpha) + \cos(\beta) = 2 \cos\left(\frac{\alpha - \beta}{2}\right) \cos\left(\frac{\alpha + \beta}{2}\right) \quad (8.49)$$

4 Inverse Trigonometry

For each quantity $\sin(\theta)$, $\cos(\theta)$, $\tan(\theta)$, $\csc(\theta)$, $\sec(\theta)$, $\cot(\theta)$, there exists an *inverse* trigonometric quantity that does the job of ‘solving for’ θ . These are called the arc-sine, arc-cosine, arc-tangent, and so on, defined as follows:

$$\arcsin(\sin(\theta)) = \theta \quad (8.50)$$

$$\arccos(\cos(\theta)) = \theta \quad (8.51)$$

$$\arctan(\tan(\theta)) = \theta \quad (8.52)$$

$$\operatorname{arccsc}(\csc(\theta)) = \theta \quad (8.53)$$

$$\operatorname{arcsec}(\sec(\theta)) = \theta \quad (8.54)$$

$$\operatorname{arccot}(\cot(\theta)) = \theta \quad (8.55)$$

Inverse Trig Notation

Confusingly enough, there is another way to write $\arcsin(\theta)$, $\arccos(\theta)$, etc., using the nomenclature

$$\sin^{-1}(\theta) = \arcsin(\theta)$$

$$\cos^{-1}(\theta) = \arccos(\theta) ,$$

and so on. This overloading of notation does not mean at all, for instance, that the $\arcsin(\theta)$ is equal to the reciprocal of $\sin(\theta)$.

4.1 Inverse Reciprocal Identities

Some handy identities we can establish early are:

$$\arcsin(1/x) = \operatorname{arccsc}(x) \quad (8.56)$$

$$\operatorname{arccsc}(1/x) = \arcsin(x) \quad (8.57)$$

$$\arccos(1/x) = \operatorname{arcsec}(x) \quad (8.58)$$

$$\operatorname{arcsec}(1/x) = \arccos(x) \quad (8.59)$$

$$\arctan(1/x) = \operatorname{arccot}(x) \quad (8.60)$$

$$\operatorname{arccot}(1/x) = \arctan(x) \quad (8.61)$$

To prove any of the above will demonstrate how to handle the rest. Choosing the $\arctan$ case,

$$A = \arctan(1/x)$$

$$B = \operatorname{arccot}(x) ,$$

and then

$$\tan(A) = 1/x$$

$$\cot(B) = x .$$

From this we see $\tan(A) = \tan(B)$, meaning $A = B$, and the proof is done.

4.2 Inverse Triangle Identities

Arccosine

Consider a right triangle with hypotenuse 1, adjacent side x , and opposite side $\sqrt{1-x^2}$. (This is just the unit circle centered at the origin.) From this, we can gather

$$x = \cos(\theta)$$

$$\arccos(x) = \theta$$

$$\sin(\arccos(x)) = \sin(\theta) ,$$

resulting in:

$$\sin(\arccos(x)) = \sqrt{1-x^2} \quad (8.62)$$

Divide through by x to get a second result:

$$\tan(\arccos(x)) = \frac{\sqrt{1-x^2}}{x} \quad (8.63)$$

Arcsine

Now we modify the triangle slightly. Suppose the hypotenuse of another right triangle is 1, and the opposite side is x , making the adjacent equal to $\sqrt{1-x^2}$. From this, we can gather

$$x = \sin(\theta)$$

$$\arcsin(x) = \theta$$

$$\cos(\arcsin(x)) = \cos(\theta) ,$$

resulting in:

$$\cos(\arcsin(x)) = \sqrt{1-x^2} \quad (8.64)$$

Similarly we can establish:

$$\tan(\arcsin(x)) = \frac{x}{\sqrt{1-x^2}} \quad (8.65)$$

Arctangent

Consider a new right triangle with hypotenuse $\sqrt{x^2+1}$, adjacent side 1, and opposite side x . For this case, we have

$$x = \tan(\theta)$$

$$\arctan(x) = \theta ,$$

implying

$$\cos(\arctan(x)) = \cos(\theta)$$

$$\sin(\arctan(x)) = \sin(\theta) .$$

From these, conclude:

$$\cos(\arctan(x)) = \frac{1}{\sqrt{x^2 + 1}} \quad (8.66)$$

$$\sin(\arctan(x)) = \frac{x}{\sqrt{x^2 + 1}} \quad (8.67)$$

Arccosecant

The reciprocal trig quantities are a little harder to analyze. For the arc-cosecant, consider a right triangle with hypotenuse x , opposite side 1, and adjacent side $\sqrt{x^2 - 1}$. Following this, we find:

$$\begin{aligned} x &= \csc(\theta) \\ \operatorname{arccsc}(x) &= \theta, \end{aligned}$$

implying

$$\begin{aligned} \sin(\operatorname{arccsc}(x)) &= \sin(\theta) \\ \cos(\operatorname{arccsc}(x)) &= \cos(\theta). \end{aligned}$$

From these, conclude:

$$\sin(\operatorname{arccsc}(x)) = \frac{1}{x} \quad (8.68)$$

$$\cos(\operatorname{arccsc}(x)) = \frac{\sqrt{x^2 - 1}}{x} \quad (8.69)$$

$$\tan(\operatorname{arccsc}(x)) = \frac{1}{\sqrt{x^2 - 1}} \quad (8.70)$$

Arcsecant

To handle the arc-secant case, swap the role of the opposite and adjacent sides in the right triangle used for the arc-cosecant case:

$$\sin(\operatorname{arcsec}(x)) = \frac{\sqrt{x^2 - 1}}{x} \quad (8.71)$$

$$\cos(\operatorname{arcsec}(x)) = \frac{1}{x} \quad (8.72)$$

$$\tan(\operatorname{arcsec}(x)) = \sqrt{x^2 - 1} \quad (8.73)$$

Arccotangent

To complete the ensemble, consider a right triangle with hypotenuse $\sqrt{x^2 + 1}$, opposite side 1, and adjacent side x . Running through the standard exercise gives three new results:

$$\tan(\operatorname{arccot}(x)) = \frac{1}{x} \quad (8.74)$$

$$\sin(\operatorname{arccot}(x)) = \frac{1}{\sqrt{x^2 + 1}} \quad (8.75)$$

$$\cos(\operatorname{arccot}(x)) = \frac{x}{\sqrt{x^2 + 1}} \quad (8.76)$$

5 Trigonometry Tables

the data generated by a trip around the unit circle.

Trigonometry tables are lists of data containing key values of $\sin(\theta)$, $\cos(\theta)$. Contained in the tables that follow are

5.1 Standard Trigonometry Tables

First Quadrant

Angle (rad)	Angle ($^{\circ}$)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
0	0	0	1	0	$\mp\infty$	1	$\mp\infty$
$\pi/16$	11.25	0.195	0.981	0.198	5.142	1.020	5.081
$\pi/8$	22.5	0.383	0.924	0.414	2.610	1.086	2.414
$3\pi/16$	33.75	0.556	0.831	0.671	1.795	1.202	1.486
$\pi/4$	45	0.707	0.707	1	1.414	1.414	1
$5\pi/16$	56.25	0.831	0.556	1.496	1.202	1.795	0.671
$3\pi/8$	67.5	0.924	0.383	2.414	1.086	2.610	0.414
$7\pi/16$	78.75	0.981	0.195	5.081	1.020	5.142	0.198
$\pi/2$	90	1	0	$\pm\infty$	1	$\pm\infty$	0

Second Quadrant

Angle (rad)	Angle ($^{\circ}$)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
$9\pi/16$	101.25	0.981	-0.195	-5.081	1.020	-5.142	-0.198
$5\pi/8$	112.5	0.924	-0.383	-2.414	1.086	-2.610	-0.414
$11\pi/16$	123.75	0.831	-0.556	-1.496	1.202	-1.795	-0.671
$3\pi/4$	135	0.707	-0.707	-1	1.414	-1.414	-1
$13\pi/16$	146.25	0.556	-0.831	-0.671	1.795	-1.202	-1.486
$7\pi/8$	157.5	0.383	-0.924	-0.414	2.610	-1.086	-2.414
$15\pi/16$	168.75	0.195	-0.981	-0.198	5.142	-1.020	-5.081
π	180	0	-1	0	$\pm\infty$	-1	$\mp\infty$

Third Quadrant

Angle (rad)	Angle ($^{\circ}$)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
$17\pi/16$	191.25	-0.195	-0.981	0.198	-5.142	-1.020	5.081
$9\pi/8$	202.5	-0.383	-0.924	0.414	-2.610	-1.086	2.414
$19\pi/16$	213.75	-0.556	-0.831	0.671	-1.795	-1.202	1.486
$5\pi/4$	225	-0.707	-0.707	1	-1.414	-1.414	1
$21\pi/16$	236.25	-0.831	-0.556	1.496	-1.202	-1.795	0.671
$11\pi/8$	247.5	-0.924	-0.383	2.414	-1.086	-2.610	0.414
$23\pi/16$	258.75	-0.981	-0.195	5.081	-1.020	-5.142	0.198
$3\pi/2$	270	-1	0	$\pm\infty$	-1	$\mp\infty$	0

Fourth Quadrant

Angle (rad)	Angle (°)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
$25\pi/16$	281.25	-0.981	0.195	-5.081	-1.020	5.142	-0.198
$13\pi/8$	292.5	-0.924	0.383	-2.414	-1.086	2.610	-0.414
$27\pi/16$	303.75	-0.831	0.556	-1.496	-1.202	1.795	-0.671
$7\pi/4$	315	-0.707	0.707	-1	-1.414	1.414	-1
$29\pi/16$	326.25	-0.556	0.831	-0.671	-1.795	1.202	-1.486
$15\pi/8$	337.5	-0.383	0.924	-0.414	-2.610	1.086	-2.414
$31\pi/16$	348.75	-0.195	0.981	-0.198	-5.142	1.020	-5.081
2π	360	0	1	0	$\mp\infty$	1	$\mp\infty$

5.2 Generating Trigonometry Tables

A scientific calculator should be able to generate values for $\sin(\theta)$, $\cos(\theta)$, and $\tan(\theta)$ in degrees or radians. Sticking just with tables for a moment, there arises the question of, what if we need information for a value not explicitly listed, i.e., what's the cosine of 83 degrees? Although slightly ahead of the standard trigonometry regimen, we can explore two answers to this question.

for very small angles θ , the sine and cosine become:

$$\sin(\theta) \approx \theta - \frac{\theta^3}{3!} + \dots$$

$$\cos(\theta) \approx 1 - \frac{\theta^2}{2!} + \dots,$$

or more concisely,

$$\begin{aligned} \theta \text{ small:} \quad \sin(\theta) &\approx \theta \\ \theta \text{ small:} \quad \cos(\theta) &\approx 1. \end{aligned}$$

Trigonometry from Polynomials

As it turns out, it's possible to show that the sine and cosine can be calculated *exactly* for any θ using the expansions:

$$\begin{aligned} \sin(\theta) &= \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \dots \\ \cos(\theta) &= 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \dots \end{aligned}$$

The expansions on the right obey all of the rules of the sine and cosine, respectively. From these, all other quantities can be generated. Note that θ must occur in radians, not degrees.

With the key terms $\sin(\theta)$, $\cos(\theta)$ in hand, all of the others can be derived from these as a matter of definition.

Small-Angle Approximation

To prepare for a second means of generating trigonometry tables, particularly intermediate values in an existing table, we'll need to borrow ahead from calculus and write the so-called 'small-angle approximation'. In essence, note that

The small angle approximation can be discovered a number of ways, not necessarily from calculus. However, the notation surely couches best in a calculus framework.

Trig Tables by Interpolation

Now, recall the so-called *angle-sum formulas* via Equations (8.31), (8.32), namely

$$\begin{aligned} \sin(\alpha + \beta) &= \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \\ \cos(\alpha + \beta) &= \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta), \end{aligned}$$

and put the following restrictions on α, β :

$$\begin{aligned} 0 &\leq \alpha < 2\pi \\ |\beta| &\ll \alpha \end{aligned}$$

In other words, α is treated like a regular angle, and β is a very small angle. Using the small-angle approximation on β , the above can be approximately restated:

$$\sin(\alpha + \beta) \approx \sin(\alpha) + \cos(\alpha)\beta \quad (8.77)$$

$$\cos(\alpha + \beta) \approx \cos(\alpha) - \sin(\alpha)\beta \quad (8.78)$$

To run through an example, suppose we want the cosine of 83 degrees. Step one is to look at the closest entry in the existing trigonometry table, where we find

$$\cos\left(\frac{7\pi}{16}\right) = \cos(78.75^\circ) = 0.195.$$

The difference between 83 degrees and 78.75 degrees is assigned to β :

$$\beta = 83^\circ - 78.75^\circ = 4.25^\circ = 0.0742 \text{ rad}$$

With β in radians, we can plug into Equation (8.78) straightforwardly:

$$\begin{aligned}\cos(83^\circ) &\approx \cos(78.75^\circ) - \sin(78.75^\circ)(0.0742) \\ \cos(83^\circ) &\approx 0.1223\end{aligned}$$

For a sanity check the ‘exact’ value of $\cos(83^\circ)$ is about 0.1219. Of course, had we chosen a smaller β to begin with, the approximation would be more accurate. In this same spirit, by choosing β in small increments, trigonometry tables of any size can be calculated with enough patience or resources.

Problem 2

Given $\sin(20^\circ) = 0.342$ and $\cos(20^\circ) = 0.940$, estimate $\sin(22^\circ)$ and $\cos(22^\circ)$.

Answer: Let

$$\begin{aligned}\alpha &= 20^\circ = 20^\circ \left(\frac{\pi}{180^\circ}\right) = 0.349 \\ \beta &= 2^\circ = 2^\circ \left(\frac{\pi}{180^\circ}\right) = 0.0349,\end{aligned}$$

so then:

$$\begin{aligned}\sin(22^\circ) &\approx 0.342 + 0.0349(0.940) \approx 0.375 \\ \cos(22^\circ) &\approx 0.940 + 0.0349(0.342) \approx 0.928\end{aligned}$$

Insanity Check

As a matter of brutal curiosity, it’s worth checking (once, if ever) that the polynomial expansions for $\sin(\theta)$, $\cos(\theta)$ obey the fundamental identity

$$\sin^2(\theta) + \cos^2(\theta) = 1.$$

This chore involves squaring two infinite polynomials, a technically impossible task, but the idea is to spot a pattern in the algebra so the work doesn’t go forever.

For a shorthand notation let $S = \sin(\theta)$, $C = \cos(\theta)$, and square each of these separately:

$$\begin{aligned}S^2 &= S\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots\right) \\ &= (1)x^2 - \left(\frac{2}{3!}\right)x^4 + \left(\frac{2}{5!} + \frac{1}{3!3!}\right)x^6 \\ &\quad - \left(\frac{2}{7!} + \frac{2}{3!5!}\right)x^8 + \dots\end{aligned}$$

$$\begin{aligned}C^2 &= C\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots\right) \\ &= 1 - \left(\frac{2}{2!}\right)x^2 + \left(\frac{2}{4!} + \frac{1}{2!2!}\right)x^4 \\ &\quad - \left(\frac{2}{6!} + \frac{2}{2!4!}\right)x^6 \\ &\quad + \left(\frac{2}{8!} + \frac{2}{2!6!} + \frac{1}{4!4!}\right)x^8 - \dots\end{aligned}$$

This is a mess, but the results for S^2 , C^2 share a few similarities. First, there is a 1 on the right side of the C^2 quantity, which means all other terms in the sum must cancel the entire right side of the S^2 sum. Looking at the powers in x , we see both sums have only even powers (not surprisingly), and moreover their signs are equal and opposite.

For the fundamental trig identity to hold, it must be that all of the coefficients attached to similar power of x are equal. Checking this, we indeed find:

$$\begin{aligned}(1) &= \left(\frac{2}{2!}\right) = 1 \\ \left(\frac{2}{3!}\right) &= \left(\frac{2}{4!} + \frac{1}{2!2!}\right) = \frac{1}{3} \\ \left(\frac{2}{5!} + \frac{1}{3!3!}\right) &= \left(\frac{2}{6!} + \frac{2}{2!4!}\right) = \frac{2}{45} \\ \left(\frac{2}{7!} + \frac{2}{3!5!}\right) &= \left(\frac{2}{8!} + \frac{2}{2!6!} + \frac{1}{4!4!}\right) = \frac{1}{315}\end{aligned}$$

Finally, piece it all together to write

$$\begin{aligned}\sin^2(\theta) &= \theta^2 - \frac{1}{3}\theta^4 + \frac{2}{45}\theta^6 - \frac{1}{315}\theta^8 + \dots \\ \cos^2(\theta) &= 1 - \theta^2 + \frac{1}{3}\theta^4 - \frac{2}{45}\theta^6 + \frac{1}{315}\theta^8 - \dots,\end{aligned}$$

summing together to one.

5.3 Trigonometry Plots

Using trigonometry tables as a database allows for graphing the values of $\sin(\theta)$, $\cos(\theta)$, etc, in near-continuous fashion. The following plots in the Cartesian plane represent the continuous limit of trigonometry tables:

Sine, Cosine, Tangent

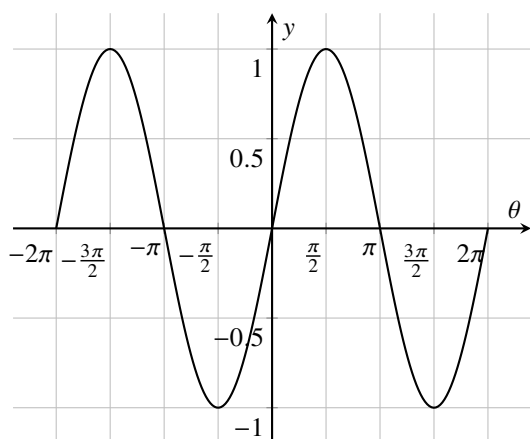


Figure 8.23. $y = \sin(\theta)$.

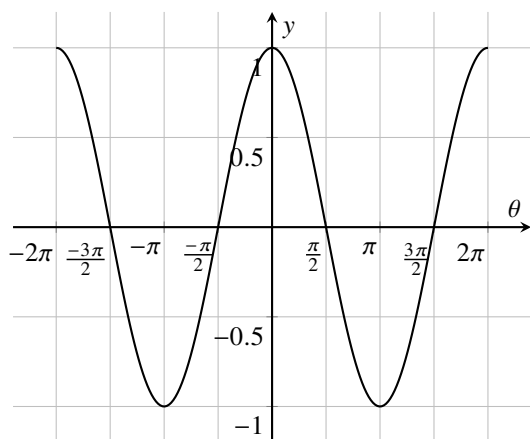


Figure 8.24. $y = \cos(\theta)$.

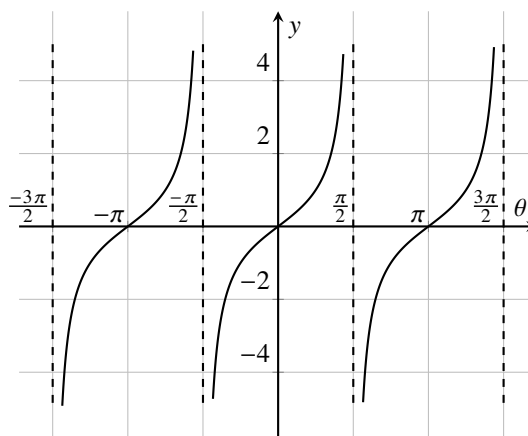


Figure 8.25. $y = \tan(\theta)$.

Cosecant, Secant, Cotangent

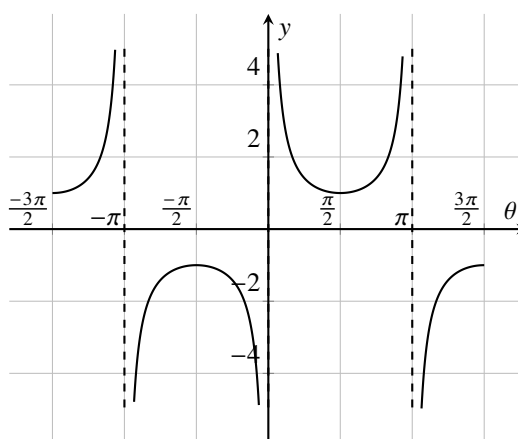


Figure 8.26. $y = \csc(\theta) = 1/\sin(\theta)$.

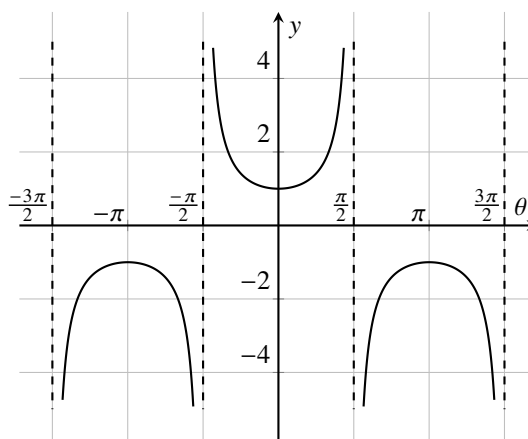
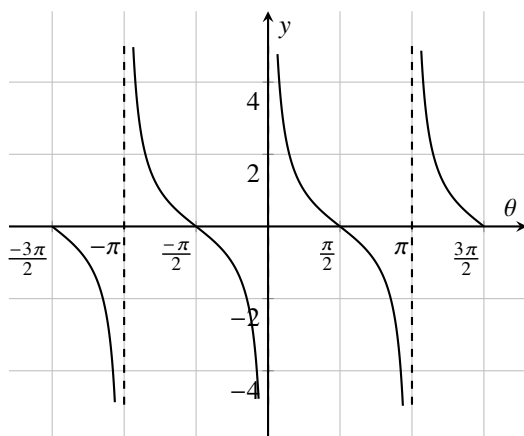


Figure 8.27. $y = \sec(\theta) = 1/\cos(\theta)$.

Figure 8.28. $y = \cot(\theta) = 1/\tan(\theta)$.

x	$\arcsin(x)$	$\arccos(x)$
-1.0	$-\pi/2$	π
-0.8	-0.9273	2.4981
-0.6	-0.6435	2.1859
-0.4	-0.4115	2.9845
-0.2	-0.2014	2.3698
0.0	0	$\pi/2$
0.2	0.2014	1.3694
0.4	0.4115	1.1593
0.6	0.6435	0.9273
0.8	0.9273	0.6435
1.0	$\pi/2$	0

5.4 Inverse Trigonometry Analysis

The inverse trigonometric quantities are a little more awkward to deal with, i.e. to generate corresponding inverse trigonometry tables. For reasons that ultimately come from computational efficiency arguments, none of which we'll repeat here, it makes sense to get everything in terms of the arctangent.

Arcsine, Arccosine

Recall Equations (8.66), (8.67) and invert these to solve for the arcsine and arccosine, respectively:

$$\arcsin(x) = \arctan\left(\frac{x}{\sqrt{1-x^2}}\right), \quad x^2 \leq 1$$

$$\arccos(x) = \arctan\left(\frac{\sqrt{1-x^2}}{x}\right), \quad 0 < x \leq 1$$

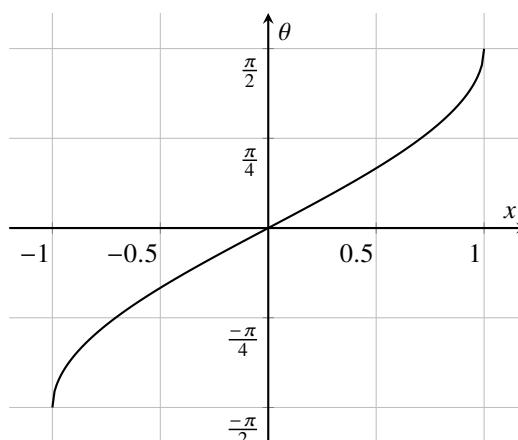
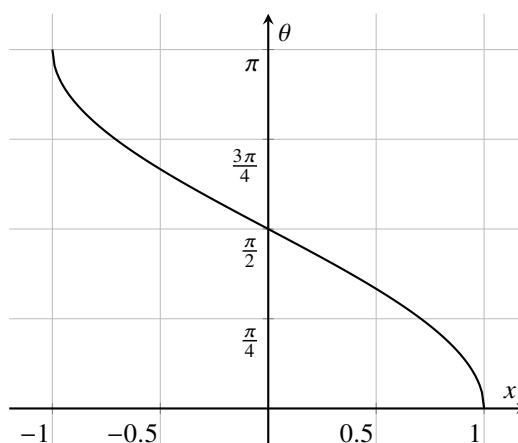
The arcsine-equation is valid $|x| \leq 1$, which happens to be the cover all cases one may throw at the arcsine.

The arccosine equation is valid as-is for $|x| > 0$, and needs a correction to cover the whole domain. It turns out that

$$\arccos(x) = \pi + \arctan\left(\frac{\sqrt{1-x^2}}{x}\right), \quad -1 \leq x < 0$$

does the job for $|x| < 0$, which is easy to verify. Neither equation for the arccosine handles the exact case $x = 0$, which by definition corresponds to $\pi/2$.

The summary of our findings is listed in the table below and also in Figures 8.29, 8.30.

Figure 8.29. $\theta = \arcsin(x)$, $|x| \leq 1$.Figure 8.30. $\theta = \arccos(x)$, $|x| \leq 1$.

Arccosecant, Arcsecant

For the arccosecant and arcsecant we repeat a similar analysis to the above starting with Equations (8.70), (8.73). This exercise results in:

$$\operatorname{arccsc}(x) = \arctan\left(\frac{1}{\sqrt{x^2 - 1}}\right), x > 0$$

$$\operatorname{arccsc}(x) = -\arctan\left(\frac{1}{\sqrt{x^2 - 1}}\right), x < 0$$

$$\operatorname{arcsec}(x) = \arctan\left(\sqrt{x^2 - 1}\right), x > 1$$

$$\operatorname{arcsec}(x) = \pi - \arctan\left(\sqrt{x^2 - 1}\right), x < -1$$

The summary of our findings is listed in the table below and also in Figures 8.29, 8.30.

x	$\operatorname{arccsc}(x)$	$\operatorname{arcsec}(x)$
$-\infty$	0	$\pi/2$
-100	-0.0100	1.5808
-3.2	-0.3178	1.8886
-1.6	-0.6751	2.2459
-1.2	-0.9851	2.5559
-1.0	$-\pi/2$	π
0.0		
1.0	$\pi/2$	0
1.2	0.9851	0.5857
1.6	0.6751	0.8957
3.2	0.3178	1.2530
100	0.0100	1.5808
∞	0	$\pi/2$

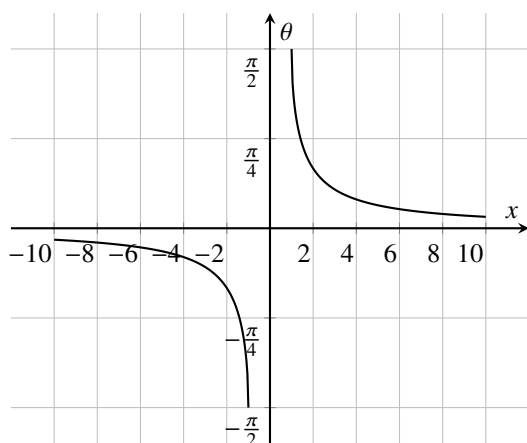


Figure 8.31. $\theta = \operatorname{arccsc}(x)$, $|x| \geq 1$.

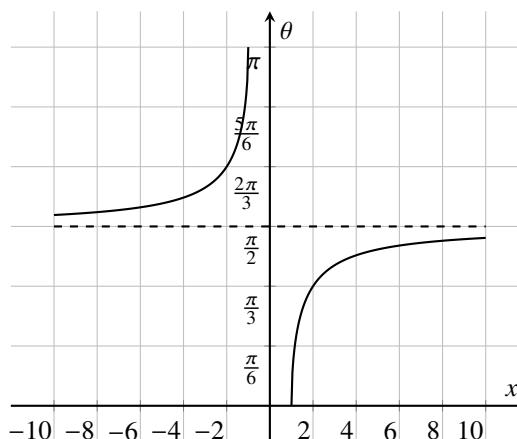


Figure 8.32. $\theta = \operatorname{arcsec}(x)$, $|x| \geq 1$.

Arctangent, Arccotangent

Finally we get to the case of arctangent and its reciprocal. There are numerous methods for grinding out the arctangent of any angle, one being called the Taylor expansion, a trick from calculus, which works in the domain $x^2 < 1$:

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

The Taylor expansion still leaves the question of what to do about the case $x^2 > 1$. For this, it's straightforward to show using trigonometric identities that

$$\arctan(x) = \frac{\pi}{2} - \arctan\left(\frac{1}{x}\right)$$

holds for any x , which turns the problem of say, calculating the arctangent of 4 into a problem of calculating the arctangent of 1/4. Taken together, the pair of above equations can be used to calculate any arctangent value.

For the arccotangent, invert Equation (8.74) to write

$$\operatorname{arccot}(x) = \arctan\left(\frac{1}{x}\right),$$

which can make direct use of the work previously done with the arctangent.

The summary of our findings is listed in the table below and also in Figures 8.33, 8.34.

x	$\arctan(x)$	$\operatorname{arccot}(x)$
$-\infty$	$-\pi/2$	0
-100	-1.5608	-0.0099
-4.0	-1.3258	-0.2450
-1.6	-1.0122	-0.5586
-0.4	-0.3805	-1.1903
0.0	0	$\mp\pi/2$
0.4	0.3805	1.1903
1.6	1.0122	0.5586
4.0	1.3258	0.2450
100	1.5608	0.0099
∞	$\pi/2$	0

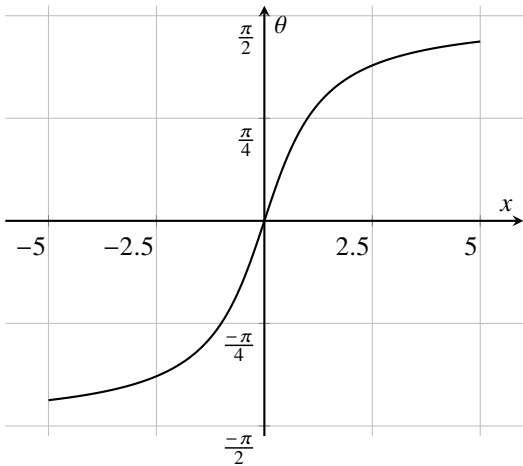


Figure 8.33. $\theta = \arctan(x)$

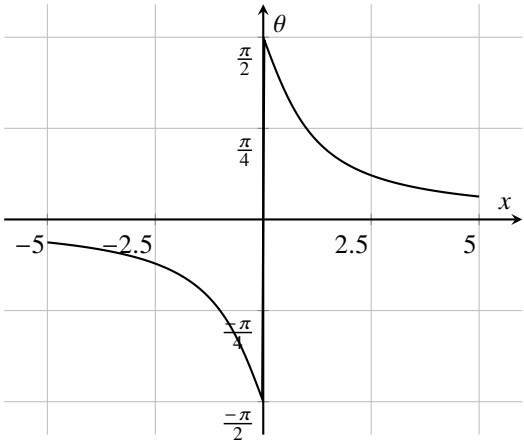


Figure 8.34. $\theta = \operatorname{arccot}(x)$

6 Trigonometry and Geometry

6.1 Law of Cosines

Trigonometry allows one to answer an age-old question from geometry which seeks to find a Pythagorean-like theorem for any arbitrarily-shaped triangle. This is answered by a special formula called the *law of cosines*.

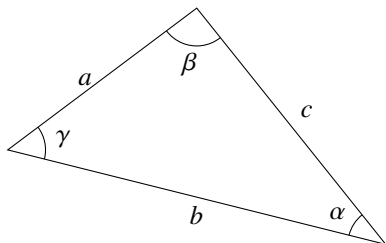


Figure 8.35. Arbitrary triangle.

Consider a triangle with three sides and three angles as labeled in Figure 8.35. Proceed by choosing any vertex, such as where sides a , c come together, and draw a line that intersects the third side, i.e. side b , at a ninety-degree angle. Note that β is not changed despite being omitted from Figure 8.36. Note too that side b is now broken into the sum $b_1 + b_2 = b$.

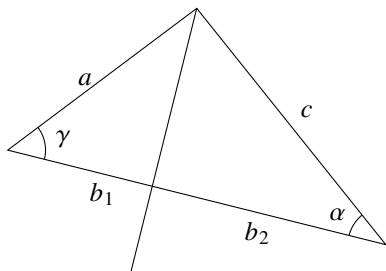


Figure 8.36. Arbitrary triangle with line intersecting side b at a right angle.

From standard trigonometry analysis, we can write a few true statements from the latter Figure:

$$\begin{aligned} a \cos(\gamma) &= b_1 \\ c \cos(\alpha) &= b_2 \\ a \sin(\gamma) &= c \sin(\alpha) \end{aligned}$$

Proceed by reconstructing the sum $b_1 + b_2$:

$$a \cos(\gamma) + c \cos(\alpha) = b,$$

and square both sides:

$$a^2 \cos^2(\gamma) + c^2 \cos^2(\alpha) + 2ac \cos(\gamma) \cos(\alpha) = b^2$$

Next make the replacements

$$\cos^2(\gamma) = 1 - \sin^2(\gamma)$$

$$\cos^2(\alpha) = 1 - \sin^2(\alpha)$$

and simplify like mad.

At the end, arrive at the all-powerful law of cosines:

$$a^2 + c^2 - 2ac \cos(\beta) = b^2 \quad (8.79)$$

By symmetry, since we could have sliced the triangle two more ways, there two more expressions of the same law:

$$b^2 + c^2 - 2bc \cos(\alpha) = a^2 \quad (8.80)$$

$$a^2 + b^2 - 2ab \cos(\gamma) = c^2 \quad (8.81)$$

6.2 Inscribed Triangle

A tricky analysis starts with a circle of any radius R with a diameter $AB = 2R$ as shown in Figure 13.5. Choose a point C on the perimeter and draw lines from C to A , B to form an inscribed triangle. (Ignore all interior labels in the Figure until they're invoked.)

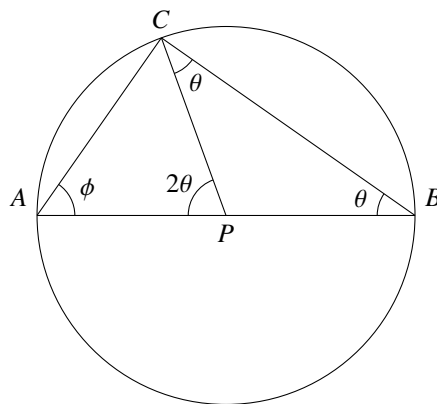


Figure 8.37. Inscribed triangle.

ACB is a Right Angle

First, it's possible to prove that $\angle ACB$ is always a right angle. Place the origin at A so that point C is located at (x, y) , and denote the line AC as a variable r such that:

$$\begin{aligned} r \cos(\phi) &= x \\ r \sin(\phi) &= y \\ r^2 &= x^2 + y^2 \end{aligned}$$

Using these variables, the circle itself obeys

$$(x - R)^2 + y^2 = R^2,$$

readily simplifying as:

$$\begin{aligned} r^2 &= 2xR \\ r &= 2R \cos(\phi) \\ r &= 2R \sin\left(\frac{\pi}{2} - \phi\right) \end{aligned}$$

To see what we're shooting for, let's momentarily assume the thing we want to prove, i.e. that $\angle ACB = \pi/2$, and make sure no contradiction arises. Going with this, observe from the Figure that

$$r = 2R \sin(\theta) ,$$

and eliminating r gives

$$2R \sin\left(\frac{\pi}{2} - \phi\right) = 2R \sin(\theta) ,$$

which can only mean

$$\theta + \phi = \frac{\pi}{2} ,$$

and no contradiction arises. This result is also known as *Thales' theorem*.

For a proper proof, use the law of cosines to discern from the Figure that

$$\begin{aligned} (\overline{BC})^2 &= (\overline{PC})^2 + (\overline{PB})^2 \\ &\quad - 2(\overline{PC})(\overline{PB}) \cos(\angle BPC) \\ (\overline{AC})^2 &= (\overline{PA})^2 + (\overline{PC})^2 \\ &\quad - 2(\overline{PA})(\overline{PC}) \cos(2\theta) , \end{aligned}$$

simplifying to

$$\begin{aligned} (\overline{BC})^2 &= 2R^2 - 2R^2 \cos(\angle BPC) \\ (\overline{AC})^2 &= 2R^2 - 2R^2 \cos(2\theta) . \end{aligned}$$

Sum the two equations to get:

$$\begin{aligned} (\overline{AC})^2 + (\overline{BC})^2 &= (2R)^2 \\ &\quad - 2R^2 (\cos(2\theta) + \cos(\pi - 2\theta)) \end{aligned}$$

The remaining cosine terms cancel exactly, and we find that the sides of the triangle obey the Pythagorean theorem:

$$(\overline{AC})^2 + (\overline{BC})^2 = (2R)^2 = (\overline{AB})^2$$

Thus, ACB is a right triangle.

PCB Equals Theta

If the center of the circle is located at P , then the length PC is equal to length PB , both of which equal the radius R . This qualifies PCB as an isosceles triangle, having two equal sides, which also means two equal angles:

$$\angle PCB = \theta = \angle PBC$$

APC Equals Twice Theta

Using the properties of the isosceles triangle, the (unlabeled) angle BPC obeys

$$\angle BPC + 2\theta = \pi .$$

Being a straight line, the total angle across APB needs to be π , which means

$$\angle APC + \angle BPC = \pi .$$

Eliminating $\pi - \angle BPC$ from each equation yields the result we want:

$$\angle APC = 2\theta$$

Area of ACB

The area of triangle ACB is

$$A = \frac{1}{2} (\overline{AB}) (\overline{PC}) \sin(2\theta) .$$

In terms of the radius, the area is

$$A = R^2 \sin(2\theta) .$$

With the base fixed, we see that the height is maximized at $\theta = \pi/4$, and the maximal area is

$$A_{\max} = \frac{1}{2} (\overline{AB}) (\overline{PC}) = R^2 .$$

6.3 Law of Sines

There is another relationship called the *law of sines* that is obeyed by all triangles.

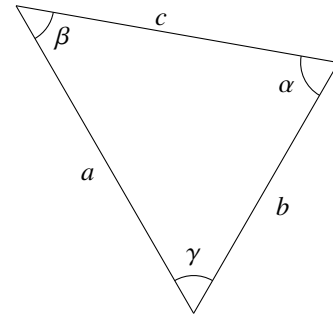


Figure 8.38. Arbitrary triangle.

Quick Derivation

Consider a triangle with three sides and three angles as labeled in Figure 8.38. Using the same process that led to Figure 8.36, choose any vertex and draw a line perpendicular to the side opposite the vertex. We already did this once to yield

$$a \sin(\gamma) = c \sin(\alpha)$$

in deriving the law of cosines. This can be repeated for each vertex to yield two more similar relations:

$$c \sin(\beta) = b \sin(\gamma)$$

$$b \sin(\alpha) = a \sin(\beta)$$

Taking all three of the above equations together yields the (weakest statement of) the law of sines:

$$\frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \quad (8.82)$$

Area-Based Derivation

A second derivation of the law of sines writes the total area T of triangle ABC three different ways. Analyzing similarly as above, we can write:

$$T = \frac{1}{2}ac \sin(\beta) = \frac{1}{2}ab \sin(\gamma)$$

$$T = \frac{1}{2}ba \sin(\gamma) = \frac{1}{2}bc \sin(\alpha)$$

$$T = \frac{1}{2}ca \sin(\beta) = \frac{1}{2}cb \sin(\alpha)$$

With these, we can not only re-derive Equation (8.82) but we can also interpret the law of sines as it relates to the area of the triangle:

$$\frac{2T}{abc} = \frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \quad (8.83)$$

Circumcircle

Imagine the arbitrary triangle being enclosed by a carefully-placed circle so that all three vertices lie on the circle's perimeter. This is called a *circumcircle*, silly enough, but is nonetheless shown in Figure 8.83. Generally, such a circle has a radius R with a center point that may or may not lie within the confines of the triangle.

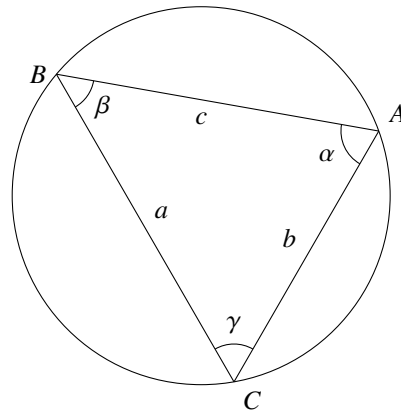


Figure 8.39. Arbitrary triangle with circumcircle.

Circumcircle Analysis

It turns out that the radius R of the circumcircle inscribing an arbitrary triangle relates to the law of sines. To prove this, start with any vertex, such as B , and draw a line through the circle's center until it intercepts the other side at point Q , i.e. draw a diameter $2R$ of the circle. Also, draw lines of length R from points A and C to the center of the circle as shown in Figure 8.40.

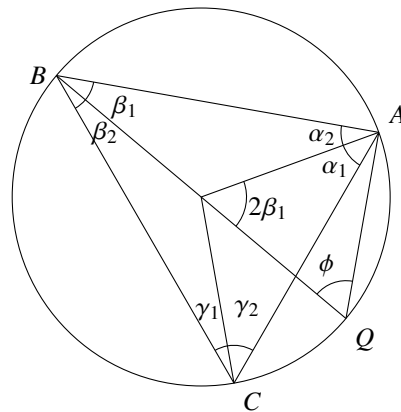


Figure 8.40. Arbitrary triangle with detailed circumcircle.

The angles α, β, γ are now partitioned (not bisected) according to

$$\alpha = \alpha_1 + \alpha_2$$

$$\beta = \beta_1 + \beta_2$$

$$\gamma = \gamma_1 + \gamma_2$$

By doing this work, the Figure highlights three isosceles triangles, each having two sides of length R and two identical

angles:

$$\begin{aligned}\alpha_1 &= \gamma_2 \\ \beta_1 &= \alpha_2 \\ \gamma_1 &= \beta_2\end{aligned}$$

Phi Equals Gamma

To proceed, recall from inscribed angle analysis to notice that $\angle BAQ$ is exactly a right angle, and thus the angle $\phi = \angle BQA$ relates to β_1 by

$$\phi + \beta_1 = \frac{\pi}{2}.$$

This is enough to establish an important relationship between γ and ϕ via

$$\begin{aligned}\gamma &= \gamma_1 + \gamma_2 = \beta_2 + \alpha_1 = (\beta - \beta_1) + (\alpha - \alpha_2) \\ &= \alpha + \beta - 2\beta_1 = \pi - \gamma - 2\beta_1 \\ \gamma &= -\gamma + 2\phi,\end{aligned}$$

finally revealing

$$\phi = \gamma.$$

Third Derivation

With $\phi = \gamma$ known, refer back to Figures 8.39, 8.40 to notice we can now write

$$2R \sin(\gamma) = c,$$

or

$$\frac{1}{2R} = \frac{\sin(\gamma)}{c}.$$

By symmetry, we could do the entire analysis twice more to land at yet another expression for the law of sines:

$$\frac{1}{2R} = \frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \quad (8.84)$$

Taking Equations (8.83), (8.84) together, we get a nifty formula for the area for a circle in terms of its sides and circumcircle radius:

$$\text{Area} = T = \frac{abc}{4R}$$

6.4 Triangle Inequality

There is an important property of triangles called the *triangle inequality*, stating that the sum of two side lengths is

always greater than or equal to the remaining length. That is, for a triangle of sides x, y, z , it follows that

$$\begin{aligned}z &\leq x + y \\ x &\leq y + z \\ y &\leq z + x\end{aligned}$$

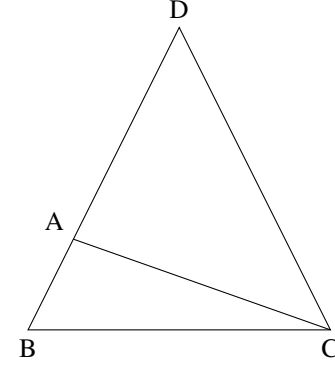


Figure 8.41. Triangle inequality.

To prove this, consider the triangle ABC depicted in Figure 8.41. Choose point D on the line through AB such that an isosceles triangle is formed with two equal lengths AC and AD . Then, angle BCD has greater measure than ACD , which means BD is greater than BC :

$$BD > BC$$

Next, note that

$$BD = AB + AD,$$

which means

$$BD = AB + AC.$$

Reading right to left, we have that the sum $AB + AC$ equals BD , which we found is greater than BC . Thus we have

$$AB + AC > BC,$$

completing the proof.

6.5 Adventitious Angles

A puzzle invented in 1922 called *Langley's Adventitious Angles* is posed as follows:

- ABC is an isosceles triangle with:

$$\angle CBA = \angle ACB = 80^\circ$$

- $\overline{CF}$ at 30° cuts $\overline{AB}$ at F .
- $\overline{BE}$ at 20° cuts $\overline{AC}$ at E .

- Prove $\angle BEF = 30^\circ$.

Figure 8.42 shows the adventitious angles puzzle as described.

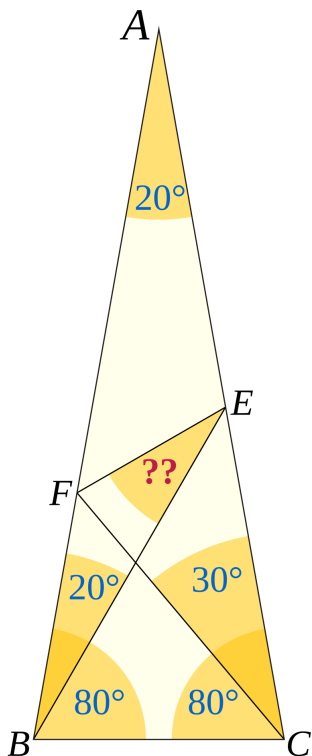


Figure 8.42. Langley's Adventitious Angles puzzle^a.

^aLangley's Adventitious Angles'. *Wikipedia*. August 11 2025.
https://en.wikipedia.org/wiki/Langley's_Adventitious_Angles

Geometric Analysis

On analyzing the Figure, one might draw two lines parallel to BC while introducing angles w, x, y, z, q . The unknown x is what we're solving for as shown in Figure 8.43.

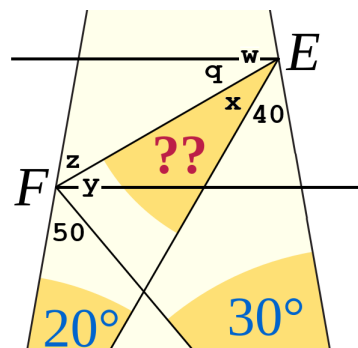


Figure 8.43. Langley's Adventitious Angles puzzle, zoomed and analyzed.

Using the declared variables, the puzzle yields an $N = 4$ system of equations and unknowns, namely

$$\begin{aligned} x + y &= 110^\circ \\ y + z &= 130^\circ \\ x + w &= 140^\circ \\ z + w &= 160^\circ \end{aligned}$$

Algebraic Analysis

Sadly, the above $N = 4$ system is not sufficiently constrained. One way to quickly see this is to subtract second and fourth equations from the first and third, respectively, to find two copies of $z - x = 20^\circ$.

Involving the angle q seems hopeful at first, but the following is still not enough to specify each angle uniquely:

$$\begin{aligned} x &= 60^\circ - q \\ y &= 50^\circ + q \\ z &= 80^\circ - q \\ w &= 80^\circ + q \end{aligned}$$

To foreshadow the solution, we can write the above in a way that begs the result $q = 30^\circ$ while simultaneously teasing all other results:

$$\begin{aligned} x &= 30^\circ + (30^\circ - q) \\ y &= 80^\circ - (30^\circ - q) \\ z &= 50^\circ + (30^\circ - q) \\ w &= 110^\circ - (30^\circ - q) \end{aligned}$$

Trigonometric Analysis

Finally we get to the proof. Using the law of sines, the puzzle yields three relationships relating length and angles.

$$\frac{\overline{FE}}{\sin(20^\circ)} = \frac{\overline{BE}}{\sin(y + 50^\circ)}$$

$$\frac{\overline{FE}}{\sin(30^\circ)} = \frac{\overline{EC}}{\sin(y)}$$

$$\frac{\overline{BE}}{\sin(80^\circ)} = \frac{\overline{EC}}{\sin(60^\circ)}$$

The rest is algebra. Eliminate $\overline{EC}$ between the last two equations to find

$$\overline{FE} \frac{\sin(y)}{\sin(30^\circ)} = \overline{BE} \frac{\sin(60^\circ)}{\sin(80^\circ)},$$

and then eliminate $\overline{BE}$ using the first equation:

$$\overline{FE} \frac{\sin(y)}{\sin(30^\circ)} = \overline{FE} \frac{\sin(50^\circ + y)}{\sin(20^\circ)} \frac{\sin(60^\circ)}{\sin(80^\circ)}$$

Cancel the factor of $\overline{FE}$ and put all y -dependence on one side of the equation to have

$$\frac{\sin(50^\circ + y)}{\sin(y)} = \frac{\sin(80^\circ) \sin(20^\circ)}{\sin(60^\circ) \sin(30^\circ)}.$$

The above equation for y is satisfied by:

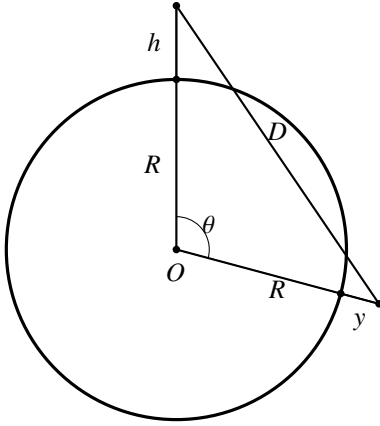
$$y = 80^\circ$$

With y on hand, all solutions for w, x, z, q follow.

6.6 Visibility on a Sphere

Problem 3

On a spherical planet of radius R , an observation tower of height h above the surface has line of sight to an antenna at height y above the surface. The tower and the antenna are separated by total distance $M = R\theta$ at the base, i.e. the geodesic separation between the two. The line of sight (length D) from the tower to the antenna may be interrupted by the sphere. Write the criteria for whether this occurs. See the Figure that follows.



Answer: Center at the origin a circle of radius R . Locate the top of the tower at

$$\begin{aligned} x_t &= 0 \\ y_t &= R + h, \end{aligned}$$

so the tower is situated vertically along the y -axis.

The line through the antenna to the origin makes angle θ with respect to the vertical. It follows that the top of the antenna has position

$$\begin{aligned} x_a &= (R + y) \cos(\pi/2 - \theta) \\ y_a &= (R + y) \sin(\pi/2 - \theta), \end{aligned}$$

which contains all known information.

Next, a straight line $z(x)$ connecting the top of the tower to the top of the antenna is given by

$$z(x) = \left(\frac{y_a - y_t}{x_a - x_t} \right) x + y_t,$$

where for shorthand, write the slope term as m such that $z = mx + y_t$.

It's straightforward to show that the point $(x_\star, y_\star)$ on $z(x)$ that is closest to the origin is given by

$$0 = (x_\star - 0) + (mx_\star + y_t - 0)m,$$

telling us:

$$x_\star = \frac{-my_t}{1 + m^2} \quad y_\star = mx_\star + y_t$$

We now make the final argument. If the distance

$$L = \sqrt{x_\star^2 + y_\star^2}$$

is greater than R , then the sphere does not interrupt the line of sight from the tower to the antenna. If L is less than R , then at least some of the line of sight passes through the sphere. That is, we look for:

$$L > R$$

Simplifying the equation for L leads to

$$L = \frac{y_t}{\sqrt{1 + m^2}}.$$

For shorthand, let $Y = R + y$ and we further find

$$1 + m^2 = \frac{Y^2 \sin^2(\theta) + (Y \cos(\theta) - y_t)^2}{Y^2 \sin^2(\theta)},$$

leading to

$$(1 + m^2) Y^2 \sin^2(\theta) = Y^2 + y_t^2 - 2Yy_t \cos(\theta).$$

Notice the right side of the above is the law of cosines applied to the triangle formed by the origin, the top of the tower, and the top of the antenna. This identifies

$$D = \sqrt{Y^2 + y_t^2 - 2Yy_t \cos(\theta)}$$

as the straight-line distance between the tower and the antenna, and we rewrite L in terms of D :

$$L = \frac{Yy_t}{D} \sin(\theta)$$

Thus, the criteria for visibility reads

$$Yy_t \sin(\theta) > RD,$$

or:

$$(R + h)(R + y) \sin(\theta) > RD$$

This has a convenient geometric interpretation. Reading left to right, the area of the parallelogram having sides $R + h$, $R + y$ must be greater than the area of the rectangle with sides R , D .



Answer: This is the $y_a = 0$ -case of the sphere visibility analysis, thus the distance to the horizon is exactly

$$D = \sqrt{2Rh + h^2}.$$

The equation of the line on which D lies is written

$$z(x) = mx + (R + h),$$

where

$$\begin{aligned} m &= \left(\frac{y_a - y_t}{x_a - x_t} \right) = \frac{R \cos(\theta) - (R + h)}{R \sin(\theta) - 0} \\ &= \frac{-D}{R} = -\tan(\theta). \end{aligned}$$

We also need the line $w(x)$ that is perpendicular to the line of sight $\overline{TH}$ and intersecting it at point P :

$$w(x) = \left(\frac{R}{D} \right) x - w_0$$

To determine w_0 and r , we notice based on the position of point P that

$$\begin{aligned} (D + X) \cos(\theta) &= (r + R) \sin(\theta) \\ (R + h) - (D + X) \sin(\theta) &= -w_0 + (r + R) \cos(\theta), \end{aligned}$$

leading to:

$$\begin{aligned} r &= \cot(\theta) (D + X) - R = \frac{XR}{D} \\ w_0 &= \frac{D + X}{\sin(\theta)} - (R + h) = \frac{X(R + h)}{D} \end{aligned}$$

With the equation for $w(x)$ in hand, we calculate the intersection with the sphere's surface $(x_\star, y_\star)$ (not labeled in the Figure). This is the base of distance q that terminates at P . For this, we solve

$$w(x_\star) = \frac{Rx_\star}{D} - w_0 = \sqrt{R^2 - x_\star^2},$$

leading to a quadratic equation for $x_\star$, letting $\alpha = R + h$:

$$x_\star^2 \alpha^2 + x_\star (-2RX\alpha) + (X^2 \alpha^2 - D^2 R^2) = 0$$

After a bit of algebra, one finds, selecting the positive branch of the root,

$$x_\star = \frac{RX + D\sqrt{R^2 - X^2}}{R + h},$$

simultaneously determining $y_\star$.

Next list the horizontal location of point P

$$x_P = (D + X) \cos(\theta) = \frac{R(D + X)}{(R + h)},$$

and write

$$\Delta x = x_P - x_\star = \frac{D}{R + h} \left(R - \sqrt{R^2 - X^2} \right).$$

Since we already know the slope of q , namely $\cot(\theta)$, we calculate the corresponding Δy by

$$\Delta y = \Delta x \frac{R}{D} = \frac{R}{R + h} \left(R - \sqrt{R^2 - X^2} \right).$$

The length q is exactly

$$q = \sqrt{\Delta x^2 + \Delta y^2}.$$

After substituting from the above, this simplifies nicely to

$$q = R - \sqrt{R^2 - X^2},$$

which we may take as a final result. Keep in mind the h -dependence in the general solution is embedded from where we start measuring X , namely point H . To use such formulas in practice, one uses either D or $M = R\theta$ to determine the position of the horizon.

We can keep going, though. Define a parameter $0 \leq \phi \leq \pi/2$ as the central angle from the horizon H to the surface point $(x_\star, y_\star)$. This tracks the distance from H to the point 'under' P via $S = R\phi$. With this, we write

$$\begin{aligned} X &= R \sin(\phi) \\ \sqrt{R^2 - X^2} &= R \cos(\phi), \end{aligned}$$

so finally

$$q = R(1 - \cos(\phi)),$$

or:

$$q = R \left(1 - \cos \left(\frac{S}{R} \right) \right)$$

Amazingly, the form for q is identical to the result we get from the $h = 0$ -case, with the replacement $\theta \rightarrow \phi$.

Obstructed View of a Building

Problem 8

Continuing the previous problem, suppose point P is a location on a tall building whose base is at the intersection of the surface and the line $\overline{OP}$. According to the observer at T , what vertical portion of the building Q is obscured by the sphere?

Answer: From the previous problem, we know:

$$x_P = (D + X) \cos(\theta)$$

$$y_P = (R + h) - (D + X) \sin(\theta)$$

We further observe that

$$Q = \sqrt{x_P^2 + y_P^2} - R$$

is the portion of the building that lies below the horizon and invisible to T . Use what we know from the previous problem to simplify:

$$Q = \sqrt{R^2 + X^2} - R$$

Or, we can prepare for the small- X approximation by writing

$$Q = R \left(\sqrt{1 + \frac{X^2}{R^2}} - 1 \right),$$

so then for $X \ll R$:

$$Q \approx \frac{X^2}{2R}$$

7 Polar Coordinate System

Recall momentarily the Cartesian coordinate system is the lattice on which all points in the plane are hung. There is no point in the plane that does not have a unique coordinate, and every coordinate corresponds to some point in the plane. As it turns out, there is another system called *polar coordinates* that can do the same job as Cartesian coordinates - to cover the plane completely.

7.1 Motivation

The polar coordinate system is built from the apparatus of trigonometry. Consider the unit circle centered at the origin represented by

$$\begin{aligned}x &= \cos(\theta) \\ y &= \sin(\theta),\end{aligned}$$

or

$$x^2 + y^2 = 1.$$

By choosing any θ , (even those outside the standard domain), the point (x, y) lands somewhere in the plane a distance 1 from the origin.

Suppose next that the unit circle is replaced by any other circle of radius r , also centered at the origin. In the same way that θ allows freedom in the *angular* dimension, the varying radius allows for freedom in the *radial* dimension. For this reason, one can see that every point (x, y) in the Cartesian plane corresponds to some ordered pair (r, θ) .

The mapping from (x, y) to (r, θ) defines the *polar coordinate system*:

$$x(r, \theta) = r \cos(\theta) \quad (8.85)$$

$$y(r, \theta) = r \sin(\theta) \quad (8.86)$$

The $x(r, \theta)$, $y(r, \theta)$ notation is there to remind us that x and y each depend on two variables as suggested in Figure 8.44.

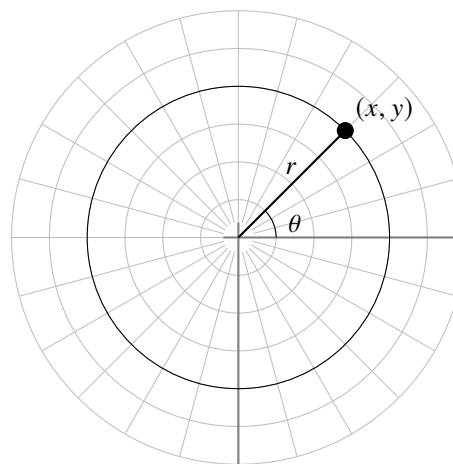


Figure 8.44. Polar coordinate system.

Radial and Angular Coordinates

In terms of x and y , the r - and θ -variables are straightforwardly isolated. Squaring Equations (8.85), polarcoordsy and taking their sum and square root gives the formula for r

$$r = \sqrt{x^2 + y^2}. \quad (8.87)$$

Note that the $\pm$ symbol is omitted from the front of the square root symbol. This is to mean that there is no such thing as negative distance from the origin when using polar coordinates.

Solve for θ by taking the ratio of the x - and y -equations, and then make use of the arctangent:

$$\theta = \arctan\left(\frac{y}{x}\right) \quad (8.88)$$

As a consistency check, we should be able to apply $\sin()$ or $\cos()$ to both sides of Equation (8.88) to recover the x , y equations. Using the trig identities (8.66), (8.67), we have

$$\cos(\theta) = \cos\left(\arctan\left(\frac{y}{x}\right)\right) = \frac{1}{\sqrt{y^2/x^2 + 1}} = \frac{x}{r}$$

and

$$\sin(\theta) = \sin\left(\arctan\left(\frac{y}{x}\right)\right) = \frac{y/x}{\sqrt{y^2/x^2 + 1}} = \frac{y}{r}$$

as expected.

7.2 Lines

Navigating the plane in polar coordinates works out differently than when using Cartesian coordinates. For example,

the Cartesian system makes trivial work out of straight lines, but things get ugly when it comes to tracing curves, such as circles, i.e.

$$y_{\text{circ}} = \pm \sqrt{R^2 - x^2} .$$

On the other hand, straight lines are a bit of a headache in polar coordinates, whereas y_{circ} simply $r = R$.

For the equation of a straight line in Cartesian coordinates

$$y = mx + b ,$$

use Equations (8.85), (8.86) for polar coordinates and the same line becomes

$$r = \frac{b}{\sin(\theta) - m \cos(\theta)} . \quad (8.89)$$

We can keep going, though. Express the slope m as the tangent of some new angle, say ϕ :

$$m = \tan(\phi)$$

With this, r becomes

$$r = \frac{b \cos(\phi)}{\sin(\theta) \cos(\phi) - \cos(\theta) \sin(\phi)} = \frac{b \cos(\phi)}{\sin(\theta - \phi)} .$$

Replace $\cos(\phi)$ using

$$\cos(\phi) = \frac{1}{1 + \tan^2(\phi)} = \frac{1}{\sqrt{1 + m^2}} ,$$

and arrive at another equation for the straight line:

$$r = \frac{b}{\sqrt{1 + m^2}} \csc(\theta - \arctan(m)) . \quad (8.90)$$

7.3 Scale and Rotation

In a similar way that the x - and y -directions in Cartesian coordinates are independent, i.e. a change in x is not a change in y , is also true in polar coordinates r, θ . Instead of vertical and horizontal changes, there are instead radial changes we'll call *scaling*, and angular changes addressed as *rotation*.

Scaling

The easy case is the radial one, where starting at some position (x_0, y_0) in the plane such that

$$\begin{aligned} x_0 &= r_0 \cos(\theta_0) \\ y_0 &= r_0 \sin(\theta_0) , \end{aligned}$$

we may multiply through by a positive constant λ to scale each coordinate and move to a new location (x, y) :

$$\begin{aligned} x &= \lambda x_0 = (\lambda r_0) \cos(\theta_0) \\ y &= \lambda y_0 = (\lambda r_0) \sin(\theta_0) \end{aligned}$$

Inspecting this result, we see that the effect of scaling each coordinate by λ simply modifies the radius via

$$r = \lambda r_0$$

while leaving the angle the same.

Rotations

A way to move from (x_0, y_0) to a new location (x, y) independent of r is to rotate about the origin, which is to change the θ -variable only. Suppose some angle ϕ is added to θ such that

$$\begin{aligned} (x_0, y_0) &= (r \cos(\theta), r \sin(\theta)) \\ (x, y) &= (r \cos(\theta + \phi), r \sin(\theta + \phi)) , \end{aligned}$$

where the top pair is the $\phi = 0$ case of the bottom pair.

The trigonometry terms can be expanded using the angle-sum formulas (8.31), (8.32), resulting in

$$\begin{aligned} x &= r \cos(\theta) \cos(\phi) - r \sin(\theta) \sin(\phi) \\ y &= r \sin(\theta) \cos(\phi) + r \cos(\theta) \sin(\phi) , \end{aligned}$$

simplifying to a set of equations (some may recognize as a rotation matrix):

$$\begin{aligned} x &= x_0 \cos(\phi) - y_0 \sin(\phi) \\ y &= x_0 \sin(\phi) + y_0 \cos(\phi) \end{aligned}$$

As a sanity check, one can check the sum

$$x^2 + y^2 = x_0^2 + y_0^2 ,$$

which assures the radius doesn't change under rotations.

The Involute

A string is wrapped around a circle of radius a . Keeping the string tight, unwind the string and keep track of its endpoint. The shape traced out is called the *involute* as shown in Figure 8.45.

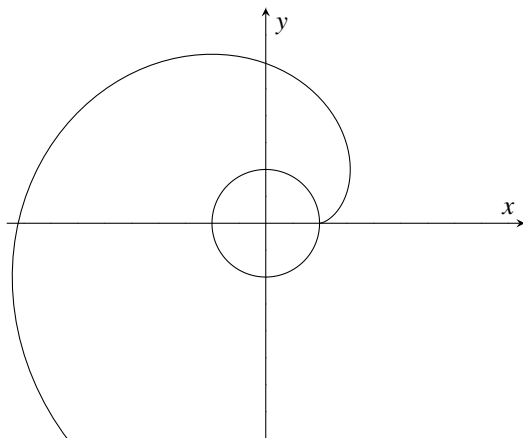


Figure 8.45. The involute.

Parameterize the unwinding using an angle ϕ , where $\phi = 0$ corresponds to the fully-wrapped string. In terms of ϕ , the endpoint of the string is given by

$$\begin{aligned} x(\phi) &= a \cos(\phi) + a\phi \sin(\phi) \\ y(\phi) &= a \sin(\phi) - a\phi \cos(\phi) . \end{aligned}$$

With x , y on hand, we can determine the absolute distance from the origin, i.e. the r -parameter for polar coordinates:

$$r = \sqrt{x^2 + y^2} = a\sqrt{1 + \phi^2}$$

The θ -parameter is a little more tricky. Using polar coordinates, start with

$$\begin{aligned} r \cos(\theta) &= a \cos(\phi) + a\phi \sin(\phi) \\ r \sin(\theta) &= a \sin(\phi) - a\phi \cos(\phi) . \end{aligned}$$

Proceed by letting

$$\phi = \tan(u)$$

for some new parameter u , which leads to

$$\begin{aligned} \cos(u) &= \pm a/r \\ \sin(u) &= \pm a\phi/r . \end{aligned}$$

Simplifying the above gives a tight relationship between the variables on hand:

$$\theta = \phi - u = \tan(u) - u$$

Note that u is confined to the domain $(\pi/2 : \pi/2)$.

7.4 Offset Circle

A circle offset from the origin is a bit messy in both Cartesian and polar coordinates. Consider a circle of radius a centered at the point (x_0, y_0) , i.e.

$$(x - x_0)^2 + (y - y_0)^2 = a^2 ,$$

where (x, y) locates any point on the perimeter as shown in Figure 8.46.

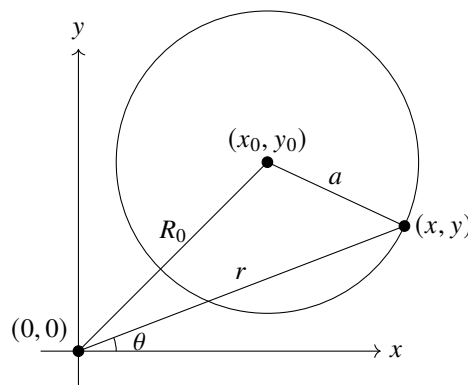


Figure 8.46. Offset circle.

Using polar coordinates, we have that the center of the circle is located by

$$\begin{aligned} x_0 &= R_0 \cos(\theta_0) \\ y_0 &= R_0 \sin(\theta_0) , \end{aligned}$$

and, of course, a the point (x, y) on the perimeter is at

$$\begin{aligned} x &= r \cos(\theta) \\ y &= r \sin(\theta) . \end{aligned}$$

Substituting the polar representations for x , y , x_0 , y_0 into the equation of the offset circle and letting the algebra cook down results in something reminiscent of the law of cosines:

$$r^2 + R_0^2 - 2rR_0 \cos(\theta - \theta_0) = a^2$$

Note that $\theta - \theta_0$ is the angle formed between r and R_0 . We can keep going, though. Isolate r using the quadratic formula and simplify again:

$$r = R_0 \cos(\theta - \theta_0) \pm \sqrt{a^2 - R_0^2 \sin^2(\theta - \theta_0)} \quad (8.91)$$

8 Circular Motion

Consider a circular ring of radius R on which an object can slide freely. Using polar coordinates, we can write the Cartesian position (x, y) in terms of R and the angle θ characterizing the placement of the object:

$$\begin{aligned}x &= R \cos(\theta) \\ y &= R \sin(\theta)\end{aligned}$$

Kinematics on a Circle

Motion on such a circle is completely characterized by $\theta(t)$. The rate at which θ changes in time is called the *angular frequency*, denoted ω (Greek *omega*) as:

$$\omega = \frac{\Delta\theta}{\Delta t}$$

In the case that ω is a constant ω_0 , we can write an equation for θ that is reminiscent of kinematics:

$$\theta = \theta_0 + \omega_0 t$$

Moreover, one can have the case where ω changes with time such that

$$\alpha = \frac{\Delta\omega}{\Delta t},$$

where α (Greek *alpha*) is the *angular acceleration*. For constant angular acceleration, we can write

$$\omega = \omega_0 + \alpha_0 t,$$

and moreover,

$$\theta = \theta_0 + \omega_0 t + \frac{1}{2} \alpha_0 t^2.$$

As it turns out, the entire splay of one-dimensional kinematics applies to circular motion by a swap of variables. You are encouraged to work out the following identities:

Identity:	Missing:
$\omega = \omega_0 + \alpha_0 t$	$\theta - \theta_0$
$\theta = \theta_0 + \omega_0 t + \alpha_0 t^2/2$	ω
$\theta = \theta_0 + (\omega_0 + \omega) t/2$	α_0
$\theta = \theta_0 + \omega t - \alpha_0 t^2/2$	ω_0
$\omega^2 = \omega_0^2 + 2\alpha_0 (\theta - \theta_0)$	t

Velocity on a Circle

Given the position (x, y) for circular motion, we can work out the velocity v_x, v_y by first writing

$$\begin{aligned}x + \Delta x &= R \cos(\theta + \Delta\theta) \\ y + \Delta y &= R \sin(\theta + \Delta\theta),\end{aligned}$$

where $\Delta x, \Delta y, \Delta\theta$ are all ‘small’ parameters that update the position.

The trigonometric terms can be simplified using the small-angle approximations (8.77), (8.78) to write

$$\begin{aligned}x + \Delta x &= R \cos(\theta) - R \sin(\theta) \Delta\theta \\ y + \Delta y &= R \sin(\theta) + R \cos(\theta) \Delta\theta,\end{aligned}$$

simplifying to

$$\begin{aligned}\Delta x &= -R \sin(\theta) \Delta\theta \\ \Delta y &= R \cos(\theta) \Delta\theta.\end{aligned}$$

Per convenience of this setup, it is not required that ω is constant. Proceed by dividing through by Δt and simplify to arrive at

$$\begin{aligned}v_x(t) &= \frac{\Delta x}{\Delta t} = -y(t) \omega(t) \\ v_y(t) &= \frac{\Delta y}{\Delta t} = x(t) \omega(t),\end{aligned}$$

and furthermore

$$v(t) = \sqrt{v_x^2 + v_y^2} = R\omega(t).$$

To remind, the arc length traced on the circle for a given angular interval is written

$$\Delta S = R\Delta\theta,$$

and dividing through by Δt delivers the same result $v = R\omega$.

8.1 Uniform Circular Motion

A special case called *uniform circular motion* has zero angular acceleration, characterized by $\omega(t) = \omega_0$. Importantly though, the linear acceleration $a(x, y)$ does not vanish. To see this, use the same ‘small parameter’ trick on the pair of velocity equations to write

$$\begin{aligned}v_x + \Delta v_x &= -(y + \Delta y) \omega_0 \\ v_y + \Delta v_y &= (x + \Delta x) \omega_0,\end{aligned}$$

expanding to

$$\begin{aligned}v_x + \Delta v_x &= v_x - R \cos(\theta) \Delta\theta \omega_0 \\ v_y + \Delta v_y &= v_y - R \sin(\theta) \Delta\theta \omega_0,\end{aligned}$$

or

$$\begin{aligned}\Delta v_x &= -R \cos(\theta) \Delta\theta \omega_0 \\ \Delta v_y &= -R \sin(\theta) \Delta\theta \omega_0.\end{aligned}$$

Divide through by Δt and simplify once more to attain the result:

$$\begin{aligned}a_x(t) &= \frac{\Delta v_x}{\Delta t} = -x(t) \omega_0^2 \\ a_y(t) &= \frac{\Delta v_y}{\Delta t} = -y(t) \omega_0^2\end{aligned}$$

Remarkably, the acceleration always points to the center of the circle. This should remind that the definition of acceleration is not only a change in speed, but a change in *direction* as well. The total linear acceleration is given by

$$a = -\frac{v^2}{R}.$$

Problem 9

A wheel of radius R is rotating with no angular acceleration. If a point on the rim of the wheel has tangential velocity v_0 , what angular acceleration is required to bring the wheel to rest in 10 seconds? Answer:

$$\alpha_0 = -\frac{v_0/R}{10 \text{ s}}$$

Problem 10

A bicycle with 700 mm diameter wheels is coasting by at 15 meters per second. The rider applies the handbrake that causes a constant angular deceleration $\alpha_0 = -0.15 \text{ rad/s}^2$

over an interval of 4 seconds. What is the final speed of the bicycle? Answer:

$$v = 15 \text{ m/s} + (700 \text{ mm}) (-0.15 \text{ rad/s}^2) (4 \text{ s})$$

Problem 11

Starting from rest, a wheel is given a constant angular acceleration $\alpha_0 = 0.250 \text{ rad/s}^2$ for the first 10 seconds of its motion. At $t = 10 \text{ s}$, the angular acceleration is switched to $\alpha_0 = 0$. How many revolutions has the wheel made in 20 seconds? Answer:

$$\begin{aligned}\omega_{10 \text{ s}} &= \alpha_0 (10 \text{ s}) \\ \theta_{10 \text{ s}} &= \frac{1}{2} \alpha_0 (10 \text{ s})^2 \\ \theta_{20 \text{ s}} &= \theta_{10 \text{ s}} + \omega_{10 \text{ s}} (10 \text{ s}) \\ N_{\text{rev}} &= \frac{\theta_{20 \text{ s}}}{2\pi}\end{aligned}$$

Problem 12

A traffic ‘roundabout’ or ‘rotary’ is approximately a circle. Suppose a car has an initial speed of 20 meters per second, and then the driver holds the brake for 3 seconds to decelerate at $-0.15 g$ before entering a roundabout with a 20 meter radius. Calculate the linear acceleration of the car while in the roundabout. Answer:

$$\begin{aligned}v_R &= 20 \text{ m/s} + (-0.15 \times 9.8 \text{ m/s}^2) (3 \text{ s}) \\ a &= -\frac{v_R^2}{20 \text{ m}}\end{aligned}$$

9 Lissajous Curves

Chapter 9

Conic Sections

1 Ellipse

1.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labeled *directrix*. Now, let us seek the set of points $\{s\} = \{(x, y)\}$ that satisfy the following rule: the distance R to the focus divided by the (purely horizontal) distance Q to the directrix equals a constant $e < 1$. In algebraic terms, this means

$$\frac{R}{Q} = e < 1. \quad (9.1)$$

Sketched in Figure 9.1 are some of the points that obey such a rule.

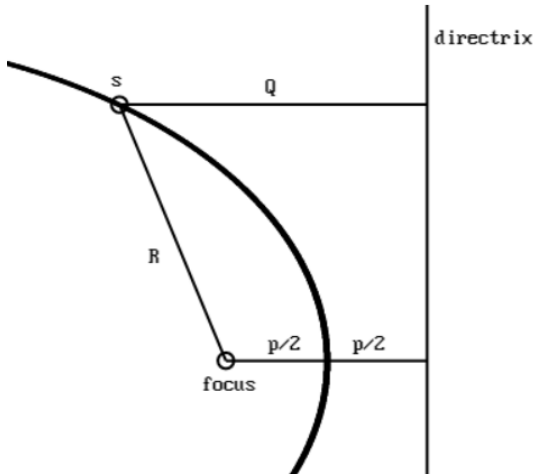


Figure 9.1. Points obeying $R/Q = e < 1$ as measured from a directrix and a focus separated by distance p . The origin is at the focus.

To determine the proper shape defined by the rule, begin with Equation (9.1) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \quad (9.2)$$

$$Q = p - x \quad (9.3)$$

Inserting the above into Equation (9.1) and completing the square in x , one finds

$$\frac{(x + c)^2}{a^2} + \frac{y^2}{b^2} = 1, \quad (9.4)$$

describing an *ellipse* centered at $x = -c$. The *semimajor axis* a , *semiminor axis* b , and offset c relate to e, p by:

$$a = \frac{ep}{1 - e^2} \quad (9.5)$$

$$b = \frac{ep}{\sqrt{1 - e^2}} \quad (9.6)$$

$$c = ae \quad (9.7)$$

The ellipse has two *vertex* points at $(-c \pm a, 0)$, and two *covertex* points at $(-c, \pm b)$.

Problem 1

Derive Equation (9.4) simultaneously with Equations (9.5), (9.6), (9.7).

1.2 Eccentricity

The constant e is called the *eccentricity* of the ellipse, and characterizes the proportions of the semimajor and minor axes. The special case $e = 0$ reduces the ellipse to a circle of radius $r = a = b$.

1.3 Internal Relations

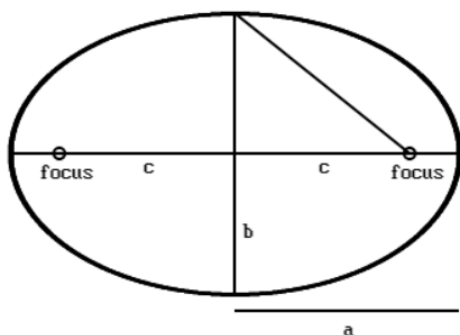


Figure 9.2. Ellipse displaying internal relations between a , b , c .

Having established that the set of points obeying $R/Q = e < 1$ forms an ellipse, we label the semimajor and semiminor axes, along with the center-to-focus distance as shown in Figure 9.2. Note that the minor axis b is *never* greater than the major axis a .

Problem 2

Derive the internal relations:

$$a^2 - b^2 = c^2 \quad (9.8)$$

$$e = \sqrt{1 - \frac{b^2}{a^2}} \quad (9.9)$$

Problem 3

Determine the length of the line segment connecting the upper vertex (height b from the center) to the focus (horizontal distance c from the center).

Semilatus Rectum

Problem 4

The *semilatus rectum* is the vertical distance from the focus to the ellipse. Show this is equal to b^2/a .

1.4 Symmetry

Reflected Origin

We decided by writing Equation (9.2) that the origin is placed at the focus of the ellipse, which is to say the origin is not at the ellipse's geometric center. Due the vertical symmetry of our construction, there also exists a complementary focus with its own directrix in the mirror image of the ellipse as shown in Figure 9.3. Should we wish to choose to rebuild using the 'left' focus as the origin, the

resulting equation is complimentary to Equation (9.4), with the sign on c reversing:

$$\frac{(x - c)^2}{a^2} + \frac{y^2}{b^2} = 1$$

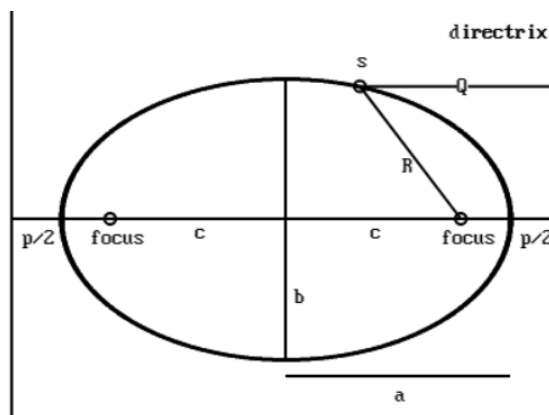


Figure 9.3. Vertical symmetry of the ellipse implies another directrix and focus.

1.5 Translations

Centered Origin

Placing the origin at the geometric center, the most symmetric equation of the ellipse reads

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1. \quad (9.10)$$

Having no offset term, the foci are located symmetrically at $x = \pm c$.

Shifted Origin

An ellipse centered at the point (x_0, y_0) is represented by

$$\frac{(x - x_0)^2}{a^2} + \frac{(y - y_0)^2}{b^2} = 1. \quad (9.11)$$

Problem 5

For the ellipse

$$5x^2 + y^2 - 20x = 0,$$

find the major and minor axes, the center, the eccentricity, the vertex points, the covertex points, and the foci. Answer: major = $2\sqrt{5}$, minor = 2, center = $(2, 0)$, $e = 2/\sqrt{5}$, vertices = $(2, \pm 2\sqrt{5})$, covertices = $(2 \pm 2, 0)$, foci = $(2, \pm 4)$

Problem 6

For the ellipse

$$x^2 + 2y^2 + 4y - 6 = 0,$$

find the major and minor axes, the center, the eccentricity, the vertex points, the covertex points, and the foci. Answer: major = $2\sqrt{2}$, minor = 2, center = $(0, -1)$, $e = 1/\sqrt{2}$, vertices = $(\pm 2\sqrt{2}, -1)$, covertices = $(0, -1 \pm 2)$, foci = $(\pm 2, -1)$

Problem 7

Find the equation of the ellipse with vertices at $(3, 1)$ and $(-1, 1)$ and eccentricity $e = 2/3$. Answer: $(x-1)^2/4 + 9(y-1)^2/20 = 1$

1.6 Polar Representation

In polar coordinates, a point (x, y) in the Cartesian plane is represented by

$$\begin{aligned} x &= r \cos(\theta) \\ y &= r \sin(\theta), \end{aligned}$$

where r is the distance to the origin and θ is the angular parameter. These can be inverted to solve for r , θ with respect to x , y :

$$\begin{aligned} r &= \sqrt{x^2 + y^2} \\ \theta &= \arctan\left(\frac{y}{x}\right) \end{aligned}$$

The definition (9.1) combined with Equations (9.2), (9.3) lends naturally to polar coordinates:

$$e = \frac{R}{Q} = \frac{r}{p-x} = \frac{r}{p-r\cos(\theta)}$$

Solving for $r(\theta)$, one finds

$$r = \frac{pe}{1 + e\cos(\theta)}. \quad (9.12)$$

Equation (9.12) traces an ellipse in the plane from an origin placed at the 'right' focus ($x = c$). To trace the ellipse from the 'left' focus ($x = -c$), we change the sign on the cosine term:

$$r = \frac{pe}{1 - e\cos(\theta)}$$

Problem 8

Show that the polar representation (focus on the left)

$$r = \frac{b^2/a}{1 - e\cos(\theta)}$$

is equivalent to the Cartesian version centered at $(ae, 0)$:

$$\frac{(x - ae)^2}{a^2} + \frac{y^2}{b^2} = 1$$

1.7 Parametric Representation

Consider an ellipse centered at the origin described by Equation (9.10):

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

By comparing the above to the fundamental identity of trigonometry, namely

$$\cos^2(\phi) + \sin^2(\phi) = 1$$

for any angle ϕ , we cannot help but make the association

$$x = a \cos(\phi) \quad (9.13)$$

$$y = b \sin(\phi). \quad (9.14)$$

Equations (9.13), (9.14) constitute a *parametric representation* of the ellipse.

Problem 9

Check that Equations (9.13), (9.14) combine to recover Equation (9.10).

1.8 Interior Identities

Sum of Radii

Consider a point (x, y) on an ellipse centered on the origin, and let

$$\begin{aligned} r_1 &= \sqrt{(x+c)^2 + y^2} \\ r_2 &= \sqrt{(x-c)^2 + y^2} \end{aligned}$$

be the distance from (x, y) to each respective focus as shown in Figure 9.4. By brute force, we can show that the sum of r_1 and r_2 is a *constant*. Proceed by writing

$$A = r_1 + r_2,$$

and square both sides to get

$$\begin{aligned} A^2 &= 2(x^2 + y^2 + c^2) \\ &\quad + 2\sqrt{(x^2e^2 + a^2 + 2cx)(x^2e^2 + a^2 - 2cx)}, \end{aligned}$$

simplifying further to

$$A^2 = 2(x^2e^2 + a^2) + 2(a^2 - x^2e^2) .$$

Performing the final cancelation, we find $A^2 = 4a^2$, or

$$r_1 + r_2 = 2a . \quad (9.15)$$

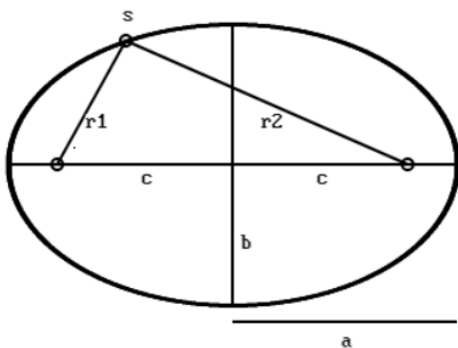


Figure 9.4. Line segments $r_{1,2}$ connect each focus to the point s on the ellipse. The sum $r_1 + r_2$ always equals the constant $2a$.

Drawing an Ellipse

The interior length identity (9.15) teaches how to draw an ellipse in the plane. Fix two pins a distance $2c$ apart, and then place a closed loop of string around the pins. Use a pen to pull the string tight, and trace around the pins while maintaining tension. The resultant shape is an ellipse with a pin at each focus.

Problem 10

Derive Equation (9.15).

Problem 11

An ellipse with eccentricity $e = 0.5$ is traced in the plane using two pins and a string. In terms of the semimajor axis a , how far apart are the pins and how long is the string?

Difference of Radii

We learned from Equation (9.15) that the sum of the interior radii $r_1 + r_2$ in the ellipse always yields the constant $2a$. Naturally one wonders if the difference of radii $r_2 - r_1$ simplifies in any nice way. Recycling most of the work done previously, we quickly find

$$r_1 - r_2 = 2xe . \quad (9.16)$$

Problem 12

Derive Equation (9.16).

¹A trick from calculus.

Decoupled Identities

Having Equations (9.15) and (9.16) in hand, we can isolate each of $r_{1,2}$ to yield a pair of tight formulas representing the ellipse:

$$r_1 = a + xe \quad (9.17)$$

$$r_2 = a - xe \quad (9.18)$$

1.9 Tangent Line to the Ellipse

At a point $s = (x, y)$ on an ellipse, there exists a *tangent line* AB that represents the instantaneous slope m_s of the ellipse as shown in Figure 9.6. The value of m_s is straightforwardly attained by implicit differentiation¹ of (9.10), which comes out to

$$m_s = \frac{-b^2 x}{a^2 y} . \quad (9.19)$$

Problem 13

At a point $(\tilde{x}, \tilde{y})$ on the ellipse $x^2/a^2 + y^2/b^2 = 1$, show that the tangent line is

$$\frac{x\tilde{x}}{a^2} + \frac{y\tilde{y}}{b^2} = 1 .$$

Derivation of Slope

If ‘implicit differentiation’ sounds foreign, we must calculate m_s the hard way by asking about the slope m_c of a chord connecting two points $s_1 = (x_1, y_1)$, $s_2 = (x_2, y_2)$ on the ellipse as sketched in Figure 9.5. Jotting down the rise-over-run formula for the slope, we have

$$m_c = \frac{y_2 - y_1}{x_2 - x_1} ,$$

and by placing the origin at the center of the ellipse, we use Equation (9.10) to eliminate y_1, y_2 such that

$$m_c = \frac{b\sqrt{1 - x_2^2/a^2} - b\sqrt{1 - x_1^2/a^2}}{x_2 - x_1} .$$

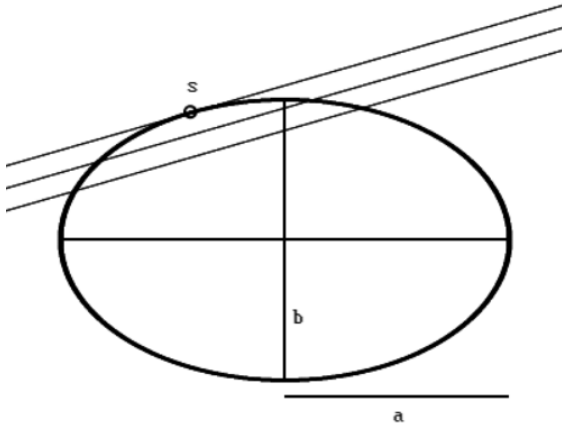


Figure 9.5. A chord cutting through the ellipse is constructed on the ellipse's edge so the endpoints share a common location s . The extension of the chord at s is the tangent line to the ellipse having slope m_s . The origin is at the center.

Our next move is to suppose that the points s_1, s_2 are very close to each other such that

$$h = x_2 - x_1$$

is a small number. In such a case, the chord connecting s_1 to s_2 becomes shorter and moves closer to the edge of the ellipse. In the limit that s_1 is approximately equal to s_2 , the chord has essentially zero length, but has a slope equal to that of the tangent line to the ellipse at $s_{1,2}$. To summarize, the slope is written in 'limit' notation:

$$m_s = \lim_{h \text{ small}} \frac{b\sqrt{1 - (x_1 + h)^2/a^2} - b\sqrt{1 - x_1^2/a^2}}{h}$$

To simplify this, we exploit the 'smallness' of h by realizing that h^2 is so small that it can be ignored altogether such that

$$(x_1 + h)^2 \approx x_1^2 + 2x_1h + h^2.$$

The argument inside the square root shall be handled by the approximation

$$\lim_{z \text{ small}} \sqrt{1 + z} \approx 1 + \frac{z}{2} - \frac{z^2}{8} + \dots, \quad (9.20)$$

where any terms of power 2 or higher can be ignored. For our problem, this means

$$b\sqrt{1 - \frac{x_1^2 + 2x_1h}{a^2}} \approx y_1 - \frac{x_1b^2h}{y_1a^2},$$

allowing the formula for m_s to simplify with all factors of h canceling out:

$$m_s = \lim_{h \text{ small}} \frac{1}{h} \left(y_1 - \frac{x_1b^2h}{y_1a^2} - y_1 \right) = -\frac{b^2x_1}{a^2y_1}$$

Note that the 1-subscript becomes redundant, as the points s_1 and s_2 become the same point s (with no subscript). Finally then, we recover Equation (9.19) for the slope of the ellipse at point s .

1.10 Reflection Property

Consider an ellipse centered on the origin as shown in Figure 9.6, with respective foci labeled $f_{1,2}$. The radii extending to a point $s = (x, y)$ on the ellipse are labeled $r_{1,2}$, and the tangent line AB is indicated. The *reflection property* of the ellipse states that *a ray emerging from one focus will reflect from the ellipse to the other focus*. To prove this, let us define:

$$\text{Angle } Asf_1 = \theta$$

$$\text{Angle } Bs f_2 = \phi$$

$$\text{Slope of } AB = m_s$$

$$\text{Slope of } r_1 = m_1$$

$$\text{Slope of } r_2 = m_2$$

The slopes m_1, m_2 are straightforward to write by inspection of Figure 9.6:

$$m_1 = \frac{y}{x + c}$$

$$m_2 = \frac{y}{x - c}$$

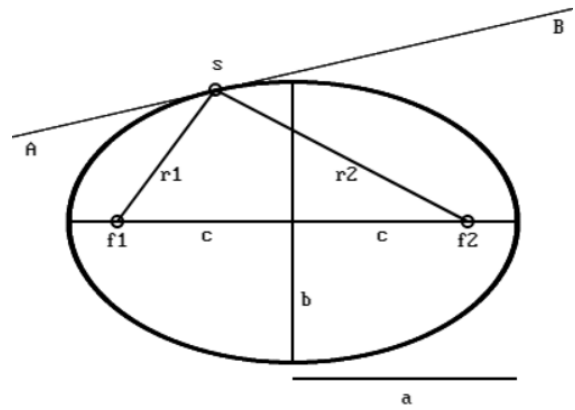


Figure 9.6. A ray emerging from one focus will reflect from the ellipse to the other focus.

Next, we'll need to use the angle-sum identity for tangent, namely

$$\tan(\alpha + \beta) = \frac{\tan(\alpha) \pm \tan(\beta)}{1 \mp \tan(\alpha) \tan(\beta)}, \quad (9.21)$$

and observe again from the Figure that

$$\begin{aligned} \tan(\theta) &= \frac{m_1 - m_s}{1 + m_1 m_s} \\ \tan(\phi) &= \frac{m_2 - m_s}{1 + m_2 m_s}. \end{aligned}$$

Simplifying each expression delivers

$$|\tan(\theta)| = |\tan(\phi)| = \frac{b^2}{cy}, \quad (9.22)$$

telling us that $\theta = \phi$ and the proof is done.

Problem 14

Prove Equation (9.22).

1.11 Normal Line to the Ellipse

Consider a *normal line* q that is perpendicular to the tangent line at point $s = (x, y)$ on the ellipse as shown in Figure 9.7. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by (9.19). The normal line q can thus be written

$$y_q = -x_q/m_s + b_q,$$

with $b_q = y + x/m_s$. Such a line is more conveniently expressed as

$$y_q = y + (x - x_q)/m_s. \quad (9.23)$$

The normal line intersects the x -axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$\begin{aligned} q_0 &= x + m_s y = x + \left(\frac{-b^2 x}{a^2} \frac{x}{y} \right) y \\ q_0 &= x \left(1 - \frac{b^2}{a^2} \right) = x e^2 \end{aligned} \quad (9.24)$$

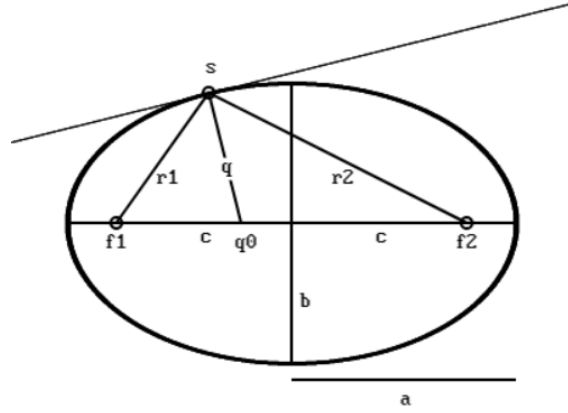


Figure 9.7. A point s on the ellipse implies a normal line q that intersects the x -axis at $x = q_0$. The origin is at the focus.

Problem 15

Determine where the normal line intersects the y -axis.

Problem 16

At a point $(\tilde{x}, \tilde{y})$ on the ellipse $x^2/a^2 + y^2/b^2 = 1$, show that the normal line is

$$\frac{x}{\tilde{x}} - \frac{y}{\tilde{y}} (1 - e^2) = e^2.$$

1.12 Vector Analysis (Optional)

Three Vectors

Referring again to Figure 9.7, let us construct three vectors from straight lines:

$$\text{Line } f_1 s = \mathbf{r}_1$$

$$\text{Line } f_2 s = \mathbf{r}_2$$

$$\text{Line } q_0 s = \mathbf{q}$$

Introducing the unit vector $\hat{\mathbf{x}}$ that points horizontally to the right, we can jot down three ways to get from the origin to point s :

$$\mathbf{s} = -c\hat{\mathbf{x}} + \mathbf{r}_1 \quad (9.25)$$

$$\mathbf{s} = c\hat{\mathbf{x}} + \mathbf{r}_2 \quad (9.26)$$

$$\mathbf{s} = q_0\hat{\mathbf{x}} + \mathbf{q} \quad (9.27)$$

Solving (9.25), (9.26) for $\mathbf{r}_1, \mathbf{r}_2$ respectively, we can divide each by its own magnitude to write unit vectors:

$$\hat{\mathbf{r}}_1 = \frac{\mathbf{s} + c\hat{\mathbf{x}}}{r_1}$$

$$\hat{\mathbf{r}}_2 = \frac{\mathbf{s} - c\hat{\mathbf{x}}}{r_2}$$

Recovering Reflection Property

In terms of the vectors $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{q}$, the reflection property of the ellipse can be proposed by writing

$$\hat{\mathbf{r}}_1 \cdot \mathbf{q} = \hat{\mathbf{r}}_2 \cdot \mathbf{q}, \quad (9.28)$$

which is to claim that the angle formed between either $\hat{\mathbf{r}}_j$ and $\mathbf{q}$ is the same. Proceeding carefully, one can simplify the above down to

$$\frac{b^2}{a} \left(\frac{a + xe}{r_1} \right) = \frac{b^2}{a} \left(\frac{a - xe}{r_2} \right), \quad (9.29)$$

and the claim is proven. The parenthesized terms cancel due to Equations (9.17), (9.18). Evidently, we discover that the dot product between either unit vector $\hat{\mathbf{r}}_{1,2}$ and the normal vector $\mathbf{q}$ equals a constant:

$$\hat{\mathbf{r}}_{1,2} \cdot \mathbf{q} = \frac{b^2}{a} \quad (9.30)$$

Problem 17

Derive Equation (9.29) from (9.28).

Recovering Interior Length Identities

Next, we make use of the reflection property of the ellipse to realize that the sum of $\hat{\mathbf{r}}_1$ and $\hat{\mathbf{r}}_2$ must be parallel to (i.e. proportional to) the vector $\mathbf{q}$. Following this lead, we first take the sum

$$\hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 = s \left(\frac{1}{r_1} + \frac{1}{r_2} \right) + c \left(\frac{1}{r_1} - \frac{1}{r_2} \right) \hat{\mathbf{x}},$$

simplifying down to

$$\hat{\mathbf{r}}_1 + \hat{\mathbf{r}}_2 = \left(\frac{r_1 + r_2}{r_1 r_2} \right) \left(s + \frac{e}{2} (r_2 - r_1) \right) \hat{\mathbf{x}} \quad (9.31)$$

Note that the interior length identity $r_1 + r_2 = 2a$ was used along the way. Comparing the above to Equation (9.27), we must have

$$q_0 = \frac{e}{2} (r_1 - r_2). \quad (9.32)$$

Problem 18

Derive Equations (9.31), (9.32) and then recover Equation (9.16).

2 Hyperbola

2.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labeled *directrix*. Now, let us seek the set of points $\{s\} = \{(x, y)\}$ that satisfy the following rule: the distance R to the focus divided by the (purely horizontal) distance Q to the directrix equals a constant $e > 1$. In algebraic terms, this means

$$\frac{R}{Q} = e > 1. \quad (9.33)$$

Sketched in Figure 9.8 are some of the points that obey such a rule.

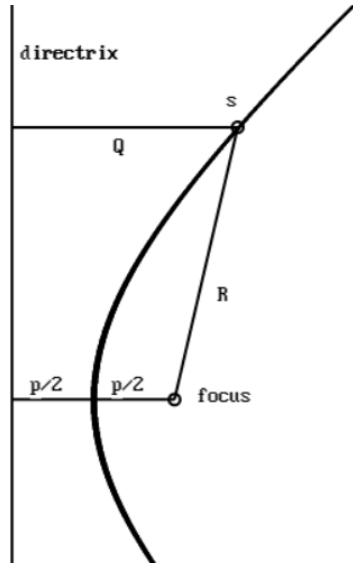


Figure 9.8. Points obeying $R/Q = e > 1$ as measured from a directrix and a focus separated by distance p . The origin is at the focus.

To determine the proper shape defined by the rule, begin with (9.33) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \quad (9.34)$$

$$Q = p + x \quad (9.35)$$

Inserting the above into (9.33) and completing the square in x , one finds

$$\frac{(x + c)^2}{a^2} - \frac{y^2}{b^2} = 1, \quad (9.36)$$

describing a *hyperbola*. The constants a, b, c relate to e, p

by:

$$a = \frac{ep}{e^2 - 1} \quad (9.37)$$

$$b = \frac{ep}{\sqrt{e^2 - 1}} \quad (9.38)$$

$$c = ae \quad (9.39)$$

The hyperbola has two vertex points occurring at $x = \pm a - c$.

2.2 Asymptotes

The hyperbola is not a closed curve like a circle or ellipse, and it's worthwhile to inquire about the hyperbola far from the origin. Supposing we let x and y be very large, especially such that $x \gg c$, Equation (9.36) roughly reads

$$\frac{x^2}{a^2} \approx \frac{y^2}{b^2},$$

implying a pair of straight lines having slope

$$m_{\pm} = \pm \frac{b}{a} \quad (9.40)$$

that are *asymptotes* to the hyperbola.

To find an exact equation for each asymptote, begin with the equation of the hyperbola (9.36) and let $y = 0$ to determine the value x^* at which the curve touches the x -axis. Doing so, we find

$$x^* = \begin{cases} a - c \\ -a - c \end{cases}.$$

Each x -intercept is negative, i.e. to the left of the focus, however the curve sketched in Figure 9.8 has a focus at $a - c$ and 'opens up' to the right. Evidently, a second x -intercept occurs at $-a - c$, implying a mirror image hyperbola opening up to the left.

The line of vertical symmetry between each copy of the hyperbola is defined by the average of each x^* -value, or

$$x_{\text{ave}}^* = \frac{(a - c) + (-a - c)}{2} = -c$$

By horizontal symmetry, we argue that each asymptote crosses the x -axis at x_{ave}^* . This is enough to determine the equation of each asymptote, coming out to

$$y = \pm \frac{b}{a} (x + c). \quad (9.41)$$

Our findings are summarized in Figure 9.9.

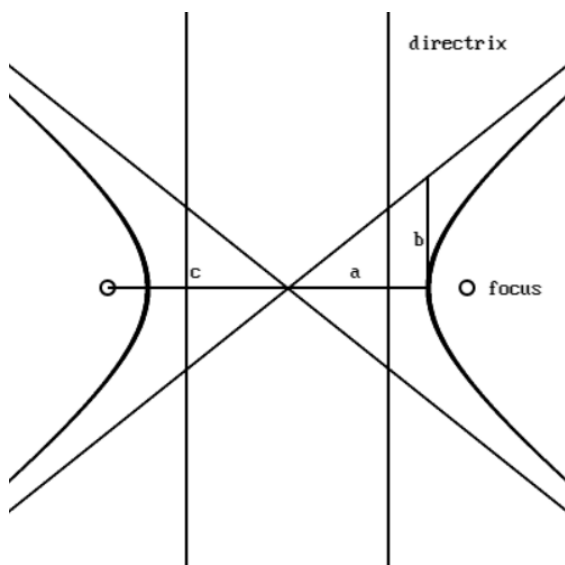


Figure 9.9. Vertical symmetry of the hyperbola implies another directrix and focus.

2.3 Internal Relations

Problem 19

Derive the internal relations:

$$a^2 + b^2 = c^2 \quad (9.42)$$

$$e = \sqrt{\frac{b^2}{a^2} + 1} \quad (9.43)$$

Problem 20

Show that a is the distance from the vertex of the hyperbola to the intersection of the asymptotes. Show that b is the vertical distance from the focus to the asymptote. Show that c is the distance from the focus to the intersection of the asymptotes.

2.4 Symmetry

Reflected Origin

We decided by wiring Equation (9.34) that the origin is placed at the right focus of the hyperbola, which is to say the origin is not at the hyperbola's geometric center. Due the vertical symmetry of our construction, there also exists a complimentary focus with its own directrix in the mirror image of the hyperbola as shown in Figure 9.9. Should we wish to choose to rebuild using the 'left' focus as the origin,

the resulting equation is complimentary to (9.36), with the sign on c reversing:

$$\frac{(x - c)^2}{a^2} - \frac{y^2}{b^2} = 1$$

2.5 Translations

Centered Origin

Placing the origin at the geometric center, the most symmetric equation of the hyperbola reads

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \quad (9.44)$$

Having no offset term, the foci are located symmetrically at $x = \pm c$.

Shifted Origin

A hyperbola centered at the point (x_0, y_0) is represented by

$$\frac{(x - x_0)^2}{a^2} - \frac{(y - y_0)^2}{b^2} = 1. \quad (9.45)$$

Problem 21

For the hyperbola

$$16y^2 - 9x^2 = 144,$$

find the major and minor axes, the center, the eccentricity, the vertex points, the asymptotes, and the foci. Answer: major = 4, minor = 3, center = $(0, 0)$, $e = 5/4$, vertices = $(0, \pm 3)$, $y = \pm 3x/4$, foci = $(0, \pm 5)$

Problem 22

For the hyperbola

$$12x^2 - 32y^2 - 12x + 96y + 27 = 0,$$

find the major and minor axes, the center, the eccentricity, the vertex points, the asymptotes, and the foci. Answer: major = $\sqrt{3}$, minor = $2\sqrt{2}$, center = $(1/2, 3/2)$, $e = \sqrt{11/3}$, vertices = $(1/2, 3/2 \pm \sqrt{3})$, $y = \pm \sqrt{3/8}x \mp \sqrt{3/32} + 3/2$, foci = $(1/2, 3/2 \pm \sqrt{11})$

Problem 23

Find the equation of the hyperbola with vertices at $(0, \pm 2)$ with asymptotes $y = \pm x/2$. Answer: $y^2/4 - x^2/16 = 1$

Problem 24

Find the equation of the hyperbola with focus points $(7, 0)$ and $(-1, 0)$ passes through $(6, \sqrt{15})$. Answer: $(x - 3)^2/4 - y^2/12 = 1$

2.6 Polar Representation

In polar coordinates, recall that a point (x, y) in the Cartesian plane is represented by

$$\begin{aligned}x &= r \cos(\theta) \\y &= r \sin(\theta),\end{aligned}$$

where r is the distance to the origin and θ is the angular parameter. These can be inverted to solve for r , θ with respect to x , y :

$$\begin{aligned}r &= \sqrt{x^2 + y^2} \\ \theta &= \arctan\left(\frac{y}{x}\right)\end{aligned}$$

The definition (9.33) combined with (9.34), (9.35) lends naturally to polar coordinates:

$$e = \frac{R}{Q} = \frac{r}{p+x} = \frac{r}{p+r \cos(\theta)}$$

Solving for $r(\theta)$, one finds

$$r = \frac{pe}{1 - e \cos(\theta)}. \quad (9.46)$$

Equation (9.46) traces a hyperbola in the plane from an origin placed at the ‘right’ focus ($x = c$). To trace the hyperbola from the ‘left’ focus ($x = -c$), we change the sign on the cosine term:

$$r = \frac{pe}{1 + e \cos(\theta)}$$

2.7 Parametric Representation

Consider a hyperbola centered at the origin described by Equation (9.44):

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

By comparing the above to the fundamental identity of hyperbolic trigonometry, namely

$$\cosh(\phi)^2 - \sinh(\phi)^2 = 1$$

for any value of ϕ , we cannot help but make the association

$$x = a \cosh(\phi) \quad (9.47)$$

$$y = b \sinh(\phi). \quad (9.48)$$

Equations (9.47), (9.48) constitute a parametric representation of the hyperbola.

Problem 25

Check that Equations (9.47), (9.48) combine to recover Equation (9.44).

2.8 Interior Identities

Difference of Radii

Consider a point (x, y) on a hyperbola centered on the origin, and let

$$\begin{aligned}r_1 &= \sqrt{(x+c)^2 + y^2} \\ r_2 &= \sqrt{(x-c)^2 + y^2}\end{aligned}$$

be the distance from (x, y) to each respective focus as shown in Figure 9.10. By brute force, we can show that the difference between r_1 and r_2 is a *constant*. Proceed by writing

$$A = r_1 - r_2,$$

and square both sides to get

$$\begin{aligned}A^2 &= 2(x^2 + y^2 + c^2) \\ &\quad - 2\sqrt{(x^2e^2 + a^2 + 2cx)(x^2e^2 + a^2 - 2cx)},\end{aligned}$$

simplifying further to

$$A^2 = 2(x^2e^2 + a^2) + 2(a^2 - x^2e^2).$$

Performing the final cancellation, we find $A^2 = 4a^2$, or

$$r_1 - r_2 = 2a. \quad (9.49)$$

Problem 26

Derive Equation (9.49).

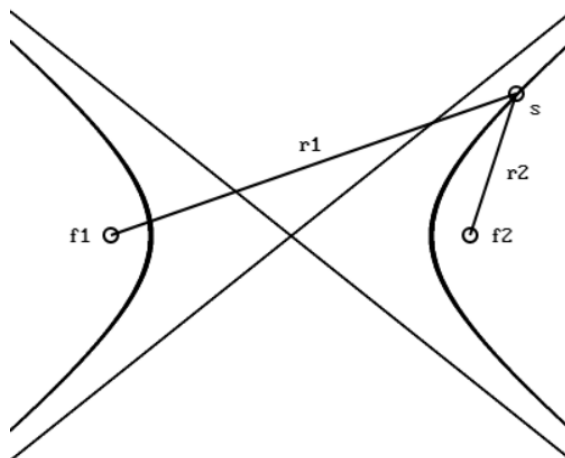


Figure 9.10. Line segments $r_{1,2}$ connect each focus to the point s on the hyperbola. The difference $r_1 - r_2$ always equals a constant $2a$.

Locating Ships

The interior length identity (9.49) teaches how to locate ships at sea by noticing that a signal emitted from any point on a hyperbola will reach each focus a fixed time apart. If the signal propagation speed (the speed of light for radar) is v , then the time interval Δt is $2a/v$. Supposing two receiver stations are separated by distance $d = 2c$ on land, we use the internal relation $a^2 + b^2 = c^2$ to write

$$b = \pm \sqrt{c^2 - a^2} = \pm \frac{1}{2} \sqrt{d^2 - v^2 \Delta t^2},$$

telling us the ship is somewhere on a known hyperbola. The ship's exact location can be discerned using a third station and the intersection of two hyperbolas.

Sum of Radii

We learned from Equation (9.49) that the difference of the interior radii $r_1 - r_2$ in the hyperbola always yields the constant $2a$. Naturally one wonders if the sum of radii $r_2 + r_1$ simplifies in any nice way. Recycling most of the work done previously, we quickly find

$$r_1 + r_2 = 2xe. \quad (9.50)$$

Problem 27

Derive Equation (9.50).

Decoupled Identities

Having Equations (9.49) and (9.50) in hand, we can isolate each of $r_{1,2}$ to yield a pair of tight formulas representing the hyperbola:

$$r_1 = a + xe \quad (9.51)$$

$$r_2 = -a + xe \quad (9.52)$$

2.9 Tangent Line to the Hyperbola

At a point $s = (x, y)$ on a hyperbola, there exists a tangent line AB that represents the instantaneous slope m_s of the hyperbola as shown in Figure 9.11. The value of m_s is straightforwardly attained by implicit differentiation of Equation (9.44), which comes out to

$$m_s = \frac{b^2 x}{a^2 y}. \quad (9.53)$$

Problem 28

At a point $(\tilde{x}, \tilde{y})$ on the hyperbola $x^2/a^2 - y^2/b^2 = 1$, show that the tangent line is

$$\frac{x\tilde{x}}{a^2} - \frac{y\tilde{y}}{b^2} = 1.$$

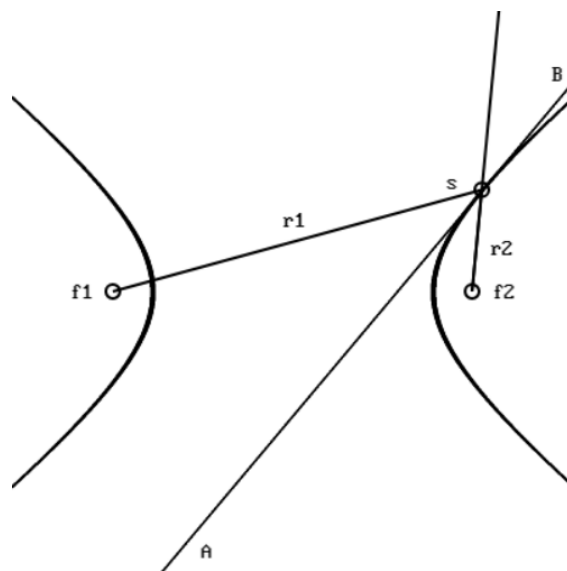


Figure 9.11. An incoming ray aimed at f_2 intersects the hyperbola at $s = (x, y)$. The reflected ray goes toward f_1 . The angle formed between the incoming ray and the tangent line AB is identical to angle Asf_2 . The origin is at the focus.

2.10 Reflection Property

Consider a hyperbola centered on the origin as shown in Figure 9.11, with respective foci labeled $f_{1,2}$. The radii extending to a point $s = (x, y)$ on the hyperbola are labeled $r_{1,2}$, and the tangent line AB is indicated. The reflection property of the hyperbola states that *an external ray aimed at a focus will be reflected by the hyperbola to the other focus*. To prove this, let us define:

$$\text{Angle } Asf_1 = \theta$$

$$\text{Angle } Asf_2 = \phi$$

$$\text{Slope of } AB = m_s$$

$$\text{Slope of } r_1 = m_1$$

$$\text{Slope of } r_2 = m_2$$

The slopes m_1, m_2 are straightforward to write by inspection of Figure 9.11:

$$m_1 = \frac{y}{x + c}$$

$$m_2 = \frac{y}{x - c}$$

Next, we'll need to use the angle-sum identity for tangent, namely

$$\tan(\alpha + \beta) = \frac{\tan(\alpha) \pm \tan(\beta)}{1 \mp \tan(\alpha) \tan(\beta)},$$

and observe again from the Figure that

$$\begin{aligned}\tan(\theta) &= \frac{m_1 - m_s}{1 + m_1 m_s} \\ \tan(\phi) &= \frac{m_2 - m_s}{1 + m_2 m_s}.\end{aligned}$$

Simplifying each expression delivers

$$|\tan(\theta)| = |\tan(\phi)| = \frac{b^2}{cy}, \quad (9.54)$$

telling us that $\theta = \phi$ and the proof is done.

Problem 29

Prove Equation (9.54).

2.11 Normal Line to the Hyperbola

Consider a normal line q that is perpendicular to the tangent line at point $s = (x, y)$ on the hyperbola as shown in Figure 9.12. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by Equation (9.53). The normal line q can thus be written

$$y_q = -x_q/m_s + b_q,$$

with $b_q = y + x/m_s$. Such a line is more conveniently expressed as

$$y_q = y + (x - x_q)/m_s. \quad (9.55)$$

The normal line intersects the x -axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$\begin{aligned}q_0 &= x + m_s y = x + \left(\frac{b^2}{a^2} \frac{x}{y}\right) y \\ q_0 &= x \left(1 + \frac{b^2}{a^2}\right) = x e^2\end{aligned} \quad (9.56)$$

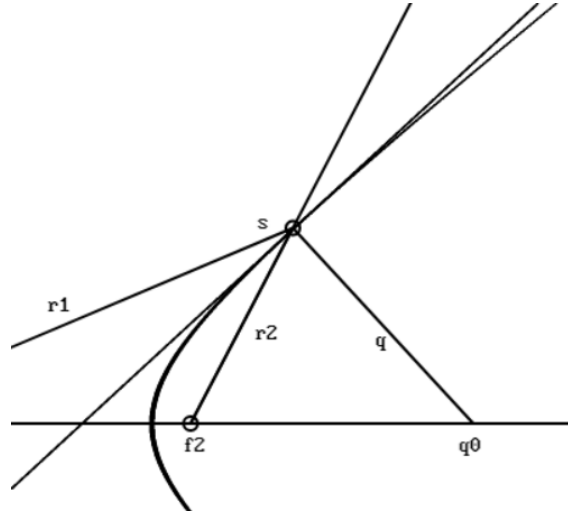


Figure 9.12. A point s on the hyperbola implies a normal line q that intersects the x -axis at $x = q_0$. The origin is at the focus.

Problem 30

Determine where the normal line intersects the y -axis.

Problem 31

At a point $(\tilde{x}, \tilde{y})$ on the hyperbola $x^2/a^2 - y^2/b^2 = 1$, show that the normal line is

$$\frac{x}{\tilde{x}} + \frac{y}{\tilde{y}} (e^2 - 1) = e^2.$$

3 Parabola

3.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labeled *directrix*. Now, let us seek the set of points $\{s\} = \{(x, y)\}$ that satisfy the following rule: the distance R to the focus divided by the (purely horizontal) distance Q to the directrix equals a constant $e = 1$. In algebraic terms, this means

$$\frac{R}{Q} = 1. \quad (9.57)$$

Sketched in Figure 9.13 are some of the points that obey such a rule.

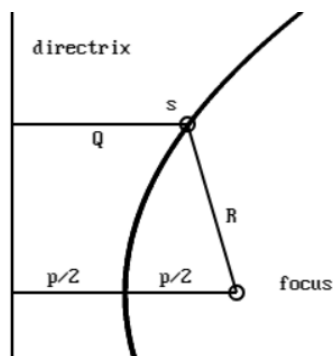


Figure 9.13. Points obeying $R/Q = 1$ as measured from a directrix and a focus separated by distance p . The origin is at the focus.

To determine the proper shape defined by the rule, begin with (9.57) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \quad (9.58)$$

$$Q = p + x \quad (9.59)$$

Inserting the above into (9.57), one finds

$$y^2 = p^2 + 2px, \quad (9.60)$$

describing a *parabola* that ‘opens up’ to the right. The parabola has one focus point (the second one is at infinity, if you like). If we want the parabola to open up to the left, place the focus on the other side of the directrix. This has the effect of reversing the sign on p . The vertex of the parabola occurs at $x = -p/2$. The line of symmetry halving the parabola is called the *axis*.

3.2 Opening Direction

A parabola is often found in the wild opening in the up- or down-direction (as opposed to left or right). To generate the

parabola that opens ‘upward’, let the directrix run horizontally and place the focus above it, effectively swapping the x - and y - variables in the equation of the parabola as shown in Figure 9.14:

$$x^2 = p^2 + 2py \quad (9.61)$$

The sign on the p -term determines the opening direction in either case.

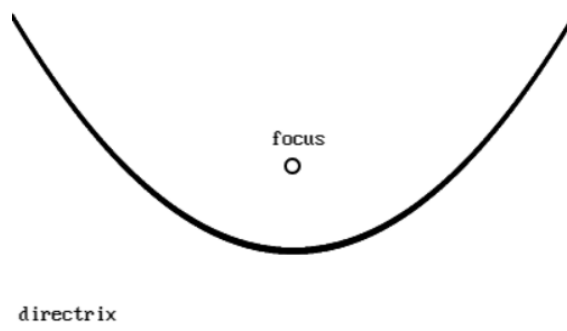


Figure 9.14. The directrix-focus construction rotated by ninety degrees produces an up- (or down-) opening parabola.

3.3 Parabolic Expressions

In Equations (9.60), (9.61), the value p^2 acts to translate the parabola vertically or horizontally. The term $2p$ is a scaling factor that stretches or skews the overall parabola. If we let scaling be handled by a new variable a , and let the translation vector (x_0, y_0) take care of the absolute placement of the parabola, Equation (9.60) is more generally written as

$$x = x_0 + a(y - y_0)^2. \quad (9.62)$$

By the same token, Equation (9.61) is more generally written as

$$y = y_0 + a(x - x_0)^2. \quad (9.63)$$

The sign on a dictates whether the parabola opens up right-left or up-down, respectively.

Problem 32

Sketch the parabola $x^2 - 2y - 6x = 0$, and show that the focus is located at $(3, -4)$ and that the directrix is located at $y = -5$.

Problem 33

The parabola $y^2 - 2ay + 2x - a^2 = 0$ has its focus on the y -axis above the origin. Find the number a and sketch the graph. Answer: $a = 1/\sqrt{2}$

Problem 34

An up- or down-opening parabola can be generally expressed as $y = ax^2 + bx + c$. In terms of a, b, c , find the vertex and the focus. Answer: $(-b/2a, c - b^2/4a)$, $1/4a$ above the vertex

3.4 Polar Representation

In polar coordinates, recall that a point (x, y) in the Cartesian plane is represented by

$$\begin{aligned}x &= r \cos(\theta) \\ y &= r \sin(\theta),\end{aligned}$$

where r is the distance to the origin and θ is the angular parameter. These can be inverted to solve for r, θ with respect to x, y :

$$\begin{aligned}r &= \sqrt{x^2 + y^2} \\ \theta &= \arctan\left(\frac{y}{x}\right)\end{aligned}$$

The definition (9.57) combined with Equations (9.58), (9.59) lends naturally to polar coordinates:

$$1 = \frac{R}{Q} = \frac{r}{p+x} = \frac{r}{p+r \cos(\theta)}$$

Solving for $r(\theta)$, one finds

$$r = \frac{p}{1 - \cos(\theta)}. \quad (9.64)$$

Equation (9.64) traces a parabola in the plane.

3.5 Internal Relations

Right Focal Chord

Problem 35

Show that in a parabola the length of the chord passing through the focus making an angle θ with the axis is equal to $L/\sin^2 \theta$, where L is the length of the *right focal chord*, the line that passes through the focus and is perpendicular to the axis. Hint: Use $x^2 = p^2 + 2py$ and then focal chords are given by $y = \cot(\theta)x$.

Problem 36

A parabolic segment (i.e. the area bounded by a parabola and a chord perpendicular to the axis) is 32 cm high and its base is 16 cm. How far is the focus from the directrix? Answer: 1 cm

3.6 Tangent Line to the Parabola

At a point $s = (x, y)$ on a parabola, there exists a tangent line AB that represents the instantaneous slope m_s of the parabola as shown in Figure 9.15. The value of m_s is straightforwardly attained by implicit differentiation of Equation (9.60), which comes out to

$$m_s = \frac{p}{y}. \quad (9.65)$$

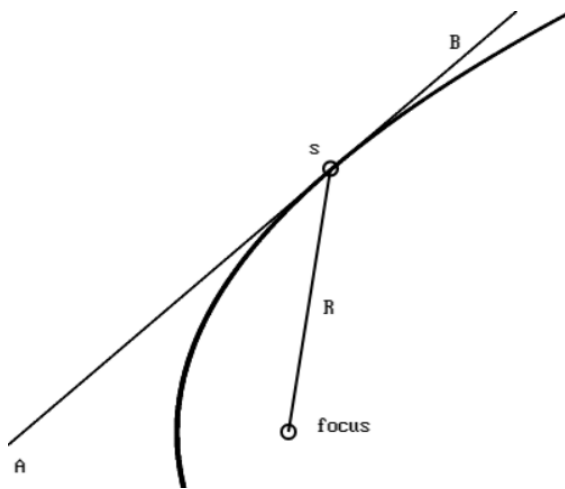


Figure 9.15. A point s on the parabola implies a tangent line AB that represents the instantaneous slope of the parabola. The origin is at the focus.

Problem 37

At a point $(\tilde{x}, \tilde{y})$ on the parabola $y^2 = p^2 + 2px$, show that the tangent line is

$$y\tilde{y} = p^2 + p(x + \tilde{x}).$$

3.7 Normal Line to the Parabola

Consider a normal line q that is perpendicular to the tangent line at point $s = (x, y)$ on the parabola as shown in Figure 9.16. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by (9.65). The normal line q can thus be written

$$y_q = -x_q/m_s + b_q,$$

with $b_q = y + x/m_s$. Such a line is more conveniently expressed as

$$y_q = y + (x - x_q)/m_s. \quad (9.66)$$

The normal line intersects the x -axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$q_0 = x + m_s y = x + p \quad (9.67)$$

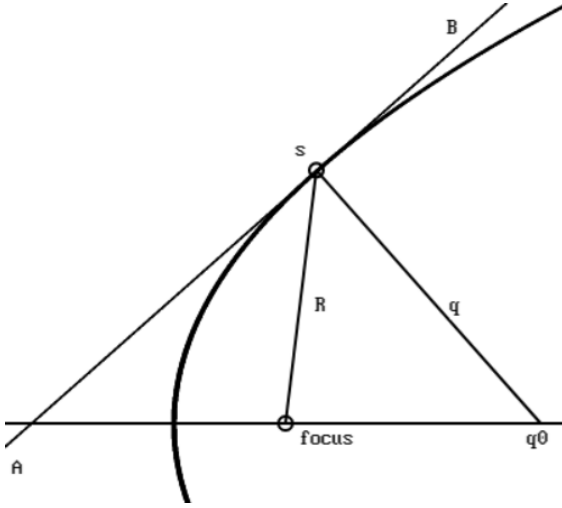


Figure 9.16. A point s on the parabola implies a normal line q that intersects the x -axis at $x = q_0$. The origin is at the focus.

Problem 38

Determine where the normal line intersects the y -axis.

Problem 39

At a point $(\tilde{x}, \tilde{y})$ on the parabola $y^2 = p^2 + 2px$, show that the normal line is

$$\frac{x - \tilde{x}}{p} + \frac{y}{\tilde{y}} = 1.$$

3.8 Reflection Property

Consider the parabola described by $y^2 = p^2 + 2px$ with the origin at the focus. As shown in Figure 9.17, a point $s = (x, y)$ on the parabola implies a tangent line AB , along with a normal line q . The reflection property of the parabola states that *a ray from the focus to the parabola is reflected parallel to the axis*. Reading this backwards, we can also say that *incoming rays parallel to the axis are reflected by the parabola to the focus*.

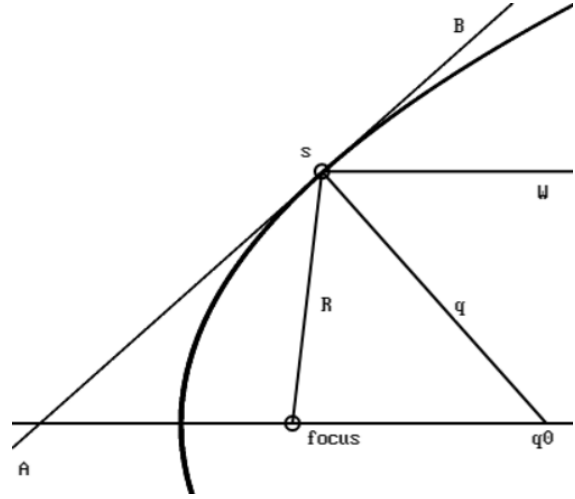


Figure 9.17. A ray emitted from the focus will intersect the parabola at $s = (x, y)$ and reflect parallel to the axis along line sW .

To proceed, refer to Figure 9.17 to define:

$$\text{Angle } Asf = \alpha_1$$

$$\text{Angle } BsW = \alpha_2$$

$$\text{Angle } fsq_0 = \beta_1$$

$$\text{Angle } Wsq_0 = \beta_2$$

$$\text{Line } sW \propto \hat{x}$$

$$\text{Slope of } AB = m_s$$

Note that the line sW is parallel to the axis. To establish the reflection property, we must show that either $\alpha_1 = \alpha_2$ or that $\beta_1 = \beta_2$.

Vector Analysis

With the construction on hand, let us write a vector $\mathbf{q}$ that points from q_0 to s :

$$\mathbf{q} = s - q_0 \hat{x} \quad (9.68)$$

Note that s is equivalent to the position vector $\mathbf{r}$, which is itself composed of a magnitude r and unit vector $\hat{r}$. To prove that $\beta_1 = \beta_2$, we observe that the unit vector representing s projected onto $\mathbf{q}$ has to equal that of the unit vector $-\hat{x}$ projected onto $\mathbf{q}$:

$$q \cos(\beta_1) = \hat{r} \cdot \mathbf{q} = -\hat{x} \cdot \mathbf{q} = q \cos(\beta_2) \quad (9.69)$$

Substituting $\mathbf{q}$ from Equation (9.68), we have

$$\hat{r} \cdot (s - q_0 \hat{x}) = -\hat{x} \cdot (s - q_0 \hat{x}),$$

boiling down to

$$r - \left(\frac{x+p}{r} \right) x = -x + q_0 .$$

On the right, the quantity $-x + q_0$ is simply the constant p according to Equation (9.67). The parenthesized quantity on the left resolves to *one* according to definition (9.57), bringing the result to

$$r - x = p . \quad (9.70)$$

To check that Equation (9.70) is true we employ polar coordinates, $x = r \cos(\theta)$, and the above quickly resolves to the polar representation (9.64) of the parabola, finishing the proof.

Problem 40

Derive Equation (9.70) from Equation (9.69).

Slope Analysis of the Parabola

The proof that that $\alpha_1 = \alpha_2$ using pure slope analysis is slightly tricky. We first note that the angle formed between R and the x -axis, i.e. θ as used in polar coordinates, is equal to two times α_1 , leading us to write

$$\tan(2\alpha_1) = \frac{y}{x} .$$

Meanwhile, the slope m_s of line AB is the tangent of α_2 :

$$m_s = \tan(\alpha_2)$$

From this point, let us cautiously assume that α_1, α_2 are equal to a common value α and make sure no contradictions arise.

Next, use the angle-sum identity (9.21) for tangent to write

$$\tan(2\alpha) = \frac{2 \tan(\alpha)}{1 - \tan^2(\alpha)} ,$$

and replace all trigonometric terms with factors of x, y , and m_s :

$$0 = m_s^2 + 2 \frac{x}{y} m_s - 1 \quad (9.71)$$

Complete the square in m_s and then solve for m_s . The result boils down to

$$m_s = \frac{p}{y} ,$$

the formula (9.65) for the slope at the point (x, y) on the parabola, validating the assumption $\alpha_1 = \alpha_2$ and completes the proof.

Problem 41

Derive Equation (9.65) from Equation (9.71).

Differential Analysis

Starting from Equation (9.71), multiply through by y^2 and make the substitution

$$y^2 = r^2 - x^2 ,$$

where implicit differentiation tells us

$$ym_s = r \frac{dr}{dx} - x .$$

With this, Equation (9.71) reduces to

$$0 = r^2 - r^2 \left(\frac{dr}{dx} \right)^2 ,$$

telling us

$$\frac{dr}{dx} = 1 ,$$

which is integrated to give us r as a function of x up to a constant:

$$r = x + \text{const}$$

Comparing the above to Equation (9.70), the integration constant is essentially p . Arriving at a familiar result without contradiction, we assure again that $\alpha_1 = \alpha_2$.

4 Slicing the Cone

4.1 Conic Sections

Now we address why the ellipse, hyperbola, and parabola are called *conic sections*. It turns out that each of these curves can be produced from the intersection of a cone and a plane. The cone can slice the plane in any way, at any angle, and *only* conic sections are produced. Consider a cone in three dimensions represented by

$$x^2 + y^2 = \alpha z^2, \quad (9.72)$$

where α is a dimensionless parameter controlling the ‘sharpness’ of the cone. Next we’ll need a plane to slice the cone, which we represent by

$$1 = \frac{x}{x_0} + \frac{z}{z_0}. \quad (9.73)$$

Conspicuously absent from Equation (9.73) is any representation of the y -variable. Due to the axial symmetry of the cone, one can always choose a coordinate system where the horizontal coordinate on the plane aligns perfectly with a cartesian axis. In this case, the plane’s intersection with the cone is defined by the angle θ formed between the plane and the horizontal. The special case $\theta = 0$ means the plane slices through the cone’s vertex (an infinitely small ellipse).

To proceed, suppose the coordinate system embedded on the plane is labeled u, v (in analog to x, y). From geometry, we can write several observations about this system:

$$u \cos(\theta) = x_0 - x \quad (9.74)$$

$$u \sin(\theta) = z \quad (9.75)$$

$$v = y \quad (9.76)$$

$$\tan(\theta) = \frac{z_0}{x_0} \quad (9.77)$$

In other words, the u -coordinate on the plane corresponds to locations mixing x and z . The v -coordinate is equivalent to the y -coordinate.

Using everything we have on hand, we write an equation

$$z_0 = \tan(\theta) \sqrt{\alpha u^2 \sin^2(\theta) - v^2} + u \sin(\theta),$$

implying

$$1 = \frac{u^2}{z_0^2} \sin^2(\theta) (\alpha \tan^2(\theta) - 1) + 2 \frac{u}{z_0} \sin(\theta) - \tan^2(\theta) \frac{v^2}{z_0^2} \quad (9.78)$$

and furthermore, after a page of algebra:

$$1 = \left(\frac{u}{z_0} \cos(\theta) \frac{(1 - \alpha \tan^2(\theta))}{\sqrt{\alpha}} - \frac{\cot(\theta)}{\sqrt{\alpha}} \right)^2 + \frac{v^2}{z_0^2} \left(\frac{1 - \alpha \tan^2(\theta)}{\alpha} \right) \quad (9.79)$$

Problem 42

Derive Equation (9.78) and Equation (9.79).

Gamma Coefficient

To help tame the symbolic explosion that has occurred, let us introduce yet another symbol γ such that

$$\gamma = 1 - \alpha \tan^2(\theta). \quad (9.80)$$

4.2 Parabolic Case

If the quantity $\gamma = 1 - \alpha \tan^2(\theta)$ resolves to zero for some special choice of α, θ , then Equation (9.78) reduces to that of a parabola:

$$1 = 2 \frac{u}{z_0} \sin(\theta) - \tan^2(\theta) \frac{v^2}{z_0^2}$$

To assure we’re looking at a parabola, note the v -coordinate occurs as v^2 , whereas the u -coordinate occurs to the first power.

4.3 Ellipse vs. Hyperbola

Looking again at Equation (9.79), it turns out that the γ coefficient also dictates whether we’re looking at an ellipse versus a hyperbola. For small angles θ , and/or for small stretch factors α , the quantity γ will always remain positive. This corresponds to an elliptical conic section. On the other hand, for large angles θ and/or large stretch factors α , the quantity γ becomes negative, giving rise to the hyperbolic conic section. In either case, we have:

$$1 = \left(\frac{u}{z_0} \cos(\theta) \frac{\gamma}{\sqrt{\alpha}} - \frac{\cot(\theta)}{\sqrt{\alpha}} \right)^2 + \frac{v^2}{z_0^2} \left(\frac{\gamma}{\alpha} \right) \quad (9.81)$$

With u, v playing analogous roles to x, y , we can immediately pick out the major and minor axes a, b which come out to:

$$a = \frac{z_0 \sqrt{\alpha}}{\gamma \cos(\theta)}$$

$$b = z_0 \sqrt{\frac{\alpha}{\gamma}}$$

Eccentricity

Going by definition, the eccentricity of the ellipse is

$$\begin{aligned} e &= \sqrt{1 - \frac{b^2}{a^2}} = \sqrt{1 - \gamma \cos^2(\theta)} \\ &= \sin(\theta) \sqrt{1 + \alpha} . \end{aligned} \tag{9.82}$$

Similarly, the eccentricity of the hyperbola is

$$\begin{aligned} e &= \sqrt{1 + \frac{b^2}{a^2}} = \sqrt{1 + \gamma \cos^2(\theta)} \\ &= \sqrt{2 - \sin^2(\theta) (1 + \alpha)} . \end{aligned} \tag{9.83}$$

Problem 43

Derive Equation (9.82) and Equation (9.83).

5 General Conic Sections

5.1 Review

By playing certain games with a directrix (line) and a focus (point) in the plane, three distinct species of curve emerge, namely the ellipse, the hyperbola, and the parabola. Each curve has a distinct shape and at least one focus as detailed:

	equation	focus
ellipse	$x^2/a^2 + y^2/b^2 = 1$	$(c, 0)$
hyperbola	$y^2/a^2 - x^2/b^2 = 1$	$(0, c)$
parabola	$y = ax^2 + bx + c$	$1/4a$

Eccentricity

A single number called eccentricity, denoted e , classifies whether the curve is an ellipse ($e < 1$), a hyperbola ($e > 1$), or a parabola ($e = 1$), as summarized:

	eccentricity
ellipse	$e = \sqrt{1 - b^2/a^2} < 1$
hyperbola	$e = \sqrt{1 + b^2/a^2} > 1$
parabola	$e = 1$

Polar Representation of Conics

The ellipse, hyperbola, and parabola are siblings in polar coordinates when the origin is at a focus. Remarkably, all three curves are represented by one single equation tuned by the eccentricity:

$$r = \frac{pe}{1 - e \cos(\theta)}$$

5.2 Generalized Conics

In the most general case, a conic section in the Cartesian plane can come to you in the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = F. \quad (9.84)$$

The coefficients A through F not only determine the curve species, but also the placement and *rotation* of the curve via the Bxy term.

5.3 Rotated Coordinates

To analyze Equation (9.84), it helps to use a second coordinate system uv that is rotated with respect to the original xy coordinate system so there is no mixed ‘rotation’ term. The uv system shares the same origin as the xy system, but is

tuned by θ to align with the curve’s principal axes. Such a rotated coordinate system can be written

$$\begin{aligned} u &= x \cos(\theta) + y \sin(\theta) \\ v &= -x \sin(\theta) + y \cos(\theta), \end{aligned}$$

which can be inverted to read

$$\begin{aligned} x &= u \cos(\theta) - v \sin(\theta) \\ y &= u \sin(\theta) + v \cos(\theta). \end{aligned}$$

Note that a positive rotation in θ corresponds to counter-clockwise progression of the uv system with respect to the xy system, which makes the curve itself appear to progress clockwise in the uv frame.

To proceed, take the quantity $Ax^2 + Bxy + Cy^2$ and substitute the x - and y -equations above to find

$$\begin{aligned} Ax^2 + Bxy + Cy^2 &= u^2 \left(A \cos^2(\theta) + C \sin^2(\theta) + B \sin(\theta) \cos(\theta) \right) \\ &+ v^2 \left(A \sin^2(\theta) + C \cos^2(\theta) - B \sin(\theta) \cos(\theta) \right) \\ &+ uv \left(-A \sin(2\theta) + C \sin(2\theta) + B \cos(2\theta) \right) \end{aligned}$$

In order to eliminate the ‘mixed’ uv -term in the rotated coordinate system, we find the restriction on θ to be given by

$$B' = -A \sin(2\theta) + C \sin(2\theta) + B \cos(2\theta) = 0,$$

or

$$\tan(2\theta) = \frac{B}{A - C}. \quad (9.85)$$

So far then, we can write

$$A'u^2 + C'v^2 + D'u + E'v = F, \quad (9.86)$$

where the primed coefficients A' through E' can be traced back to the original coefficients.

Problem 44

Write explicit formulas for A' , B' , C' , D' , E' in terms of θ and the unprimed coefficients.

Problem 45

The equation

$$21x^2 + 31y^2 - \sqrt{300}xy = 144$$

describes a tilted ellipse centered at the origin. Determine the angle θ required to un-tilt the ellipse and write the new equation. (Answer: $2\theta = \tan^{-1}(\sqrt{3})$, $u^2/9 + v^2/4 = 1$)

Problem 46

The axis of the hyperbola $x^2/a^2 - y^2/b^2 = 1$ is tilted by 45 degrees using the origin as a pivot so that the new axis lies along the line $y = x$. (The axis cuts through both focus points and the rotated hyperbola lives strictly in the first and third quadrants.) Prove that the equation of the tilted hyperbola is

$$v = \left(\frac{a^2 + b^2}{a^2 - b^2} \right) u \pm \frac{ab}{a^2 - b^2} \sqrt{4u^2 - 2(a^2 - b^2)}.$$

Problem 47

The axis of the hyperbola $x^2 - y^2 = 2$ is tilted by 45 degrees using the origin as a pivot so that the new axis lies along the line $y = x$. Show that the new equation is $v = 1/u$.

Problem 48

The axis of the parabola $y = x^2 - 1/4$ is tilted by 45 degrees using the focus as a pivot so that the new axis lies along the line $y = x$. Prove that the equation of the tilted parabola is

$$v = u + \frac{1}{\sqrt{2}} \pm \sqrt{2\sqrt{2}u + 1}.$$

Classifying Rotated Conics

With the mixed uv -term squelched out, the type of curve described by Equation (9.86) is indicated by the signs on the A' and C' terms. If either A' or C' is zero, the curve is parabolic. The curve is elliptical if A and C agree each positive, and so on. If both A and C are zero, the curve at best linear.

5.4 Discriminant of a Conic

It is possible to determine the type of curve described by the general Equation (9.84) without manually rotating coordinates. To do this, we must calculate the *discriminant* of the conic, defined by

$$\mathcal{D} = B^2 - 4AC. \quad (9.87)$$

It turns out that $\mathcal{D}$ resolves to the same value regardless of the rotation angle of the coordinate system, making $\mathcal{D}$ an *invariant* quantity. To prove this, let us calculate

$$\mathcal{D}' = (B')^2 - 4A'C'$$

and check if $\mathcal{D}' = \mathcal{D}$ in the general case. To get started, we'll calculate the square of B' and the product $-4A'C'$

separately:

$$\begin{aligned} (B')^2 &= (A^2 + C^2 - 2AC) \sin^2(2\theta) \\ &\quad + B^2 \cos^2(2\theta) + B(C - A) \sin(4\theta) \\ -4A'C' &= (-A^2 - C^2 + B^2) \sin^2(2\theta) \\ &\quad - B(C - A) \sin(4\theta) \\ &\quad - 4AC + 2AC \sin^2(2\theta) \end{aligned}$$

Taking the sum of the two results, we see that all of the ugly terms cancel out and the form $B^2 - 4AC$ emerges:

$$(B')^2 - 4A'C' = B^2 - 4AC \quad (9.88)$$

More succinctly, we see $\mathcal{D}' = \mathcal{D}$ for any angle θ . Not surprisingly, the coefficients D, D', E, E', F , are not involved in the discriminant or the classification of the curve.

Problem 49

Derive Equation (9.88).

5.5 Classifying General Conics

Having shown that the discriminant $B^2 - 4AC$ of a general conic section

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = F$$

is invariant with respect to coordinate system rotations, we are free to choose the system that tunes for $B' = 0$ in accordance with Equation (9.85) to write

$$\mathcal{D} = B^2 - 4AC = -4A'C'. \quad (9.89)$$

Recall that the parabolic case corresponds to either of A' or C' being zero, causing $\mathcal{D} = 0$. If A' and C' agree in sign, the curve is elliptical and $\mathcal{D}$ remains negative. If A' and C' disagree in sign, the curve is hyperbolic and $\mathcal{D}$ is positive.

	discriminant
ellipse	$B^2 - 4AC < 0$
hyperbola	$B^2 - 4AC > 0$
parabola	$B^2 - 4AC = 0$

Problem 50

Consider the hyperbola given by $xy = 1$. Use Equation (9.89) to express the same hyperbola with no mixing term.

Using the Discriminant

It's possible to show using calculus that the area of the ellipse is given by

$$\text{Area} = \pi ab ,$$

where a, b are the major and minor axes. With this information, we can determine the area of the ellipse

$$Ax^2 + Bxy + Cy^2 = 1 .$$

Choose a second uv -coordinate system whose orientation satisfies Equation (9.85), and the same ellipse takes the form

$$A'u^2 + C'v^2 = 1 .$$

Comparing this to the usual equation of an ellipse, it seems that A' is the inverse square of the major axis, and similarly for C' and the minor axis. The area of this ellipse is thus

$$\text{Area} = \frac{\pi}{\sqrt{A'C'}} .$$

Now involve the discriminant via Equation (9.89)

$$B^2 - 4AC = -4A'C'$$

to replace the primed terms with the original variables and the problem is finished:

$$\text{Area} = \frac{\pi}{\sqrt{4AC - B^2}}$$

Chapter 10

Taylor Polynomial

1 Time-Shifted Kinematics

Students studying constant-acceleration kinematics are greeted early by the core equations for the position $x(t)$ and velocity $v(t)$ of an object as a function of time:

$$\begin{aligned}x(t) &= x_0 + v_0 t + \frac{1}{2} a t^2 \\v(t) &= v_0 + a t\end{aligned}$$

The initial values x_0, v_0 are the position and velocity at $t = 0$. From the above, the myriad kinematic identities are straightforwardly built.

Notation

In writing these equations, two notational conventions are employed. First, $x(t)$ explicitly denotes position as a function of time, and, correspondingly, the initial value should properly be written $x(t = 0)$. For convenience, this is abbreviated to x_0 . Extending this logic, $x(t)$ itself may be compactly expressed as x_t . Thus, the kinematic equations can be written more economically as:

$$\begin{aligned}x_t &= x_0 + v_0 t + \frac{1}{2} a t^2 \\v_t &= v_0 + a t\end{aligned}$$

Uniform Jerk

A natural extension of basic kinematics is to consider what happens when acceleration itself varies linearly with time - a phenomenon known as *jerk*. For the case of constant (uniform) jerk, denoted j , analogous equations can be formulated,

with v and a taking the place of x and v :

$$\begin{aligned}v_t &= v_0 + a_0 t + \frac{1}{2} j t^2 \\a_t &= a_0 + j t\end{aligned}$$

However, when attempting to write a corresponding equation for position $x(t)$, the pattern alone suggests the next term should involve t^3 , but the exact coefficient is unclear. Introduce an unknown constant A to capture this uncertainty:

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3$$

Time-Shift Analysis

To resolve this, introduce a shift of the time variable such that

$$t \rightarrow t + h,$$

where h is arbitrary. Substituting into the expression for $x(t)$ yields

$$\begin{aligned}x(t + h) &= x_0 + v_0(t + h) \\&\quad + \frac{1}{2} a_0(t + h)^2 + \frac{1}{A} j(t + h)^3,\end{aligned}$$

and now the job is to fully expand the right side. Doing so, and then combining like terms in powers of h results in:

$$\begin{aligned}x(t + h) &= \left(x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3 \right) \\&\quad + h \left(v_0 + a_0 t + \frac{3}{A} j t^2 \right) \\&\quad + \frac{1}{2} h^2 \left(a_0 + \frac{6}{A} j t \right) + \frac{1}{6} h^3 (j)\end{aligned}$$

The first group corresponds to the original expression for x_t . The subsequent terms resemble v_t and a_t , extended by the effects of jerk. Comparing these to the previously established expressions for v_t and a_t , consistency is achieved only if $A = 6$. Thus, the correct position equation for constant jerk is:

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{6} j t^3$$

Time-Shifted Kinematics

Returning to $x(t+h)$, the parenthesized expressions can now be identified as x_t , v_t , etc., yielding a generalized *time-shifted* kinematic formula:

$$x(t+h) = x_t + v_t h + \frac{1}{2} a_t h^2 + \frac{1}{6} j h^3 \quad (10.1)$$

In this expression, any moment t can serve as the effective initial condition, with h controlling the subsequent evolution.

Inverse Relations

It's worthwhile to invert the time-shifted kinematic relations, expressing x_0 , v_0 , and a_0 in terms of x_t , v_t , and a_t :

$$\begin{aligned} a_0 &= a_t - jt \\ v_0 &= v_t - a_t t + \frac{1}{2} j t^2 \\ x_0 &= x_t - v_t t + \frac{1}{2} a_t t^2 - \frac{1}{6} j t^3 \end{aligned}$$

Qualitatively, the inverted system can be regarded as the a time-reversal of the original with $t \rightarrow -t$.

Problem 1

An experimental rocket accelerates along a linear track with constant jerk J (neglecting gravity). At time $t = T$, measurements show its position X , velocity V , and acceleration A . Determine the initial position, velocity, and acceleration at $t = 0$. Additionally, express the rocket's position at arbitrary time t .

Answer: The rocket's motion is given by

$$\begin{aligned} X &= x_0 + v_0 T + \frac{1}{2} a_0 T^2 + \frac{1}{6} J T^3 \\ V &= v_0 + a_0 T + \frac{1}{2} J T^2 \\ A &= a_0 + J T, \end{aligned}$$

straightforwardly inverted:

$$\begin{aligned} x_0 &= X - VT + \frac{1}{2} AT^2 - \frac{1}{6} JT^3 \\ v_0 &= V - AT + \frac{1}{2} JT^2 \\ a_0 &= A - JT \end{aligned}$$

For the path of motion, Equation (10.1) tells us

$$x(T+h) = X + Vh + \frac{1}{2} Ah^2 + \frac{1}{6} Jh^3,$$

where h is the time elapsed since T . The time elapsed since $t = 0$ is

$$T + h = t,$$

so finally

$$x(t) = X + V(t-T) + \frac{A(t-T)^2}{2} + \frac{J(t-T)^3}{6}.$$

For two sanity checks, let $t = T$ to recover $x(T) = X$. Also let $t = 0$ to recover $x(0) = x_0$.

Uniform Snap

Problem 2

When the jerk is allowed to vary linearly with time, one speaks of *snap* of the motion, represented by

$$j(t) = j_0 + kt,$$

where k is the snap coefficient. Following similar arguments, one proposes

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{6} j_0 t^3 + \frac{1}{B} k t^4$$

for the path of motion, where B is a numeric constant. Use time-shift analysis to figure out $B = 24$, and then appropriately append x_t , x_0 , v_t , v_0 , etc.

Numeric Coefficients

Examining the pattern of coefficients in $x(t)$, a clear factorial structure emerges. Each coefficient results from expanding powers of $(t+h)^n$, readily derived via Pascal's triangle or binomial expansion.

Recognizing this, the factorial operator (!) provides a compact expression:

$$N! = N(N-1)(N-2) \cdots (2)(1),$$

with the limit case $0! = 1$. Thus, the position equation generalizes as:

$$x(t) = x_0 + v_0 t + \frac{1}{2!} a_0 t^2 + \frac{1}{3!} j_0 t^3 + \frac{1}{4!} k t^4$$

1.1 Change of Basepoint

To illustrate time-shift analysis in practice, consider the quartic polynomial

$$x(t) = -3 + t - 3t^2 - \frac{1}{2}t^3 + \frac{5}{16}t^4.$$

Identify the polynomial's coefficients as:

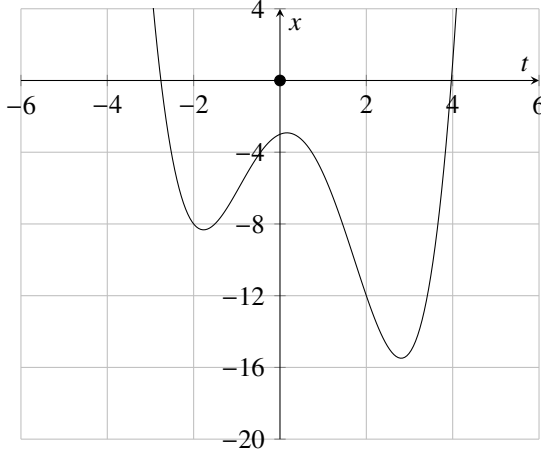
$$x_0 = -3$$

$$v_0 = 1$$

$$\frac{a_0}{2} = -3$$

$$\frac{j_0}{3!} = -\frac{1}{2}$$

$$\frac{k}{4!} = \frac{5}{16}$$



As written, the equation for $x(t)$ is suited such that the 'origin in time' is $t = 0$.

This expression is centered at $t = 0$. A *change of basepoint* involves selecting a new reference time, for instance $t_p = 2$ to do the origin's job, and expressing the polynomial relative to this point using h as the evolution variable.

To do so, evaluate the kinematic quantities at $t = 2$:

$$x_t = \left(x_0 + v_0 t + \frac{a_0}{2} t^2 + \frac{j_0}{6} t^3 + \frac{k}{24} t^4 \right) \Big|_2 = -12$$

$$v_t = \left(v_0 + a_0 t + \frac{j_0}{2} t^2 + \frac{k_0}{6} t^3 \right) \Big|_2 = -7$$

$$a_t = \left(a_0 + j_0 t + \frac{k_0}{2} t^2 \right) \Big|_2 = 3$$

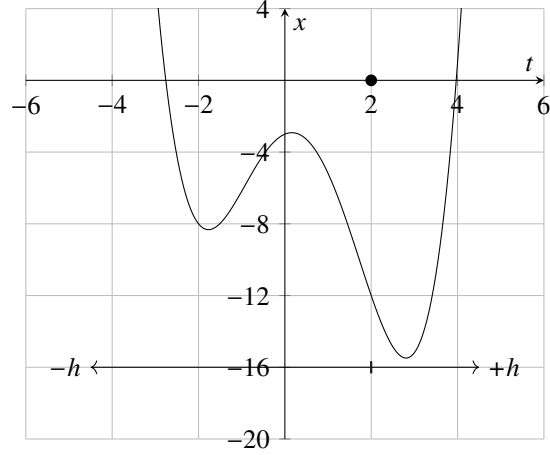
$$j_t = (j_0 + k_0 t) \Big|_2 = 12$$

$$k = \frac{15}{2}$$

A new equation is written

$$x(t_p + h) = -12 - 7h + \frac{3}{2}h^2 + 2h^3 + \frac{5}{16}h^4,$$

were the t_p -dependence is hidden in the numeric coefficients. The plot of this is in all ways identical to the original as shown:



Restoring Coefficients

For a sanity check, one may reverse-work the previous example, which is to start with the $x(h)$ -equation and recover the original version $x(t)$. A brutal way to do this is to let $h = t - 2$, and then simplify. Alternatively, one may use the inverse relations

$$j_0 = (j_t - kt) \Big|_2$$

$$a_0 = \left(a_t - j_t t + \frac{1}{2} k t^2 \right) \Big|_2$$

$$v_0 = \left(v_t - a_t t + \frac{1}{2} j_t t^2 - \frac{1}{6} k t^3 \right) \Big|_2$$

$$x_0 = \left(x_t - v_t t + \frac{1}{2} a_t t^2 - \frac{1}{6} j_t t^3 + \frac{1}{24} k t^4 \right) \Big|_2,$$

and the original numbers pop back out.

2 Taylor Polynomial

Now we keep pushing the idea of time-shifted kinematics. Supposing we choose a fixed point in time $t = t_p$ and use h to track time, the quantity $t_p + h$ becomes the *effective* time t :

$$t = t_p + h$$

With this, a general kinematic equation for $x(t_p + h)$ can be written:

$$x(t) = x_{t_p} + v_{t_p}(t - t_p) + \frac{1}{2}a_{t_p}(t - t_p)^2 + \frac{1}{6}j_{t_p}(t - t_p)^3 + \dots$$

Introducing a generalized notation to represent the velocity, acceleration, jerk, and so on, let us make the associations

$$\begin{aligned} x_{t_p} &\rightarrow x_{t_p}^{(0)} \\ v_{t_p} &\rightarrow x_{t_p}^{(1)} \\ a_{t_p} &\rightarrow x_{t_p}^{(2)} \\ j_{t_p} &\rightarrow x_{t_p}^{(3)} \\ k_{t_p} &\rightarrow x_{t_p}^{(4)}, \end{aligned}$$

and so on. On the left we've run out of 'named' items after snap, thus the general symbol $x_{t_p}^{(q)}$ is utilized to denote the q th coefficient.

2.1 Generalized Kinematics

In condensed form, $x(t)$ can be written in a most general way as the sum

$$x(t) = x_{t_p} + \sum_{q=1}^n \frac{1}{q!} x_{t_p}^{(q)} (t - t_p)^q, \quad (10.2)$$

which we'll call the *Taylor polynomial*. The upper limit n can be any number, depending on the total number of motion coefficients in play.

Generalized Coefficients

In the Taylor polynomial, note that t_p can be taken as *any* point in the domain of $x(t)$, and the equation 'adjusts' accordingly. To pay for this, for any given t_p , we have to

calculate all of the 'slope terms', which looks like

$$\begin{aligned} \left(x_0 + v_0 t + \frac{a_0}{2} t^2 + \frac{j_0}{6} t^3 + \frac{k_0}{24} t^4 + \dots \right) \Big|_{t_p} &= x_{t_p} \\ \left(v_0 + a_0 t + \frac{j_0}{2} t^2 + \frac{k_0}{6} t^3 + \dots \right) \Big|_{t_p} &= v_{t_p} \\ \left(a_0 + j_0 t + \frac{k_0}{2} t^2 + \dots \right) \Big|_{t_p} &= a_{t_p} \\ (j_0 + k_0 t + \dots) \Big|_{t_p} &= j_{t_p}, \end{aligned}$$

remembering that each of the terms x_{t_p} , v_{t_p} , etc., represent $x_{t_p}^{(0)}$, $x_{t_p}^{(1)}$, etc.

Inverse Coefficients

The inverse relations to the above, namely the structure that isolates x_0 , v_0 , etc., can be expressed by:

$$\begin{aligned} j_0 &= (j_t - k_t t + \dots) \Big|_{t_p} \\ a_0 &= \left(a_t - j_t t + \frac{k_t}{2} t^2 - \dots \right) \Big|_{t_p} \\ v_0 &= \left(v_t - a_t t + \frac{j_t}{2} t^2 - \frac{k_t}{6} t^3 + \dots \right) \Big|_{t_p} \\ x_0 &= \left(x_t - v_t t + \frac{a_t}{2} t^2 - \frac{j_t}{6} t^3 + \frac{k_t}{24} t^4 - \dots \right) \Big|_{t_p} \end{aligned}$$

Velocity

Consistent with the way $x(t)$ is written, we can write a similar formula for the velocity $v(t)$:

$$v(t) = v_{t_p} + \sum_{q=2}^n \frac{1}{(q-1)!} x_{t_p}^{(q)} (t - t_p)^{q-1}$$

By a shift of index $q - 1 = r$, this reads

$$v(t) = v_{t_p} + \sum_{r=1}^n \frac{1}{r!} x_{t_p}^{(r+1)} (t - t_p)^r,$$

which can be shortened once more by making the association

$$v_{t_p}^{(r)} = x_{t_p}^{(r+1)}.$$

2.2 Slope of a Polynomial

The relationship between $x(t)$ and $v(t)$ applies, in a sense, to any polynomial. Given a polynomial $y(t)$ with arbitrary coefficients

$$y(t) = A + Bt + Ct^2 + Dt^3 + \dots,$$

we're still free to interpret A , B , etc., as kinematic coefficients, i.e.

$$A = x_0$$

$$B = v_0$$

$$C = a_0/2!$$

$$D = j_0/3!$$

$$E = k_0/4!,$$

etc., and suddenly the problem looks like kinematics again.

If the term $y(t)$ is in all respects equivalent to a position $x(t)$, then the slope of $y(t)$ must be equivalent to the velocity $v(t)$. Using the Taylor polynomial, the velocity is trivial to write:

$$v(t) = v_0 + a_0t + \frac{1}{2!}j_0t^2 + \frac{1}{3!}k_0t^3 + \dots$$

Restoring the original coefficients, we find a formula for the slope $y_t^{(1)}$ of the function $y(t)$:

$$y_t^{(1)} = B + 2Ct + 3Dt^2 + 4Et^3 + \dots$$

2.3 Area Under a Polynomial

Recalling the uniform-acceleration regime, the plot of the velocity $v(t)$ is a straight line in time with initial value v_0 .

Inevitably, one should become curious about the *area* contained above the t -axis and under $v(t)$. Doing this exercise using geometry, we find, at time t , the area A to be the sum of two parts, a rectangle and a triangle, having respective areas

$$A_{\text{rectangle}} = tv_0$$

$$A_{\text{triangle}} = \frac{1}{2}t\Delta v,$$

where

$$\Delta v = v(t) - v_0.$$

Taking the sum of each area, and also replacing $v(t)$ with $v_0 + at^2/2$, we find

$$A_{\text{total}} = v_0t + \frac{1}{2}at^2,$$

which is exactly equal to the displacement $x(t) - x_0$. Evidently, for uniform acceleration at least, the area under the velocity plot equals the displacement:

$$A_{\text{total}} = x(t) - x_0$$

Exploiting Taylor Polynomial

Using the results painfully gained in this study by plain algebra, we can step around the calculus-based method for calculating the area under a polynomial curve. Going for the general case, suppose you're handed a polynomial with arbitrary coefficients:

$$y(t) = A + Bt + Ct^2 + Dt^3 + \dots$$

The key is to make the association $y(t) \leftrightarrow v(t)$, which means to interpret $y(t)$ as the velocity of some so-far undetermined curve $x(t)$. The coefficients A , B , C , etc., must be put into familiar terms, where borrowing the whole kinematics apparatus, we have

$$A = v_0$$

$$B = a_0$$

$$C = j_0/2!$$

$$D = k_0/3!,$$

and so on.

Then, without any new thinking at all, we already know what $x(t)$ should look like in terms of kinematic coefficients, which is

$$x(t) - x_0 = v_0t + \frac{1}{2!}a_0t^2 + \frac{1}{3!}j_0t^3 + \frac{1}{4!}k_0t^4 + \dots$$

Replacing the kinematic coefficients with the original unknowns, the result

$$x(t) - x_0 = At + \frac{1}{2}Bt^2 + \frac{1}{3}Ct^3 + \frac{1}{4}Dt^4 + \dots$$

emerges, and like magic, the problem is solved.

Problem 3

Calculate the area under the curve

$$y(t) = (2 + 3t)^2.$$

Answer: Begin by expanding $y(t)$ to get

$$y(t) = 4 + 12t + 9t^2,$$

from which we discern

$$v_0 = 4$$

$$a_0 = 12$$

$$\frac{1}{2}j = 9.$$

Assembling the quantity $x(t) - x_0$ from these coefficients,

we simply write

$$x(t) - x_0 = 4t + 6t^2 + 3t^3,$$

and the problem is solved.

3 Euler's Constant

From the Taylor polynomial given as Equation (10.2), it's interesting to conceive of the situation where all 'slope terms' are the same number, i.e.

$$x_{t_p}^{(q)} = x_{t_p}^{(0)} = x_{t_p},$$

which assumes (and without loss of generality) that time is a *dimensionless* variable. This has similar consequence for x_0 , v_0 , a_0 , and so on, for these are now identical after adjusting for units of time.

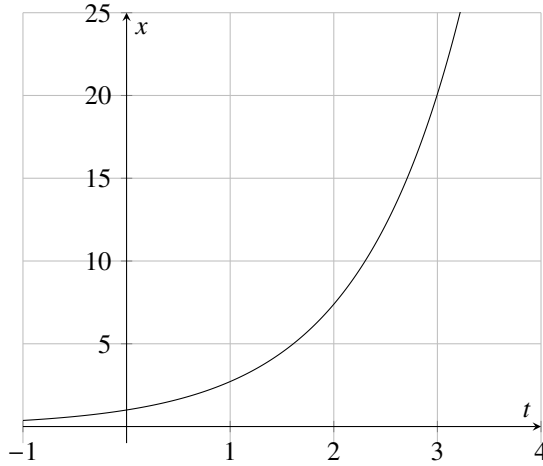
Proceeding, we have

$$x(t) = x_{t_p} \sum_{q=0}^{\infty} \frac{1}{q!} (t - t_p)^q,$$

and momentarily let $x_{t_p} = 1$ and $t_p = 0$ to yield

$$x(t) = \sum_{q=0}^{\infty} \frac{1}{q!} t^q,$$

as sketched in the following plot:



The case $t = 0$ yields a special construction

$$e = x(0) = \sum_{q=0}^{\infty} \frac{1}{q!} \quad (10.3)$$

named e for *Euler's constant*. With enough patience, one may estimate the value of e by brute force:

$$e = 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots$$

Euler's Constant

A derivation of Euler's constant begins with the binomial theorem

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k,$$

where letting

$$\begin{aligned} x &= 1 \\ y &= 1/n \end{aligned}$$

gives

$$\left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^n \frac{n!}{k! (n-k)!} \frac{1}{n^k}.$$

In the large- n limit, one readily finds

$$\lim_{n \rightarrow \infty} \frac{n!}{(n-k)!} \approx n^k,$$

leaving

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!}.$$

On the right is the now-familiar Euler's constant. Evidently, e is also given by the limit on the left:

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \quad (10.4)$$

To get a sense of this number, try a few increasing n -values:

$$n = 1 : \left(1 + \frac{1}{1}\right)^1 = 2.0$$

$$n = 2 : \left(1 + \frac{1}{2}\right)^2 = 2.25$$

$$n = 3 : \left(1 + \frac{1}{3}\right)^3 \approx 2.370$$

$$n = 10 : \left(1 + \frac{1}{10}\right)^{10} \approx 2.5937$$

$$n = 100 : \left(1 + \frac{1}{100}\right)^{100} \approx 2.7048$$

$$n = 1000000 : \left(1 + \frac{1}{1000000}\right)^{1000000} \approx 2.71828$$

Evidently the Euler constant with $n \rightarrow \infty$ is settling to

$$e \approx 2.7182818284590 \dots$$

Exponential Growth

Let us quickly re-derive Equation (10.4) but instead let $y = t/n$ (as it occurs in the binomial theorem). In this variation we find

$$\lim_{n \rightarrow \infty} \left(1 + \frac{t}{n}\right)^n = \sum_{k=0}^{\infty} \frac{1}{k!} t^k,$$

where the right side is the same $x(t)$ graphed above.

Meanwhile, directly from Equation (10.4) one may establish

$$e^t = \lim_{n \rightarrow \infty} \left(1 + \frac{t}{n}\right)^n.$$

Eliminating the common term, we discover the formula for natural exponential growth:

$$e^t = \sum_{q=0}^{\infty} \frac{1}{q!} t^q \quad (10.5)$$

Rescaling

To restore time as a quantity of proper dimension, start again with Equation (10.2) and assume that the ratio of ascending coefficients is a constant α :

$$x_{t_p} = x_{t_p}^{(0)} = \alpha^q x_{t_p}^{(q)}$$

Proceeding as before, a factor of α joins the t -variable

$$x(t) = x_{t_p} \sum_{q=0}^{\infty} \frac{1}{q!} \alpha^q (t - t_p)^q,$$

and this propagates down to:

$$e^{\alpha t} = \lim_{n \rightarrow \infty} \left(1 + \frac{\alpha t}{n}\right)^n \quad (10.6)$$

Note that α has inverse time units.

Reflecting

From the above we export the result:

$$e^{\alpha t} = \sum_{q=0}^{\infty} \frac{(\alpha t)^q}{q!} \quad (10.7)$$

Setting $\alpha t \rightarrow -\alpha t$, we have another relation to handle backward evolution in time:

$$e^{-\alpha t} = \sum_{q=0}^{\infty} \frac{(-\alpha t)^q}{q!} \quad (10.8)$$

Offsetting

Also, one may introduce a shift of variables such that the above becomes:

$$e^{\alpha(t-t_0)} = \sum_{q=0}^{\infty} \frac{(\alpha(t-t_0))^q}{q!} \quad (10.9)$$

Application: Dark Integer

The above construction may help clear some mental residue from earlier lessons. It's straightforward to show using geometric series that

$$0.999\bar{9} = 1,$$

provided there are an infinite number of 9's on the right.

With or without geometric series, we can generate another convincing argument by letting $A = 0.999\bar{9}$, and then

$$10A = 9.999\bar{9}$$

$$10A - A = 9.999\bar{9} - 0.999\bar{9}$$

$$9A = 9.000\bar{0}$$

$$A = 1.000\bar{0}.$$

Now we've expressed A two different ways, verifying the claim.

A student could possibly challenge whether A is exactly one, or if there's not some equivalent to a floating-point error after an infinite string of zeros following the decimal. Putting this more formally consider a quantity

$$A = 1.0000 \dots 0000\rho,$$

where ρ is some nonzero digit, a *dark integer* between 0 and 9 occurring after a large number of zeros, say, N of them.

Separating ρ from its larger counterpart, one may write

$$A = 1 + \frac{\rho}{10^N}.$$

Next, define a variable n such that $n = 10^N$, and we have

$$A = 1 + \frac{\rho}{n}.$$

Raise each side to the n power and apply the large- n limit:

$$\lim_{n \rightarrow \infty} A^n = \lim_{n \rightarrow \infty} \left(1 + \frac{\rho}{n}\right)^n = e^\rho$$

Now, if A is slightly greater than one, then the quantity A^n will diverge to infinity. Meanwhile, the greatest possible

return from the right is e^9 , which is not infinity, and we have a contradiction.

Similar goes for if A is slightly less than one, in where the left would converge to zero. The only way to get this on the right is to have $\rho = -\infty$, disqualifying ρ as a dark integer.

The only valid solution to the above is $A = 1$, which forces $\rho = 0$:

$$\lim_{n \rightarrow \infty} 1^n = e^0 = 1$$

With ρ confined to zero and only zero, we conclude there is no dark integer.

3.1 Hyperbolic Curves

Two noteworthy combinations of the the Euler exponential equations can be constructed, namely the *hyperbolic cosine* and the *hyperbolic sine*, given by, respectively:

$$\cosh(\alpha t) = \frac{e^{\alpha t} + e^{-\alpha t}}{2} \quad (10.10)$$

$$\sinh(\alpha t) = \frac{e^{\alpha t} - e^{-\alpha t}}{2} \quad (10.11)$$

Simultaneous to these we can get the originals back by taking the sum and difference of each:

$$\begin{aligned} e^{\alpha t} &= \cosh(\alpha t) + \sinh(\alpha t) \\ e^{-\alpha t} &= \cosh(\alpha t) - \sinh(\alpha t) \end{aligned}$$

Furthermore, given the infinite expansion of e^t , the hyperbolic functions can be written in open form:

$$\cosh(t) = 1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \cdots \quad (10.12)$$

$$\sinh(t) = t + \frac{t^3}{3!} + \frac{t^5}{5!} + \frac{t^7}{7!} + \cdots \quad (10.13)$$

It's straightforward to show from the above that the pair of hyperbolic functions obey, for a dimensionless variable t ,

$$(\cosh(t))^2 - (\sinh(t))^2 = 1. \quad (10.14)$$

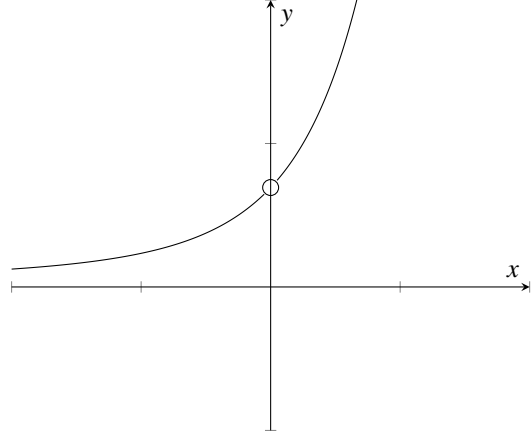
This is somewhat like the standard identity from trigonometry if it weren't for the negative sign. In fact, a whole slew of 'hyperbolic trigonometry' identities can be derived that are analogous to the 'ordinary' trig identities.

3.2 Natural Logarithm

Consider the curious quantity

$$y(x) = \lim_{x \rightarrow 0} \frac{n^x - 1}{x}, \quad (10.15)$$

plotted as follows:



From the plot of Equation (10.15), we see the point at $x = 0$ is probably a removeable singularity, which is to say we can come up with an answer for $y(0)$ given how y behaves in that neighborhood.

Proceed by solving for n to write

$$n = \lim_{x \rightarrow 0} (1 + xy)^{1/x},$$

which is starting to look like the derivation of Euler's constant. By Equation (10.6), the right side evaluates to e^y , telling us

$$y(0) = \log_e(n),$$

and for a final answer we take:

$$\ln(n) = \lim_{x \rightarrow 0} \frac{n^x - 1}{x}$$

The missing point in the plot of $y(x)$ is the natural log of n .

4 Periodic Curve

Consider the generalized (infinite) Taylor polynomial given as Equation (10.2) with $t_p = 0$:

$$x(t) = \sum_{q=0}^{\infty} \frac{1}{q!} x_0^{(q)} t^q$$

Next, consider a shift of variables $t \rightarrow t + w$ such that

$$x(t + w) = \sum_{q=0}^{\infty} \frac{1}{q!} x_0^{(q)} (t + w)^q,$$

which, up to w replacing h to avoid confusion, resembles the setup for time-shifted kinematics that led to Equation (10.2) originally.

For a familiar but nontrivial exercise, one can expand the right side in powers of w to write a generalization of Equation (10.1), particularly

$$x(t + w) = \sum_{q=0}^{\infty} \frac{1}{q!} x_t^{(q)} w^q. \quad (10.16)$$

A similarly-familiar exercise involves solving for the coefficients $x_t^{(q)}$, which in this case turn out as

$$\begin{aligned} x_0^{(0)} + x_0^{(1)} t + x_0^{(2)} \frac{t^2}{2} + x_0^{(3)} \frac{t^3}{6} + \dots &= x_t^{(0)} \\ x_0^{(1)} + x_0^{(2)} t + x_0^{(3)} \frac{t^2}{2} + x_0^{(4)} \frac{t^3}{6} + \dots &= x_t^{(1)} \\ x_0^{(2)} + x_0^{(3)} t + x_0^{(4)} \frac{t^2}{2} + x_0^{(5)} \frac{t^3}{6} + \dots &= x_t^{(2)}, \end{aligned}$$

and so on, which, just to remind, are the same as x_t , v_t , a_t , etc in kinematics-style notation.

From the above list, multiply each equation by ascending powers of w such that each has the same units of time, and then sum all terms vertically to get a curious identity:

$$\begin{aligned} x_t^{(0)} + w x_t^{(1)} + w^2 x_t^{(2)} + \dots &= \\ + x_0^{(0)} + w x_0^{(1)} + w^2 x_0^{(2)} + \dots & \\ + t \left(x_0^{(1)} + w x_0^{(2)} + w^2 x_0^{(3)} + \dots \right) & \\ + \frac{t^2}{2} \left(x_0^{(2)} + w x_0^{(3)} + w^2 x_0^{(4)} + \dots \right) & \\ + \frac{t^3}{6} \left(x_0^{(3)} + w x_0^{(4)} + w^2 x_0^{(5)} + \dots \right) + \dots & \end{aligned}$$

Periodicity Condition

Suppose, for all times t in the domain of $x(t)$, the property

$$x(t + w) = x(t)$$

always holds, called the *periodicity condition*. Immediately true, too, is the stronger statement for all orders of slope terms:

$$x_{t+w}^{(q)} = x_t^{(q)}$$

All we need is the special case of the periodicity condition

$$x(w) = x(0) = x_0$$

and subsequently

$$x_w^{(q)} = x_0^{(q)}.$$

To proceed, take that messy identity we wrote above and set $t = w$. The zero-order terms cancel right away due to periodicity:

$$\begin{aligned} 0 &= w \left(x_0^{(1)} + w x_0^{(2)} + w^2 x_0^{(3)} + \dots \right) \\ &+ \frac{w^2}{2} \left(x_0^{(2)} + w x_0^{(3)} + w^2 x_0^{(4)} + \dots \right) \\ &+ \frac{w^3}{6} \left(x_0^{(3)} + w x_0^{(4)} + w^2 x_0^{(5)} + \dots \right) + \dots, \end{aligned}$$

For this to be true in the general case, it *must* be that the parenthesized terms cannot all be positive, and cannot all be negative, else divergence would occur. Whatever the above is trying to say, let us call it the *periodicity constraint*. After a bit of algebra, it's possible to cook the periodicity constraint down to a double sum

$$0 = \sum_{k=1}^{\infty} x_0^{(k)} w^k J_k,$$

where

$$J_k = \sum_{j=1}^k \frac{1}{j!}$$

for brevity.

To anticipate the next move, break the outer k -sum into two parts: one sum for even k , denoted k_e , and another sum for odd k , denoted k_o :

$$0 = \left(\sum_{\text{even } k} x_0^{(k_e)} w^{k_e} J_e \right) + \left(\sum_{\text{odd } k} x_0^{(k_o)} w^{k_o} J_o \right)$$

Now, the periodicity condition must also work when w is swapped with $-w$, or any integer multiple of w for that matter. Going with $w \rightarrow -w$, we see that the 'even' sum on the left would be completely unchanged by this, whereas the

‘odd’ sum on the right would gain a global minus sign. In other words, using generic labels, we have a situation with

$$\text{Even} + \text{Odd} = 0$$

$$\text{Even} - \text{Odd} = 0,$$

which is only true if

$$\text{Even} = \text{Odd} = 0.$$

Each parenthesized sum above, ‘odd’ and ‘even’, must separately equal zero.

4.1 Cosine and Sine

Separated into even and odd terms, the periodicity constraint encourages, but does not outright demand, that the $x_0^{(k)}$ -terms have alternating signs and all the same magnitude. To be definitive, let us have

$$1 = x_0^{(0)} = x_0^{(4)} = x_0^{(8)} = \dots$$

$$-1 = x_0^{(2)} = x_0^{(6)} = x_0^{(10)} = \dots$$

for the even terms, and

$$1 = x_0^{(1)} = x_0^{(5)} = x_0^{(9)} = \dots$$

$$-1 = x_0^{(3)} = x_0^{(7)} = x_0^{(11)} = \dots$$

for the odd terms.

These results let us write two cases for the resulting $x(t)$, namely

$$\cos(t) = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \dots \quad (10.17)$$

$$\sin(t) = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \dots, \quad (10.18)$$

where of course, the variable t can be replaced by the combination αt as done previously.

Partial Sum Plots

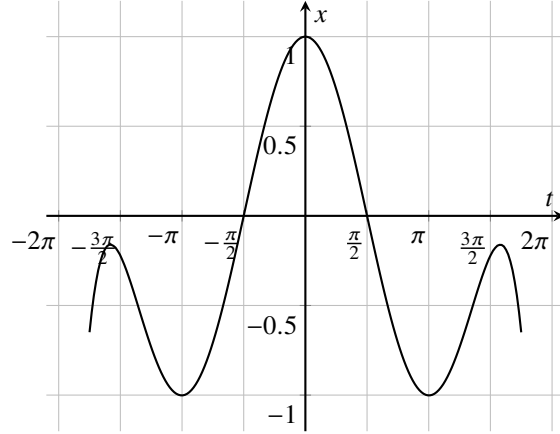


Figure 10.1. $x \approx \cos(t)$ to six terms.

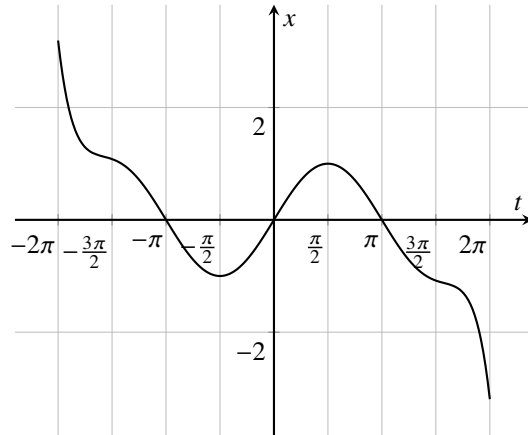


Figure 10.2. $x \approx \sin(t)$ to six terms.

5 Laws of Motion

Consider a pair of polynomials, one called $U(x)$ depending on $x(t)$, and the other called $T(v)$ depending on $v(t)$. These symbols may ring familiar as energy terms, which will indeed turn out true. As Taylor polynomials, the formulae for U and T are:

$$U(x) = U_{x_p} + \sum_{q=1}^n \frac{1}{q!} U_{x_p}^{(q)} (x - x_p)^q$$

$$T(v) = T_{v_p} + \sum_{q=1}^n \frac{1}{q!} T_{v_p}^{(q)} (v - v_p)^q$$

Conservation of Energy

Now suppose we are interested in the sum of T and U , a quantity labeled E such that

$$E = T(v) + U(x) .$$

To make things interesting, suppose E is a constant in time, which would mean

$$T(v) + U(x) = T_{v_p} + U_{x_p} .$$

By doing this, we have just enforced a powerful notion called *conservation of energy*.

With the above, take the sum of the T - and U -equations to get

$$0 = \sum_{q=1}^n \frac{1}{q!} U_{x_p}^{(q)} \Delta x^q + \sum_{q=1}^n \frac{1}{q!} T_{v_p}^{(q)} \Delta v^q ,$$

where:

$$\Delta x = x - x_p$$

$$\Delta v = v - v_p$$

Writing out the first term in each sum and rearranging a bit gives

$$0 = \left(U_{x_p}^{(1)} \Delta x + T_{v_p}^{(1)} \Delta v \right) + \sum_{q=2}^n \frac{1}{q!} \left(U_{x_p}^{(q)} \Delta x^q + T_{v_p}^{(q)} \Delta v^q \right) .$$

First-order Equations

Now we explore the regime where both Δx and Δv are ‘small intervals’ such that higher powers of these quantities

tend to diminish. From this we may ignore the remaining summation and keep the low-order terms already plucked out. If x and v are to be related by kinematics, it should follow that

$$\Delta x \approx v_p \Delta t$$

must hold for a similarly-small interval

$$\Delta t = t - t_p .$$

Boiling all this down, we transform the above down to

$$0 = U_{x_p}^{(1)} v_p \Delta t + T_{v_p}^{(1)} \Delta v .$$

From the first-order energy equation we may write

$$U_{x_p}^{(1)} v_p = -T_{v_p}^{(1)} \frac{\Delta v}{\Delta t} ,$$

which is suggestive of two proportionality relationships

$$-U_{x_p}^{(1)} \propto \frac{\Delta v}{\Delta t}$$

$$T_{v_p}^{(1)} \propto v_p .$$

Mass

Introducing a proportionality constant m while maintaining the negative sign between the two terms, we conclude from the above that:

$$-U_{x_p}^{(1)} = m \frac{\Delta v}{\Delta t} \quad (10.19)$$

$$T_{v_p}^{(1)} = m v_p \quad (10.20)$$

In order for the quantities E , T , U to have units of energy, the constant m can only have units of mass.

Newton’s Second Law

Equation (10.19) is a special case of *Newton’s second law*, which is concisely written:

$$-U_{x_p}^{(1)} = m x_{t_p}^{(2)}$$

In general, the left side represents the *force*, denoted F . In the general case, *force is mass times acceleration*:

$$F = m \frac{\Delta v}{\Delta t}$$

Potential Energy

The relationship

$$F = -U_{x_p}^{(1)}$$

is a special case of energy-conserving systems, where $U(x)$ is the *potential energy* of the system:

$$U(x) = \text{potential energy}$$

Kinetic Energy

The second result $T_{v_p}^{(1)} = mv_p$ relates the linear momentum to the slope of $T(v)$, which we identify as the *kinetic energy*. Knowing the slope of $T(v_p)$ is simply mv_p , it's easy to see that the kinetic energy is generally given by

$$T(v) = \frac{1}{2}mv^2.$$

Working the same result in the other direction, we further deduce

$$T_{v_p}^{(2)} = m,$$

and all higher $T_{v_p}^{(q)}$ are zero.

Total Energy

To summarize, we have that the total energy of a body in a so-far unspecified potential is constant:

$$E = \frac{1}{2}mv^2 + U_{x_p} + \sum_{q=1}^n \frac{1}{q!} U_{x_p}^{(q)} (x - x_p)^q$$

5.1 Mechanical Equilibrium

Since $U(x)$ is arbitrarily-shaped, one can imagine locating a special x_p that corresponds to an extreme of U , i.e. a local peak or a valley in its profile. In such a case, the slope of U is zero at that point

$$U_{x_p}^{(1)} = 0,$$

corresponding to *mechanical equilibrium*.

Small Oscillations

Small displacements from x_p are characterized by $x - x_p$ being a small quantity. As before, we argue that higher-order terms in the above sum are negligible, however truncating the series too soon leaves a tautology supporting no motion at all:

$$E = \cancel{\frac{1}{2}mv^2} + U_{x_p} + \sum_{q=1}^n \cancel{\frac{1}{q!} U_{x_p}^{(q)} (x - x_p)^q}$$

In light of $U_{x_p}^{(1)}$ being zero, the lowest nonzero term in the sum corresponds to $q = 2$, thus we have, for small displacements from x_p :

$$E - U_{x_p} = \frac{1}{2}mv^2 + \frac{1}{2}U_{x_p}^{(2)} (x - x_p)^2$$

The sign on $U_{x_p}^{(2)}$ determines the stability of motion around x_p . For $U_{x_p}^{(2)} < 0$, displacement from x_p causes v to grow extremely quickly and the approximation breaks down.

Hooke's Law

When x_p corresponds to an extreme point with $U_{x_p}^{(2)} > 0$, the system exhibits small oscillations centered on x_p . The potential energy term

$$U(x) = \frac{1}{2}U_{x_p}^{(2)} (x - x_p)^2$$

can be treated as arising by a spring force centered centered at x_p

$$f(x) = -U_{x_p}^{(2)} (x - x_p),$$

with $U_{x_p}^{(2)}$ being the effective spring constant. Using Newton's second law on this situation gives an equation for the subsequent motion:

$$mx_{t_p}^{(2)} = -U_{x_p}^{(2)} (x - x_p)$$

This particular form is also called *Hooke's law* for springs.

5.2 Freefall in Gravity

Using energy considerations, we can make sense of the standard kinematic identity

$$v^2 = v_0^2 + 2a\Delta x.$$

Supposing the motion represented is for a body of mass m , multiply through by $m/2$ and also expand $\Delta x = x - x_0$ to get, after simplifying:

$$\frac{1}{2}mv_0^2 - max_0 = \frac{1}{2}mv^2 - max$$

By letting $a = -g$ for freefall acceleration, we conclude that the gravitational potential energy for a body near Earth's surface is given by

$$U_{\text{grav}}(x) = mgx,$$

where x is measured vertically from the ground.

Chapter 11

Limits, Functions, Sequences

1 Limits

In mathematics, one speaks of *limit* when a starting value, call it A , is ‘becoming’ another value B :

$$A \rightarrow B$$

The use of the right-arrow ($\rightarrow$) is to remind that using an equal sign, i.e. simply writing $A = B$ is a hasty, possibly illegal jump through the real numbers.

1.1 Notion of Limit

The ‘notion of limit’ probably seems pedantic and perhaps mundane at first sight, but the ancient Greeks, particularly Zeno, famously failed to reconcile limits as anything but mathematical barriers in the universe and deemed the whole thing a paradox. The understanding of limits eluded popular thinking until the days of Newton and Leibniz.

Precisely ‘how’ a limit is traversed has significance, and is usually facilitated by the assumption that there is a smooth continuum of values, usually real numbers represented by x , between the initial and final points of the limit, i.e.

$$A \leq x < B.$$

Notice the mismatch in comparison symbols in the above, as we have $\leq$ on the A -side and $<$ on the B -side. This is because we shouldn’t outright assume that x ever reaches B precisely. The condition $x = B$ is considered a special case.

1.2 Differential Limit

Compressing the picture of ‘limit’, suppose the gap between two values were extremely small, or ‘vanishingly’ small, or ‘arbitrarily’ small. There are synonyms for a *differential limit*.

To continue, suppose we are interested in the distance from any fixed point x_0 to a neighboring point x . To capture this one may begin with

$$\Delta x = x - x_0,$$

where on the left is the usual ‘delta x ’ symbol for representing net displacement.

If the distance Δx is to be very small, i.e. if x is close to x_0 , we can write the above as a differential limit:

$$dx = \lim_{x \rightarrow x_0} x - x_0 \quad (11.1)$$

The term Δx is replaced by dx , which means ‘differential x ’. On the right, there is a new ‘limit’ term slipped in to remind that $x - x_0$ is close to zero.

Note that the differential limit does not apply to integers, whole numbers, or natural numbers. Only real (and complex, but never mind) numbers make sense in light of the differential limit.

Very Small Numbers

In the differential limit, the interval dx refers to a ‘small’ number. It should make sense that things like dx^2 and dx^3 are absurdly small numbers, and can often be omitted when they pop up in calculations.

To illustrate, suppose we have a number A that is the sum of a close-by number A_0 and a differential quantity dx

$$A = A_0 + dx,$$

and let us become interested in the quantity A^2 , attained by squaring both sides of the above:

$$A^2 = A_0^2 + 2A_0dx + dx^2$$

Now, we know already that A^2 is approximately equal to A_0^2 plus some small correction having to do with dx , namely $2A_0dx + dx^2$.

One can see that, especially in the infinite limit, that dx^2 will vanish toward zero much ‘faster’ than the middle term, thus dx^2 can be omitted outright in this case:

$$A^2 \approx A_0^2 + 2A_0dx$$

Problem 1

Let $A = A_0 + dx$ and $B = B_0 + dy$. Calculate the product AB and rank resulting terms by magnitude.

1.3 Infinite Limit

Another use for limits is pushing a variable toward infinity, which occurs in the formula for the natural exponential named after Euler:

$$e^x = \lim_{h \rightarrow \infty} \left(1 + \frac{x}{h}\right)^h \quad (11.2)$$

In the above, we see a curious struggle arise with increasing h . The $1/h$ -term inside the parentheses tries to drag the result downward, but the exponent h tries to drag the result back upward. The end result is $e \approx 2.71828 \dots$

Problem 2

Show that:

$$\lim_{x \rightarrow 1} x^{1/(1-x)} = e$$

Infinite Sum

An infinite limit can take the form of an *infinite sum*, as is the case with geometric series:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots \quad (11.3)$$

As long as $|x| < 1$, the above always holds, quite astonishingly.

On the left is the simple fraction $1/(1-x)$, where on the right is all whole-number powers of x , all the way to x^∞ , summed together.

We can see there is some kind of limit at play in the geometric series even though x is fixed. Rather, the number of terms in the series, which can only be a whole number, is the variable being pushed.

Sigma Notation

An equivalent statement of the above that uses the ‘lim’ nomenclature goes like

$$\frac{1}{1-x} = \lim_{n \rightarrow \infty} \sum_{j=0}^n x^j,$$

where j is an index variable and n is its upper limit (a positive integer). Keep in kind $|x| < 1$ for the formula to work.

The uppercase E-like symbol (Greek *Sigma*) is the clue for a summation. The variable j is called the *index*., which in this case, assumed all integer values 0, 1, 2, ..., n .

One

A psychological trick used in marketing is to price an item that is, say, \$7.00 (U.S. dollars, but currency doesn’t matter) as \$6.99 instead. The idea is that the shopper only sees the 6 and forgets that the price is closer to \$7. It’s typical for fuel prices to be advertised with a third digit after the decimal, always a 9, so drivers are tricked out of an extra $\approx$ \$0.01 per gallon purchased.

Suppose the price of an item for sale is $p = \$0.99999\bar{9}$, with the 9s carrying on *forever*. You hand the clerk a \$1 note to pay for the item. Did you overpay?

The instinct that says ‘yes, the clerk owes you some change for the transaction, even if it’s a small amount’ is overturned by geometric series. Observe that an infinite string of 9s following a decimal can be converted to an infinite sum of fractions:

$$p = 0.9999\bar{9} \dots = \frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \frac{9}{10^4} + \dots$$

Factor 9/10 from the right side

$$p = \frac{9}{10} \left(1 + \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^3} + \frac{1}{10^4} + \dots \right),$$

and then see that the parenthesized quantity is precisely the geometric series as the right side of Equation (11.3) with $x = 1/10$. This means the infinite sum can be replaced by $1/(1-x)$:

$$p = \frac{9}{10} \left(\frac{1}{1 - 1/10} \right) = \frac{9}{10} \left(\frac{10}{9} \right) = 1$$

See what just happened? We have two expressions for the variable p , which can only mean they are equal:

$$0.99999\bar{9} = p = 1$$

Thus the clerk owes you nothing back, and the item price is exactly \$1 despite the advertising.

1.4 Convergence and Divergence

When a limit ‘settles on’ a reasonable answer, i.e., a finite real number, one speaks of *convergence*. For instance, Equation (11.2) converges to Euler’s constant as h goes to infinity. The larger we make h , the slower the sum changes. Convergence also applies to the geometric series as Equation (11.3). When $|x| < 1$, the terms on the right decrease in magnitude until making vanishingly small contribution to the overall sum.

When convergence does not occur, it is likely that the object in hand exhibits *divergence*, which means tending toward $\pm\infty$. This is easy to see with the geometric series when purposefully choosing $|x| > 1$. On the right, there are ever-increasing powers of x , and for every new term added the sum changes drastically, refusing to settle down.

1.5 Sidedness

An important subtlety regarding limits pertains to the ‘direction’ in which a value is approached. For a variable x that is to approach, or ‘limit to’ a fixed value x_0 , it could very well be that the initial case $x < x_0$ produces one answer, and the other case $x > x_0$ produce an entirely different answer.

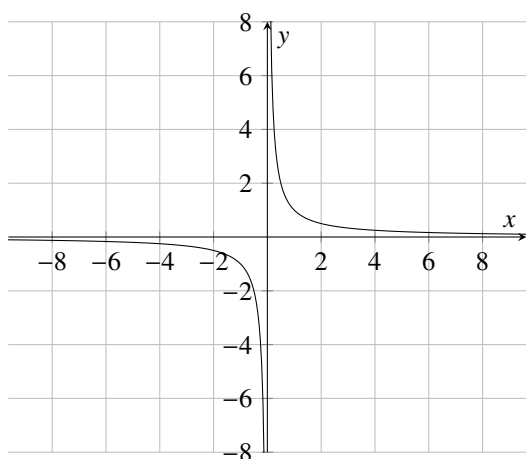


Figure 11.1. Reciprocal curve $y = 1/x$.

To demonstrate, consider the reciprocal curve $y = 1/x$ shown in Figure 11.1, which occupies the first and third quadrants and is symmetric about the origin. Notice that the point $x = 0$ is troublesome for this curve, as $y(0)$ itself is undefined. However, we can set up a limit to approach $x = 0$ from either the left- or the right-hand side. For the left-side limit, we have

$$\lim_{x \rightarrow 0^-} \frac{1}{x} = -\infty,$$

whereas on the right, we get a similar statement with all – signs as + signs

$$\lim_{x \rightarrow 0^+} \frac{1}{x} = \infty,$$

which couldn’t possibly be more different.

A new superscript has been slipped into each of the above equations, but the supporting term is already a subscript so it’s easy to miss. Reiterating, we denote an approach from the negative or positive direction (left or right) using

$$(x < x_0) \text{ or Left-Sided: } \lim_{x \rightarrow x_0^-},$$

$$(x > x_0) \text{ or Right-Sided: } \lim_{x \rightarrow x_0^+},$$

and the results need not be the same.

One-Sidedness

It is possible for only the left- or right-sided limit to exist, but not both. For instance, the square root curve $y = \sqrt{x}$ can only handle positive inputs, with the lowest allowable being $x = 0$, giving $y = 0$. Meanwhile, negative inputs lead to imaginary numbers or worse. Near the point $(0, 0)$, the best we can say is

$$\begin{aligned} \lim_{x \rightarrow 0^-} \sqrt{x} &= \text{Undefined} \\ \lim_{x \rightarrow 0^+} \sqrt{x} &= 0. \end{aligned}$$

Zero to the Zero

For another example, amusing debate arises among students and hobbyists when discussing the quantity 0^0 . Basic laws of exponents tell us that any number x raised to the zero-power equals one, however if we start with zero and raise it to any power x , the answer ought to be zero. So when with both numbers are zero, i.e. 0^0 , what happens?

Letting x be a real number, we set up the problem by writing a one-sided limit

$$\lim_{x \rightarrow 0^+} x^x = A,$$

where A stands for ‘Answer’. Next, explicitly check descending x values (starting from a reasonable guess) and inspect for an emerging trend or pattern:

$$0.1^{0.1} = 0.794328 \dots$$

$$0.01^{0.01} = 0.9549926 \dots$$

$$0.001^{0.001} = 0.99311605 \dots$$

$$0.0001^{0.0001} = 0.999079390 \dots$$

Evidently, the answer is getting closer to $A = 1$ as x is going to zero. This is motivation enough to write

$$\lim_{x \rightarrow 0^+} x^x = 1 ,$$

begging the conclusion

$$0^0 = 1 .$$

As a note of caution, not all numerical systems will treat 0^0 this way, particularly those that deal only in integers or a subset of them.

To affirm this result, one can write

$$\lim_{x \rightarrow 0^+} x \ln(x) = \ln(A) ,$$

which implies

$$0 \ln(0) \rightarrow \ln(A) .$$

Formally, the above means

$$0 \times (-\infty) \rightarrow \ln(A) ,$$

where the left side is undefined. However, taking the limit implies we get close to $x = 0$ without arriving there, thus the left side is arguably zero and the right yields $\ln(A) = 0$, or $A = 1$.

Two-Sidedness

A two-sided limit is more ‘well-behaved’ than a one-sided limit. In the two-sided case, the left- and right-sided limits are in agreement at a given x_0 :

$$\lim_{x \rightarrow x_0^-} \leftrightarrow \lim_{x \rightarrow x_0^+}$$

Two-sidedness applies to ‘unbroken’ curves (we’ll refine this word later), such as $y = \sqrt{x}$ at any point except $x = 0$, or instead $y = \cos(x)$ at any x at all.

1.6 Limits in Geometry

Limits have their place in the Cartesian plane as well as on the number line.

Circle and Line Problem

In Figure 11.2, a unit circle is centered at $(1, 0)$, and then a point (x, y) on the perimeter is chosen. If the distance from $(0, 0)$ to (x, y) is r , consider a line from $(0, r)$ through (x, y) that intersects the x -axis at $(p, 0)$.

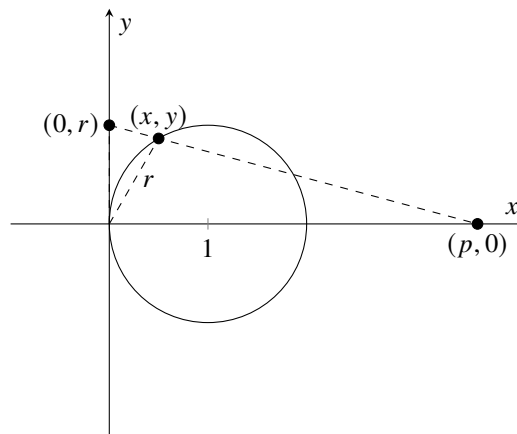


Figure 11.2. Unit circle intersecting straight line.

And now for the problem: Express the distance p in terms of r , and what happens to p as $r \rightarrow 0$? Right away, your mind’s eye should ‘animate’ the diagram and watch p slide along the x -axis as (x, y) moves on the circle.

To crack this, we’ll need the equation of the offset circle

$$(x - 1)^2 + y^2 = 1 ,$$

along with the Pythagorean theorem relating x , y , and r

$$x^2 + y^2 = r^2 ,$$

which after eliminating x tells us

$$y = \sqrt{r^2 - \frac{1}{4}r^4} .$$

Meanwhile, the line from $(0, r)$ to $(p, 0)$ has a slope given by $m = -r/p$, but the slope can also be written $m = (y - r)/x$. Eliminating m between these two equations and getting p in terms of r results in

$$p = \frac{r^2}{2 - \sqrt{4 - r^2}} .$$

Halfway there, but what happens to p as r goes to zero? Instinct may tell you the answer is $p \rightarrow 0$, but this would be *wrong*. Instead, try decreasing values in r and watch for

a pattern:

$$p(1) = \frac{1}{2 - \sqrt{3}} \approx 2.73205$$

$$p(.5) = \frac{(.5)^2}{2 - \sqrt{4 - (.5)^2}} \approx 3.93649$$

$$p(.25) = \frac{(.25)^2}{2 - \sqrt{4 - (.25)^2}} \approx 3.98431$$

$$p(.125) = \frac{(.125)^2}{2 - \sqrt{4 - (.125)^2}} \approx 3.99609$$

$$p(.0625) = \frac{(.0625)^2}{2 - \sqrt{4 - (.0625)^2}} \approx 3.99902$$

Evidently, the limit of p as r goes to zero is tending to 4.

2 Functions

2.1 Definition

A *function* is a ‘black box’ that takes an input value x and returns an output value $f(x)$. In the most general sense, the term x can represent *anything*, and $f(x)$ can be anything else.

Since algebra and calculus deal primarily in real numbers, we’ll narrow our idea of functions to those that work with real numbers. In this sense, we can think of a function as something that takes a number and gives back a new, presumably different number.

Domain and Range

The set of all possible x -values that a function can receive is called the *domain*, and the set of all possible outputs $f(x)$ is called the *range*. For instance, the domain of the function $f(x) = \cos(x)$ is all real numbers, as the cosine refuses no inputs. The range, on the other hand, is confined between -1 and 1 . For another example, the domain of the square root function $f(x) = \sqrt{x}$ is all non-negative real numbers, and so too is the range.

Uniqueness

For a given function f that takes input x , there can only be one output $f(x)$. For the case of curves $y(x)$ in the Cartesian plane, this amounts to the so-called ‘vertical line test’. If a curve $y(x)$ ever has two y -values for a given x , then y is not a function.

A common demonstration of this takes the equation of the unit circle

$$x^2 + y^2 = 1,$$

where the entire circle cannot be generated by a single Cartesian function for failing the vertical line test. On the other hand though, breaking the equation into two functions is valid, where we write

$$y_{\text{top}}(x) = \sqrt{1 - x^2}$$

$$y_{\text{bottom}}(x) = -\sqrt{1 - x^2},$$

where y_{top} and y_{bottom} are both functions of x .

2.2 Cartesian Functions

Functions of the form $y = f(x)$ that play nicely in Cartesian coordinates are the standard play things of calculus. Every

item in the following list qualifies as a function in Cartesian coordinates:

Line	$y = mx + b$
Parabola	$y = ax^2 + bx + c$
Cubic	$y = ax^3 + bx^2 + cx + d$
Polynomial	$y = a_0 + a_1x + a_2x^2 + \dots$
Square Root	$y = \sqrt{x}$
Exponential	$y = a^x$
Natural Exponential	$y = e^x$
Hyperbolic Cosine	$y = (e^x + e^{-x})/2$
Hyperbolic Sine	$y = (e^x - e^{-x})/2$
Logarithm	$y = \log(x)$
Natural Log	$y = \ln(x)$
Factorial	$y = x(x-1)(x-2)\dots 2 \cdot 1$
Cosine	$y = 1 - x^2/2! + x^4/4! - \dots$
Sine	$y = x - x^3/3! + x^5/5! - \dots$
Tangent	$y = x + x^3/3 + 2x^5/15 + \dots$

For most of the functions listed above, the valid domain is all real numbers $\mathbb{R}$. Special care is needed with some cases, such as the square root $y = \sqrt{x}$ by rejecting negative inputs.

The curious ‘factorial’ function, sometimes denoted

$$x! = x(x-1)(x-2)\dots,$$

is only valid for whole-number inputs, and has the curious property that

$$0! = 1.$$

This is certainly not reinforced by a traditional limit, but is instead a convention.

Periodicity

Functions that obey

$$f(x) = f(x + nL)$$

where L is a constant and n is an integer are *periodic*, and L is the period. Of the examples listed above, only the cosine and sine are periodic as written. The polynomial expression for tangent only works for one period.

Asymptotes

When a curve $y = f(x)$ ‘disappears off’ to infinity, whether it be in the horizontal direction or the vertical direction or otherwise, there may be an invisible line that the curve clearly does not cross. Such a line is called an *asymptote*, and the curve is said to be ‘asymptotic’ to said line.

One curve we’ve seen exhibiting asymptotic behavior is the reciprocal function $y = 1/x$ depicted in Figure 11.1. The curve never crosses the lines $x = 0$ and $y = 0$, despite being arbitrary close to doing so.

Not every curve that disappears off to infinity has an asymptote, though. For instance the parabolic curve $y = x^2$, despite the enormous U-shape, keeps extending horizontally as it does vertically.

2.3 Classifying Functions

Injective

A function is *one-to-one*, also known as *injective*, when every element in the domain $\{x\}$ maps to a unique element in the range $\{y\}$.

To illustrate quickly, a straight line $y = mx + b$ is injective over all real numbers. However, the parabola $y = x^2$ is not injective unless we confine the domain to say, $x > 0$.

Surjective

A function is *surjective*, also called ‘onto’, when every point in the range $\{y\}$ can be reached by some input in $\{x\}$.

Many curves we encounter are not surjective. To cook up an easy example anyway, consider the function $y = 2x$ over the domain of natural numbers. The domain is explicitly

$$\{x\} = \{1, 2, 3, 4, 5, 6, 7, 8, \dots\},$$

and the range is

$$\{y\} = \{2, 4, 6, 8, \dots\}.$$

This setup is deemed surjective because every member in $\{y\}$ is present in $\{x\}$.

Bijjective

A curve that is both injective and surjective is called *bijjective*. That is, a function is bijective if every member in $\{x\}$ uniquely leads to a member in $\{y\}$, and vice versa.

The square root $y = \sqrt{x}$ qualifies as a bijective function, as every x points to a unique y and vice-versa. Conversely, we can square both sides to have $x = y^2$, which is another

bijjective function $x = f(y)$ in the domain of non-negative real numbers.

2.4 Inverse Functions

Any bijective function $f(x)$ implies the existence of its *inverse function*, denoted $f^{-1}(x)$, also bijective, that interchanges the role of the domain and range. That is, you pretend y is the working variable and $x(y)$ is the inverse to $y(x)$.

The inverse of $y = f(x)$ can be formally defined as

$$f^{-1}(f(x)) = x \quad (11.4)$$

for bijective functions. Applying f to both sides gives a similar statement

$$f(f^{-1}(y)) = y.$$

Symmetry

On the Cartesian plane, it turns out that a function $y = f(x)$ and its inverse $x = f^{-1}(y)$ are symmetric about the line $y = x$.

To show this, suppose we have a function f and its inverse f^{-1} represented as follows:

$$y_1 = f(x)$$

$$y_2 = f^{-1}(x)$$

(To be concrete, one could imagine $y_1 = x^2$ and $y_2 = \sqrt{x}$ in the domain of non-negative reals, but this analysis will stay general.) Next, choose a point x_1 in the domain of y_1 so that

$$y_2 = f^{-1}(f(x_1)) = x_1,$$

and similarly choose a point x_2 in the domain of y_2 :

$$y_1 = f(f^{-1}(x_2)) = x_2$$

By now we have two locations in the Cartesian plane, namely (x_1, y_1) , (x_2, y_2) . The line connecting these has slope m and is given by

$$m = \frac{y_2 - y_1}{x_2 - x_1}.$$

Using $y_2 = x_1$ and $y_1 = x_2$, we quickly find

$$m = \frac{x_1 - x_2}{x_2 - x_1} = -1,$$

which says the slope of the line connecting our two test points is -1 . Perpendicular to this line is the line with slope $m_{\perp} = 1$, and the claim is proven.

2.5 Even and Odd Functions

Any function that obeys

$$f(-x) = f(x)$$

is called *even*, and is symmetric about the line $x = 0$ in the Cartesian plane. Any function that obeys

$$f(-x) = -f(x)$$

is called *odd*, and is anti-symmetric about the line $x = 0$.

While most functions are not even or odd exclusively, it is true that any function can be conceived as being the sum of an even part plus an odd part:

$$f(x) = f_{\text{even}}(x) + f_{\text{odd}}(x)$$

In terms of $f(x)$ and $f(-x)$, the even and odd components of a function are written

$$f_{\text{even}} = \frac{f(x) + f(-x)}{2}$$

$$f_{\text{odd}} = \frac{f(x) - f(-x)}{2}.$$

2.6 Functions and Limits

Functions and limits obey rules somewhat analogous to those of algebra, such as the distributive property. To flesh these out, consider a pair of functions $f(x)$, $g(x)$ where, for a special value x_0 in the domain, we also have

$$\lim_{x \rightarrow x_0} f(x) = A$$

$$\lim_{x \rightarrow x_0} g(x) = B.$$

Multiplication by a Constant

Consider a constant C . If $f(x)$ is multiplied by C , it should follow that C can be pulled outside of a limit:

$$\lim_{x \rightarrow x_0} (Cf(x)) = C \lim_{x \rightarrow x_0} f(x) = CA \quad (11.5)$$

The proof for this begins with a mouthful of mathematical jargon. Starting from the definition

$$\lim_{x \rightarrow x_0} f(x) = A,$$

we discern that for any given positive value ϵ (Greek *epsilon*), there exists another positive quantity δ (Greek *delta*), such that if

$$0 < |x - x_0| < \delta$$

then

$$|f(x) - A| < \epsilon.$$

All in one line, we write:

$$0 < |x - x_0| < \delta \implies |f(x) - A| < \epsilon$$

Fair enough, but for this proof we need the same thing for $|Cf(x) - CA|$. Using the language on hand, this means we need to show:

$$0 < |x - x_0| < \delta \implies |Cf(x) - CA| < \epsilon$$

Next, choose a different $\epsilon_1 > 0$, which means there exists a different $\delta_1 > 0$. We are free to let $\epsilon_1 = \epsilon/|C|$ to establish:

$$0 < |x - x_0| < \delta_1 \implies |f(x) - A| < \frac{\epsilon}{|C|}$$

With this setup, take a look at $|Cf(x) - CA|$ and factor out the $|C|$ -term:

$$|Cf(x) - CA| = |C| |f(x) - A|$$

The right-most quantity $|f(x) - A|$ can be replaced provided we let $\delta = \delta_1$, which leads to

$$|Cf(x) - CA| < |C| \frac{\epsilon}{|C|},$$

and the proof is done.

Distribution into Sum

The ‘limit’ construct distributes freely into the sum of two functions:

$$\lim_{x \rightarrow x_0} (f(x) \pm g(x)) =$$

$$\lim_{x \rightarrow x_0} f(x) \pm \lim_{x \rightarrow x_0} g(x) = A \pm B \quad (11.6)$$

To prove this, define an $\epsilon > 0$ and a pair of terms $\delta_1 > 0$, $\delta_2 > 0$ such that:

$$0 < |x - x_0| < \delta_1 \implies |f(x) - A| < \frac{\epsilon}{2}$$

$$0 < |x - x_0| < \delta_2 \implies |g(x) - B| < \frac{\epsilon}{2}$$

Choosing the positive channel in Equation (11.6), take sum of $f(x)$ and $g(x)$ and consider the quantity

$$|f(x) + g(x) - (A + B)|,$$

and use the triangle inequality to write

$$|f(x) + g(x) - (A + B)| = |f(x) + A| + |g(x) + B| ,$$

and replace the right side to finish:

$$|f(x) + g(x) - (A + B)| = \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

For the difference of $f(x)$ and $g(x)$, there is no need to introduce new epsilons and deltas. Exploit the previous result along with Equation (11.5) to find

$$\lim_{x \rightarrow x_0} (f(x) - g(x)) = A - B .$$

Distribution into Product

The ‘limit’ construct distributes also into the product of two functions:

$$\begin{aligned} \lim_{x \rightarrow x_0} (f(x) \cdot g(x)) &= \\ \lim_{x \rightarrow x_0} f(x) \cdot \lim_{x \rightarrow x_0} g(x) &= A \cdot B \end{aligned} \quad (11.7)$$

The proof of this requires some more epsilon-delta work. Define an $\epsilon > 0$ and a pair of terms $\delta_1 > 0$, $\delta_2 > 0$ such that:

$$\begin{aligned} 0 < |x - x_0| < \delta_1 &\implies |f(x) - A| < \sqrt{\epsilon} \\ 0 < |x - x_0| < \delta_2 &\implies |g(x) - B| < \sqrt{\epsilon} \end{aligned}$$

Choose δ to be the smaller of $\delta_{1,2}$, and then for $0 < |x - x_0| < \delta$, we find

$$\begin{aligned} |(f(x) - A) \cdot (g(x) - B)| &= \\ |(f(x) - A)| \cdot |(g(x) - B)| &, \end{aligned}$$

and replace the right side to establish

$$|(f(x) - A) \cdot (g(x) - B)| \leq \sqrt{\epsilon} \sqrt{\epsilon} .$$

With this, we’ve proved

$$\lim_{x \rightarrow x_0} (f(x) - A) \cdot (g(x) - B) = 0 ,$$

which helps toward the final result.

To proceed, expand the quantity

$$|(f(x) - A) \cdot (g(x) - B)|$$

using the distributive property:

$$\begin{aligned} |(f(x) - A) \cdot (g(x) - B)| &= \\ f(x)g(x) - Bf(x) - Ag(x) + AB \end{aligned}$$

Impose the limit $x \rightarrow x_0$ on every term. Right away, the left side is zero, and we’re left with

$$\begin{aligned} \lim_{x \rightarrow x_0} f(x)g(x) &= \\ B \lim_{x \rightarrow x_0} f(x) + A \lim_{x \rightarrow x_0} g(x) - AB &= \\ = BA + AB - AB &= \\ = AB , \end{aligned}$$

finishing the proof.

Distribution into Quotient

The ‘limit’ construct distributes also into the quotient of two functions:

$$\lim_{x \rightarrow x_0} \left(\frac{f(x)}{g(x)} \right) = \frac{\lim_{x \rightarrow x_0} f(x)}{\lim_{x \rightarrow x_0} g(x)} = \frac{A}{B} \quad (11.8)$$

Note of course that B cannot be zero.

For this proof, we need to first spend some effort to establish

$$\lim_{x \rightarrow x_0} \frac{1}{g(x)} = \frac{1}{B} .$$

Whatever B is, there exists a $\delta_1 > 0$ such that

$$0 < |x - x_0| < \delta_1 \implies |g(x) - B| < \frac{|B|}{2}$$

Now for a few algebraic manipulations. Start with $|B| = |B|$, and add zero to write

$$|B| = |B - g(x) + g(x)| .$$

By the triangle equality, we can proceed with

$$|B| < |B - g(x)| + |g(x)| ,$$

which, assuming the above, is replaced by

$$|B| < \frac{|B|}{2} + |g(x)| .$$

From this we discern

$$\frac{1}{|g(x)|} < \frac{2}{|B|} .$$

To continue, introduce a second $\delta_2 > 0$ such that

$$0 < |x - x_0| < \delta_2 \implies |g(x) - B| < \frac{|B|^2}{2} \epsilon .$$

Choose δ to be the smaller of $\delta_{1,2}$, and then examine the quantity $|1/g(x) - 1/B|$ in the regime

$$0 < |x - x_0| < \delta < \delta_{1,2}$$

so

$$\begin{aligned} \left| \frac{1}{g(x)} - \frac{1}{B} \right| &= \left| \frac{B - g(x)}{Bg(x)} \right| \\ &= \frac{1}{|B|} \frac{1}{|g(x)|} |g(x) - B| \\ &< \frac{1}{|B|} \frac{2}{|B|} \frac{|B|^2}{2} \epsilon. \end{aligned}$$

Everything except ϵ cancels on the right, and we have shown

$$\lim_{x \rightarrow x_0} \frac{1}{g(x)} = \frac{1}{B}.$$

From here, we have

$$\lim_{x \rightarrow x_0} \left(\frac{f(x)}{g(x)} \right) = \lim_{x \rightarrow x_0} \left(f(x) \frac{1}{g(x)} \right),$$

which by Equation (11.7) is also

$$\lim_{x \rightarrow x_0} \left(\frac{f(x)}{g(x)} \right) = \left(\lim_{x \rightarrow x_0} f(x) \right) \left(\lim_{x \rightarrow x_0} \frac{1}{g(x)} \right),$$

and this resolves to A/B , completing the proof.

Integer Powers

The ‘limit’ construct distributes also into exponents. We’ll establish this for integers n only, however:

$$\lim_{x \rightarrow x_0} (f(x))^n = \left(\lim_{x \rightarrow x_0} f(x) \right)^n = A^n \quad (11.9) \quad \text{also implying}$$

To begin, begin with

$$\lim_{x \rightarrow x_0} (f(x))^n = \lim_{x \rightarrow x_0} \left(f(x)^{n-1} f(x) \right),$$

which decouples by Equation (11.7) to

$$\lim_{x \rightarrow x_0} (f(x))^n = A \cdot \lim_{x \rightarrow x_0} \left(f(x)^{n-1} \right).$$

From here, we can use Equation (11.7) again to find

$$\lim_{x \rightarrow x_0} (f(x))^n = A^2 \cdot \lim_{x \rightarrow x_0} \left(f(x)^{n-2} \right),$$

and repeat this recursively for m steps

$$\lim_{x \rightarrow x_0} (f(x))^n = A^m \cdot \lim_{x \rightarrow x_0} (f(x)^{n-m}),$$

stopping when $n = m$.

Note that it’s possible to show that Equation (11.9) also holds when n is any real number.

X to the X

The quantity $y = x^x$ is satisfied by both $x = 1/2$ and $x = 1/4$ (same y). This is interesting because $1/2 \neq 1/4$, but $(1/2)^{1/2} = (1/4)^{1/4}$. Apart from the pair $(1/2, 1/4)$, let us find all pairs that satisfy $a^a = b^b$.

Let either member of the pair be written $1/z$, and the other is $1/z$ divided by a constant λ such that

$$\left(\frac{1}{z} \right)^{1/z} = \left(\frac{1}{\lambda z} \right)^{1/\lambda z}.$$

Take the natural log of each side and simplify to solve for z to get

$$\begin{aligned} z &= \lambda^{1/(\lambda-1)} \\ \lambda z &= \lambda^{\lambda/(\lambda-1)}, \end{aligned}$$

and similarly:

$$\begin{aligned} \frac{1}{z} &= \lambda^{-1/(\lambda-1)} \\ \frac{1}{\lambda z} &= \lambda^{-\lambda/(\lambda-1)} \end{aligned}$$

While the above is a workable solution, proceed by substituting $q = \lambda/(\lambda - 1)$ to find

$$\begin{aligned} \frac{1}{z} &= \lambda^{-q/\lambda} \\ \frac{1}{\lambda z} &= \lambda^{-q}, \\ z^\lambda &= \lambda z. \end{aligned}$$

The pair $(1/2, 1/4)$ corresponds to $\lambda = 2, z = 2$ satisfied by the above. For another test, let $q = 3$ to find $z = 9/4, \lambda = 3/2$, implying:

$$\left(\frac{4}{9} \right)^{4/9} = \left(\frac{8}{27} \right)^{8/27}$$

Let $q = 4$ to get $\lambda = 4/3, z = 64/27$ to discover the pair $(27/64, 81/256)$.

Testing a ‘large’ value such as $\lambda = 100$ gives $z = 100^{1/99} \approx 1.048$. Going further with $\lambda = 1000$ gives $z = 1000^{1/999} \approx 1.0069$. It should be clear that the upper bound on λ is infinity, corresponding to $z = 1$. To prove this, verify

$$z_\infty = \lim_{\lambda \rightarrow \infty} \lambda^{1/(\lambda-1)} = 1$$

as expected.

As constructed, we know $\lambda = 2$ is an allowed λ -value, but what is the minimum λ ? Trying $\lambda = 1$, calculate

$$z_1 = \lim_{\lambda \rightarrow 1} \lambda^{1/(\lambda-1)} = e.$$

This is reassuring as $\lambda = 1$ corresponds to the solution $(1/e, 1/e)$, which happens to be the minimum of $y = x^x$.

2.7 Continuity and Smoothness

Piecewise Functions

Consider the curious function

$$f(x) = \begin{cases} 2x - 1 & x < 1 \\ \sqrt{x} + 1 & x \geq 1 \end{cases},$$

which takes different form on either side of the point $x = 1$ as shown in Figure 11.3.

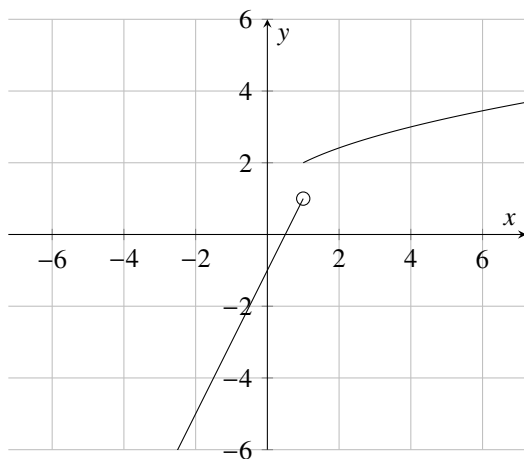


Figure 11.3. Piecewise function.

Continuity

While the above qualifies as a function in every sense, there's still something 'off' about such a piecewise function, namely the abrupt jump at $x = 1$. This type of phenomenon is called a *jump discontinuity*. Or, the function is said to be *discontinuous* at $x = 1$. The remainder of the curve for all points $x \neq 1$ is *continuous*.

Endpoint Discontinuity

Functions that threaten to return imaginary or complex results, such as $f(x) = \sqrt{x}$ when $x < 0$, have to terminate somewhere in the domain before this occurs. For the example on hand, the point $x = 0$ is an *endpoint discontinuity*.

Smoothness

Modifying the piecewise function above, suppose we nixed the $+1$ from the definition to write

$$f(x) = \begin{cases} 2x - 1 & x < 1 \\ \sqrt{x} & x \geq 1 \end{cases},$$

which has the effect of joining the two separate curves at the point $x = 1$. By doing this, the piecewise function is now continuous, but there is still a sense of something amiss, as there is an abrupt kink in $f(x)$ at $x = 1$. For this reason, the function lacks *smoothness* at $x = 1$. Every other place on the curve is both continuous and smooth.

Essential Singularity

A more severe type of discontinuity is called *essential* or *infinite* singularity, which occurs when one or both results of a two-sided limit around the discontinuity reach toward $\pm\infty$. This is the kind of behavior we see at $x = 0$ in the reciprocal curve $y = 1/x$ depicted in Figure 11.1.

2.8 Removable Singularity

There are plenty of functions in the wild that contain a singularity at first sight, but after some analysis, the singularity can be removed. Naturally, these are called *removable* singularities. A singularity x_0 is formally removable from a function $f(x)$ when the left- and right-sided limits near x_0 are in agreement, but $f(x_0)$ is singular or undefined:

$$\lim_{x \rightarrow x_0^-} f(x) = \lim_{x \rightarrow x_0^+} f(x) = \lim_{x \rightarrow x_0} f(x)$$

Factorable Numerator

For an example, consider the function

$$f(x) = \frac{x^2 - 3x + 2}{x - 1},$$

which clearly has a problem at $x = 1$. Looking at a few values around $x = 1$ though, we find something interesting:

$$\begin{aligned} f(0.8) &= -1.2 \\ f(0.9) &= -1.1 \\ f(0.99) &= -1.01 \\ f(0.999) &= -1.001 \\ f(1) &=? \\ f(1.001) &= -0.999 \\ f(1.01) &= -0.99 \\ f(1.1) &= -0.9 \\ f(1.2) &= -0.8 \end{aligned}$$

Given how $f(x)$ behaves near $x = 1$, it seems that $f(1)$ is tantalizingly close to -1 . In the language of limits, this would mean

$$\lim_{x \rightarrow x_0^-} f(x) = -1 = \lim_{x \rightarrow x_0^+} f(x),$$

the signature of a removable singularity.

In fact, there is something fishy about the example function, because $x - 1$ can be factored out of the numerator as

$$f(x) = \frac{(x-1)(x-2)}{x-1} = x-2,$$

like there was never a singularity at all.

Problem 3

Classify all singularities in

$$y = f(x) = \frac{x-1}{x^2-1}.$$

Answer: Removable singularity at $x = 1$ and infinite singularity at $x = -1$.

Natural Logarithm

For a less trivial example, consider the function

$$y = f(x) = \frac{n^x - 1}{x}, \quad (11.10)$$

and let's be interested in the quantity $f(0)$. Immediately we see that $x = 0$ makes Equation (11.10) blow up, but can we remove this point? Consulting a plot of $f(x)$ shown in Figure 11.4, it seems the singularity is removable.

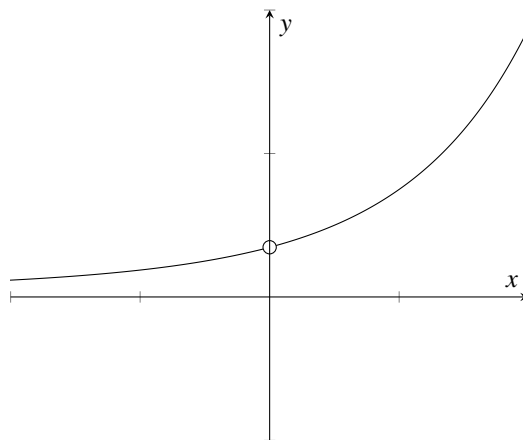


Figure 11.4. Function with removable singularity at $x = 0$.

To proceed, solve Equation (11.10) for n , and set up a limit that pushes x to zero:

$$n = \lim_{x \rightarrow 0} (1 + xy)^{1/x}$$

This setup is looking almost familiar, but even more so if we make the substitution

$$h = 1/x,$$

because the limit of x going to zero is the same as the limit of h going to infinity:

$$n = \lim_{h \rightarrow \infty} \left(1 + \frac{y}{h}\right)^h$$

By Equation (11.2), the right side of the above is identical to e^y , so we find $n = e^y$. Solving for y , we finally have

$$f(0) = \ln(n).$$

Evidently, the singular point in $f(x)$ is the natural log of n . For completeness, this result can also be stated in a way complimentary to Equation (11.2):

$$\ln(x) = \lim_{h \rightarrow 0} \frac{x^h - 1}{h} \quad (11.11)$$

Sinc Function

An interesting singularity arises in the so-called 'sinc' function, which is defined as $\sin(x)$ divided by x ,

$$\text{sinc}(x) = \frac{\sin(x)}{x}, \quad (11.12)$$

sketched in Figure 11.5. As a product of odd functions, the sinc function is even, i.e. symmetric about $x = 0$.

The period of oscillation is still 2π like the sine function. The amplitude wiggles pathetically under an envelope of $y = 1/x$.

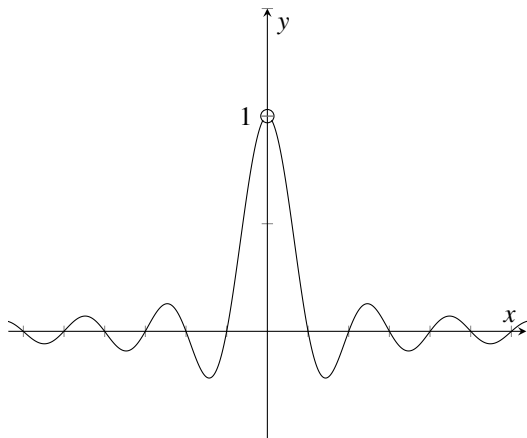


Figure 11.5. Plot of $\sin(x)/x$.

The point $x = 0$ invokes division by zero, however the plot of the sinc function begs the singularity be removed, as the plot seems to ‘want’ to pass through the point $(0, 1)$. Framing the result as a limit, this is

$$\text{sinc}(0) = \lim_{x \rightarrow 0} \frac{\sin(x)}{x}.$$

Evaluating the right side of the above can be done in a number of ways. One method is to expand $\sin(x)$ as an infinite polynomial, i.e.

$$\sin(x) = x - x^3/3! + x^5/5! - \dots,$$

and then $\sin(x)/x$ must be

$$\frac{\sin(x)}{x} = 1 - x^2/3! + x^4/5! - \dots,$$

and now the right side has no singularity at all. We can forget about limits and simply set $x = 0$ to find

$$\text{sinc}(0) = \lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1, \quad (11.13)$$

and the problem is finished.

2.9 Squeeze Theorem

Consider a two functions $f(x)$, $h(x)$ in the Cartesian plane such that

$$f(x) \geq h(x)$$

in the whole domain. Suppose also that there is a point x_0 where the two functions are equal to the same value L . This takes the form of a limit since the curves can never intersect:

$$\lim_{x \rightarrow x_0} f(x) = L = \lim_{x \rightarrow x_0} h(x)$$

To so-called *squeeze theorem* states that, if we have a third function $g(x)$ defined such that

$$f(x) \geq g(x) \geq h(x),$$

which is ‘between’ the first two curves, then the value of $g(x)$ at x_0 is also approaching L :

$$\lim_{x \rightarrow x_0} f(x) = \lim_{x \rightarrow x_0} g(x) = \lim_{x \rightarrow x_0} h(x) = L$$

Problem 4

Use the squeeze theorem to show that

$$\lim_{x \rightarrow 0} x^2 \sin\left(\frac{1}{x}\right) = 0.$$

Answer: From the properties of the sine function, we can write for sure that

$$-1 \leq \sin\left(\frac{1}{x}\right) \leq 1$$

for any x . Then, multiply the entire statement through by x^2 :

$$-x^2 \leq \sin\left(\frac{1}{x}\right) \leq x^2$$

In the limit $x \rightarrow 0$, the terms $-x^2$ and x^2 clearly both go to zero, so the quantity squeezed between them must also go to zero.

2.10 A Strange Beast

Fun Problem

Try to solve for x in the equation

$$\sqrt{x + \sqrt{x + \sqrt{x + \sqrt{x + \sqrt{\dots}}}}} = 2,$$

where the left side contains an infinite nesting of $\sqrt{x + \dots}$ as shown.

This is actually easier than it seems (spoiler alert). Square both sides to write

$$x + 2 = 2^2,$$

and easily find $x = 2$. (Many who encounter this for the first time can’t resist trying it on a calculator.)

Disaster

Now, try to solve for x in this version:

$$\sqrt{x + \sqrt{x + \sqrt{x + \sqrt{x + \sqrt{\dots}}}}} = 1$$

Doing the same steps, i.e. squaring both sides and so on, the above boils down to

$$x^2 + 1 = 1,$$

solved only by $x = 0$. Hang on though, because substituting $x = 0$ tells us

$$\sqrt{0 + \sqrt{0 + \sqrt{0 + \sqrt{0 + \sqrt{\dots}}}}} = 1,$$

but $0 = 1$ *can't* be right, so where is the error?

Analysis

To see what went wrong, generalize the problem by writing an 'open' equation

$$y = \sqrt{x + \sqrt{x + \sqrt{x + \sqrt{x + \sqrt{\dots}}}}},$$

where y surely describes a curve, but hesitate to call y a function just yet. Squaring both sides now gives

$$y^2 = x + y,$$

and solve for y again:

$$y = \frac{1}{2} \pm \frac{\sqrt{1 + 4x}}{2}$$

From here, we see that setting $x = 0$ gives two results, namely $y_1 = 1$ and $y_2 = 0$. These are two perfectly legal solutions to $y^2 = x + y$.

However, if y is to be a proper function, it follows that one of y_1 or y_2 must be thrown out. To avoid getting $0 = 1$, discard the y_1 solution and keep y_2 . The proper way to write $y(x)$ as a closed function must be done in piecewise fashion:

$$y(x) = \begin{cases} 0 & x = 0 \\ 1/2 + \sqrt{1 + 4x}/2 & x > 0 \end{cases}$$

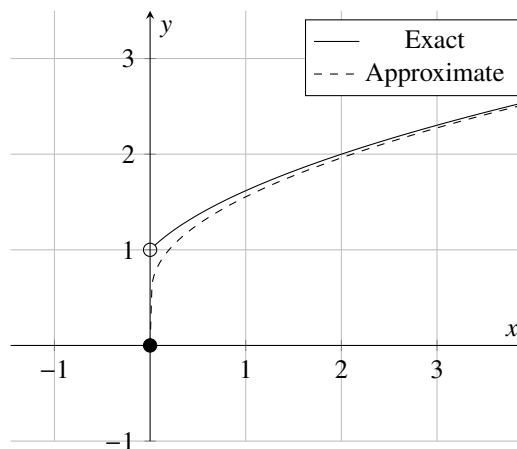


Figure 11.6. A strange beast.

The function $y(x)$ is plotted as a solid curve in Figure 11.6. Note the 'hole' removed from $(0, 1)$ is placed instead at $(0, 0)$, and no solution exists for $y = 1$.

Also in the Figure is a dashed-line representation of the approximate answer generated from the 'open' form of $y(x)$ out to three square roots. Truncating the radical this way, one sees the dashed curve starting from $(0, 0)$ and trying to reach $(0, 1)$ in a continuous manner. In the infinite limit of nested roots, the piecewise function needs to take over.

Golden Ratio

An interesting solution to the so-called strange beast equation occurs at $x = 1$. From this we have

$$y(1) = \frac{1 + \sqrt{5}}{2} = 1.618034 \dots,$$

a celebrated number called the *golden ratio*, denoted ϕ . Given the open version for $y(x)$, we derive a nifty expression for ϕ :

$$\sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{\dots}}}}} = \phi$$

3 Sequences

3.1 Definition of a Sequence

A *sequence* is an ordered list of numbers or other information. An entire sequence can be represented by a letter or symbol, such as A , or more explicitly, $\{A\}$. Each member of a sequence is assigned a unique whole number index, typically written as a subscript j called an *index*, starting with $j = 1$ unless otherwise specified.

For instance, the numbers 3, 6, 9, in that order, can be assigned

$$\begin{aligned} A_1 &= 3 \\ A_2 &= 6 \\ A_3 &= 9, \end{aligned}$$

and written out via

$$\{A\} = \{3, 6, 9\}.$$

A sequence with a countable number of members n is called a *finite* sequence, or *closed* sequence. All finite sequences can be represented by the form

$$\{A\} = \{A_1, A_2, A_3, \dots, A_n\}.$$

Note that the exact representation of a sequence can vary among sources and authors. For instance, the above sometimes appears as:

$$\{A\} = \{A_j\} = \{A_j\}_1^n$$

The starting index need not be $j = 1$.

3.2 Infinite Sequences

A sequence with an infinite number of members $n \rightarrow \infty$ can also be conceived, called an *infinite sequence*. The ‘last’ term L in an infinite sequence can be expressed as the limit

$$L = \lim_{j \rightarrow \infty} A_j.$$

Convergent Sequence

For finite L , we may require that for any positive ϵ , there exists some integer m such that

$$j > m \rightarrow |A_j - L| < \epsilon.$$

This is only supported by sequences where each A_j is finite. Such a sequence is called *convergent*. The terms of a

convergent sequence approach a single value for increasing j .

By this criteria, the following infinite sequences are convergent (assume the patterns go forever):

$$\{A\} = \{10, 20, 1, 2, 0.1, 0.2, 0.01, 0.02, \dots\}$$

$$\{B\} = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \dots\right\}$$

$$\{C\} = \left\{\frac{9}{10}, \frac{99}{100}, \frac{999}{1000}, \frac{9999}{10000}, \dots\right\}$$

Divergent Sequence

A *divergent* sequence is one that has $L \rightarrow \pm\infty$. This is characterized by a growing trend in the late-stage A_j for which there is no finite ceiling (or floor). The following sequences diverge (assume the patterns go forever):

$$\{Q\} = \{1, 10, 100, 1000, 10000, \dots\}$$

$$\{R\} = \{-1, -2, -3, -4, -5, -6, \dots\}$$

Periodic Sequence

A *periodic sequence* is an infinite sequence that neither converges nor diverges, but instead repeats after a certain index. A sequence such as

$$\{A\} = \{1, 2, 3, 1, 2, 3, 1, 2, 3, \dots\}$$

is periodic.

Recursive Sequence

By studying polynomial division, one can argue into existence a sequence $\{F\}$ obeying the *recursion relation*,

$$F_j = pF_{j-1} + F_{j-2},$$

where p is a parameter. The case $F_1 = 1$ with $p = 1$ gives the Fibonacci sequence:

$$\{F_{p=1}\} = \{1, 1, 2, 3, 5, 8, 13, 21, \dots\}$$

Nonexistent Limits

Sequences that have a random nature or are otherwise ‘not going anywhere’ imply a nonexistent limit. For instance, alternating values of ± 1 via

$$\{B\} = \{5, -5, 2, -2, 7, -7, 3, -3, \dots\}$$

implies a nonexistent limit. Even though the terms A_n seem to be oscillating around zero, there is no sense of a strict limit.

Similar comments go for the digits of any irrational number such as π :

$$\{P\} = \{3, 1, 4, 1, 5, 9, 2, 6, 5, \dots\}$$

3.3 Analogy to Functions

A particular comparison between sequences and functions can be framed in terms of a limit. Consider a sequence $\{A\}$, along with a function $f(x)$ such that $f(n) = A_n$. With this, the following is always true:

$$\lim_{x \rightarrow \infty} f(x) = L = \lim_{j \rightarrow \infty} A_j \quad (11.14)$$

Squeeze Theorem

In the same way that functions obey the so-called squeeze theorem, so too is the case for sequences. Consider three sequences $\{A\}$, $\{B\}$, $\{C\}$. If

$$A_j < B_j < C_j$$

and

$$\lim_{j \rightarrow \infty} A_j = L = \lim_{j \rightarrow \infty} C_j$$

then

$$L = \lim_{j \rightarrow \infty} B_j .$$

Absolute Value

Using the fact that any variable x obeys $\pm x \leq |x|$, it follows that member x_j in a sequence obeys

$$-|x_j| \leq x_j \leq |x_j| .$$

For an infinite sequence $\{x_j\}$, we can further write (see scalar multiplication below):

$$\lim_{j \rightarrow \infty} (-|x_j|) = L = - \lim_{j \rightarrow \infty} |x_j|$$

In the special case

$$\lim_{j \rightarrow \infty} |x_j| \rightarrow 0 ,$$

thereby making $L \rightarrow 0$, the squeeze theorem leads to a stronger conclusion:

$$\lim_{j \rightarrow \infty} x_j \rightarrow 0$$

3.4 Algebraic Properties

The members of two convergent sequences $\{A\}$, $\{B\}$, obey algebraic properties you would expect. Each of the following is a consequence of Equation (11.14):

Addition

$$\lim_{j \rightarrow \infty} (A_j \pm B_j) = \lim_{j \rightarrow \infty} A_j \pm \lim_{j \rightarrow \infty} B_j$$

Scalar Multiplication

$$\lim_{j \rightarrow \infty} \lambda A_j = \lambda \lim_{j \rightarrow \infty} A_j$$

Product

$$\lim_{j \rightarrow \infty} (A_j B_j) = \left(\lim_{j \rightarrow \infty} A_j \right) \left(\lim_{j \rightarrow \infty} B_j \right)$$

Ratio

As long as the denominator is not zero:

$$\lim_{j \rightarrow \infty} \frac{A_j}{B_j} = \frac{\lim_{j \rightarrow \infty} A_j}{\lim_{j \rightarrow \infty} B_j}$$

Exponent

As long as $A_j > 0$:

$$\lim_{j \rightarrow \infty} A_j^q = \left(\lim_{j \rightarrow \infty} A_j \right)^q$$

3.5 Geometric Sequence

Consider the geometric sequence

$$\{x_j\} = \{x^j\}_0^\infty .$$

Intuitively, it would make sense that the sequence converges only in the interval $-1 < x \leq 1$, but some subtleties arise when proving this. Starting with $f(x) = x^j$, write

$$\lim_{x \rightarrow \infty} x^j = L = \lim_{n \rightarrow \infty} A_n .$$

Handling the easy cases first, we see that $x > 1$ leads to $L \rightarrow \infty$, and the sequence diverges. The exact case $x = 1$ gives $L = 1$, which is convergent. The trivial case $x = 0$ gives $L = 0$, also convergent.

For $0 < x < 1$, the limit of $f(x)$ goes to zero, so L is also going to zero, and the sequence converges. The case $-1 < x < 0$ is also convergent. To show this, let $r^j = |x^j|$

so we're checking the domain $0 < r < 1$. This reproduces the previous case, and we have

$$\lim_{j \rightarrow \infty} r^j = 0 = \lim_{j \rightarrow \infty} |x^j| ,$$

thus the sequence converges.

The case $x = -1$ attempts a limit that does not exist. Writing this case out, one finds the periodic sequence

$$\{(-1)^j\} = \{1, -1, 1, -1, 1, -1, \dots\} ,$$

exhibiting neither convergence nor divergence.

The ugliest case occurs at $x < -1$. For instance, choosing $x = -2$ leads to

$$\{(-2)^j\} = \{1, -2, 4, -8, 16, -32, 64, \dots\} ,$$

which is not convergent, not divergent, and non-periodic.

In conclusion, we say the geometric sequence $\{x_j\}$ converges only if $-1 < x \leq 1$, and furthermore:

$$\lim_{j \rightarrow \infty} x^j = \begin{cases} 0 & -1 < x < 1 \\ 1 & x = 1 \end{cases} \quad (11.15)$$

3.6 Terminology

Now comes some obligatory terminology to be less verbose when talking about sequences. The following comments apply to both finite and infinite sequences.

Increasing

A sequence $\{A\}$ is *increasing* if, for all j :

$$A_{j+1} > A_j$$

Decreasing

A sequence $\{A\}$ is *decreasing* if, for all j :

$$A_{j+1} < A_j$$

Monatonic Sequence

A sequence $\{A\}$ that is increasing or decreasing is called *monatonic*.

Bounded Below

For a sequence $\{A\}$, if there exists a *lower bound* M such that $M < A_j$ for all j , the sequence is called *bounded below*.

Bounded Above

For a sequence $\{A\}$, if there exists an *upper bound* M such that $M > A_j$ for all j , the sequence is called *bounded above*.

Bounded

A sequence $\{A\}$ that is bounded below and bounded above is *bounded*. A sequence that converges is bounded and monatonic.

4 Series

4.1 Partial Sum

Given a sequence $\{A_j\}_1^n$, one can imagine packing the members A_j into a *partial sum*:

$$\begin{aligned}s_1 &= A_1 \\ s_2 &= A_1 + A_2 \\ s_3 &= A_1 + A_2 + A_3 \\ s_m &= A_1 + A_2 + A_3 + \cdots + A_m\end{aligned}$$

For the last sum, it's assumed that $m \leq n$.

Finite Series

In the special case $m = n$, i.e. when the partial sum adds all terms in a sequence, the sum is called the *series*. For finite n , the series has a finite number of terms.

4.2 Sigma Notation

Any sum

$$s_m = A_1 + A_2 + A_3 + \cdots + A_m$$

can be written using the so-called sigma notation:

$$s_m = \sum_{j=1}^m A_j$$

The variable j , most often an integer, is called the *index*. The initial and final values for j appear as the respective subscript and superscript on the Σ symbol. The index increases by one with each iteration of the sum.

Infinite Series

For an infinite sequence $n \rightarrow \infty$, the sum

$$S = \lim_{m \rightarrow \infty} s_m = \sum_{j=1}^{\infty} A_j$$

is the *infinite series*.

Perhaps the most accessible infinite series is the infinite geometric series, which converges for $|x| < 1$:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{j=0}^{\infty} x^j$$

4.3 Converging Series

Some of the groundwork and terminology developed for sequences also applies to infinite series. If $\{A\}$, $\{B\}$ are both convergent sequences, things like addition and scalar multiplication carry directly to the series:

$$\begin{aligned}\sum_{j=1}^{\infty} (A_j + B_j) &= \sum_{j=1}^{\infty} A_j + \sum_{j=1}^{\infty} B_j \\ \sum_{j=1}^{\infty} \lambda A_j &= \lambda \sum_{j=1}^{\infty} A_j\end{aligned}$$

On the other hand, the product of $\{A\}$, $\{B\}$ is not a single series over the product of the components:

$$\left(\sum_{j=1}^{\infty} A_j \right) \left(\sum_{j=1}^{\infty} B_j \right) \neq \sum_{j=1}^{\infty} (A_j B_j)$$

Even worse, it's not clear that the product of two convergent series is itself convergent.

Criteria for Convergence

A sequence $\{A\}$ converges if and only if

$$\lim_{j \rightarrow \infty} A_j = 0.$$

To prove this, write two partials sums s_j and s_{j-1} , which only differ by the coefficient A_j :

$$A_j = s_j - s_{j-1}$$

If the sequence converges, both s_j and s_{j-1} converge, namely because j and $j-1$ both tend to infinity. Since the above equation takes their difference, we get zero on the right, finishing the proof.

Absolute Convergence

A sequence $\{A\}$ *converges absolutely* if

$$\sum_{j=1}^{\infty} |A_j|$$

converges.

Conditional Convergence

A sequence $\{A_j\}$ converges conditionally if

$$\sum_{j=1}^{\infty} A_j$$

converges but

$$\sum_{j=1}^{\infty} |A_j|$$

diverges.

Rearranging Terms

When a series is converging, the terms in the series can be rearranged without consequence. This is explicitly untrue for divergent series.

For an example from the geometric series, it's easy to show that

$$\frac{1}{2} = \frac{1}{4} - \frac{1}{8} + \frac{1}{16} - \frac{1}{32} + \frac{1}{64} - \frac{1}{128} + \cdots,$$

where the series on the right converges. Grouping positives and negatives together, which should be done with caution in general, gives

$$\begin{aligned} \frac{1}{2} &= \left(\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots \right) \\ &\quad - \left(\frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots \right), \end{aligned}$$

simplifying to

$$\frac{1}{2} = \left(\frac{1}{4} - \frac{1}{8} \right) \left(\frac{1}{1 - 1/4} \right) = \left(\frac{3}{8} \right) \left(\frac{4}{3} \right) = \frac{1}{2},$$

as expected. The answer will be same regardless of how the terms in the series are arranged.

The story is different for a conditionally converging series. In fact, it turns out that the terms in such a diverging series can be arranged to give *any* real number.

4.4 Finite Sums

Fixed Exponent

Going back to the finite sequence $\{A_j\}_1^n$, there are a few cases that arise often in calculations, so let's understand

these ahead of time. Consider the sums:

$$S_1 = \sum_{j=1}^n j = 1 + 2 + 3 + \cdots + n$$

$$S_2 = \sum_{j=1}^n j^2 = 1 + 2^2 + 3^2 + \cdots + n^2$$

$$S_3 = \sum_{j=1}^n j^3 = 1 + 2^3 + 3^3 + \cdots + n^3$$

The sum S_1 can be evaluated by counting matching pairs of numbers that sum to n . The number n itself is trivially paired with zero. The number 1 is paired with $n-1$, summing to n . This continues for any pair j and $n-j$, and there are $n/2$ of these. The 'middle' number in the series, namely $n/2$, gets no matching partner. Therefore we have:

$$S_1 = \frac{n}{2}n + \frac{n}{2} = \frac{n(n+1)}{2}$$

In fact, this identity is worth memorizing:

$$\sum_{j=1}^n j = \frac{n(n+1)}{2} \quad (11.16)$$

Calculating S_2 is a bit harder. To get started, guess the solution as being a polynomial involving powers of n and unknown coefficients:

$$S_2(n) = \alpha n^3 + \beta n^2 + \gamma n.$$

If Greek characters are unfamiliar, these are 'alpha' (α), 'beta' (β), 'gamma' (γ).

For $n=1$, the sum is simply 1, and we find

$$S_2(1) = 1 = \alpha + \beta + \gamma.$$

For $n=2$, we may write

$$S_2(2) = 1 + 2^2 = 8\alpha + 4\beta + 2\gamma,$$

and similarly for $n=3$,

$$S_2(3) = 1 + 2^2 + 3^2 = 27\alpha + 9\beta + 3\gamma.$$

What we now have is a system of three equations and three unknowns, which is enough to solve for α, β, γ . Doing so by hand or by using matrix methods, we end up with:

$$\alpha = 1/6$$

$$\beta = 1/2$$

$$\gamma = 1/6$$

Evidently then, we have

$$S_2(n) = \frac{n^3}{3} + \frac{n^2}{2} + \frac{n}{6}$$

and the hard work is done. The right side can be factored to deliver the final result:

$$\sum_{j=1}^n j^2 = \frac{n(n+1)(2n+1)}{6} \quad (11.17)$$

For completeness, it turns out the result for S_3 is:

$$\sum_{j=1}^n j^3 = \left(\frac{n(n+1)}{2} \right)^2 \quad (11.18)$$

The proof can be done with the method of unknown coefficients.

Powers of Two

Consider the finite sum

$$I_n = \sum_{j=0}^n 2^j = 1 + 2 + 2^2 + 2^3 + \cdots + 2^n$$

for finite integer n .

To evaluate the sum, divide both sides by two to write

$$\begin{aligned} \frac{I_n}{2} &= \frac{1}{2} + 1 + 2 + 2^2 + \cdots + 2^{n-1} \\ &= \frac{1}{2} + I_{n-1}. \end{aligned}$$

Note from the definition of I_n that

$$I_{n-1} = I_n - 2^n,$$

and thus

$$\frac{I_n}{2} = \frac{1}{2} + I_n - 2^n.$$

Solving for I_n , one finds

$$I_n = 2^{n+1} - 1.$$

In summary:

$$\sum_{j=0}^n 2^j = 2^{n+1} - 1 \quad (11.19)$$

4.5 Shift of Index

The series

$$S_n = \sum_{j=1}^n A_j$$

has one index variable j running over the integers from 1 to n . Sometimes it's useful to perform a shift of index to a new letter k such that

$$k = j + \alpha,$$

where α is any integer:

$$S_n = \sum_{k=1+\alpha}^{n+\alpha} A_{k-\alpha}$$

For a striking example of this, consider the infinite series

$$C = \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots = \sum_{j=1}^{\infty} \frac{j}{(j+1)!}.$$

Substitute $k = j + 1$ and the above becomes

$$C = \sum_{k=2}^{\infty} \frac{k-1}{k!} = \sum_{k=2}^{\infty} \frac{1}{(k-1)!} - \sum_{k=2}^{\infty} \frac{1}{k!}.$$

Let $m = k - 1$ in the first sum, and simply relabel $k \rightarrow m$ in the second:

$$C = \sum_{m=1}^{\infty} \frac{1}{m!} - \sum_{m=2}^{\infty} \frac{1}{m!}$$

Pluck the first term from the first sum and the rest cancels out:

$$C = 1 + \sum_{m=2}^{\infty} \frac{1}{m!} - \sum_{m=2}^{\infty} \frac{1}{m!} = 1$$

Chapter 12

Probability and Statistics

1 Events and Probability

Probability theory is a branch of mathematics for studying systems with inherent randomness or uncertainty. It works closely with *statistics*, another branch of mathematics concerned with the organization and interpretation of data, along with *combinatorics*, a formal method of counting.

Stochastic Process

Systems that exhibit random or pattern-less behavior contain a *stochastic* component. Typical stochastic processes may include flipping a coin, drawing a card from a deck, rolling a dice, or playing darts while blindfolded. A *stochastic event* is any data generated by a stochastic process, and the set of all possible stochastic events is called the system's *sample space*.

1.1 Events

Elementary Events

Events that are considered *elementary* carry one 'unit' of information, loosely speaking. A coin landing on 'heads', or a dice landing on 3 qualify as elementary events.

Compound Events

Simple events that occur in groups are called *compound events*. Drawing a Queen of Hearts from a deck of cards carries two units information, and may be interpreted in several ways: 'draw a Queen AND a heart', or 'draw a Queen OR a Heart', or perhaps 'draw NOT a Diamond'. Such events are compound for this reason.

Compound Event Notation

Borrowing the familiar symbols from elementary logic, we denote the word 'AND' with the 'cap' symbol $\cap$, equivalent to multiplication ($\cdot$). Meanwhile, the word 'OR' uses the 'cup' symbol $\cup$, or sometimes just a plus sign ($+$). The 'NOT' operator is abbreviated by a dash above the symbol, as in 'NOT' $A = \bar{A}$. Any event that is infinitely improbable, impossible, or undefined is denoted by the 'Empty set' symbol, $\emptyset$. In summary:

$$A \text{ AND } B = A \cdot B = A \cap B$$

$$A \text{ OR } B = A + B = A \cup B$$

$$\text{NOT } A = \bar{A}$$

$$\text{Empty set} = \emptyset$$

The logic of probabilistic analysis is the same as 'ordinary' logic. For instance, the philosophical axiom 'nothing can be and not be simultaneously' is contained in the statement:

$$A \cap \bar{A} = \emptyset$$

State

The *state* of a system, loosely defined, is any particular configuration of the variables used to describe that system. For instance, a snapshot of a chessboard contains the present state of the game. Any event taking place in a system usually changes its state. If the system is to evolve in time, as would a game of chess, then future states evolves from the present state according to some rules or model of evolution.

1.2 Probability

Statistical Probability

A stochastic process that iterates over a very large or infinite number of trials will produce data points randomly distributed among the space of all possible data points for that process. For all events of type A , the ratio of occurrences N_A over all N events is called the *statistical probability* of event A , defined as:

$$P(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N} \quad (12.1)$$

$P(A)$ strictly has values between 0 and 1, inclusive.

Normalization Conditions

All other events B , C , etc., are represented by the symbol $\bar{A}$ ('NOT' A), and obey:

$$N_A + N_{\bar{A}} = N \quad (12.2)$$

$$P(\bar{A}) + P(A) = 1 \quad (12.3)$$

Classical Probability

A definition that skirts around the invocation of $N \rightarrow \infty$ is called the *classical probability*. For instance, it does not require an infinite number of rolls on a six-sided dice to know the chances of landing on 3 are one out of six, as this quality is built into the dice itself. Classical systems like dice or playing cards are most succinctly analyzed using classical probability.

Counting States

In probability and statistics, it's often necessary to know the total number of states available to a system, sometimes requiring rigorous combinatorial consideration.

Problem 1

The last four digits of a phone number have the format ABCD, where each letter represents any integer from 0 to 9, inclusive. What is the probability of randomly guessing the number 7766?

Answer: Right away, we know how to list all possible states of the password, starting from 0000 and ending at 9999 in numerical order. With $N = 1000$ passwords, the probability of randomly choosing the correct password $N_A = 7766$ is:

$$P(7766) = \frac{1}{N} = \frac{1}{10000}$$

Problem 2

A bank account password has format ABCD, where each letter represents any integer from 0 to 3, inclusive. What is the probability of randomly guessing the password?

Answer: All two-digit arrangements solved by AB are contained in:

$$\begin{aligned} \omega = & 00, 01, 02, 03, \\ & 10, 11, 12, 13, \\ & 20, 21, 22, 23, \\ & 30, 31, 32, 33 \end{aligned}$$

From here, observe that all four-digit arrangements are contained on an $\omega \times \omega$ grid having $N = 16^2 = 256$ total members, or

$$P(N_A) = \frac{1}{N} = \frac{1}{256}.$$

1.3 Mutually Exclusive Events

A pair of *mutually exclusive events* A and B are those that cannot occur simultaneously. Their coincidence can only belong to the *empty set* as

$$A \cap B = \emptyset.$$

If two events are mutually exclusive, the probability of either event occurring is the sum of the individual probabilities:

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) \quad (12.4)$$

Problem 3

Calculate the probability of rolling a 3 or a 4 on a six-sided dice.

Answer: As mutually exclusive events, we simply have

$$P = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Problem 4

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts. Answer:

$$P = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{3}{52}$$

1.4 Non-Exclusivity

Non-mutually exclusive events are those that cause ‘double counting’ in $P(A \cup B)$, and are adjusted by subtracting the probability that both occur:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (12.5)$$

Or, in street terms, the above reads:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Problem 5

From a 52-card deck, calculate the probability of drawing a Heart or a Face card, or one that is both. Answer:

$$P = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

Problem 6

A class of 30 students is in session. 16 are studying French, and 21 are studying Spanish. Choosing a student at random, find the probability that they:

- study French
- study Spanish
- study French and Spanish
- study only French
- study only Spanish
- study French or Spanish

Answer: Denote F for French and S for Spanish. Then the easy ones can be listed off:

$$P(F) = 16/30$$

$$P(S) = 21/30$$

If the number of ‘multilingual’ students studying both French and Spanish is denoted M , then

$$(16 - M) + M + (21 - M) = 30$$

must hold, telling us $M = 7$, or

$$P(F \cap S) = 7/30.$$

With M known, the the number of students studying just one subject can be written:

$$P(\text{French only}) = P(F) - P(F \cap S) = 9/30$$

$$P(\text{Spanish only}) = P(S) - P(F \cap S) = 14/30$$

Finally, the number of students studying French or Spanish should equal the total, which is the sum of those studying French only, Spanish only, or both. The probability should equal one:

$$P(F \cup S) = \frac{9}{30} + \frac{14}{30} + \frac{7}{30} = 1$$

1.5 Independent Events

Two events A and B that occur simultaneously as the compound event $A \cap B$ are *independent* if not causally connected.

In general, the statistical probability for the compound event $A \cap B$ reads

$$P(A \cap B) = \lim_{N \rightarrow \infty} \frac{1}{N} N_{A \cap B},$$

where in the $N \rightarrow \infty$, limit the quantity $N_{A \cap B}$ becomes $N_A \cdot P(B)$. We deduce that, for independent events, the compound probability is the product of the individual probabilities:

$$P(A \cap B) = P(A) P(B) \quad (12.6)$$

Problem 7

Calculate the probability of two fair coin tosses each landing on ‘tails’. Answer:

$$P(T \cap T) = P(T) \cdot P(T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Problem 8

From a 52-card deck, what is the probability of randomly drawing (i) a Queen, (ii) a Heart, (iii) the Queen of Hearts? Answer:

$$P(Q) = 1/4$$

$$P(H) = 1/13$$

$$P(Q \cap H) = P(Q) \cdot P(H) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}$$

Coworker Problem

Two people are neighbors and travel to the same job. Person A owns car A , which has a 70% chance of starting in the morning, and a 30% chance of stalling (not starting). Person B owns car B with an 80% chance of starting. If one or both cars start, both people arrive at work. If neither car starts, they both miss work. In a span of 100 workdays, how many days are missed?

Each morning, one of four things happen:

$$P_1 = \text{A starts, B starts}$$

$$P_2 = \text{A starts, B stalls}$$

$$P_3 = \text{A stalls, B starts}$$

$$P_4 = \text{A stalls, B stalls}$$

As independent events, we further have

$$P_1 = (0.7) (0.8) = 0.56$$

$$P_2 = (0.7) (1 - 0.8) = 0.14$$

$$P_3 = (1 - 0.7) (0.8) = 0.24$$

$$P_4 = (1 - 0.7) (1 - 0.8) = 0.06,$$

which passes the sanity check

$$\sum_{j=1}^4 P_j = 1.$$

The answer to the question is the combined probability of at least one car starting. For this, we simply have

$$P = P_1 + P_2 + P_3 = 0.56 + 0.14 + 0.24 = 0.94,$$

or 94%. Six days are missed of every hundred.

1.6 Conditional Probability

In contrast to independent events, systems may bear a notion of ‘dependent events’, meaning that event B can occur only if event A occurs. This is called a *conditional probability*, denoted $P(B|A)$, enunciated ‘ B given A ’. By definition, the probability of event B occurring given condition A is

$$P(B|A) = \lim_{N \rightarrow \infty} \frac{1}{N_A} N_{A \cap B}.$$

The term $N_{A \cap B}$ is the number of events B that occur given event A , which shows up again in the equation for $P(A \cap B)$:

$$P(A \cap B) = \lim_{N \rightarrow \infty} \frac{N_{A \cap B}}{N}.$$

Divide the two equations and simplify to derive the statement of conditional probability:

$$P(A \cap B) = P(B|A) P(A) \quad (12.7)$$

Note that the above generalizes the case of independent events, for if events A and B are independent, this result reduces to $P(A \cap B) = P(A) P(B)$ again.

Problem 9

In a 52-card deck, calculate the probability that the first three cards are Kings. Answer:

$$\begin{aligned} P(KKK) &= P(K) P(K|K) P(K|(K|K)) \\ &= \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \approx 0.000181 \end{aligned}$$

Problem 10

In a 52-card deck, calculate the probability that the first three cards are KQJ , in that order, with mixed suits allowed. Answer:

$$\begin{aligned} P(KQJ) &= P(K) P(Q|K) P(J|(Q|K)) \\ &= \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} \approx 0.000483 \end{aligned}$$

Problem 11

Suppose a pair of six-sided dice are rolled, landing on faces X and Y , respectively. What is the probability that $X = 2$ given that $X + Y \leq 5$?

Answer: Of the 36 possible outcomes for a pair of dice, 10 of them obey $X + Y \leq 5$. Listing these, find that only $(2, 1)$, $(2, 2)$, $(2, 3)$ satisfy $X = 2$, a total of three outcomes. The final answer is the ratio of these counts:

$$P(X = 2 | X + Y \leq 5) = \frac{3}{10}$$

Inversion Trick

Making use of the normalization condition

$$P(\bar{A}) + P(A) = 1,$$

it often helps in problem solving to use the negated event $\bar{A}$ as the working variable.

Problem 12

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts.

Answer: For this, define the event $\bar{A}$ of *not* drawing the Queen of Hearts and use the inversion trick as follows:

$$\bar{A} = \bar{A}_1 \cdot \bar{A}_2 \cdot \bar{A}_3$$

$$\begin{aligned} P(\bar{A}_1 \cap \bar{A}_2) &= P(\bar{A}_1) P(\bar{A}_2 | \bar{A}_1) \\ &= \frac{51}{52} \cdot \frac{50}{51} = \frac{50}{52} \end{aligned}$$

$$\begin{aligned} P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) &= P(\bar{A}_1 \cap \bar{A}_2) P(\bar{A}_3 | \bar{A}_1 \cap \bar{A}_2) \\ &= \frac{50}{52} \cdot \frac{49}{50} = \frac{49}{52} \end{aligned}$$

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{49}{52} = \frac{3}{52}$$

Radioactive Decay

An unstable atom is one that expels energy by ejecting a sub-atomic particle or photon. Having no internal time-keeping mechanism, an unstable atom is entirely ‘unaware’ of its absolute age, and the its decay occurs at a random moment after becoming unstable.

Supposing the observation of an unstable atom begins at $t = 0$, the conditional probability of the atom decaying in a small time window Δt after time $t > 0$ is

$$P_{\Delta t/t}^{\text{decay}} = \tau^{-1} \Delta t ,$$

where τ^{-1} is a constant related to (but not precisely equal to) the statistical half-life of the element, defined such that $\Delta t \ll \tau$.

The probability of the atom being ‘still alive’ in the interval Δt is

$$P_{\Delta t/t}^{\text{alive}} = 1 - \Delta t/\tau .$$

Decompose the entire ‘alive’ state into a product of conditional probabilities by slicing the time t into n identical copies of the short interval Δt as:

$$P^{\text{alive}}(t) = P_{\Delta t/t_1}^{\text{alive}} \cdot P_{\Delta t/t_2}^{\text{alive}} \cdots P_{\Delta t/t_n}^{\text{alive}} = \left(1 - \frac{t}{n\tau}\right)^n$$

Letting $n \rightarrow \infty$ permits use of the identity

$$\lim_{n \rightarrow \infty} \left(1 + \frac{A}{n}\right)^n = e^A .$$

It follows that the probability that a single unstable atom will still be ‘alive’ obeys

$$P(t) = e^{-t/\tau} . \quad (12.8)$$

One can work out the so-called half life $\tau_{1/2}$ of the atom by inquiring when $P(t)$ reduces to $1/2$.

Missing Face Problem

A six-sided dice is rigged to keep rolling if it lands on 2. Prove that the statistical probability of rolling a 3 is $1/5$.

Denote B as the event 2, and denote A as event 3. With a single roll, the probabilities of A or B occurring are easy to write down:

$$\begin{aligned} p_1(A) &= 1/6 \\ p_1(B) &= 1/6 \end{aligned}$$

Of course, event B is unstable and induces a re-roll, which has a $1/6$ chance of generating event A again, and the same chance for event B :

$$\begin{aligned} p_2(A|B) &= (1/6)(1/6) = (1/6)^2 \\ p_2(B|B) &= (1/6)(1/6) = (1/6)^2 \end{aligned}$$

With event $B|B$ comes another re-roll, and we stack on the conditional probabilities as

$$\begin{aligned} p_3(A|B|B) &= (1/6)^3 \\ p_3(B|B|B) &= (1/6)^3 , \end{aligned}$$

and the pattern is clear.

It follows that A could occur after any number of rolls, or potentially occur after an infinite string of B -events, and the total probability for A occurring is

$$\begin{aligned} P(A) &= p_1(A) + p_2(A|B) \\ &\quad + p_3(A|B|B) + p_4(A|B|B|B) + \cdots , \end{aligned}$$

simplifying to:

$$P(A) = \frac{1}{6} \cdot \left(1 + \left(\frac{1}{6}\right) + \left(\frac{1}{6}\right)^2 + \left(\frac{1}{6}\right)^3 + \cdots\right)$$

In the infinite limit, the geometric series in parentheses converges to $6/5$. The probability of event A occurring is therefore:

$$P(A) = \frac{1}{6} \cdot \frac{6}{5} = \frac{1}{5}$$

1.7 Bayes’ Theorem

For two events A and B , recall that the statement of conditional probability reads

$$P(A \cap B) = P(B|A) P(A) ,$$

which gives the likelihood events A and B simultaneously occurring. It’s equally valid to write the statement with A and B swapped, giving a complimentary statement for ‘ A given B ’:

$$P(B \cap A) = P(A|B) P(B)$$

Now, since $A \cap B$ is logically equivalent to $B \cap A$, we immediately know $P(A \cap B) = P(B \cap A)$, allowing the two conditional equations to be combined to arrive at *Bayes’ theorem*:

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)} \quad (12.9)$$

Laptop Repair Shop

Source: CMPSCI240 UMass Amherst 2013

Problem 13

You work in a laptop repair shop. 80% of laptops brought in have been dropped, 15% of laptops have had a drink spilled on them, and 5% of laptops have a variety of other problems. A customer drops off a laptop and doesn't tell you what happened to it. You notice the laptop is emitting a slight coffee-like smell. Based on your knowledge of broken laptops, you estimate that 20% of dropped laptops have a slight coffee-like smell, 65% of laptops that have had something spilled on them have a slight coffee-like smell, and 5% of laptops that have some other problem have a slight coffee-like smell.

Provide labels for the events described in the problem. Find the probability that the laptop had something spilled on it given that it has a slight coffee-like smell.

Part 2:

After closer inspection, you note that the laptop has no cracks on the case. Based on your knowledge of broken laptops, you estimate that 80% of dropped laptops have cracked cases, 11% of laptops that have had something spilled on them have cracked cases, and 9% of laptops that have some other problem have cracked cases.

If the probability that a laptop smells like coffee and the probability that a laptop has a cracked case are conditionally independent of each other given the cause of the damage (drop, spill, or other), what is the probability that the laptop had something spilled on it if it has a slight coffee-like smell and no cracks in the case?

Solution 13

Denote D for 'drop', S for 'spill', and O for 'other'. Let the letter F denote 'slight coffee-like smell'. The information provided in the problem may be written:

$$P(D) = .80$$

$$P(S) = .15$$

$$P(O) = .05$$

$$P(F|D) = .20$$

$$P(F|S) = .65$$

$$P(F|O) = .05$$

We further deduce:

$$\begin{aligned} P(F) &= P(F|D)P(D) + P(F|S)P(S) \\ &\quad + P(F|O)P(O) = .26 \end{aligned}$$

To answer the question, we need to compute $P(S|F)$, which is the inversion of $P(F|S)$. Applying Bayes' theorem, we easily find:

$$P(S|F) = \frac{P(S)P(F|S)}{P(F)} = .375$$

Part 2:

The problem asks us to evaluate $P(S|F \cap Z^C)$, where Z denotes 'crack' and Z^C denotes 'no crack,' using the information

$$P(Z|D) = .80$$

$$P(Z|S) = .11$$

$$P(Z|O) = .09,$$

or equivalently:

$$P(Z^C|D) = 1 - .80 = .20$$

$$P(Z^C|S) = 1 - .11 = .89$$

$$P(Z^C|O) = 1 - .09 = .91$$

Proceed by applying Bayes' theorem directly to write

$$P(S|F \cap Z^C) = \frac{P(F \cap Z^C|S)P(S)}{P(F \cap Z^C)}.$$

Due to the independence between F and Z^C , the term

$$P(F \cap Z^C|S)$$

decouples into

$$P(F|S)P(Z^C|S).$$

The denominator term $P(F \cap Z^C)$ can be recast as a sum that factors in a similar way:

$$\begin{aligned} P(F \cap Z^C) &= \sum_{i=D,S,O} P(F \cap Z^C|X_i)P(X_i) \\ &= \sum_{i=D,S,O} P(F|X_i)P(Z^C|X_i)P(X_i) \end{aligned}$$

Evaluating the final answer is now straightforward:

$$P(S|F \cap Z^C) = .7169$$

2 Applications

2.1 Copernicus Method

Source: Futility Closet
<https://futilitycloset.com/>

Princeton astrophysicist J. Richard Gott was visiting the Berlin Wall in 1969 when a curious thought occurred to him. His visit occurred at a random moment t years after the wall was created. Dividing the total lifespan T of the wall into four equal intervals, Gott reasoned there is a 50% chance that t lands within the middle two quarters of the wall's lifespan. Built in 1961, the wall was $t = 8$ years old at the time of his visit.

With this setup, we see at minimum that t is one quarter of T . At the other extreme, t could be three quarters of T . We therefore write

$$T_{\max} = 4t \quad T_{\min} = t \times \frac{4}{3}$$

to establish upper and lower estimates of the wall's lifespan. Inserting $t = 8$ years, we find

$$T_{\max} = 32 \text{ yr} \quad T_{\min} \approx 10.67 \text{ yr},$$

which, when these are added to 1961, produces the pair of results

$$\text{Year}_{\max} \approx 1993 \quad \text{Year}_{\min} \approx 1972.$$

That is, Gott found a 50% chance that the Berlin Wall would fall between the years 1972 and 1993. The wall came down in 1989.

Generalization

Now generalize the above method using N intervals instead of four. Doing so, we begin with

$$T_{\max} = Nt \quad T_{\min} = \frac{Nt}{N-1}.$$

Of course, the window defined by $T_{\max} - T_{\min}$ no longer corresponds to a probability of 50%, but must be adjusted to

$$p(N) = \frac{N-2}{N},$$

or

$$N(p) = \frac{2}{1-p},$$

which must be an integer.

An interesting exercise is one that calculates $\Delta T = T_{\max} - T_{\min}$ and expresses the result all in terms of $p(N)$. Figuring this out, one finds

$$\Delta T = \frac{4tp}{1-p^2},$$

which can be inverted via the quadratic formula:

$$p(\Delta T) = \frac{-2t}{\Delta T} + \sqrt{\left(\frac{2t}{\Delta T}\right)^2 + 1}$$

Problem 14

Suppose you encounter a man for the first time in the 42nd year of his life. Determine the upper and lower bounds of the interval in which he has a 33% chance of expiring (in total years). Repeat for 66% and 75%. Answer:

$$T_{\max(33\%)} = 3 \times 42 = 126$$

$$T_{\min(33\%)} = \frac{3}{2} \times 42 = 63$$

$$T_{\max(66\%)} = 6 \times 42 = 252$$

$$T_{\min(66\%)} = \frac{6}{5} \times 42 = 50.4$$

$$T_{\max(75\%)} = 8 \times 42 = 336$$

$$T_{\min(75\%)} = \frac{8}{7} \times 42 = 48$$

2.2 Dice Stacking

Suppose you are given a black box containing a pair of three-sided dice, abbreviated $2d3$. Denoting the outcome of a given dice as 1, 2, 3, we can list the 'internal outcomes' if the whole box is rolled:

$$\omega = 11, 12, 21, 13, 22, 31, 23, 32, 33$$

As a black box, there is no way to look inside. Instead, the box contains a small creature who announces the sum of the digits after a roll. Summing pairs of digits in ω , we write the corresponding black box outcomes as

$$\sigma = 2, 3, 3, 4, 4, 4, 5, 5, 6,$$

having five unique members.

In a more visual notation, the above is equivalent to:

$$\sigma_2 = \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}$$

$$\sigma_3 = \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}, \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}$$

$$\sigma_4 = \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}, \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}, \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}$$

$$\sigma_5 = \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}, \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}$$

$$\sigma_6 = \begin{array}{|c|c|} \hline \cdot & \cdot \\ \hline \end{array}$$

Effective d5?

The question on hand is, can we use the black box as an effective five-sided dice? There are five unique members in σ , but the density of outcomes is not uniform.

From the information contained in σ , we can work toward an effective $d5$ so long as the nonuniform distribution problem can be dealt with. Proceed by listing the allowed outcomes divided by the respective density, i.e.

$$\gamma = \frac{1}{1} (2), \frac{1}{2} (3), \frac{1}{3} (4), \frac{1}{2} (5), \frac{1}{1} (6) .$$

Multiply by $3 \cdot 2$ to get rid of all denominators:

$$\gamma = 6 (2), 3 (3), 2 (4), 3 (5), 6 (6)$$

Explicitly, γ is a list with twenty items:

$$\gamma = 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 6$$

Outcomes that are under-represented in σ are over-represented in γ , and vice-versa.

The d5 Gamma Ray

The item γ is the ‘gamma-array’, or ‘gamma ray’ for short. Taking any gamma ray γ , assign the items in γ to an index via $\gamma(k)$, where k is an integer between 1 and 20, inclusive.

In the above example, one can see there is nothing particularly special about the arrangement of items 2, 3, 4, etc.

in the array. These are listed by group in ascending order for mere convenience.

Since the intent is to build a uniformly-behaving $d5$ dice, it’s prudent to conceive of the set $\{\gamma\}$ of *all* possible reshuffles of the items in γ . Of course, one wouldn’t attempt writing down the full set $\{\gamma\}$, or stepping through its members in any systematic way. It suffices instead to have a function that randomly shuffles an existing gamma ray to produce another.

The d5 Algorithm

An effective $d5$ is achieved with the following steps:

1. Choose a random gamma ray from $\{\gamma\}$ and let $k = 1$.
 2. Let x equal the random sum of a $2d3$ roll:

$$x = \text{dice}(3) + \text{dice}(3)$$
 3. If $x = \gamma(k)$, record $x - 1$ as a valid result.
 4. Let $k \rightarrow k + 1$.
 5. If $k > 20$ then Goto 1.
 6. Goto 2.
-

3 Combinatorics

3.1 Arrangements

The number A_n of all *arrangements* of n distinguishable (non-repeated) elements is equal to the factorial of the total number of elements:

$$A_n = n!$$

If the set of n elements contains any number m identical members, the number of arrangements over-counts by a factor of m -factorial, which must be divided out:

$$A_n^m = \frac{n!}{m!} \quad (12.10)$$

Problem 15

Consider the set of twelve elements *ACEFGILMNTUX*. What is the probability that a random arrangement of the elements will spell out *MAGNETICFLUX*? Answer:

$$P(\text{MAGNETICFLUX}) = \frac{1}{12!}$$

Problem 16

Consider the word *FLUXELECTRIC*. What is the probability that a random arrangement of the letters will spell out *ELECTRICFLUX*? Answer:

$$P(\text{ELECTRICFLUX}) = \frac{2!2!2!}{12!}$$

3.2 Permutations

For a set of width n , partition each of the $n!$ arrangements into two bins such that one bin contains the first m elements in the arrangement, and the other bin contains the remaining $n - m$ elements. For each of the m elements in the first bin, the unused elements in the second bin are subject to $(n - m)!$ arrangements. Dividing out this factor yields the *permutation* number:

$$P_n^m = \frac{n!}{(n - m)!} \quad (12.11)$$

Qualitatively, the permutation number tells how many ways there are to choose m unique elements from a set of n total elements.

Problem 17

A door keypad is unlocked by a code of four different integers between 0 to 9, inclusive. The same integer cannot be used twice. How any possible passwords are there?

Answer: For a $k = 4$ digit password drawing (and consuming) from $N = 10$ integers, observe that N of them are available for the first digit A . $N - 1$ of the digits are available for the second digit B , and so on, with the k^{th} digit selecting from $N - k + 1$ unused integers. In general, we can intuitively write

$$P_N^k = N(N - 1) \dots (N - k + 1) = \frac{N!}{(N - k)!},$$

which builds the permutation formula:

$$P_{10}^4 = \frac{10!}{(10 - 4)!} = \frac{10!}{6!} = 5040$$

Problem 18

In a 52-card deck, calculate the probability that the first three cards are Kings. (This is a repeat of an earlier problem.)

Answer: The total number of ways to draw any three cards from 52 is

$$P_{52}^3 = \frac{52!}{(52 - 3)!} = 52 \cdot 52 \cdot 50.$$

Meanwhile, the number of ways to draw any three Kings from four total Kings is

$$P_4^3 = \frac{4!}{(4 - 3)!} = 4 \cdot 3 \cdot 2.$$

The ratio of these is the answer:

$$P(KKK) = \frac{P_4^3}{P_{52}^3} = \frac{4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50} \approx 0.000181$$

Problem 19

In a 52-card deck, calculate the probability that the first three cards are *KQJ*, in that order, with mixed suits allowed. (This is a repeat of an earlier problem.)

Answer: Consider the three cards *KQJ* in that order. Listing off all ways this could occur, we see there are 4^3 possibilities in total, as each card has four suits to choose from. The total number of ways to draw any three cards from 52 is

$$P_{52}^3 = \frac{52!}{(52 - 3)!} = 52 \cdot 52 \cdot 50.$$

The ratio of these is the answer:

$$P(KQJ) = \frac{4^3}{52 \cdot 51 \cdot 50} \approx 0.000483$$

Birthday Problem

Consider a room populated by N people. What is the probability that any two people were born on the same day? (Ignore leap year.)

Begin with the trivial case $N = 2$, in where there is a $1/365$ chance of a common birthday:

$$P(2) = \frac{1}{365} = 1 - \frac{364}{365}$$

The result is written in the form $1 - X$ so we may focus on X , the probability of *no* common birthday.

A third person entering the system, making $N = 3$, has $365 - 2 = 363$ available days to avoid a common birthday. The probability becomes

$$P(3) = 1 - \frac{364}{365} \cdot \frac{363}{365},$$

and the pattern becomes clear. For total population N , the probability that some pair of people share a birthday ought to be:

$$\begin{aligned} P(N) &= 1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{(365 - N + 1)}{365} \\ &= 1 - \frac{365!}{365^N (365 - N)!} \end{aligned}$$

On the right, note that X has been expressed as a recursion of conditional probabilities:

$$X(n|n-1) = \frac{365 - (n-1)}{365}$$

$$X(N) = \prod_{n=2}^N X(n|n-1)$$

The same result can be written directly by the permutation formula. Choosing N people of 365, we have

$$P_{365}^N = \frac{365!}{(365 - N)!}.$$

Meanwhile, the number of ways to assign birthdays to N people is 365^N without worrying about sharing. The ratio of these is the same $X(N)$ calculated above:

$$X(N) = \frac{1}{365^N} P_{365}^N = \frac{365!}{365^N (365 - N)!}.$$

Following is a list of various populations N with their

corresponding $P(N)$:

N	P(N)
5	2.71%
10	11.7%
20	41.1%
23	50.7%
30	70.6%
50	97.0%

Remarkably, the population need only be 23 in order for there to be a 50% chance that any two people share a birthday.

3.3 Combinations

Extending the derivation of the permutation formula, it may happen that the precise order of elements in the ' m ' bin does not matter, meaning the list of permutations is overpopulated by a factor of $m!$. Dividing out this factor, we attain the number of *combinations* in the system:

$$C_n^m = \frac{n!}{m! (n - m)!}$$

The numbers C_n^m are none other than the binomial coefficients:

$$C_n^m = \binom{n}{m}$$

Problem 20

From a 52-card deck, a five-card hand is drawn at random. How many five-card hands are possible? Answer:

$$C_{52}^5 = \frac{52!}{5! (52 - 5)!}$$

Problem 21

From a 52-card deck, calculate the probability of drawing a royal flush (A-K-Q-J-10) in any order in any one suit. Answer:

$$P(RF) = \frac{4}{C_{52}^5} = \frac{1}{649740} \approx 0.00000154$$

Lottery Game

In a lottery game, the winning numbers are five non-repeating integers between 1 and 75, inclusive, along with one bonus integer between 1 and 15, inclusive. Guessing the five winning numbers at random, let us calculate the probability $P = (n, b)$ that n of the guessed numbers match the winning numbers, with or without the bonus b also being correctly guessed.

As an application of combinatoric analysis, it follows that there are $C_{75}^5 = 17,259,390$ ways to guess the winning five numbers, and $C_{15}^1 = 15$ choices for the bonus number. Following are the probabilities of guessing partial winning numbers, with and without the bonus.

$$P(5, 1) = \frac{1}{C_{75}^5 \cdot C_{15}^1} = \frac{1}{258,890,850}$$

$$P(4, 1) = \frac{C_5^4 \cdot C_{70}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{739,688}$$

$$P(3, 1) = \frac{C_5^3 \cdot C_{70}^2}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{10,720}$$

$$P(2, 1) = \frac{C_5^2 \cdot C_{70}^3}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{473}$$

$$P(1, 1) = \frac{C_5^1 \cdot C_{70}^4}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{56}$$

$$P(0, 1) = \frac{C_5^0 \cdot C_{70}^5}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{21}$$

$$P(5, 0) = \frac{C_5^5 \cdot C_{70}^0 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{18,492,204}$$

$$P(4, 0) = \frac{C_5^4 \cdot C_{70}^1 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{52,835}$$

$$P(3, 0) = \frac{C_5^3 \cdot C_{70}^2 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{766}$$

$$P(2, 0) = \frac{C_5^2 \cdot C_{70}^3 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{34}$$

$$P(1, 0) = \frac{C_5^1 \cdot C_{70}^4 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{4}$$

$$P(0, 0) = \frac{C_5^0 \cdot C_{70}^5 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{2}{3}$$

4 Variables and Data

4.1 Normalization

For an event A , the probability $P(A)$ of the event occurring has a trivial yet important relationship to $P(\bar{A})$, via the normalization condition

$$1 = P(A) + P(\bar{A}) .$$

In words, normalization means there is a 100% chance that event A either occurs or does not occur.

For n repeated events, also called trials, A is replaced by A_k , where the index k tracks the event number such that $1 \leq k \leq n$. Using this notation, we lump $\bar{A}$ and all subsequent $\bar{A}_k$ into the coefficients A_k to write a more general normalization condition:

$$1 = \sum_{k=1}^n P(A_k) \quad (12.12)$$

4.2 Statistical Average

Expanding out the normalization condition above, we have a sequence with n terms on the right:

$$1 = P(A_1) + P(A_2) + \cdots + P(A_n)$$

By multiplying A_k into each respective $P(A_k)$ term, the equation becomes the *statistical average*, or *weighted average* $\langle A \rangle$ of the events A_k . To denote this, we write:

$$\langle A \rangle = A_1 \cdot P(A_1) + A_2 \cdot P(A_2) + \cdots + A_n \cdot P(A_n)$$

In summation notation, the above result reads:

$$\langle A \rangle = \sum_{k=1}^n A_k \cdot P(A_k) \quad (12.13)$$

4.3 Expectation Value

A function f that depends on any event A_k can also be averaged using this apparatus. Generalizing the above, we can easily write an equation for the *expectation value* of f :

$$\langle f \rangle = \sum_{k=1}^n f(A_k) \cdot P(A_k) \quad (12.14)$$

With the above, we may also calculate $\langle f^2 \rangle$ without hesitation:

$$\langle f^2 \rangle = \sum_{k=1}^n (f(A_k))^2 \cdot P(A_k) \quad (12.15)$$

Problem 22

A six-sided dice that chooses a random number $1 \leq A_k \leq 6$ is tossed in succession to produce $n \gg 1$ events. Calculate the average outcome. Answer:

$$\langle A \rangle = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = 3.5$$

Problem 23

A six-sided dice that is missing the 2-face but has an extra 4-face is tossed in succession to produce $n \gg 1$ events. Calculate the average outcome. Answer:

$$\langle A \rangle = \frac{1}{6} + \frac{0}{6} + \frac{3}{6} + \frac{4 \cdot 2}{6} + \frac{5}{6} + \frac{6}{6} = \frac{23}{6} = 3.83$$

4.4 Standard Deviation

Further insight into f can be gained by inserting $(f(A_k) - \langle f \rangle)^2$ as the argument in Equation (12.15). By doing so, and then taking the square root of the entire result, we arrive at the *standard deviation* in the system:

$$\sigma_f = \sqrt{\sum_{k=1}^n (f(A_k) - \langle f \rangle)^2 \cdot P(A_k)} \quad (12.16)$$

Using only the definitions above, it's easy to show that the standard deviation is equivalent to

$$\begin{aligned} \sigma_f &= \sqrt{\langle f^2 \rangle - 2\langle f \rangle \langle f \rangle + \langle f \rangle^2} \\ \sigma_f &= \sqrt{\langle f^2 \rangle - \langle f \rangle^2} . \end{aligned} \quad (12.17)$$

4.5 Mean, Median, Mode

Mean

For a given set of numerical data, a synonym for the statistical average is the *mean*.

Median

The *median* in a sorted numerical set is the number directly in the 'middle' of the list. If the set has an odd number of members, simply pick the middle number. If the set has an even number of members, the median is the average of the two numbers competing for the middle slot. Note that the median is not the same as the midpoint value.

Mode

The *mode* is the member or set of members that occurs most often in a given set. For example, the set

$$\{1, 2, 2, 2, 3, 3, 4, 4, 4\}$$

has two modes, namely 2 and 4. On the other hand, the set

$$\{2, 4, 6, 8, 10\}$$

has no mode.

Problem 24

Suppose:

- x is the average of m and 9
- y is the average of $2m$ and 15
- z is the average of $3m$ and 18

What is the average of x, y, z in terms of m ? Answer: $m + 7$

4.6 Experimental Error Analysis

In any experiment, a measured quantity A should always be accompanied by a window of uncertainty ΔA such that the working value is $A \pm \Delta A$. The same goes for any other measured quantity B , which is alone meaningless without writing $B \pm \Delta B$.

Supposing we're interested in the product $P = AB$ of two measured values, we write

$$\begin{aligned} P + \Delta P &= (A \pm \Delta A)(B \pm \Delta B) \\ &= AB + A\Delta B + B\Delta A + \Delta A\Delta B. \end{aligned}$$

If the uncertainties $\Delta A, \Delta B$ are much less than the principal values A, B , then the product $\Delta A\Delta B$ is very small compared to the other terms and is effectively zero. Thus, the uncertainty in P is

$$\pm \Delta P = A\Delta B + B\Delta A,$$

simplifying to

$$\pm \frac{\Delta P}{P} = \left(\frac{\Delta A}{A} + \frac{\Delta B}{B} \right),$$

or, again because $\Delta A\Delta B$ is negligible:

$$\frac{\Delta P}{P} = \sqrt{\left(\frac{\Delta A}{A} \right)^2 + \left(\frac{\Delta B}{B} \right)^2}$$

Interestingly, the same form arises from the first-order Taylor's theorem in two dimensions. For a general function $F = F(A \pm \Delta A, B \pm \Delta B)$, we find

$$(\Delta F)^2 \approx \left(\frac{\partial F}{\partial A} \right)^2 (\Delta A)^2 + \left(\frac{\partial F}{\partial B} \right)^2 (\Delta B)^2,$$

which readily generalizes for more variables. In the following we bloom out the expansion for ΔF for common scenarios.

Sums and Differences

Consider three measured quantities A, B, C such that

$$F = A + B - C.$$

The above trivially delivers

$$\Delta F = \sqrt{\Delta A^2 + \Delta B^2 + \Delta C^2}.$$

Products

For the product $P = AB$, we already found that the uncertainty in P is given by

$$\frac{\Delta P}{P} = \sqrt{\left(\frac{\Delta A}{A} \right)^2 + \left(\frac{\Delta B}{B} \right)^2}.$$

For example, if the dimensions of a box are given by $L \pm \Delta L, W \pm \Delta W, H \pm \Delta H$, then the uncertainty in the volume $V = LWH$ is

$$\Delta V = V \sqrt{\frac{\Delta L^2}{L^2} + \frac{\Delta W^2}{W^2} + \frac{\Delta H^2}{H^2}}.$$

The surface area of one L -face is approximated to first order via

$$L^2 + 2L\Delta L = A_L + \Delta A_L.$$

and similar for the W - and H -faces. Thus the uncertainty in the total surface area

$$S = 2(A_L + A_W + A_H)$$

is

$$\Delta S = 4\sqrt{L^2\Delta L^2 + W^2\Delta W^2 + H^2\Delta H^2}.$$

Ratios

As it turns out, the same uncertainty formula applies to the product $P = AB$ and to the ratio $R = A/B$. In detail:

$$\begin{aligned}(\Delta R)^2 &= \frac{1}{B^2} (\Delta A)^2 + \frac{A^2}{B^4} (\Delta B)^2 \\ \frac{(\Delta R)^2}{R^2} &= \frac{(\Delta A)^2}{A^2} + \frac{(\Delta B)^2}{B^2} \\ \frac{\Delta R}{R} &= \frac{\Delta P}{P}\end{aligned}$$

Exponents

For the quantity $F = A^p B$, where p is an exponent with no uncertainty, it's straightforwardly shown that

$$\frac{\Delta F}{F} = \sqrt{\left(\frac{p\Delta A}{A}\right)^2 + \left(\frac{\Delta B}{B}\right)^2}.$$

Electrical Resistors

Two electrical resistors R_1, R_2 can be arranged in series to form an effective resistor

$$R_S = R_1 + R_2,$$

or arranged in parallel such that

$$R_P = \frac{1}{1/R_1 + 1/R_2}.$$

Of course, electrical resistors are always imperfect, thus we should introduce uncertainties $\Delta R_1, \Delta R_2$ to get a sense of $\Delta R_S, \Delta R_P$.

For ΔR_S , we simply write

$$\Delta R_S = \sqrt{\Delta R_1^2 + \Delta R_2^2}$$

without any hard work. For ΔR_P , we eventually find

$$\begin{aligned}\Delta R_P &= \sqrt{\left(\frac{\partial R_P}{\partial R_1}\right)^2 \Delta R_1^2 + \left(\frac{\partial R_P}{\partial R_2}\right)^2 \Delta R_2^2} \\ &= R_P^2 \sqrt{\frac{\Delta R_1^2}{R_1^4} + \frac{\Delta R_2^2}{R_2^4}}.\end{aligned}$$

Chapter 13

Vectors

1 Introduction to Vectors

1.1 Taxonomy of Vectors

Definition

A *vector* is an ordered list of numbers or variables. One example of a vector is the sequence of numbers 2, 5, and 7, which can be written in *vector notation*:

$$\mathbf{V} = \langle 2, 5, 7 \rangle$$

Vector notation requires a *label* for the vector, $\mathbf{V}$ in our example, specially marked by an arrow (). On the right, the so-called *vector literal* is enclosed by left-and right-angle brackets $\langle \rangle$ as shown. Each number in the vector is separated by a comma.

Vector Components

The individual elements in a vector are formally called *components*, and the total number of components is the *dimension* of the vector. The order in which the components of a vector are listed *does* matter. For example, the three-dimensional vector $\mathbf{V} = \langle 2, 5, 7 \rangle$ is completely different from its reversed version $\langle 7, 5, 2 \rangle$.

Component Subscripts

In a vector, any given component is represented using the vector's symbol without the arrow, but including an *index subscript*. For instance, we could represent $\mathbf{V}$ as

$$\mathbf{V} = \langle V_a, V_b, V_c \rangle$$

with $V_a = 2$, $V_b = 5$, $V_c = 7$, but the letters a , b , c could easily have been x , y , z , or perhaps 1, 2, 3. Vector component labels are, after the dust settles, purely for bookkeeping.

1.2 Representing Vectors

Vectors of two dimensions are suited for visualization on the Cartesian plane. Given a vector $\mathbf{V} = \langle V_a, V_b \rangle$, we plot $\mathbf{V}$ with the following recipe:

- Choose any *base point* for the vector on the plane. From the base point:
- Measure V_a units horizontally, measure V_b units vertically.
- Plot the vector *tip point*. Connect base and tip with an arrow.

Plotted in Fig. 13.1 are two *equivalent* representations of the vector $\mathbf{A} = \langle 2, 6 \rangle$. Note that the vector doesn't 'care' about the choice of base point.

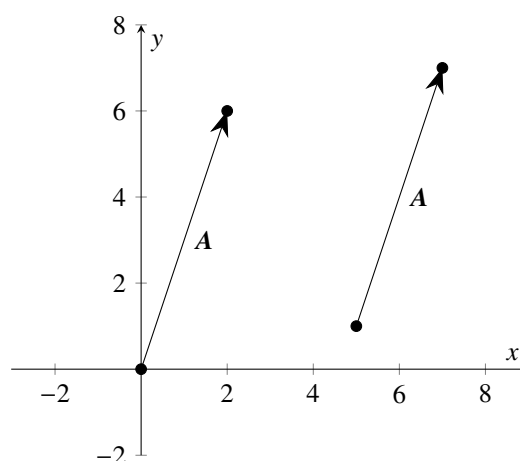


Figure 13.1. Vector $\mathbf{A} = \langle 2, 6 \rangle$ plotted from two different base points $(0, 0)$ and $(5, 1)$.

It should also follow that the above construction extends to dimensions beyond two. For instance, vectors of three dimensions can be visualized in a three-dimensional coordinate system, and so on.

Position Vector

A vector whose base point is the origin $(0, 0)$ is called a *position vector*, often denoted $\mathbf{R}$ or $\mathbf{X}$. A position vector

$$\mathbf{R} = \langle R_x, R_y \rangle$$

is equivalent to the ordered pair (x, y) , denoting a unique point in the Cartesian plane.

Magnitude and Direction

Given the ‘arrow’ representation of a vector, we notice two important features:

- Vectors have a *magnitude*, i.e. the total arrow length.
- Vectors have a *direction*, i.e. a notion of pointing somewhere.

The ‘information’ in a vector is completely represented by its magnitude and its direction. (This may grant some relief as to why we can be so loose about the choice of base point.)

Calculating the Magnitude

A vector $\mathbf{A}$ of dimension N has a magnitude given by

$$A = |\mathbf{A}| = \sqrt{A_1^2 + A_2^2 + \cdots + A_N^2}. \quad (13.1)$$

Intuitively, the magnitude of a vector can be thought of the hypotenuse of an N -dimensional triangle. For the special case $N = 2$, the above reduces to the Pythagorean theorem.

Calculating the Direction

The direction of an N -dimensional vector $\mathbf{A}$ is always implied by the components A_j , but an explicit formula for the ‘angle’ of the vector is only trivial for small N . Working the $N = 2$ case, the direction in which a vector $\mathbf{A} = \langle A_x, A_y \rangle$ is pointing is given by

$$\phi = \arctan \left(\frac{A_y}{A_x} \right) \quad (13.2)$$

To justify (13.2), assume A_x and A_y are two sides of a right triangle such that

$$A_x = A \cos(\phi)$$

$$A_y = A \sin(\phi),$$

and eliminate the magnitude A .

2 Vector Addition

Definition

Two vectors $\mathbf{A}$, $\mathbf{B}$ of equal dimension N can be added by combining like components, resulting in a vector $\mathbf{C}$ with N components:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \quad (13.3)$$

The j th component is given by

$$C_j = A_j + B_j \quad (13.4)$$

$$j = 1, 2, 3, \dots, N.$$

In the special case that the vectors being added are position vectors, also known as *displacement vectors*, the final position is called the *net displacement*.

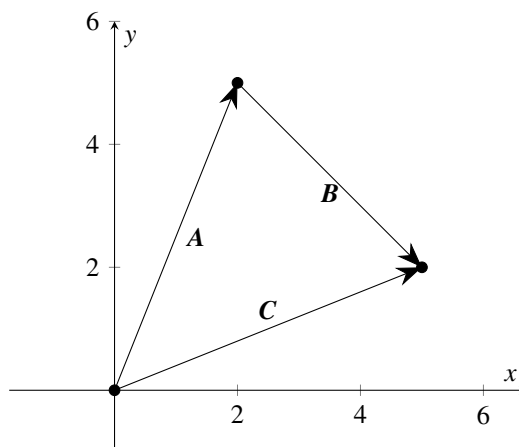


Figure 13.2. Vector addition $\mathbf{C} = \mathbf{A} + \mathbf{B}$.

Commutativity of Vector Addition

Following immediately from (13.3)-(13.4) is the *commutativity of addition*:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \quad (13.5)$$

In particular, (13.5) tells us that the order in which two vectors are added does not affect the result.

Associativity of Vector Addition

The sum of three vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ involves two addition operations. Also following from (13.3)-(13.4) is the *associativity of addition*, telling us that the order of the two addition operations does not effect the result:

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C} \quad (13.6)$$

Arrow Trick

The ‘arrow’ representation of a vector avails a beautiful shortcut for vector addition. Given a pair of two-dimensional vectors $\mathbf{A}$, $\mathbf{B}$, recall that each vector can be drawn *anywhere* in the Cartesian plane. *By arranging the two vectors in tip-to-tail fashion, the vector sum goes from the tail of the first to the tip of the second.*

Fig. 13.2 demonstrates the ‘arrow trick’ on two example vectors $\mathbf{A} = \langle 2, 5 \rangle$, $\mathbf{B} = \langle 3, -3 \rangle$, whose sum easily comes out to $\mathbf{C} = \langle 5, 2 \rangle$. By plotting $\mathbf{A}$, $\mathbf{B}$ as suggested, the sum $\mathbf{C}$ is visually represented by an arrow beginning at the base of $\mathbf{A}$ and ending at the tip of $\mathbf{B}$.

Additive Inverse

Given any vector $\mathbf{A} = \langle A_1, A_2, A_3, \dots, A_N \rangle$, the *additive inverse* is another vector that reverses the sign on all components in $\mathbf{A}$, denoted $-\mathbf{A}$, where

$$-\mathbf{A} = \langle -A_1, -A_2, -A_3, \dots, -A_N \rangle. \quad (13.7)$$

Zero Vector

The so-called *zero vector* is the vector that contains only zeros:

$$\mathbf{0} = \langle 0, 0, 0, \dots, 0 \rangle \quad (13.8)$$

For hopefully obvious reasons, turns out that the sum of any vector and its additive inverse always yields the zero vector:

$$\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$$

In practice, the zero vector is simply written 0, omitting the arrow.

An interesting corollary to the rules of vector addition is that any *closed* sequence of vectors sums to zero. For example, drawing a triangle without lifting the pen from the surface is represented by $\mathbf{A} + \mathbf{B} + \mathbf{C} = \mathbf{0}$.

Vector Subtraction

With the additive inverse established, the notion of *vector subtraction* can be framed in terms of vector addition. Given two vectors $\mathbf{A}$, $\mathbf{B}$, the difference $\mathbf{D} = \mathbf{A} - \mathbf{B}$ can be visualized with the same ‘arrow trick’, so long as we reverse the direction on $\mathbf{B}$ as shown in Fig. 13.3.

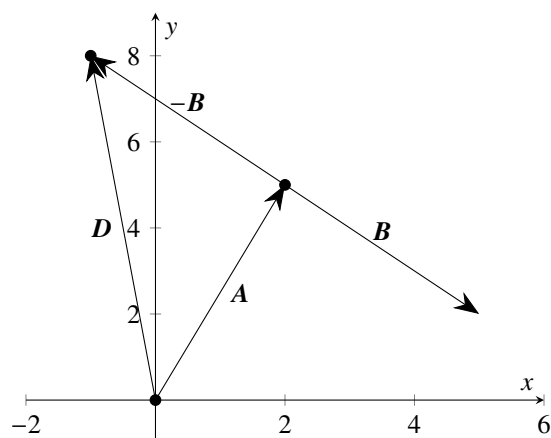


Figure 13.3. Vector subtraction $D = A - B$.

Problem 1

In a flat desert, a man walks east for one hour, northwest for two hours, and then south for two hours. Assuming he walks at constant speed v , what is the man's net displacement vector? Specify the magnitude and direction using polar coordinates.

3 Scalar Multiplication

A vector $\mathbf{A}$ can be ‘scaled’ by a number α called a *scalar*, which has the effect of multiplying the scalar into each component, yielding a new vector $\mathbf{B}$:

$$\mathbf{B} = \alpha \mathbf{A} = \langle \alpha A_1, \alpha A_2, \alpha A_3, \dots, \alpha A_N \rangle \quad (13.9)$$

Parallel Vectors

Two vectors whose components are identical up to a scale factor α are said to be *parallel*. Somewhat like parallel lines, two parallel vectors can have different magnitudes, but point in the same direction. The vectors $\mathbf{A}$, $\mathbf{B}$ in (13.9) are necessarily parallel.

Lines

Straight lines in the Cartesian plane are easily represented with vector addition and scalar multiplication. Consider the slope-intercept form of a line, namely $y = mx + b$, where m is the slope and b is the y-intercept at $(0, b)$. As a vector, the y-intercept can be written

$$\mathbf{b} = \langle 0, b \rangle .$$

Required next is a vector $\mathbf{m}$ that represents the slope of the line, which we capture by writing

$$\begin{aligned} \mathbf{m} &= \langle m_x, m_y \rangle \\ \frac{m_y}{m_x} &= m . \end{aligned}$$

Multiplying $\mathbf{m}$ by any scalar value α will lengthen, shorten, or reverse its effective placement.

Putting the two ingredients together, it follows that any point on the line $y = mx + b$ is equivalently represented as

$$\mathbf{r} = \alpha \mathbf{m} + \mathbf{b} , \quad (13.10)$$

where $\mathbf{r} = \langle x, y \rangle$ is the resulting position vector, as shown in Fig. 13.4. In case (13.10) isn’t convincing, one may resolve $\mathbf{r}$ back into components

$$\begin{aligned} x &= \alpha m_x \\ y &= \alpha m_y + b , \end{aligned}$$

where eliminating α recovers the familiar $y = mx + b$.

Perpendicular Lines

Two lines in the Cartesian plane are perpendicular one line’s slope is m , and the slope is $m_{\perp} = -1/m$. In terms of the components m_x, m_y , this means

$$\begin{aligned} m &= \frac{m_y}{m_x} \\ m_{\perp} &= \frac{-m_x}{m_y} . \end{aligned}$$

From this, the ‘perpendicular slope vector’ $\mathbf{m}_{\perp}$ is evidently $\mathbf{m}_{\perp} = \langle -m_y, m_x \rangle$.

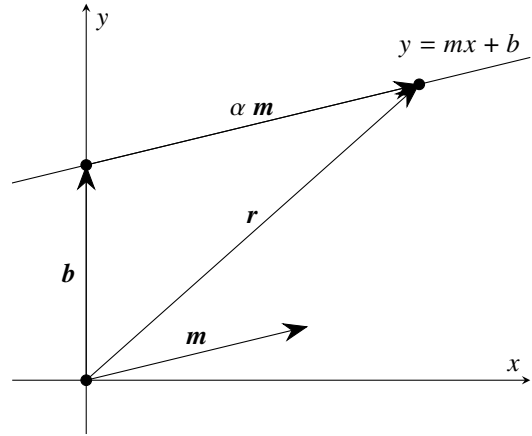


Figure 13.4. Vector construction of a straight line $y = mx + b$.

3.1 Algebraic Properties

Associativity with Scalars

If a vector is modified by two scalars, the order in which they’re applied does not matter:

$$\alpha (\beta \mathbf{A}) = (\alpha \beta) \mathbf{A} = (\beta \alpha) \mathbf{A} = \beta (\alpha \mathbf{A}) \quad (13.11)$$

Distributive Properties

Readily provable from the properties of vector addition and scalar multiplication are the distributive properties involving the sum of two scalars:

$$(\alpha + \beta) \mathbf{A} = \alpha \mathbf{A} + \beta \mathbf{A} \quad (13.12)$$

$$\alpha (\mathbf{A} + \mathbf{B}) = \alpha \mathbf{A} + \alpha \mathbf{B} \quad (13.13)$$

4 Vector Products

4.1 Dot Product

Two vectors of equal dimension can be ‘multiplied’ to form a scalar, called the *dot product*, or the *scalar product*. The dot product is an operation that tells us how much of one vector’s ‘shadow’ falls upon another vector, resulting in a scalar called a *projection*. For two vectors $\mathbf{A}$, $\mathbf{B}$ of dimension N , the dot product reads

$$\mathbf{A} \cdot \mathbf{B} = A_1 B_1 + A_2 B_2 + \cdots + A_N B_N ,$$

or, in summation notation:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{j=1}^N A_j B_j \quad (13.14)$$

Commutativity Relation

Implicit in the definition (13.14) is the commutativity of the dot product (in any number of dimensions):

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} \quad (13.15)$$

Geometric Interpretation of Vectors

The definition (13.14) becomes more intuitive by studying the $N = 2$ case. Consider two arbitrary vectors given by

$$\begin{aligned} \mathbf{A} &= \langle A \cos(\phi_A), A \sin(\phi_A) \rangle \\ \mathbf{B} &= \langle B \cos(\phi_B), B \sin(\phi_B) \rangle , \end{aligned}$$

where A and B are the respective magnitudes. Calculating $\mathbf{A} \cdot \mathbf{B}$ using the formula provided results in

$$\begin{aligned} \mathbf{A} \cdot \mathbf{B} &= AB (\cos(\phi_A) \cos(\phi_B) + \sin(\phi_A) \sin(\phi_B)) \\ \mathbf{A} \cdot \mathbf{B} &= AB \cos(\phi_B - \phi_A) , \end{aligned}$$

telling us that *the dot product is equal to the product of the magnitudes and the cosine of the angle between the vectors*. In general, this result reads

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{AB} , \quad (13.16)$$

where θ is the angle between the vectors in any number of dimensions. The special case $N = 2$ corresponds to $\theta = \phi_B - \phi_A$.

Vector Orthogonality

From the two-dimensional dot product, note that the case $\phi_A - \phi_B = \pm\pi/2$ returns $\cos(\pm\pi/2) = 0$ on the left, telling us that *the dot product between perpendicular vectors is zero*. The formal term for ‘perpendicular’ is *orthogonal*, and this notion generalizes to N dimensions:

$$\mathbf{A} \cdot \mathbf{B} = 0 \quad (13.17)$$

In the Cartesian plane, recall that the slope of a line and another perpendicular line are represented by the vectors

$$\begin{aligned} \mathbf{m} &= \langle m_x, m_y \rangle \\ \mathbf{m}_\perp &= \langle -m_y, m_x \rangle , \end{aligned}$$

respectively. We verify these vectors to be orthogonal by calculating

$$\mathbf{m} \cdot \mathbf{m}_\perp = -m_x m_y + m_y m_x = 0 .$$

Vector Magnitude

The dot product is responsible for the formula (13.1) for calculating the magnitude of a vector. Indeed, for an N -dimensional vector $\mathbf{A}$, we find the dot product with itself to be

$$\mathbf{A} \cdot \mathbf{A} = A_1^2 + A_2^2 + A_3^2 + \cdots + A_N^2 ,$$

which is the square of the magnitude of $\mathbf{A}$. More concisely:

$$A = |\mathbf{A}| = \sqrt{\mathbf{A} \cdot \mathbf{A}} \quad (13.18)$$

Distributive Property

For three vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ of equal dimension, the dot product obeys the distributive property as one may expect:

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C} \quad (13.19)$$

Law of Cosines

An important relation from trigonometry called the *law of cosines* is derived using dot products. Consider the vector sum

$$\mathbf{A} - \mathbf{B} = \mathbf{C} ,$$

and then square both sides:

$$\begin{aligned} (\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B}) &= \mathbf{C} \cdot \mathbf{C} \\ \mathbf{A} \cdot \mathbf{A} + \mathbf{B} \cdot \mathbf{B} - 2\mathbf{A} \cdot \mathbf{B} &= \mathbf{C} \cdot \mathbf{C} \end{aligned}$$

Labeling θ as the angle between vectors $\mathbf{A}$, $\mathbf{B}$, the above simplifies to the law of cosines:

$$A^2 + B^2 - 2AB \cos(\theta) = C^2 \quad (13.20)$$

Note that all right triangles have $\theta = \pi/2$, in which case (13.20) reduces to the Pythagorean theorem.

4.2 Cross Product

Two vectors of equal dimension can be ‘multiplied’ to form a new vector, called the *cross product*, or the *vector product*. The cross product is, for most purposes, a strictly three-dimensional operation. Consider the pair of vectors with $N = 3$:

$$\begin{aligned} \mathbf{A} &= \langle A_x, A_y, A_z \rangle \\ \mathbf{B} &= \langle B_x, B_y, B_z \rangle \end{aligned}$$

The cross product $\mathbf{A} \times \mathbf{B}$ is defined as

$$\mathbf{A} \times \mathbf{B} = \langle C_x, C_y, C_z \rangle, \quad (13.21)$$

where

$$\begin{aligned} C_x &= A_y B_z - A_z B_y \\ C_y &= A_z B_x - A_x B_z \\ C_z &= A_x B_y - A_y B_x, \end{aligned}$$

and is *orthogonal* to both $\mathbf{A}$ and $\mathbf{B}$.

Determinant Notation

The cross product formula (13.21) is tricky to memorize, and can be more transparently represented as a ‘block of numbers’ (*not* a matrix), sometimes called *determinant notation*:

$$\mathbf{A} \times \mathbf{B} = \begin{vmatrix} (x) & (-y) & (z) \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}$$

Without sweating the details of determinant notation, you can play a matching game between the determinant representation of $\mathbf{A} \times \mathbf{B}$ and the formula (13.21) to remember how it goes.

Orthogonality Check

To ensure that $\mathbf{A} \times \mathbf{B}$ is mutually orthogonal to $\mathbf{A}$,

$$\mathbf{A} \cdot (\mathbf{A} \times \mathbf{B})$$

to see what comes out. In detail, the former case proceeds as

$$\begin{aligned} \mathbf{A} \cdot (\mathbf{A} \times \mathbf{B}) &= A_x A_y B_z - A_x A_z B_y + A_y A_z B_x \\ &\quad - A_y A_x B_z + A_z A_x B_y - A_z A_y B_x \\ &= B_z (A_x A_y - A_y A_x) - B_y (A_x A_z - A_z A_x) \\ &\quad + B_z (A_x A_y - A_y A_x) \\ &= 0. \end{aligned}$$

This also holds true for the $\mathbf{B}$ -case.

Null Case

The cross product of a vector with itself is identically zero:

$$\mathbf{A} \times \mathbf{A} = 0 \quad (13.22)$$

Anti-Commutativity Relation

Given the definition (13.21) of the cross product, one sees that swapping $\mathbf{A}$, $\mathbf{B}$ puts a minus sign on the result. This is known as the *anti-commutativity* of the cross product:

$$\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A} \quad (13.23)$$

Right Hand Rule

There is a trick that allows one to know the direction of $\mathbf{A} \times \mathbf{B}$ known as the (oft-dreaded) *right hand rule*. To know the direction of the vector $\mathbf{A} \times \mathbf{B}$, the steps are as follows:

1. On your right hand: point your thumb, index finger, and middle finger out in perpendicular directions.
2. Let your index finger be vector $\mathbf{A}$, let your middle finger be vector $\mathbf{B}$.
3. Your thumb points along vector $\mathbf{A} \times \mathbf{B}$.

Geometric Interpretation

The definition (13.21) becomes more intuitive by studying a special case. Consider the pair of three-dimensional vectors confined to the xy -plane given by

$$\begin{aligned} \mathbf{A} &= \langle A \cos(\phi_A), A \sin(\phi_A), 0 \rangle \\ \mathbf{B} &= \langle B \cos(\phi_B), B \sin(\phi_B), 0 \rangle, \end{aligned}$$

where A and B are the respective magnitudes. Calculating $\mathbf{A} \times \mathbf{B}$ using the formula provided results in

$$\begin{aligned} \mathbf{A} \times \mathbf{B} &= \langle 0, 0, AB(\cos \phi_A \sin \phi_B - \cos \phi_B \sin \phi_A) \rangle \\ \mathbf{A} \times \mathbf{B} &= \langle 0, 0, AB \sin(\phi_B - \phi_A) \rangle, \end{aligned}$$

telling us that *the cross product is equal to the product of the magnitudes and the sine of the angle between the vectors*. In general, this result also tells us

$$\sin(\theta) = \frac{|\mathbf{A} \times \mathbf{B}|}{AB}, \quad (13.24)$$

where θ is the angle between the vectors at any relative orientation.

Area of a Parallelogram

The quantity $AB \sin(\theta)$ can be interpreted as the area of a parallelogram having base B and height $h = A \sin \phi$. For the right-angle case $\phi = \pi/2$, the parallelogram becomes a rectangle of area AB . In the language of vectors, the product $|\mathbf{A} \times \mathbf{B}|$ is the area of the parallelogram with sides A, B .

Area of a Triangle

A parallelogram sliced along the diagonal forms two triangles of the same area. The area of one such triangle is

$$\text{Area} = \frac{1}{2} |\mathbf{A} \times \mathbf{B}|.$$

Thales' Theorem

From trigonometry, Thales' theorem states that the angle ACB in Figure 13.5 is a right angle. This is readily proven using vectors and the cross product.

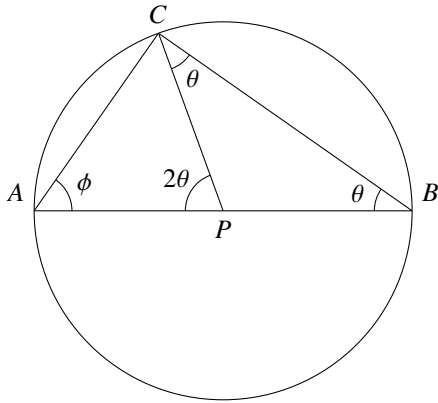


Figure 13.5. Inscribed triangle.

Placing the origin at point A , two edges of the triangle are given by the vectors:

$$\begin{aligned} \overrightarrow{AC} &= \langle \overline{AC} \cos(\phi), \overline{AC} \sin(\phi) \rangle \\ \overrightarrow{BC} &= \langle -\overline{BC} \cos(\theta), \overline{BC} \sin(\theta) \rangle \end{aligned}$$

The cross product between vectors AC and BC gives the triangular area after simplifying:

$$\text{Area} = \frac{1}{2} (\overline{AC}) (\overline{BC}) \sin(\theta + \phi)$$

Meanwhile, the same area is written

$$\text{Area} = \frac{1}{2} (\overline{AC}) (\overline{BC}) \sin(\angle ACB),$$

thus we find

$$\angle ACB = \theta + \phi.$$

Since the sum of interior angles must equal π , we can further write

$$\theta + \phi + \angle ACB = \pi,$$

simplifying to

$$\angle ACB = \frac{\pi}{2}$$

and the proof is done.

Problem 2

In Figure 13.5, show that the dot product between vector AC and vector BC is zero.

Problem 3

For three points in the plane, show that the distance from a point P to a line through the other two points A, B is:

$$q = \frac{|(\mathbf{A} - \mathbf{B}) \times (\mathbf{P} - \mathbf{B})|}{|\mathbf{A} - \mathbf{B}|}$$

4.3 Vector Identities

Consider three vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$, each of three dimensions.

Triple Product

The scalar quantity

$$V = \mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) \quad (13.25)$$

is called the *triple product*. Intuitively, the triple product describes the volume of the parallel-piped with sides $\mathbf{A}, \mathbf{B}, \mathbf{C}$. One can show by brute force that (13.25) obeys the cyclic relations:

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B})$$

BAC-CAB Formula

A useful equation known as the *BAC-CAB* identity, reads

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} (\mathbf{A} \cdot \mathbf{C}) - \mathbf{C} (\mathbf{A} \cdot \mathbf{B}) . \quad (13.26)$$

The proof of (13.26) is slightly long but straightforward, using (optional) determinant notation to contain the cross product:

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \begin{vmatrix} (x) & (-y) & (z) \\ A_x & A_y & A_z \\ B_y C_z - B_z C_y & B_z C_x - B_x C_z & B_x C_y - B_y C_x \end{vmatrix}$$

$$\begin{aligned} \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) &= \langle A_y B_x C_y - A_y B_y C_x - A_z B_z C_x + A_z B_x C_z, 0, 0 \rangle + \\ &\quad \langle 0, A_z B_y C_z - A_z B_z C_y - A_x B_x C_y + A_x B_y C_x, 0 \rangle + \\ &\quad \langle 0, 0, A_x B_z C_x - A_x B_x C_z - A_y B_y C_z + A_y B_z C_y \rangle \\ &= B_x \langle A_y C_y + A_z C_z, 0, 0 \rangle - A_x B_x \langle 0, C_y, C_z \rangle + \\ &\quad B_y \langle 0, A_z C_z + A_x C_x, 0 \rangle - A_y B_y \langle C_x, 0, C_z \rangle + \\ &\quad B_z \langle 0, 0, A_x C_x + A_y C_y \rangle - A_z B_z \langle C_x, C_y, 0 \rangle \end{aligned}$$

$$\begin{aligned} \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) &= B_x \langle \mathbf{A} \cdot \mathbf{C}, 0, 0 \rangle - A_x B_x \langle C_x, C_y, C_z \rangle + \\ &\quad B_y \langle 0, \mathbf{A} \cdot \mathbf{C}, 0 \rangle - A_y B_y \langle C_x, C_y, C_z \rangle + \\ &\quad B_z \langle 0, 0, \mathbf{A} \cdot \mathbf{C} \rangle - A_z B_z \langle C_x, C_y, C_z \rangle \end{aligned}$$

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B} (\mathbf{A} \cdot \mathbf{C}) - \mathbf{C} (\mathbf{A} \cdot \mathbf{B})$$

5 Basis Vectors

5.1 Unit Vectors

Consider a vector V of dimension N , having magnitude V . A special vector $\hat{V}$, called a *unit vector*, is defined as V divided by its own magnitude:

$$\hat{V} = \frac{1}{V} V \quad (13.27)$$

That is, a unit vector always has magnitude one, and points along the original vector. A vector of the form (13.27) is said to be *normalized*.

A more intuitive way to understand unit vectors is to rearrange (13.27) to write

$$V = V \hat{V},$$

which says that a full vector V is the product of the magnitude V and the ‘direction’ unit vector $\hat{V}$.

Problem 4

What is the vector that bisects the angle between two vectors U , V ?

5.2 Basis Vectors

Consider an arbitrary vector V of dimension N . A curious way to express

$$V = \langle V_1, V_2, V_3, \dots, V_N \rangle$$

is to fully pull apart each component so that V is the sum of N pure sub-vectors:

$$\begin{aligned} V = & \langle V_1, 0, 0, \dots \rangle \\ & + \langle 0, V_2, 0, \dots \rangle + \langle \dots, 0, V_3, 0, \dots \rangle \\ & + \dots + \langle \dots, 0, V_N \rangle \end{aligned}$$

Each sub-vector contains just one component V_j , which can be factored out of the sub-vector as a scalar. The sub-vectors that remain are called *basis vectors*, denoted $\hat{e}_j$.

$$\begin{aligned} \hat{e}_1 &= \langle 1, 0, 0, \dots, 0 \rangle \\ \hat{e}_2 &= \langle 0, 1, 0, \dots, 0 \rangle \\ \hat{e}_3 &= \langle 0, 0, 1, \dots, 0 \rangle \\ &\dots \\ \hat{e}_N &= \langle 0, 0, 0, \dots, 1 \rangle \end{aligned} \quad (13.28)$$

There is one basis vector $\hat{e}_j$ for each of the N dimensions in which the vector is situated.

Orthogonality of Basis Vectors

Basis vectors are all mutually orthogonal by necessity. For two different basis vectors $\hat{e}_j$, $\hat{e}_k$, the *orthogonality relation* is

$$\hat{e}_j \cdot \hat{e}_k = 0. \quad (13.29)$$

On the other hand, two of the same basis vector $\hat{e}_k$ obeys

$$\hat{e}_k \cdot \hat{e}_k = 1. \quad (13.30)$$

In the general case, any set of basis vectors $\{\hat{e}_j\}$ that obeys (13.29), (13.30) is said to be *orthonormal*.

Cartesian Position Vector

In the Cartesian xy -plane, a *position vector* is typically denoted $R = \langle x, y \rangle$, suggesting two basis vectors

$$\begin{aligned} \hat{e}_x &= \langle 1, 0 \rangle = \hat{x} \\ \hat{e}_y &= \langle 0, 1 \rangle = \hat{y}, \end{aligned}$$

and furthermore

$$R = x \hat{x} + y \hat{y}.$$

Note that the same notation extrapolates to three dimensions, in which case

$$\begin{aligned} \hat{e}_x &= \langle 1, 0, 0 \rangle = \hat{x} \\ \hat{e}_y &= \langle 0, 1, 0 \rangle = \hat{y} \\ \hat{e}_z &= \langle 0, 0, 1 \rangle = \hat{z} \end{aligned}$$

are the basis vectors, implying

$$R = x \hat{x} + y \hat{y} + z \hat{z}.$$

Cartesian Basis Vectors

The Cartesian basis vectors $\hat{x}$, $\hat{y}$, $\hat{z}$ obey orthogonality relations regardless of the number of dimensions

$$\begin{aligned} \hat{x} \cdot \hat{y} &= 0 \\ \hat{y} \cdot \hat{z} &= 0 \\ \hat{z} \cdot \hat{x} &= 0, \end{aligned}$$

along with cyclic relations strictly applying to three dimensions:

$$\begin{aligned} \hat{x} \times \hat{y} &= \hat{z} \\ \hat{y} \times \hat{z} &= \hat{x} \\ \hat{z} \times \hat{x} &= \hat{y} \end{aligned}$$

5.3 Linear Combinations

Having established the notion of basis vectors, we are free to express arbitrary vector $\mathbf{V}$ as a *linear combination* of each $\hat{\mathbf{e}}_j$, namely

$$\mathbf{V} = V_1 \hat{\mathbf{e}}_1 + V_2 \hat{\mathbf{e}}_2 + V_3 \hat{\mathbf{e}}_3 + \cdots + V_N \hat{\mathbf{e}}_N, \quad (13.31)$$

or in summation notation:

$$\mathbf{V} = \sum_{j=1}^N V_j \hat{\mathbf{e}}_j \quad (13.32)$$

In the above, $\mathbf{V}$ can potentially point to any ‘place’ in the N -dimensional space in which it lives. Such a place is formally called a *vector space*.

Component Isolation

One may ‘solve’ for the V_j th component in a vector $\mathbf{V}$ by exploiting the orthogonality relations (13.29), (13.30). Start with (13.31), and multiply any particular $\hat{\mathbf{e}}_k$ into both sides:

$$\hat{\mathbf{e}}_k \cdot \mathbf{V} = V_1 \hat{\mathbf{e}}_k \cdot \hat{\mathbf{e}}_1 + V_2 \hat{\mathbf{e}}_k \cdot \hat{\mathbf{e}}_2 + \cdots + V_N \hat{\mathbf{e}}_k \cdot \hat{\mathbf{e}}_N$$

Next, observe that *all except one* of the dot products on the right will cancel due to (13.30). The whole sum collapses to the term with $j = k$, namely $V_k \hat{\mathbf{e}}_k \cdot \hat{\mathbf{e}}_k$, simplifying to V_k . Formally, we have uncovered the obvious yet satisfying statement:

$$V_k = \mathbf{V} \cdot \hat{\mathbf{e}}_k \quad (13.33)$$

With an explicit formula for the V_j th component of a vector, it’s curious to see happens by replacing V_j in (13.32). Carrying this out, we can write a component-free way to reference a vector and its contents:

$$\mathbf{V} = \sum_{j=1}^N (\mathbf{V} \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j \quad (13.34)$$

Spanning the Vector Space

It’s important to notice that a linear combination $\mathbf{V}$, with appropriate values of V_j , could represent *any* point in the N -dimensional *vector space* in which the vector is embedded. This is possible because the set of basis vectors $\{\hat{\mathbf{e}}_j\}$ are said to *span* the vector space.

6 Polar Representation

In the Cartesian plane, consider a position vector

$$\mathbf{r} = \langle r_x, r_y \rangle .$$

The ‘magnitude-and-direction’ interpretation of $\mathbf{r}$ assigns the magnitude r to the hypotenuse of a right triangle, where the adjacent and opposite sides are respectively given by

$$r_x = r \cos(\phi) \quad (13.35)$$

$$r_y = r \sin(\phi) , \quad (13.36)$$

congruent with equations (13.1)-(13.2). The angle parameter ϕ is also known as the *phase* of the vector, a dimensionless argument unique on the interval $[0 : 2\pi)$.

Polar Coordinate System

Equations (13.35)-(13.36) represent a mapping from system of Cartesian coordinates to the system of polar coordinates. Any point in the plane that can be represented by the ordered pair (x, y) has an equivalent representation as the ordered pair (r, ϕ) .

Traditionally the symbol θ is used in place of ϕ in a two-dimensional polar coordinate system, thus we take the position vector to be

$$\begin{aligned} \mathbf{r} &= \langle r \cos(\theta), r \sin(\theta) \rangle \\ &= r \langle \cos(\theta), \sin(\theta) \rangle \end{aligned}$$

Polar Position Vector

Starting from the Cartesian position vector

$$\mathbf{R} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} ,$$

replace x, y with their equivalent representations in polar coordinates:

$$\mathbf{R} = r \cos(\theta) \hat{\mathbf{x}} + r \sin(\theta) \hat{\mathbf{y}}$$

From the above we can factor the radial term r , leaving a tangle of basis vectors and trig terms:

$$\mathbf{R} = r (\cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}})$$

Meanwhile, notice that the magnitude of

$$R = \sqrt{\mathbf{R} \cdot \mathbf{R}}$$

reduces to $R = r$. This simultaneously means that $\mathbf{R}$ ’s unit vector, i.e.

$$\hat{\mathbf{R}} = \frac{\mathbf{R}}{R}$$

is equal to

$$\hat{\mathbf{R}} = \cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}} .$$

As a matter of custom, the polar position vector is expressed in lowercase, i.e. $\mathbf{r}$. This is also true for the polar basis vector $\hat{\mathbf{r}}$. The tightest way to write the position vector in polar coordinates is:

$$\mathbf{r} = r \hat{\mathbf{r}}$$

6.1 Polar Basis Vectors

Starkly different from the fixed vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}$, the polar basis vector $\hat{\mathbf{r}}$ is not fixed in the plane, and instead varies with θ via

$$\hat{\mathbf{r}} = \cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}} .$$

In the same way that $\hat{\mathbf{r}}$ points in the direction of increasing r , there ought to exist a second basis vector $\hat{\boldsymbol{\theta}}$, always perpendicular to $\hat{\mathbf{r}}$, that ‘points’ in the direction of increasing θ . This can be attained by rotating $\hat{\mathbf{r}}$ by ninety degrees

$$\hat{\boldsymbol{\theta}} = \cos\left(\theta + \frac{\pi}{2}\right) \hat{\mathbf{x}} + \sin\left(\theta + \frac{\pi}{2}\right) \hat{\mathbf{y}} ,$$

simplifying to

$$\hat{\boldsymbol{\theta}} = -\sin(\theta) \hat{\mathbf{x}} + \cos(\theta) \hat{\mathbf{y}} .$$

Notice that the θ -basis vector is not needed to write the position $\mathbf{r}$.

Problem 5

Define a vector $\hat{\mathbf{z}} = \hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}$ that points in a perpendicular direction to the r - θ plane. Show explicitly that:

$$\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}} = \hat{\mathbf{z}}$$

$$\hat{\boldsymbol{\theta}} \times \hat{\mathbf{z}} = \hat{\mathbf{r}}$$

$$\hat{\mathbf{z}} \times \hat{\mathbf{r}} = \hat{\boldsymbol{\theta}}$$

7 Change of Basis

Consider a two-dimensional vector $V = \langle V_x, V_y \rangle$, naturally expressed as a linear combination in the Cartesian basis

$$\begin{aligned}\hat{e}_1 = \hat{x} &= \langle 1, 0 \rangle \\ \hat{e}_2 = \hat{y} &= \langle 0, 1 \rangle.\end{aligned}$$

By convention, the Cartesian coordinate system is usually aligned with the edges of a rectangular sheet of paper or computer screen. The orientation of the coordinate system is of course arbitrary, and we must be free to *rotate* the basis vectors without ‘physical’ consequences.

Figure 13.6 shows a two-dimensional example with two sets of basis vectors $\{\hat{e}_j\}$, $\{\hat{u}_j\}$ embedded on the Cartesian plane. In particular, the basis vector $\hat{u}$ is rotated up from $\hat{x}$ by some arbitrary angle, and similarly $\hat{v}$ corresponds to $\hat{y}$ by the same angle. Any given linear combination r has a different representation in each basis.

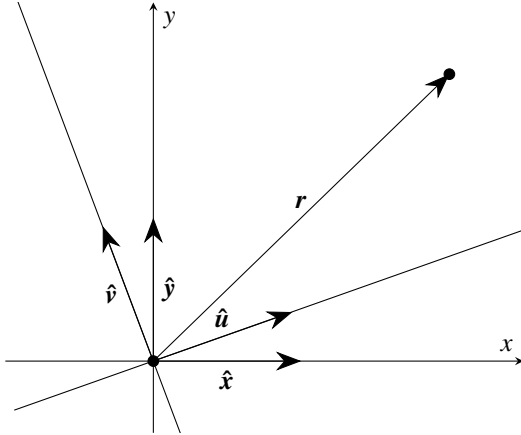


Figure 13.6. Vector r as a linear combination in two different bases.

Generalizing this idea to N dimensions, we can say that linear combinations of the form (13.32) can be re-expressed in terms of a different set of orthonormal basis vectors $\{\hat{u}_j\}$:

$$(V)' = \sum_{j=1}^N V'_j \hat{u}_j \quad (13.37)$$

Analogous to (13.33), the primed components relate to the vector by:

$$V'_k = (V)' \cdot \hat{u}_k \quad (13.38)$$

7.1 Two Dimensions

Suppose a different set basis vectors $\hat{u}, \hat{v}$ is given in terms of the original $\{\hat{e}_j\}$ basis, for instance

$$\begin{aligned}\hat{u}_1 = \hat{u} &= \left\langle \frac{\sqrt{3}}{2}, \frac{1}{2} \right\rangle \\ \hat{u}_2 = \hat{v} &= \left\langle \frac{-1}{2}, \frac{\sqrt{3}}{2} \right\rangle,\end{aligned}$$

or equivalently,

$$\begin{aligned}\hat{u}_1 &= \frac{\sqrt{3}}{2} \hat{e}_1 + \frac{1}{2} \hat{e}_2 \\ \hat{u}_2 &= -\frac{1}{2} \hat{e}_1 + \frac{\sqrt{3}}{2} \hat{e}_2.\end{aligned}$$

Note that each $\hat{u}_j$ is a linear combination of each $\hat{e}_j$. The coefficients $\sqrt{3}/2, 1/2$, etc. are carefully chosen to assure orthonormality between $\hat{u}_{1,2}$.

If a vector r is expressed in the $\{\hat{e}_j\}$ basis as the linear combination

$$r = r_1 \hat{e}_1 + r_2 \hat{e}_2,$$

the so-called ‘change of basis’ occurs if we algebraically replace all $\hat{e}_j$ with $\hat{u}_j$, which first requires inverting the above relations:

$$\begin{aligned}\hat{e}_1 &= \frac{\sqrt{3}}{2} \hat{u}_1 - \frac{1}{2} \hat{u}_2 \\ \hat{e}_2 &= \frac{1}{2} \hat{u}_1 + \frac{\sqrt{3}}{2} \hat{u}_2\end{aligned}$$

Then, the vector r can be written

$$\begin{aligned}(r)' &= r_1 \left(\frac{\sqrt{3}}{2} \hat{u}_1 - \frac{1}{2} \hat{u}_2 \right) + r_2 \left(\frac{1}{2} \hat{u}_1 + \frac{\sqrt{3}}{2} \hat{u}_2 \right) \\ (r)' &= \left(r_1 \frac{\sqrt{3}}{2} + r_2 \frac{1}{2} \right) \hat{u}_1 + \left(-r_1 \frac{1}{2} + r_2 \frac{\sqrt{3}}{2} \right) \hat{u}_2,\end{aligned}$$

where the components $r_{1,2}$ are finally readable as

$$\begin{aligned}r'_1 &= r_1 \frac{\sqrt{3}}{2} + r_2 \frac{1}{2} \\ r'_2 &= -r_1 \frac{1}{2} + r_2 \frac{\sqrt{3}}{2},\end{aligned}$$

and a form like (13.37) is attained:

$$(r)' = r'_1 \hat{u}_1 + r'_2 \hat{u}_2$$

General Case

The above example can be easily generalized such that $\hat{\mathbf{u}}$ points anywhere in the Cartesian plane, with $\hat{\mathbf{v}}$ appropriately perpendicular to $\hat{\mathbf{u}}$. To achieve this, we introduce an arbitrary parameter θ such that

$$\begin{aligned}\hat{\mathbf{u}}_1 &= \cos(\theta) \hat{\mathbf{e}}_1 + \sin(\theta) \hat{\mathbf{e}}_2 \\ \hat{\mathbf{u}}_2 &= -\sin(\theta) \hat{\mathbf{e}}_1 + \cos(\theta) \hat{\mathbf{e}}_2.\end{aligned}$$

By straightforward algebra, we find the inverted version to be

$$\begin{aligned}\hat{\mathbf{e}}_1 &= \cos(\theta) \hat{\mathbf{u}}_1 - \sin(\theta) \hat{\mathbf{u}}_2 \\ \hat{\mathbf{e}}_2 &= \sin(\theta) \hat{\mathbf{u}}_1 + \cos(\theta) \hat{\mathbf{u}}_2.\end{aligned}$$

Of course, this generalization is simply a matter of switching to polar coordinates. Notice that $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ are identical in form to $\hat{\mathbf{r}}$, $\hat{\theta}$, respectively.

7.2 Matrix Notation

The pairs of equations above are suggestive of a matrix formulation, particularly

$$\begin{bmatrix} \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}_1 \\ \hat{\mathbf{e}}_2 \end{bmatrix}, \quad (13.39)$$

and

$$\begin{bmatrix} \hat{\mathbf{e}}_1 \\ \hat{\mathbf{e}}_2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix}, \quad (13.40)$$

respectively.

Rotation Matrix

In tighter notation, we further write

$$\hat{\mathbf{u}}_j = \sum_{k=1}^2 \tilde{R}_{jk} \hat{\mathbf{e}}_k \quad (13.41)$$

$$\hat{\mathbf{e}}_j = \sum_{k=1}^2 R_{jk} \hat{\mathbf{u}}_k, \quad (13.42)$$

where R_{jk} is known as the *rotation matrix*:

$$R_{jk} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

With a convenient representation for the basis vectors $\{\hat{\mathbf{e}}_j\}$, an arbitrary linear combination

$$\mathbf{r} = r_1 \hat{\mathbf{e}}_1 + r_2 \hat{\mathbf{e}}_2$$

becomes

$$\begin{aligned}(\mathbf{r})' &= r_1 \sum_{k=1}^2 R_{1k} \hat{\mathbf{u}}_k + r_2 \sum_{k=1}^2 R_{2k} \hat{\mathbf{u}}_k \\ &= \sum_{k=1}^2 \left(\sum_{j=1}^2 R_{jk} r_j \right) \hat{\mathbf{u}}_k,\end{aligned}$$

telling us that the k th component of the vector $(\mathbf{r})'$ is given by

$$r'_k = \sum_{j=1}^2 R_{jk} r_j. \quad (13.43)$$

Problem 6

Check that:

$$\begin{bmatrix} \hat{\mathbf{r}} \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{r}} \\ \hat{\theta} \end{bmatrix}$$

7.3 N Dimensions

At the center of the change-of-basis problem is the issue of relating the two orthonormal bases $\hat{\mathbf{e}}_j$, $\hat{\mathbf{u}}_j$ to one another. In N dimensions, the basis vectors are related by linear combinations

$$\hat{\mathbf{u}}_j = \sum_{k=1}^N \tilde{U}_{jk} \hat{\mathbf{e}}_k \quad (13.44)$$

$$\hat{\mathbf{e}}_j = \sum_{k=1}^N U_{jk} \hat{\mathbf{u}}_k. \quad (13.45)$$

Having two subscripts, the terms $\tilde{U}_{jk}$, U_{jk} are not vector components, but instead *matrix* components. These typically end up being coefficients like $\sqrt{3}/2$, $1/2$, and so on.

To isolate the matrix components U_{jk} and $\tilde{U}_{jk}$, multiply (via dot product) the basis vectors $\hat{\mathbf{e}}_m$, $\hat{\mathbf{u}}_m$, respectively into (13.44), (13.45):

$$\hat{\mathbf{e}}_m \cdot \hat{\mathbf{u}}_j = \sum_{k=1}^N \tilde{U}_{jk} \hat{\mathbf{e}}_m \cdot \hat{\mathbf{e}}_k$$

$$\hat{\mathbf{u}}_m \cdot \hat{\mathbf{e}}_j = \sum_{k=1}^N U_{jk} \hat{\mathbf{u}}_m \cdot \hat{\mathbf{u}}_k$$

Due to orthonormality, the right side of each equation resolves to zero except for the case with $m = k$, allowing the components to be isolated:

$$\tilde{U}_{jm} = \hat{\mathbf{e}}_m \cdot \hat{\mathbf{u}}_j = \hat{\mathbf{u}}_j \cdot \hat{\mathbf{e}}_m \quad (13.46)$$

$$U_{jm} = \hat{\mathbf{u}}_m \cdot \hat{\mathbf{e}}_j = \hat{\mathbf{e}}_j \cdot \hat{\mathbf{u}}_m \quad (13.47)$$

From this, deduce also that

$$\tilde{U}_{jm} = U_{mj} . \quad (13.48)$$

Linear Combinations

An arbitrary linear combination

$$\mathbf{V} = \sum_{j=1}^N V_j \hat{\mathbf{e}}_j$$

can be written with all $\hat{\mathbf{e}}_j$ replaced according to (13.45):

$$(\mathbf{V})' = \sum_{j=1}^N \sum_{k=1}^N U_{jk} V_j \hat{\mathbf{u}}_k = \sum_{k=1}^N \left(\sum_{j=1}^N U_{jk} V_j \right) \hat{\mathbf{u}}_k$$

Comparing the above to (13.37) gives a formula for the k th component of the vector $(\mathbf{V})'$:

$$(V')_k = \sum_{j=1}^N U_{jk} V_j \quad (13.49)$$

Unity Condition

We can learn a bit more about the matrix components U_{jk} , $\tilde{U}_{jk}$ by eliminating $\hat{\mathbf{u}}_k$ between (13.44)-(13.45):

$$\begin{aligned} \hat{\mathbf{e}}_j &= \sum_{k=1}^N U_{jk} \left(\sum_{m=1}^N \tilde{U}_{km} \hat{\mathbf{e}}_m \right) \\ &= \sum_{k=1}^N \sum_{m=1}^N (U_{jk} \tilde{U}_{km}) \hat{\mathbf{e}}_m \end{aligned}$$

Next, use the symbol I to represent the quantity

$$I_{jm} = \sum_{k=1}^N (U_{jk} \tilde{U}_{km}) ,$$

and the above becomes

$$\hat{\mathbf{e}}_j = \sum_{m=1}^N I_{jm} \hat{\mathbf{e}}_m .$$

For the above to make sense, all terms in the right-hand sum must vanish except for that with $m = j$. Explicitly, this means I_{jm} obeys

$$I_{jm} = \begin{cases} 1 & m = j \\ 0 & m \neq j \end{cases} ,$$

reminiscent of (13.29)-(13.30).

8 Rotations

8.1 Rotated Vectors

Starting with a vector $\mathbf{r}$ in two dimensions, particularly

$$\mathbf{r} = \langle r \cos(\phi), r \sin(\phi) \rangle,$$

we may inquire what happens when we modify the phase such that $\phi \rightarrow \phi + \theta$, effectively rotating the vector in the plane.

Carrying this out, one writes

$$(\mathbf{r})' = r \langle \cos(\phi + \theta), \sin(\phi + \theta) \rangle,$$

or, expanding the trigonometry terms,

$$\begin{aligned} r'_x &= r (\cos(\phi) \cos(\theta) - \sin(\phi) \sin(\theta)) \\ r'_y &= r (\sin(\phi) \cos(\theta) + \cos(\phi) \sin(\theta)). \end{aligned}$$

Things get interesting when we keep simplifying

$$\begin{aligned} r'_x &= r_x \cos(\theta) - r_y \sin(\theta) \\ r'_y &= r_x \sin(\theta) + r_y \cos(\theta), \end{aligned}$$

allowing a tight expression

$$(r')_k = \sum_{j=1}^N R_{jk} r_j,$$

identical to Equation (13.43).

8.2 Rotated Coordinates

Consider a vector $\mathbf{A}$ living in the ‘standard’ Cartesian basis:

$$\mathbf{A} = A_x \hat{\mathbf{x}} + A_y \hat{\mathbf{y}}$$

While leaving the vector unchanged, let us rotate the coordinate system by some angle θ so that the uv -plane replaces the xy -plane, with $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ replacing the respective $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ unit vectors.

In this construction, the unit vectors $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ are totally analogous to $\hat{\mathbf{r}}, \hat{\theta}$, and we can borrow the matrix that brings us from $\hat{\mathbf{x}}, \hat{\mathbf{y}}$ to the rotated system:

$$\begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix}$$

In terms of the rotated basis vectors, the original basis vectors are the inversion of the above:

$$\begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}} \\ \hat{\mathbf{v}} \end{bmatrix}$$

Inverse Rotation Matrix

While we’re here, it’s worth noting that the inverse of R , namely

$$\tilde{R} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \tilde{R}_{11} & \tilde{R}_{12} \\ \tilde{R}_{21} & \tilde{R}_{22} \end{bmatrix}$$

is the inverse rotation matrix.

Consequence for Vectors

We’re now ready to address what happens to the components of $\mathbf{A}$ in the rotated coordinate system. We frame the rotated vector as

$$\mathbf{A} = A_u \hat{\mathbf{u}} + A_v \hat{\mathbf{v}},$$

and the job is to solve for A_u, A_v . One way to proceed is to substitute $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ and simplify:

$$\begin{aligned} \mathbf{A} &= A_u (\cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}}) \\ &\quad + A_v (-\sin(\theta) \hat{\mathbf{x}} + \cos(\theta) \hat{\mathbf{y}}) \\ \mathbf{A} &= (A_u \cos(\theta) - A_v \sin(\theta)) \hat{\mathbf{x}} \\ &\quad + (A_u \sin(\theta) + A_v \cos(\theta)) \hat{\mathbf{y}} \end{aligned}$$

So far, we’ve managed to show:

$$\begin{aligned} A_x &= A_u \cos(\theta) - A_v \sin(\theta) \\ A_y &= A_u \sin(\theta) + A_v \cos(\theta) \end{aligned}$$

Evidently, the coordinates can be related by the rotation matrix

$$\begin{bmatrix} A_x \\ A_y \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} A_u \\ A_v \end{bmatrix},$$

but we need the inverse of this to isolate the new components:

$$\begin{bmatrix} A_u \\ A_v \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} A_x \\ A_y \end{bmatrix}$$

In other words, a positive rotation in the coordinate system looks like a negative rotation of the vector components.

9 Vectors and Limits

Vectors, whose components are numbers and functions, obey all of the established properties of limits. While the technical proof for this is attainable, it's worth staving off a formal effort for multivariate calculus, and for now take it on intuition that vectors and limits get along nicely.

9.1 Velocity and Acceleration

In three dimensions, the position vector can be written

$$\mathbf{r} = \langle x, y, z \rangle .$$

For two positions $\mathbf{r}_1, \mathbf{r}_2$, one can establish the displacement vector $\Delta \mathbf{r}$ such that $\Delta \mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$.

If the position vector is a function of time, the displacement vector is also a function of time:

$$\Delta \mathbf{r}(t) = \mathbf{r}(t_2) - \mathbf{r}(t_1) .$$

Dividing through by

$$\Delta t = t_2 - t_1 ,$$

we write

$$\frac{\Delta \mathbf{r}(t)}{\Delta t} = \frac{\mathbf{r}(t_2) - \mathbf{r}(t_1)}{\Delta t} .$$

In the limit that $\Delta t \rightarrow 0$, the above becomes the *instantaneous velocity vector*:

$$\mathbf{v}(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{r}(t_2) - \mathbf{r}(t_1)}{\Delta t}$$

Similar to the displacement, we can also consider the change in velocity

$$\Delta \mathbf{v}(t) = \mathbf{v}(t_2) - \mathbf{v}(t_1) .$$

Much like the above, divide through by Δt and take the limit $\Delta t \rightarrow 0$ to write the *instantaneous acceleration vector*:

$$\mathbf{a}(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{v}(t_2) - \mathbf{v}(t_1)}{\Delta t}$$

9.2 Pi from Nested Radicals

Here we apply vectors to a curious problem that approximates the value of π by covering unit circle with triangles of known area. Figure 13.7 shows the first quadrant of the unit circle with several lines and points labeled to aid the derivation. The origin is at what would be the center of the complete circle.

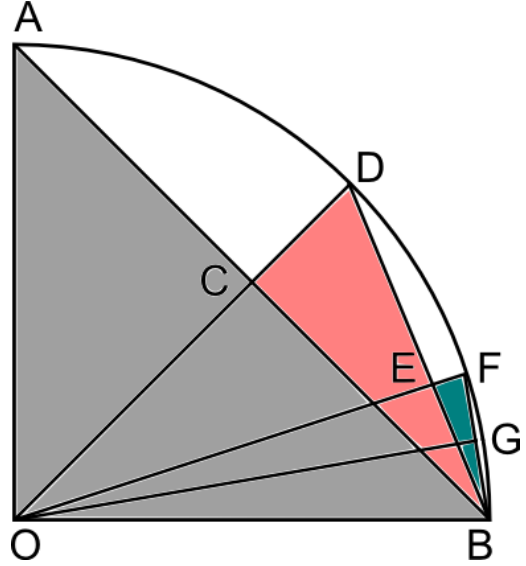


Figure 13.7. Covering a quarter circle with triangles.

To establish some notation, let the lines OA, OB define a pair of unit vectors:

$$\hat{i} = \overrightarrow{OB}$$

$$\hat{j} = \overrightarrow{OA}$$

Also, let the hypotenuse of AOB be a vector $\mathbf{h}_0$ such that

$$\mathbf{h}_0 = \hat{i} - \hat{j} ,$$

with magnitude

$$|\mathbf{h}_0| = \sqrt{2} .$$

Order-Zero Triangle (1)

The largest triangle that fits in the quarter unit circle is depicted AOB , whose area is $1/2$. Using the vectors on hand, the area of AOB shall be written

$$A_0 = \frac{1}{2} |\hat{i}| |\hat{j}| ,$$

which is a fancy way to write $1/2$.

Order-One Triangle (2)

Next, we seek two identical triangles that cover the largest uncovered portion of the quarter circle. In Figure 13.7 these are depicted DCA, DCB , respectively.

Analyzing DCB , consider a unit vector $\hat{x}_1$ and a shorter vector x_1 such that

$$x_1 = \overline{OC} = \frac{\hat{i} + \hat{j}}{2}$$

$$\hat{x}_1 = \overline{OD} = \frac{\hat{i} + \hat{j}}{\sqrt{2}},$$

whose difference in length is CD .

Then, the area of DCB is

$$A_1 = \frac{(CD)(CB)}{2}$$

$$= \frac{1}{2} |\hat{x}_1 - x_1| \frac{1}{2} |h_0|.$$

That is, the base is line CB , whose length is half the magnitude h_0 . The height CB is given by the difference in x -vectors.

Notice that, because $\hat{x}_1, x_1$ are parallel, the following simplification can be made:

$$|\hat{x}_1 - x_1| = 1 - x_1$$

Calculating out A_1 , one finds, after some algebra:

$$A_1 = \frac{1}{2} \left(-\frac{1}{2} + \frac{1}{\sqrt{2}} \right)$$

The hypotenuse of DCB is denoted h_1 and is given by

$$h_1 = \hat{i} - \hat{x}_1,$$

and has length DB . Note that the vector h_1 doesn't play into the area of DCB , rather the previous h_0 is used. Explicitly, the vector h_1 reads

$$h_1 = \left(1 - \frac{1}{\sqrt{2}} \right) \hat{i} + \frac{1}{\sqrt{2}} \hat{j},$$

having magnitude

$$|h_1| = \sqrt{2 - \sqrt{2}}.$$

Order-Two Triangles (4)

To keep covering the circle, we'll take four copies of triangle FEB as shown in the Figure. To find the area of just FEB , first notice

$$x_2 = \overline{OE} = \hat{x}_1 + \frac{1}{2} h_1$$

$$\hat{x}_2 = \overline{OF} = x_2 / |x_2|,$$

whose difference in length is EF . In detail, it's straightforward to show that

$$x_2 = \frac{1}{2} \left(1 + \frac{1}{\sqrt{2}} \right) \hat{i} + \frac{1}{2\sqrt{2}} \hat{j},$$

where

$$x_2 = \frac{1}{2} \sqrt{2 + \sqrt{2}},$$

and

$$\hat{x}_2 = \left(\frac{1 + 1/\sqrt{2}}{\sqrt{2 + \sqrt{2}}} \right) \hat{i} + \left(\frac{\sqrt{2 - \sqrt{2}}}{2} \right) \hat{j}.$$

The area of FEB is

$$A_2 = \frac{(EF)(EB)}{2}$$

$$= \frac{1}{2} |\hat{x}_2 - x_2| \frac{1}{2} |h_1|$$

$$= \frac{1}{4} (1 - x_2) |h_1|,$$

which plays much like the previous case with all indices shifted up by one. Calculating out A_2 , one finds, after a lot of algebra:

$$A_2 = \frac{1}{4} \left(-\frac{1}{\sqrt{2}} + \sqrt{2 - \sqrt{2}} \right)$$

The hypotenuse of FEB is h_2 , given by

$$h_2 = \hat{i} - \hat{x}_2,$$

which is also just like the the formula for h_1 with the indices bumped by one. Explicitly, the vector h_2 reads

$$h_2 = \left(1 - \frac{(1 + 1/\sqrt{2})}{\sqrt{2 + \sqrt{2}}} \right) \hat{i} + \left(\frac{\sqrt{2 - \sqrt{2}}}{2} \right) \hat{j},$$

having magnitude

$$|h_2| = \sqrt{2 - \sqrt{2 + \sqrt{2}}}.$$

Order-Three Triangles (8)

By now we're running out of letters in the Figure, but the pattern continues. The next step has eight total triangles. Begin with

$$\begin{aligned} \mathbf{x}_3 &= \hat{\mathbf{x}}_2 + \frac{1}{2} \mathbf{h}_2 \\ \hat{\mathbf{x}}_3 &= \mathbf{x}_3 / |\mathbf{x}_3|, \end{aligned}$$

implying the area to be

$$\begin{aligned} A_3 &= \frac{1}{2} |\hat{\mathbf{x}}_3 - \mathbf{x}_3| \frac{1}{2} |\mathbf{h}_2| \\ &= \frac{1}{4} (1 - x_3) |\mathbf{h}_2|, \end{aligned}$$

and furthermore:

$$\mathbf{h}_3 = \hat{\mathbf{i}} - \hat{\mathbf{x}}_3$$

Leaving the algebra to the dedicated reader, the area A_3 resolves to:

$$A_3 = \frac{1}{8} \left(-\sqrt{2 - \sqrt{2}} + 2\sqrt{2 - \sqrt{2 + \sqrt{2}}} \right)$$

Order-N Triangles

It will take an infinite number of iterations to cover the entire quarter circle with increasingly smaller triangles. At the n th step, it follows that

$$\begin{aligned} \mathbf{x}_n &= \hat{\mathbf{x}}_{n-1} + \frac{1}{2} \mathbf{h}_{n-1} \\ \hat{\mathbf{x}}_n &= \mathbf{x}_n / |\mathbf{x}_n| \\ \mathbf{h}_n &= \hat{\mathbf{i}} - \hat{\mathbf{x}}_n, \end{aligned}$$

with

$$A_n = \frac{1}{4} (1 - x_n) |\mathbf{h}_{n-1}|.$$

While we'll take the above as a workable result, note that

$$|\mathbf{h}_{n-1}| = \sqrt{2(1 - \hat{\mathbf{i}} \cdot \hat{\mathbf{x}}_{n-1})}.$$

With this, the area simplifies to

$$A_n = \frac{\sqrt{2}}{4} (1 - x_n) \sqrt{1 - \hat{\mathbf{i}} \cdot \hat{\mathbf{x}}_{n-1}},$$

which only works for $n \geq 2$.

Also, it's easy to derive

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{x}}_{n-1} = \cos\left(\frac{\pi}{2^n}\right)$$

from the vectors on hand. It would be bad form, however, to invoke π in the midst of trying to calculate it, so we'll leave the cosine function alone.

Working out the area A_4 from the above, which is absolutely tedious without a machine, one finds:

$$\begin{aligned} A_4 &= -\frac{2\sqrt{2 - \sqrt{2 + \sqrt{2}}}}{16} \\ &\quad + \frac{4\sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2}}}}}{16} \end{aligned}$$

N-Sided Polygon

Tallying all areas of all triangles up to the N th step gives one quarter the area of an N -sided polygon with equal angles and equal sides. Doing so, we write

$$P(N) = \sum_{n=0}^N w_n A_n,$$

where A is the total area, A_n is the area of *one* triangle of order n , and w_n is the number of triangles of order n , particularly $w_n = 2^n$.

Condensing variables again, we also write, for $n \geq 2$:

$$P(N) = \sum_{n=0}^N \frac{2^n}{4} (1 - x_n) |\mathbf{h}_{n-1}| = \sum_{n=0}^N P_n$$

Evaluating P(N)

Now the real work begins. Starting with $N = 0$ and working up, we find

$$P_0 = 2^0 A_0 = \frac{1}{2}$$

$$P_1 = 2^1 A_1 = -\frac{1}{2} + \frac{1}{\sqrt{2}}$$

$$P_2 = 2^2 A_2 = -\frac{1}{\sqrt{2}} + \sqrt{2 - \sqrt{2}}$$

$$P_3 = 2^3 A_3 = -\sqrt{2 - \sqrt{2}} + 2\sqrt{2 - \sqrt{2 + \sqrt{2}}},$$

along with

$$P_4 = 2^4 A_4 = -2\sqrt{2 - \sqrt{2 + \sqrt{2}}} + 4\sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2}}}}$$

Now an amazing simplification happens. Calculating the sum $P(N)$ requires adding all terms P_n up to the N th term. However, notice that each P_n contains a positive term and a negative term. The negative term is always the exact negative of the previous positive term. The end result is, only the positive term in P_N survives the summation.

Going from the pattern on hand, we evidently have

$$P(N) = \frac{2^N}{4} \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2 + \dots}}}}$$

with N total square roots.

Area of N-Sided Polygon

Define Π (uppercase of π) such that

$$\Pi(N) = 4P(N) ,$$

which is the area of an N -sided polygon (all four quadrants).

For the first few orders, a calculator reveals:

$$\begin{aligned} \Pi(0) &= 2 \\ \Pi(1) &= 2.8284271248 \dots \\ \Pi(2) &= 3.0614674589 \dots \\ \Pi(3) &= 3.1214451523 \dots \\ \Pi(4) &= 3.1365484905 \dots \\ \Pi(5) &= 3.1403311570 \dots \\ \Pi(10) &= 3.1415914215 \dots \\ \Pi(15) &= 3.1415926524 \dots \\ \Pi(20) &= 3.1415926536 \dots \end{aligned}$$

The area becomes suspiciously close to π as the number of iterations increases, and we get about ten digits of π after $N = 18$ iterations.

The number of triangles for the quarter-area is given by

$$W(N) = \sum_{n=0}^N w_n = \sum_{n=0}^N 2^n = 2^{N+1} - 1 ,$$

and for $N = 18$, we approximately have

$$W(18) = 2^{19} - 1 .$$

The total number of triangles is four times the above. Over two million triangles are needed to get π to ten digits:

$$4W(18) = 2^{21} - 4 = 2097148$$

Area of Unit Circle

In the limit $N \rightarrow \infty$, the triangles cover the unit circle and the area converges to π :

$$\pi = \lim_{N \rightarrow \infty} \Pi(N) ,$$

or

$$\pi = \lim_{N \rightarrow \infty} 2^N \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2 + \dots}}}} .$$

There are N square roots on the right.

This result is a bit counter-intuitive in the sense that 2^N tends to infinity while the quantity under the outermost square root goes to zero. It just happens that infinity times zero, in this particular limit, equals π .

Chapter 14

Complex Algebra

1 History of Complex Numbers

The story of *complex numbers* begins more than a century before calculus, in a time when mathematicians were still puzzling through what we would now consider high school algebra.

The issue of solving depressed cubic equations

$$x^3 + bx = c$$

was especially prescient in mid-1500s Italy, and eventually a pair of mathematicians would derive the *del Ferro-Tartaglia* formula

$$x_0 = \sqrt[3]{\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} + \sqrt[3]{\frac{c}{2} - \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}},$$

allowing a single solution to the depressed cubic to be attained.

While equation the above works for a certain class of depressed cubic problems, it is still peculiar in the sense that negative numbers may end up embedded under the radical symbols. Staying within the rules of algebra, the un-treatable quantity always boils down to $\sqrt{-1}$, making x_0 impossible to simplify as such.

Aware of this, mathematician Rafael Bombelli had the ‘wild thought’ to work with factors of $\sqrt{-1}$ anyway. This means to suppose the ‘ugliness’ of the cube-root terms in x_0 could be split away from the well-behaved part such that

$$\begin{aligned}\sqrt[3]{\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} &= U + \sqrt{-1} V \\ \sqrt[3]{\frac{c}{2} - \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} &= U - \sqrt{-1} V\end{aligned}$$

for two unknown coefficients U and V . Then, when we assemble x_0 again, the V -terms cancel,

$$x_0 = U + \sqrt{-1} V + U - \sqrt{-1} V = 2U,$$

which is guaranteed to come out to a ‘clean’ number.

Real, Imaginary, Complex

To avoid using terms like ‘clean’ numbers, it is generally meant that numbers containing no factors of $\sqrt{-1}$ are called *real*. Any real number multiplied by $\sqrt{-1}$ is *imaginary* number. The sum of a real number and an imaginary number is a *complex* number.

Imaginary Component

Synonyms for the ‘real part’ and ‘imaginary part’ of a complex number, respectively are the *components* of the number.

The present task is to solve for the previously-defined components U, V in terms of the coefficients b, c . Do this by raising each of side of the above to the third power to find

$$\begin{aligned}\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}} &= \\ U(U^2 - 3V^2) + \sqrt{-1} V(3U^2 - V^2),\end{aligned}$$

availing the connection:

$$\begin{aligned}\text{Real part of } \left(\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}\right) &= U(U^2 - 3V^2) \\ \text{Imag. part of } \left(\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}\right) &= V(3U^2 - V^2)\end{aligned}$$

Pursing Bombelli's method, we end up with a pair of two equations with two unknowns (neither contain $\sqrt{-1}$). Faced with this, Bombelli had a second wild thought: perhaps U and V need to be *positive integers* (we won't dwell much on this detail). Solving for U , V completes the recipe for cooking the first solution x_0 .

With x_0 in hand, the term $x - x_0$ can be factored out of the depressed cubic equation, yielding the form

$$x^3 + bx - c = (x - x_0) \left(x^2 + x_0x + \frac{c}{x_0} \right).$$

Since the remaining term is quadratic in x , there are at most two more solutions $x_{1,2}$ to the equation that can be found with the quadratic formula on:

$$x^2 + x_0x + \frac{c}{x_0} = 0$$

Worked Example

Putting these ideas to work, suppose we need to find all solutions to

$$x^3 - 30x - 36 = 0.$$

Identifying $b = -30$ and $c = 36$, the del Ferro-Tartaglia formula tells us

$$x_0 = \sqrt[3]{18 + \sqrt{-1} 26} + \sqrt[3]{18 - \sqrt{-1} 26}.$$

By Bombelli's reasoning, we seek integer solutions to U , V such that

$$\begin{aligned} 18 &= U(U^2 - 3V^2) \\ 26 &= V(3U^2 - V^2). \end{aligned}$$

The prime factorization of 18 is $2 \times 3 \times 3$, suggesting that $U = 3$, and thus $V = 1$. Suddenly we've got a solution:

$$x_0 = 2U = 6$$

This is quite a remarkable achievement, since by writing

$$\sqrt[3]{18 + \sqrt{-1} 26} + \sqrt[3]{18 - \sqrt{-1} 26} = 6,$$

we see undeniably that, regardless of how one may feel about $\sqrt{-1}$, it is useful for problem solving.

Problem 1

With $x_0 = 6$ as a known solution to

$$x^3 - 30x - 36 = 0,$$

show that the other two solutions are

$$\begin{aligned} x_1 &= 3 + \sqrt{3} \\ x_2 &= 3 - \sqrt{3}. \end{aligned}$$

2 Complex Numbers

2.1 Definition

Complex Numbers

Let the complex number z exist as an ordered pair of two (real) numbers called *components* a, b such that

$$z = \langle a, b \rangle . \quad (14.1)$$

Complex Conjugate

For every complex number z , there exists the *complex conjugate*, also a complex number, denoted $\bar{z}$ or z^* , by flipping the sign on b :

$$\bar{z} = z^* = \langle a, -b \rangle . \quad (14.2)$$

Relationship to Real Numbers

A complex number $z = \langle a, 0 \rangle$ is the same as the scalar $z = a$:

$$\langle a, 0 \rangle = a \quad (14.3)$$

Problem 2

Check that the complex conjugate of the complex conjugate recovers the complex number:

$$\overline{\bar{z}} = z \quad (14.4)$$

2.2 Complex Arithmetic

Scalar Multiplication

Complex numbers can be ‘scaled’ by real numbers called *scalars*:

$$\lambda z = \langle \lambda a, \lambda b \rangle \quad (14.5)$$

Complex Addition

For two complex numbers $z_1 = \langle a_1, b_1 \rangle$, $z_2 = \langle a_2, b_2 \rangle$, their sum is

$$z_1 + z_2 = \langle a_1 + a_2, b_1 + b_2 \rangle . \quad (14.6)$$

Complex Multiplication

For two complex numbers $z_1 = \langle a_1, b_1 \rangle$, $z_2 = \langle a_2, b_2 \rangle$, their product is simultaneously defined by a commutation relation and a conjugate relation:

$$z_1 \cdot z_2 = z_2 \cdot z_1 \quad (14.7)$$

$$\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2} \quad (14.8)$$

Conspicuously absent from the list of axioms is an explicit formula for complex multiplication. For completeness, the formula for complex multiplication reads

$$z_1 \cdot z_2 = \langle a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1 \rangle , \quad (14.9)$$

but this does not need to be axiomatic unless you’re in a hurry. Instead, we’ll soon derive Equation (14.9) from the equations preceding it.

2.3 Properties of Addition

Isolating Complex Components

Given a complex number $z = \langle a, b \rangle$ and its complex conjugate $\bar{z} = \langle a, -b \rangle$, take their sum and difference, respectively, to write a pair of relations that ‘solve for’ a, b :

$$\langle a, 0 \rangle = \frac{z + \bar{z}}{2} \quad (14.10)$$

$$\langle 0, b \rangle = \frac{z - \bar{z}}{2} \quad (14.11)$$

Note that a and b in isolation are each real numbers. As they appear in $z = \langle a, b \rangle$, a is the so-called ‘real part’, and b is the ‘imaginary part’.

Commutation and Conjugation

Two relationships readily verifiable from the axioms are the respective *commutation* and *conjugation* relations for addition:

$$z_1 + z_2 = z_2 + z_1 \quad (14.12)$$

$$\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2} \quad (14.13)$$

Problem 3

Verify Equations Equation (14.12) and Equation (14.13) using any of the axioms.

2.4 Properties of Multiplication

Finally we encounter the first piece of hard work, which is to derive the formula for complex multiplication. Most texts simply take Equation (14.9) as an axiom to avoid a slightly dry derivation, and you are welcome to do so now as well.

Derivation

As a starting point, propose the product $z_1 \cdot z_2$ to result in a new complex number comprised of every order-two combination of $a_{1,2}, b_{1,2}$

$$z_1 \cdot z_2 = \langle a_1, b_1 \rangle \cdot \langle a_2, b_2 \rangle = \langle q, r \rangle ,$$

where

$$\begin{aligned} q &= \alpha a_1 a_2 + \beta a_1 b_2 + \gamma a_2 b_1 + \delta b_1 b_2 \\ r &= \tilde{\alpha} a_1 a_2 + \tilde{\beta} a_1 b_2 + \tilde{\gamma} a_2 b_1 + \tilde{\delta} b_1 b_2 , \end{aligned}$$

and each Greek index α, β , etc. (eight in total) resolves to 1, 0, or -1 . We shall nail these down in several steps:

Impose the conjugation relation Equation (14.8), causing all b -terms to flip sign. The q -term must remain the same under this change, but the r -term must flip sign. This can only hold if

$$\tilde{\alpha} = \beta = \gamma = \tilde{\delta} = 0 ,$$

so we now have:

$$\begin{aligned} q &= \alpha a_1 a_2 + \delta b_1 b_2 \\ r &= \tilde{\beta} a_1 b_2 + \tilde{\gamma} a_2 b_1 , \end{aligned}$$

Impose the commutation relation Equation (14.7), causing all 1- and 2-subscripts to swap. The q -equation remains unchanged, but the r equation demands

$$\tilde{\beta} = \tilde{\gamma} \neq 0 .$$

Swap all a - and b -symbols. Doing so should completely change the results, however the r -equation is invariant with respect to the swap. It follows that α and δ in the q -equation must disagree in sign:

$$\alpha = -\delta \neq 0$$

Boiling everything down:

$$\begin{aligned} q &= \alpha (a_1 a_2 - b_1 b_2) \\ r &= \tilde{\beta} (a_1 b_2 + a_2 b_1) \end{aligned}$$

Let $b_2 = 0$. The corresponding product becomes

$$z_1 \cdot z_2 = \langle \alpha a_1 a_2, \tilde{\beta} a_2 b_1 \rangle = a_2 \langle \alpha a_1, \tilde{\beta} b_1 \rangle ,$$

which looks much like the scalar multiplication λz_1 with $\lambda = a_2$. For the sake of keeping complex multiplication consistent with scalar multiplication, let us finally set

$$\alpha = \tilde{\beta} = 1 ,$$

finishing the derivation.

Associative Property

Complex numbers $z_j = \langle a_j, b_j \rangle$ with $j = 1, 2, 3$ obey the *associative property*

$$(z_1 \cdot z_2) \cdot z_3 = z_1 \cdot (z_2 \cdot z_3) , \quad (14.14)$$

shown by brute force:

$$\begin{aligned} (z_1 \cdot z_2) \cdot z_3 &= \langle a_1 a_2 - b_1 b_2, a_1 b_2 + b_1 a_2 \rangle \cdot \langle a_3, b_3 \rangle \\ &= (a_1 a_2 a_3 - b_1 b_2 a_3 - a_1 b_2 b_3 - b_1 a_2 b_3, \\ &\quad a_1 a_2 b_3 - b_1 b_2 b_3 + a_1 b_2 a_3 + b_1 a_2 a_3) \\ &= (a_1 (a_2 a_3 - b_2 b_3) - b_1 (b_2 a_3 + a_2 b_3), \\ &\quad a_1 (a_2 b_3 + b_2 a_3) + b_1 (b_2 b_3 - a_2 a_3)) \\ &= \langle a_1, b_1 \rangle \cdot \langle a_2 a_3 - b_2 b_3, a_2 b_2 + b_2 a_3 \rangle \\ &= z_1 \cdot (z_2 \cdot z_3) \end{aligned}$$

Distributive Property

Complex numbers $z_j = \langle a_j, b_j \rangle$ with $j = 1, 2, 3$ also obey the *distributive property*:

$$z_1 \cdot (z_2 + z_3) = (z_1 \cdot z_2) + (z_1 \cdot z_3) , \quad (14.15)$$

also shown by brute force:

$$\begin{aligned} z_1 \cdot (z_2 + z_3) &= \langle a_1, b_1 \rangle \cdot \langle a_2 + a_3, b_2 + b_3 \rangle \\ &= (a_1 a_2 + a_1 a_3 - b_1 b_2 - b_1 b_3, \\ &\quad a_1 b_2 + a_1 b_3 + a_2 b_1 + a_3 b_1) \\ &= \langle a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1 \rangle + \\ &\quad \langle a_1 a_3 - b_1 b_3, a_1 b_3 + a_3 b_1 \rangle \\ &= (z_1 \cdot z_2) + (z_1 \cdot z_3) \end{aligned}$$

2.5 Complex Magnitude

The quantity

$$|z| = \sqrt{z \cdot \bar{z}} \quad (14.16)$$

is called the *magnitude* of z , and is always a non-negative real number:

$$\begin{aligned} |z| &= \sqrt{z \cdot \bar{z}} = \sqrt{\langle a, b \rangle \cdot \langle a, -b \rangle} \\ &= \sqrt{\langle a^2 + b^2, 0 \rangle} = \sqrt{a^2 + b^2} \end{aligned}$$

This result foreshadows *some* kind of geometric interpretation of complex numbers in the sense that $|z|$ is the hypotenuse of a right triangle with sides a, b .

2.6 Complex Division

The last standard arithmetic operation is complex division, which seems inoperable at face value:

$$d = \frac{z_1}{z_2} = \frac{\langle a_1, b_1 \rangle}{\langle a_2, b_2 \rangle}$$

To make a useful complex number from this, multiply the top and bottom by $\bar{z}_2$:

$$d = \frac{\langle a_1, b_1 \rangle \langle a_2, -b_2 \rangle}{\langle a_2, b_2 \rangle \langle a_2, -b_2 \rangle} = \frac{z_1 \cdot \bar{z}_2}{|z_2|^2} \quad (14.17)$$

Explicitly, the above means:

$$d = \frac{1}{a_2^2 + b_2^2} \langle a_1 a_2 + b_1 b_2, -a_1 b_2 + a_2 b_1 \rangle \quad (14.18)$$

Problem 4

Prove that the complex division operation obeys its own conjugate relation:

$$\bar{z}_1 / \bar{z}_2 = \overline{z_1 / z_2} \quad (14.19)$$

2.7 Generalized Complex Arithmetic

If we seek to write a tighter set of axioms, one could start with the general equations

$$\begin{aligned} z_1 \star z_2 &= z_2 \star z_1 \\ \overline{z_1 \star z_2} &= \overline{z_1} \star \overline{z_2} \end{aligned}$$

for a generalized operator $\star$. As we've seen, the multiplication operator results from seeking all order-two combinations of $a_{1,2}, b_{1,2}$. By the same token, the addition operator results from seeking all order-one combinations of $a_{1,2}, b_{1,2}$.

Problem 5

Derive the addition operation Equation (14.6) using only Equations Equation (14.12) and Equation (14.13).

Problem 6

Let z_1 be any complex number. Find all other complex numbers z_2 that satisfy $z_1 \cdot z_2 = 0$.

Problem 7

Let z_1 be any complex number. Find all other complex numbers z_2 that satisfy $z_1 \cdot z_2 = z_1$.

3 Complex Plane

3.1 Imaginary Unit

Having established the notion of complex multiplication, now consider the product:

$$\langle 0, 1 \rangle \cdot \langle 0, 1 \rangle = \langle 0 - 1 \cdot 1, 0 \rangle = -1$$

The left side has two instances of $\langle 0, 1 \rangle$ whose product is -1 , which can only mean

$$\langle 0, 1 \rangle = \sqrt{-1},$$

known as the *imaginary unit*: Often, the imaginary unit is denoted as i such that

$$i = \sqrt{-1}$$

and

$$i^2 = -1.$$

3.2 Complex Numbers as Operators

For a complex number $z = \langle a, b \rangle$, consider the ‘identity’ statement

$$\langle 1, 0 \rangle \cdot \langle a, b \rangle = \langle a, b \rangle. \quad (14.20)$$

Equation (14.20) is the ‘least invasive’ multiplication operation in which z_1 can participate.

Seeking a corresponding identity with $\langle 0, 1 \rangle$, we find, by complex multiplication,

$$\langle 0, 1 \rangle \cdot \langle a, b \rangle = \langle -b, a \rangle. \quad (14.21)$$

Qualitatively, the ‘operator’ $\langle 0, 1 \rangle$ swaps a with b and introduces a negative sign as shown. This amounts to another hint that the real and imaginary components of z are somehow ‘orthogonal’, meaning there could be some geometric interpretation of complex numbers.

Going on this hunch, let us think of $\langle 1, 0 \rangle$ and $\langle 0, 1 \rangle$ as ‘basis vectors’, and write a linear combination with two undetermined coefficients α, β :

$$\alpha \langle 1, 0 \rangle + \beta \langle 0, 1 \rangle = \langle \alpha, 0 \rangle + \langle 0, \beta \rangle = \langle \alpha, \beta \rangle$$

Almost obviously, such an operator is nothing more than a complex number $\langle \alpha, \beta \rangle$. Trying this operator on a different complex number $z = \langle a, b \rangle$, we have

$$\langle \alpha, \beta \rangle \cdot \langle a, b \rangle = \langle \alpha a - \beta b, \alpha b + \beta a \rangle, \quad (14.22)$$

or more concisely,

$$\langle \alpha, \beta \rangle \cdot z = z' = \langle a', b' \rangle.$$

3.3 Rotation Operator

Let us find an operator, i.e. some just a complex number $z_\phi = \langle \alpha_\phi, \beta_\phi \rangle$ that acts on $z = \langle a, b \rangle$ such that the components change but the magnitude does not. From Equation (14.22), we write

$$\begin{aligned} |z'| &= \sqrt{(a')^2 + (b')^2} \\ &= \sqrt{(\alpha_\phi a - \beta_\phi b)^2 + (\beta_\phi a + \alpha_\phi b)^2}, \end{aligned}$$

simplifying nicely to

$$|z'| = |z| \sqrt{\alpha_\phi^2 + \beta_\phi^2}. \quad (14.23)$$

In the special case that z_ϕ has $\alpha_\phi^2 + \beta_\phi^2 = 1$, then z_ϕ qualifies as a rotation operator. The locus of α_ϕ, β_ϕ describes a ‘complex unit circle’, begging the parameterization

$$\alpha_\phi = \cos(\phi) \quad (14.24)$$

$$\beta_\phi = \sin(\phi), \quad (14.25)$$

where ϕ is a real continuous parameter. As a sanity check, one can see that $\phi = 0, \phi = \pi/2$ correspond to the respective operators $\langle 1, 0 \rangle, \langle 0, 1 \rangle$. Let us therefore take the *complex rotation operator* to be the complex number

$$z_\phi = \langle \cos(\phi), \sin(\phi) \rangle. \quad (14.26)$$

3.4 Radius and Phase

The rotation operator allows us to interpret complex numbers in a curious way. Given the real number r , any complex number z with magnitude

$$|z| = r \quad (14.27)$$

is the product

$$z = r z_\phi = \langle r \cos(\phi), r \sin(\phi) \rangle = \langle a, b \rangle. \quad (14.28)$$

Borrowing terminology from polar coordinates, the magnitude of a complex number is equivalent to the *radius*, and the angle parameter is called the *phase*. In terms of these, the components of a complex number read

$$a = r \cos(\phi) \quad (14.29)$$

$$b = r \sin(\phi), \quad (14.30)$$

easily inverted:

$$r = |z| = \sqrt{a^2 + b^2} \quad (14.31)$$

$$\phi = \arctan\left(\frac{b}{a}\right) \quad (14.32)$$

As a matter of terminology, recall that the a - and b -components of z are the real and imaginary parts, i.e.

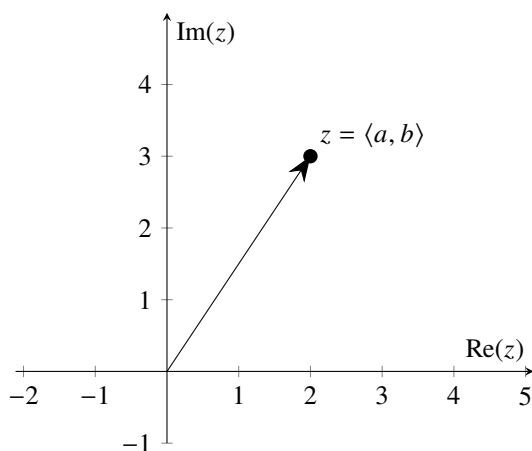
$$\begin{aligned} a &= \operatorname{Re}(z) \\ b &= \operatorname{Im}(z) . \end{aligned}$$

The phase angle ϕ is often denoted $\operatorname{Arg}(z)$, or ‘argument of z ’. All together, an equivalent statement of Equation (14.32) reads

$$\operatorname{Arg}(z) = \arctan\left(\frac{\operatorname{Im}(z)}{\operatorname{Re}(z)}\right) . \quad (14.33)$$

3.5 Real and Imaginary Axes

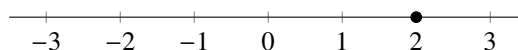
The r, ϕ interpretation of complex numbers leads to an almost-identical mathematical apparatus needed for plane polar coordinates. Complex numbers occupy a space that is analogous to the Cartesian xy -plane, except the x -axis is replaced by the real axis, and the y -axis is replaced by the imaginary axis. In this connection we speak freely of the *complex plane*:



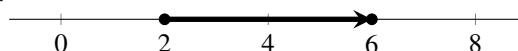
Any ‘location’ on the $\operatorname{Im}(z) = 0$ line is a purely real number $z = \langle a, 0 \rangle = a$, whereas anywhere on the $\operatorname{Re}(z) = 0$ is a purely imaginary number $z = \langle 0, b \rangle$. Any off-axis location is a complex number $z = \langle a, b \rangle$.

3.6 Number as Location

Thinking for a moment about classical arithmetic, it’s convenient to represent any real number, such as $x = 2$, on a number line:



Any operation performed on x , such as multiplying by three, can be represented as a spatial displacement on the line:

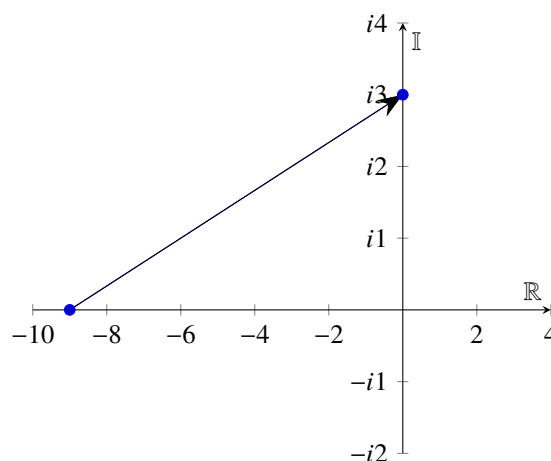


In fact, all of the ‘well-behaved’ operations that one could possibly perform on x will land *somewhere* on the number line.

The first hint that the number line is somehow incomplete arises when we ask, much like the renaissance-era mathematicians asked, ‘what happens when we take the square root of a negative number?’ While this issue was a sign for most to turn around, recall that Bombelli, in a moment of ‘wild thought’ had the idea to separate the real component from the imaginary component. He understood that numbers like

$$\begin{aligned} x^2 &= -9 \\ x &= \sqrt{-9} = \sqrt{-1} \times \sqrt{9} = 3\sqrt{-1} \end{aligned}$$

must be represented off of the number line, even without foreknowledge of the complex plane. Since we *do* know about the the complex plane though, the core of Bombelli’s insight can now be visualized without ambiguity:



3.7 Complex Numbers and Vectors

Since there is much talk of two-component objects and their relationship to a plane, it’s worthwhile to ask how closely

complex numbers resemble vectors. Begin this inquiry by considering two complex numbers z_1, z_2 , and solve Equation (14.29)-Equation (14.30) for the respective trigonometry terms:

$$\cos(\phi_1) = \frac{a_1}{|z_1|}$$

$$\sin(\phi_1) = \frac{b_1}{|z_1|}$$

$$\cos(\phi_2) = \frac{a_2}{|z_2|}$$

$$\sin(\phi_2) = \frac{b_2}{|z_2|}$$

Multiply the cos-terms and the sin-terms respectively

$$\cos(\phi_1) \cos(\phi_2) = \frac{a_1 a_2}{|z_1| |z_2|}$$

$$\sin(\phi_1) \sin(\phi_2) = \frac{b_1 b_2}{|z_1| |z_2|},$$

and combine the results:

$$\cos(\phi_1 - \phi_2) = \frac{a_1 a_2 + b_1 b_2}{|z_1| |z_2|} \quad (14.34)$$

Had z_1 and z_2 been vectors $\mathbf{z}_1$ and $\mathbf{z}_2$, the quantity $a_1 a_2 + b_1 b_2$ stands out as the dot product $\mathbf{z}_1 \cdot \mathbf{z}_2$. Evidently, we have

$$\mathbf{z}_1 \cdot \mathbf{z}_2 = |z_1| |z_2| \cos(\phi_1 - \phi_2). \quad (14.35)$$

Note that the quantity $a_1 a_2 + b_1 b_2$ can be written yet another way, namely

$$a_1 a_2 + b_1 b_2 = \frac{1}{2} (\bar{z}_1 \cdot z_2 + z_1 \cdot \bar{z}_2), \quad (14.36)$$

allowing a tight relationship to be written:

$$\mathbf{z}_1 \cdot \mathbf{z}_2 = \frac{1}{2} (\bar{z}_1 \cdot z_2 + z_1 \cdot \bar{z}_2) \quad (14.37)$$

Problem 8

Write equations analogous to Equation (14.34), Equation (14.35), and Equation (14.36) to derive the cross product analog to Equation (14.37):

$$\mathbf{z}_1 \times \mathbf{z}_2 = \frac{1}{2} (\bar{z}_1 \cdot z_2 - z_1 \cdot \bar{z}_2) \quad (14.38)$$

Problem 9

Derive:

$$\bar{z}_1 \cdot z_2 = \langle \mathbf{z}_1 \cdot \mathbf{z}_2, \mathbf{z}_1 \times \mathbf{z}_2 \rangle \quad (14.39)$$

Problem 10

Using vectors as an analogy, derive the triangle inequality for complex numbers:

$$||z_1| - |z_2|| \leq |z_1 + z_2| \leq |z_1| + |z_2| \quad (14.40)$$

3.8 Complex Numbers and Matrices

Matrix Notation

A square matrix of size 2×2 is most generally written

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

For the multiplication of two such matrices M, N , one uses the standard matrix multiplication procedure:

$$\begin{aligned} MN &= \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} \\ &= \begin{bmatrix} M_{11}N_{11} + M_{12}N_{21} & M_{11}N_{12} + M_{12}N_{22} \\ M_{21}N_{11} + M_{22}N_{21} & M_{21}N_{12} + M_{22}N_{22} \end{bmatrix} \end{aligned}$$

Note the product MN is also a size 2×2 matrix.

Supposing the second column in N wasn't present, the matrix is a column vector

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

For the multiplication MV , one writes

$$MV = \begin{bmatrix} M_{11}V_1 + M_{12}V_2 \\ M_{21}V_1 + M_{22}V_2 \end{bmatrix},$$

which is also a column vector.

Complex Representation

Let us now suppose that a complex number $z = \langle x, y \rangle = x + iy$ is equivalent to a column vector

$$z = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Let's also suppose a second complex number $w = \langle a, b \rangle = a + ib$ can be written as a matrix. We know the components of such a matrix have something to do with a, b , but for now propose the matrix of arbitrary coefficients

$$w = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}.$$

One can readily verify the above construction satisfies the rules of complex addition and matrix addition simultaneously. What we must show explicitly is that the other operations work, particularly multiplication and division, i.e. the notion of inverse.

Complex Multiplication

To determine the components of the matrix w , impose the condition that the matrix-vector multiplication wz give the *same* result as if w, z were treated as complex numbers, i.e.

$$wz = \langle ax - by, bx + ay \rangle .$$

For this, we write

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax - by \\ bx + ay \end{bmatrix} ,$$

indicating

$$\begin{aligned} \alpha &= a , & \beta &= -b \\ \gamma &= b , & \delta &= a . \end{aligned}$$

Evidently, the complex multiplication wz in matrix form is given by

$$wz = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax - by \\ bx + ay \end{bmatrix} .$$

Commutation

For commutation of the product $wz = zw$ to make sense, we express z as a matrix and w as a column vector to get the same result:

$$zw = \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} xa - yb \\ xb + ya \end{bmatrix} .$$

From this, we can reliably take

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \leftrightarrow \begin{bmatrix} x & -y \\ y & x \end{bmatrix}$$

as an adequate representation of any complex number.

Complex Conjugate

The complex conjugate $\bar{z}$ of a complex number, in any representation, reverses the sign on the imaginary component:

$$\bar{z} = \langle a, -b \rangle = \begin{bmatrix} a \\ -b \end{bmatrix} \leftrightarrow \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$$

Calculating $z\bar{z}$ in matrix notation, one finds

$$z\bar{z} = \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \begin{bmatrix} x & y \\ -y & x \end{bmatrix} = (x^2 + y^2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ,$$

which is the identity matrix scaled by the square of the magnitude of z .

Problem 11

Show that the determinant of z in matrix form is equal to the square of the magnitude of z .

Complex Matrix Inverse

For any 2×2 matrix M , it's possible to show that the inverse of M is

$$M^{-1} = \frac{1}{M_{11}M_{22} - M_{12}M_{21}} \begin{bmatrix} M_{22} & -M_{12} \\ -M_{21} & M_{11} \end{bmatrix}$$

where the product MM^{-1} yields the 2×2 identity matrix.

For a complex number $z = \langle x, y \rangle$, the inverse is, in matrix notation:

$$z^{-1} = \frac{1}{x^2 + y^2} \begin{bmatrix} x & y \\ -y & x \end{bmatrix}$$

For a sanity check, the product zz^{-1} must yield the identity matrix

$$zz^{-1} = \frac{1}{x^2 + y^2} \begin{bmatrix} x & -y \\ y & x \end{bmatrix} \begin{bmatrix} x & y \\ -y & x \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ,$$

which is indeed the case.

Problem 12

For a complex number

$$z = \langle x, y \rangle \leftrightarrow \begin{bmatrix} x & -y \\ y & x \end{bmatrix} ,$$

derive:

$$\bar{z} = z^{-1} (\det z)$$

Very Complex Multiplication

Next, we must see what happens when multiplying two complex numbers $w = \langle a, b \rangle$, $z = \langle x, y \rangle$ when both are in matrix form. Let A denote the real part of the product and B denote the imaginary part such that

$$wz = \langle ax - by, bx + ay \rangle = \langle A, B \rangle .$$

Then, the matrix multiplication of w, z gives

$$wz = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} x & -y \\ y & x \end{bmatrix} = \begin{bmatrix} A & -B \\ B & A \end{bmatrix} ,$$

which itself has the format of a complex number in matrix notation.

Complex Division

For two complex numbers $w = \langle a, b \rangle$, $z = \langle x, y \rangle$, the ratio $d = wz^{-1}$ is also complex, given by

$$d = \frac{1}{x^2 + y^2} \langle ax + by, -bx + ay \rangle .$$

For shorthand, write the vector components (excluding the $x^2 + y^2$ term) as A' , B' , respectively:

$$d = \frac{1}{x^2 + y^2} \langle A', B' \rangle$$

As a matrix product, the quantity wz^{-1} ends up containing the same information as complex division. Explicitly:

$$\begin{aligned} wz^{-1} &= \frac{1}{x^2 + y^2} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} x & y \\ -y & x \end{bmatrix} \\ &= \frac{1}{x^2 + y^2} \begin{bmatrix} A' & B' \\ -B' & A' \end{bmatrix} \end{aligned}$$

Column Vector Restoration

For two complex numbers $w = \langle a, b \rangle$, $z = \langle x, y \rangle$, the matrix product

$$wz = \begin{bmatrix} A & -B \\ B & A \end{bmatrix}$$

is straightforwardly converted to column representation by the operation

$$wz \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} .$$

For this we needed the column vector

$$p = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

which in light of complex numbers is analogous to one. The above is summarized by

$$(wz) p = \begin{bmatrix} A \\ B \end{bmatrix} = \langle A, B \rangle = \langle ax - by, bx + ay \rangle .$$

Trying the same trick with p on the ratio $d = wz^{-1}$ doesn't work. Instead, we must multiply on the left side of wz by a matrix

$$q = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

such that

$$\begin{aligned} q(wz) &= \frac{1}{x^2 + y^2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} A' & B' \\ -B' & A' \end{bmatrix} \\ &= \frac{1}{x^2 + y^2} \begin{bmatrix} A' & B' \\ -B' & -A' \end{bmatrix} , \end{aligned}$$

so then the final multiplication by p on the right finishes the job:

$$\begin{aligned} q(wz)p &= \frac{1}{x^2 + y^2} \begin{bmatrix} A' & B' \\ -B' & -A' \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \frac{1}{x^2 + y^2} \begin{bmatrix} A' \\ B' \end{bmatrix} = d \end{aligned}$$

3.9 Embedded Complex Numbers

As a matter of curiosity, one may wonder if it makes sense to work with 'embedded complex numbers' such as

$$\langle \langle A, B \rangle, C \rangle \quad \langle A, \langle B, C \rangle \rangle .$$

In the same way that Equation (14.3) permits the association $\langle A, 0 \rangle \leftrightarrow A$, let us extend this by writing

$$\langle \langle A, B \rangle, 0 \rangle \leftrightarrow \langle A, B \rangle .$$

Using the above, we find for the first case:

$$\begin{aligned} \langle \langle A, B \rangle, C \rangle &= \langle \langle A, B \rangle, 0 \rangle + \langle 0, C \rangle \\ &= \langle A, B \rangle + \langle 0, C \rangle \\ &= \langle A, B + C \rangle \end{aligned}$$

Evidently, the 'embedded' complex number flattens down to the 'ordinary' complex number:

$$\langle \langle A, B \rangle, C \rangle \rightarrow \langle A, B + C \rangle$$

Working out the second case in similar fashion:

$$\begin{aligned} \langle A, \langle B, C \rangle \rangle &= \langle A, 0 \rangle + \langle 0, \langle B, C \rangle \rangle \\ &= \langle A, 0 \rangle + \langle 0, 1 \rangle \cdot \langle \langle B, C \rangle, 0 \rangle \\ &= \langle A, 0 \rangle + \langle 0, 1 \rangle \cdot \langle B, C \rangle \\ &= \langle A, 0 \rangle + \langle -C, B \rangle \\ &= \langle A - C, B \rangle \end{aligned}$$

Once again, the embedded complex number flattens down to a (different) ordinary complex number. To summarize, we can stack the two results to write the more general statement

$$\langle \langle A, B \rangle, \langle C, D \rangle \rangle = \langle A - D, B + C \rangle .$$

Now consider two complex numbers

$$z_{1,2} = \langle a_{1,2}, b_{1,2} \rangle ,$$

and also let $A = a_1 a_2$, $D = b_1 b_2$, $B = a_1 b_2$, $C = b_1 a_2$. The right side of the above becomes identical to the formula for the complex multiplication $z_1 \cdot z_2$, giving an equivalent representation as an embedded number:

$$z_1 \cdot z_2 = \langle \langle a_1 a_2, a_1 b_2 \rangle, \langle b_1 a_2, b_1 b_2 \rangle \rangle$$

Since each component $a_{1,2}, b_{1,2}$ is scalar, the above becomes

$$\begin{aligned} z_1 \cdot z_2 &= \langle a_1 \langle a_2, b_2 \rangle, b_1 \langle a_2, b_2 \rangle \rangle \\ &= a_1 a_2 + \langle 0, a_1 b_2 \rangle + \langle 0, b_1 a_2 \rangle + \langle 0, b_1 \langle 0, b_2 \rangle \rangle \\ &= \langle a_1 a_2 - b_1 b_2, a_1 b_2 + b_1 a_2 \rangle, \end{aligned}$$

and the standard notation is recovered.

4 Euler's Formula

Now we derive the most important identity in of complex algebra, called *Euler's formula*.

Repeated Rotations

To kick things off, recall how the complex rotation operator Equation (14.26) acts on any complex number $z = \langle a, b \rangle$ to produce a new complex number z' such that

$$z' = \langle \cos(\phi), \sin(\phi) \rangle \cdot z,$$

where the parameter ϕ is an arbitrary real number.

Next, suppose that ϕ is the sum of two arbitrary angles θ_1, θ_2 , and the above becomes

$$z' = \langle \cos(\theta_1 + \theta_2), \sin(\theta_1 + \theta_2) \rangle \cdot z,$$

simplifying nicely to

$$z' = \langle \cos(\theta_1), \sin(\theta_1) \rangle \cdot \langle \cos(\theta_2), \sin(\theta_2) \rangle \cdot z.$$

Evidently, the *effective* rotation operator representing ϕ becomes the product of two rotation operators representing θ_1, θ_2 .

Generalizing this pattern, suppose instead that ϕ is the sum of n copies of the same angle θ such that

$$\phi = \theta_1 + \theta_2 + \theta_3 + \cdots = n\theta,$$

so a rotation can be written

$$z' = \left\langle \cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right) \right\rangle^n \cdot z.$$

(Don't let the presence of an exponent throw you off - this is just shorthand for $n - 1$ complex multiplications of the same number.)

Summarizing our progress so far, the rotation operator can be interpreted as a repetition of small rotations:

$$\langle \cos(\phi), \sin(\phi) \rangle = \left\langle \cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right) \right\rangle^n \quad (14.41)$$

Infinite Rotations

Given the re-interpreted rotation operator Equation (14.41), the inevitable question is, what happens when n is extremely large, or perhaps infinitely large? Looking at the cos- and sin-terms, we're left to evaluate

$$\begin{aligned} \cos(\phi/n) \\ \sin(\phi/n) \end{aligned}$$

as the argument ϕ/n goes to zero. Borrowing from elementary trigonometry and precalculus though, the relations

$$\lim_{n \rightarrow \infty} \cos\left(\frac{\phi}{n}\right) = 1 \quad (14.42)$$

$$\lim_{n \rightarrow \infty} \sin\left(\frac{\phi}{n}\right) = \frac{\phi}{n} \quad (14.43)$$

apply, letting us write

$$\lim_{n \rightarrow \infty} \left\langle \cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right) \right\rangle^n = \lim_{n \rightarrow \infty} \left\langle 1, \frac{\phi}{n} \right\rangle^n.$$

Summarizing, we now see that the rotation operator is an infinite repetition of small rotations:

$$\langle \cos(\phi), \sin(\phi) \rangle = \lim_{n \rightarrow \infty} \left\langle 1, \frac{\phi}{n} \right\rangle^n \quad (14.44)$$

Invoking Euler's Constant

Looking again at the complex number $\langle 1, \phi/n \rangle$, split the components via

$$\left\langle 1, \frac{\phi}{n} \right\rangle = \langle 1, 0 \rangle + \left\langle 0, \frac{\phi}{n} \right\rangle.$$

First, note that $\langle 1, 0 \rangle$ is equivalent to 1, so the complex notation can be dropped from the first term. Next, factor n^{-1} from the second term, and we have

$$\left\langle 1, \frac{\phi}{n} \right\rangle = 1 + \frac{1}{n} \langle 0, \phi \rangle.$$

Plugging this back into Equation (14.44) gives

$$\langle \cos(\phi), \sin(\phi) \rangle = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \langle 0, \phi \rangle \right)^n. \quad (14.45)$$

The right-side of the above should remind us of another notion from precalculus, namely Euler's constant, defined as:

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \right)^n,$$

or in more useful form,

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n} \right)^n$$

for any real number x .

Euler's Formula

Euler reasoned that the formula for e^x could be modified (it's *his* formula, after all) to receive complex arguments:

$$e^z = \lim_{n \rightarrow \infty} \left(1 + \frac{z}{n}\right)^n$$

If this is the case, the right side of Equation (14.45) can be written as an exponential

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n} \langle 0, \phi \rangle\right)^n = e^{\langle 0, \phi \rangle},$$

which is amazing, because it means the *left* side of Equation (14.45) is an exponential! Putting these results together, we write

$$\langle \cos(\phi), \sin(\phi) \rangle = e^{\langle 0, \phi \rangle}, \quad (14.46)$$

hailed as Euler's formula. At face value, Euler's formula is an economical way to write the rotation operator.

It's worth pausing to write Euler's formula in terms of the so-called imaginary unit, which is to associate

$$\langle 0, 1 \rangle = \sqrt{-1} = i,$$

thus the above becomes:

$$\cos(\phi) + i \sin(\phi) = e^{i\phi}$$

Setting $\phi = \pi$ reveals a remarkable connection between key players of mathematics:

$$0 = 1 + e^{i\pi}$$

4.1 Polar Form

Let us revisit the radius-and-phase construction of a complex number. For a complex number $z = \langle a, b \rangle$, recall that Equations Equation (14.29)-Equation (14.32)

$$\begin{aligned} a &= r \cos(\phi) \\ b &= r \sin(\phi) \\ r &= |z| = \sqrt{a^2 + b^2} \\ \phi &= \arctan(b/a) \end{aligned}$$

allow us to understand z as a rotation of the real number r 'up into' the complex plane by an angle ϕ :

$$z = \langle a, b \rangle = r \langle \cos(\phi), \sin(\phi) \rangle$$

With Euler's formula in hand, we may replace the trigonometry terms via Equation (14.46), and arrive at the

most elegant expression of a complex number in terms of radius and phase, the so-called polar form:

$$z = r e^{\langle 0, \phi \rangle} \quad (14.47)$$

Problem 13

Show that:

$$\cos(n \arccos(x)) = \sum_{k=0}^{n/2} \binom{n}{2k} x^{n-2k} (x^2 - 1)^k$$

Hint: Let $x = \cos(\theta)$ and arrive at

$$\cos(n\theta) = \operatorname{Re}((\cos(\theta) + i \sin(\theta))^n).$$

4.2 Multiplication and Division

Given the polar form Equation (14.47) of complex numbers, the multiplication formula Equation (14.9) and the division formula Equation (14.18) can be revised. For two complex numbers $z_1 \langle r_1, \phi_1 \rangle$, $z_2 \langle r_2, \phi_2 \rangle$, we have

$$\begin{aligned} z_1 \cdot z_2 &= r_1 r_2 e^{\langle 0, \phi_1 + \phi_2 \rangle} \\ &= r_1 r_2 \langle \cos(\phi_1 + \phi_2), \sin(\phi_1 + \phi_2) \rangle \end{aligned} \quad (14.48)$$

for multiplication, and for division:

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{r_1}{r_2} e^{\langle 0, \phi_1 - \phi_2 \rangle} \\ &= \frac{r_1}{r_2} \langle \cos(\phi_1 - \phi_2), \sin(\phi_1 - \phi_2) \rangle \end{aligned} \quad (14.49)$$

Problem 14

Derive Equations (14.48)-(14.49).

4.3 Trigonometric Functions

Euler's formula lends to a curious representation of the standard trigonometric functions. Start with Equation (14.46) and let $\phi \rightarrow -\phi$ to write

$$\langle \cos(\phi), -\sin(\phi) \rangle = e^{\langle 0, -\phi \rangle},$$

and then add both equations (also divide by two) to isolate the cosine:

$$\langle \cos(\phi), 0 \rangle = \frac{1}{2} (e^{\langle 0, \phi \rangle} + e^{\langle 0, -\phi \rangle}) \quad (14.50)$$

An analogous procedure isolates the sine:

$$\langle 0, \sin(\phi) \rangle = \frac{1}{2} (e^{\langle 0, \phi \rangle} - e^{\langle 0, -\phi \rangle}) \quad (14.51)$$

Recalling that the complex rotation operator can be written

$$z_\phi = \langle \cos(\phi), \sin(\phi) \rangle,$$

note that the above can be expressed more tightly in accordance with Equations (14.10)-(14.11):

$$\begin{aligned}\langle \cos(\phi), 0 \rangle &= \frac{z_\phi + \bar{z}_\phi}{2} \\ \langle 0, \sin(\phi) \rangle &= \frac{z_\phi - \bar{z}_\phi}{2}\end{aligned}$$

4.4 Hyperbolic Functions

Continuing the discussion of trigonometric functions, isolate the real number ϕ and apply the operator $\langle 0, 1 \rangle$, resulting in $\langle 0, \phi \rangle$ without question. Next take the complex number $\langle 0, \phi \rangle$ and apply the same operator to find

$$\langle 0, 1 \rangle \cdot \langle 0, \phi \rangle = -\phi.$$

The reason we do this is to start with Euler's formula Equation (14.46), and multiply each instance of ϕ by the complex number $\langle 0, 1 \rangle$, giving

$$\langle \cos(\langle 0, \phi \rangle), \sin(\langle 0, \phi \rangle) \rangle = e^{-\phi}.$$

A similar set of steps leads to a version with $-\phi \rightarrow \phi$, or

$$\langle \cos(\langle 0, \phi \rangle), -\sin(\langle 0, \phi \rangle) \rangle = e^{\phi}.$$

Notice the quantity on the right is a real number, which must mean

$$\langle 0, \sin(\langle 0, \phi \rangle) \rangle = \langle 0, 1 \rangle \cdot \sin(\langle 0, \phi \rangle)$$

is also real, thus $\sin(\langle 0, \phi \rangle)$ by itself is imaginary.

The above can be written:

$$\begin{aligned}\cos(\langle 0, \phi \rangle) + \langle 0, 1 \rangle \cdot \sin(\langle 0, \phi \rangle) &= e^{-\phi} \\ \cos(\langle 0, \phi \rangle) - \langle 0, 1 \rangle \cdot \sin(\langle 0, \phi \rangle) &= e^{\phi}\end{aligned}$$

Take the sum and the sine term cancels, allowing $\cos(\langle 0, \phi \rangle)$ to be isolated:

$$\cos(\langle 0, \phi \rangle) = \frac{1}{2} (e^{\phi} + e^{-\phi})$$

Similarly, take the difference and the cosine term vanishes:

$$-\langle 0, 1 \rangle \cdot \sin(\langle 0, \phi \rangle) = \frac{1}{2} (e^{\phi} - e^{-\phi})$$

It turns out that these 'complex trigonometry' terms above have special names, the *hyperbolic cosine* and *hyperbolic sine*, respectively:

$$\begin{aligned}\cosh(\phi) &= \cos(\langle 0, \phi \rangle) \\ \sinh(\phi) &= -\langle 0, 1 \rangle \cdot \sin(\langle 0, \phi \rangle)\end{aligned}$$

Explicitly:

$$\cosh(\phi) = \frac{1}{2} (e^{\phi} + e^{-\phi}) \quad (14.52)$$

$$\sinh(\phi) = \frac{1}{2} (e^{\phi} - e^{-\phi}) \quad (14.53)$$

The complex trigonometric and complex hyperbolic functions are a bit easier to remember in the imaginary unit notation. For a complex number $z = e^{i\phi}$, we found

$$\cos(\phi) = \frac{1}{2} (e^{i\phi} + e^{-i\phi})$$

$$\sin(\phi) = \frac{1}{2i} (e^{i\phi} - e^{-i\phi})$$

$$\cosh(\phi) = \frac{1}{2} (e^{\phi} + e^{-\phi})$$

$$\sinh(\phi) = \frac{1}{2} (e^{\phi} - e^{-\phi}),$$

and furthermore:

$$\begin{aligned}\cosh(\phi) &= \cos(i\phi) \\ \sinh(\phi) &= -i \sin(i\phi) \\ \cosh(i\phi) &= \cos(\phi) \\ \sinh(i\phi) &= i \sin(\phi)\end{aligned}$$

5 Roots and Branches

5.1 Complex Natural Logarithm

Start again with Euler's polar form Equation (14.47)

$$z = r e^{\langle 0, \phi \rangle},$$

and let us dissect this equation apart along a new seam. Recall that one way to adjust the 'position' of a point in the complex plane is to change the phase ϕ , but nothing special happens if the phase wanders outside of $[0 : 2\pi)$. Euler's formula has this fact built-in, as

$$\begin{aligned} e^{\langle 0, \phi \pm 2\pi n \rangle} &= \langle \cos(\phi \pm 2\pi n), \sin(\phi \pm 2\pi n) \rangle \\ &= \langle \cos(\phi), \sin(\phi) \rangle \\ &= e^{\langle 0, \phi \rangle} \end{aligned}$$

holds for any integer n .

Now things get interesting. The presence of e in Equation (14.47) beckons trying the natural logarithm ($\ln$) on both sides, resulting in:

$$\ln(z) = \ln(r) + \langle 0, \phi \rangle \quad (14.54)$$

Unlike $e^{\langle 0, \phi \rangle}$, the complex natural logarithm does not 'reset' at 2π . The phase term $\langle 0, \phi \rangle$ in Equation (14.54) is not modified by a trigonometry function, thus ϕ keeps accumulating value across $[0 : 2\pi)$. The complex natural logarithm is unique for *every* phase.

5.2 Complex Square Root

Consider the square root of $z(r, \phi)$, which by Euler's formula reads

$$z^{1/2} = \left(r e^{\langle 0, \phi \rangle} \right)^{1/2} = r^{1/2} e^{\langle 0, \phi/2 \rangle}. \quad (14.55)$$

The square root always seems to cause trouble in mathematics, and the complex square root is no exception. Examine two representations of the same point in the complex plane, for instance (r, ϕ) and $(r, \phi + 2\pi)$.

Calculating $z^{1/2}$ for each case gives

$$\begin{aligned} \sqrt{z(r, \phi)} &= \sqrt{r} e^{\langle 0, \phi/2 \rangle} \\ \sqrt{z(r, \phi + 2\pi)} &= -\sqrt{r} e^{\langle 0, \phi/2 \rangle}, \end{aligned}$$

and we see the phase causes an abrupt sign flip. By convention, the positive solution is called the *principal root*.

5.3 Complex nth Root

Consider the so-called n th root problem

$$z^n = a \quad (14.56)$$

for integer n , and complex numbers z and a . The question is, which value(s) of z make this statement true? The answer is made easy with Euler's formula Equation (14.47), where we recast each variable as

$$\begin{aligned} z &= |z| e^{\langle 0, \phi \rangle} \\ a &= |a| e^{\langle 0, \theta + 2\pi m \rangle}. \end{aligned}$$

To proceed most generally, a phase factor of $2\pi m$ is slipped into the phase of a , knowing full well this doesn't actually change its value, where m is any positive or negative integer. Rewriting Equation (14.56) gives

$$|z|^n e^{\langle 0, n\phi \rangle} = |a| e^{\langle 0, \theta + 2\pi m \rangle}.$$

Then, the radial and angular components on each side are equal:

$$\begin{aligned} |z|^n &= |a| \\ n\phi &= \theta + 2\pi m \end{aligned}$$

Solving the above for $|z|, \phi$, we have

$$|z| = |a|^{1/n} \quad (14.57)$$

$$\phi = \frac{\theta}{n} + \frac{2\pi m}{n}, \quad (14.58)$$

where

$$m = 0, 1, 2, \dots, (n-1).$$

Note that in order to stay on one branch, the integer m is restricted to produce unique solutions for ϕ . The principal root is in general defined as the solution with the greatest real component.

5.4 Complex Exponent

The same multi-value problem that arises with the complex logarithm and complex roots applies to complex exponents. Consider two complex numbers $z = \langle a, b \rangle$, $w = \langle c, d \rangle$. Starting with the polar expression $z = |z| e^{\langle 0, \phi \rangle}$, the complex exponent z^w calculation is slightly nontrivial:

$$\begin{aligned} z^w &= \left(|z| e^{\langle 0, \phi \rangle} \right)^{\langle c, d \rangle} \\ &= e^{\ln|z| \langle c, d \rangle} e^{\langle -d\phi, 0 \rangle} e^{\langle 0, c\phi \rangle} \\ z^w &= \exp \langle c \ln|z| - d\phi, d \ln|z| + c\phi \rangle \end{aligned} \quad (14.59)$$

6 Complex Functions

The complex exponential Equation (14.47), complex natural logarithm Equation (14.54), along with complex roots and exponents all qualify as *complex functions*, usually denoted $w(z)$ for complex numbers $z(x, y)$. In the general case, a complex function produces a complex number

$$w(z) = \langle u(x, y), v(x, y) \rangle, \quad (14.60)$$

having respective components $u(x, y)$, $v(x, y)$, each a real function.

6.1 Notion of Inverse

If a given function $w(z)$ can be inverted into an equation for $z(w)$, then we have

$$z(w) = \langle f(u, v), g(u, v) \rangle. \quad (14.61)$$

The functions $f(u, v)$, $g(u, v)$ contain all of the gritty details of actually inverting w .

6.2 Branch Cuts

The peculiar behavior of the complex natural logarithm Equation (14.54) and the complex square root Equation (14.55) suggest that care must be taken when ‘stepping across’ the boundary $0 \leftrightarrow 2\pi n$.

In general, discontinuity in $w(z)$ arises anywhere in the complex plane that involves a sudden abrupt jump in phase. Such ‘fault lines’ are curves called *branch cuts*.

Complex Natural Logarithm

For the natural logarithm, we choose the branch on the line $\phi = \pm\pi$ (not 2π , by convention).

The very small ‘wedge’ centering on $\phi = \pm\pi$ is where the phase of $\ln(z)$ jumps abruptly, i.e. the branch cut.

Complex Square Root

The square root function Equation (14.55) has two branches, characterized by $\pm \text{Im}(\sqrt{z})$, separated by the branch cut

$(-\infty, 0)$. Choosing the positive branch allows us to unambiguously associate $\sqrt{a^2}$ with $+a$, the principal root.

6.3 Riemann Surface

The multi-valued nature of certain complex functions cannot be completely represented in a two-dimensional plot on the complex plane. To deal with functions exhibiting branching behavior, a third dimension representing the phase of $w(z)$ is required.

With the notion of $z(w) = (f, g)$ on hand, generating three-dimensional plots called *Riemann surfaces* is a standard exercise in plotting. The parameters f, g become analogous to x and y in a standard plot, and the off-plane direction is often associated with $\text{Im}(w)$, but only by convention. It can be just as informative to plot $\text{Re}(w)$ in the third dimension.

Complex Natural Logarithm

To visualize the complex natural logarithm, start with $w(z) = \ln(z)$, easily inverted:

$$z(w) = e^w = e^{\langle u, v \rangle} = \langle e^u \cos(v), e^u \sin(v) \rangle$$

In this case, it makes sense to choose a polar parameterization with $r = e^u$:

$$X(r, v) = r \cos(v)$$

$$Y(r, v) = r \sin(v)$$

$$Z(r, v) = \ln(r) + v$$

The ‘plot variables’ are uppercase symbols X, Y, Z .

Plotting this system with $(r > 0, v \in [-3\pi : 3\pi])$ leads to Fig. 14.1. For constant r , the complex logarithm traces out a *helix* for varying ϕ . The family of all helices made this way comprise a ‘ramp’ spiraling around the origin as shown in Fig. 14.1. The resulting Riemann surface is most generally called a *manifold*.

It’s also possible to choose just one branch for two-dimensional plotting. For the complex natural logarithm, this amounts to cutting the ‘middle’ sheet from the stack in

Fig. 14.1, and flattening it down onto the complex plane as shown in Figure 14.2. (Color may vary.)

The bold-shaded lines indicate integer r, ϕ , respectively.

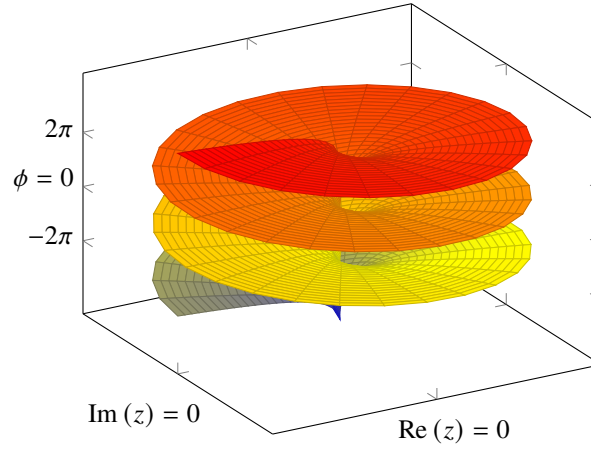


Figure 14.1. Complex natural logarithm $\ln(z) = \ln(r) + \langle 0, \phi \rangle$.

Specifically, the circle corresponds to $r = 1$ (with $r = 2$ outside of the plot), and the straight spokes correspond to

$$\phi = 0, 1, 2, 3, -3, -2, -1,$$

listing counterclockwise from $\phi = 0$.

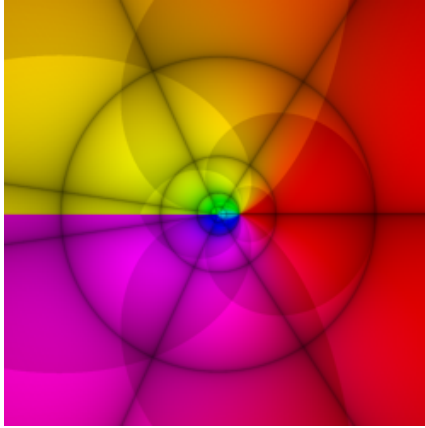


Figure 14.2. Complex logarithm on the branch $(-\pi : \pi]$.

Complex Square Root

The Riemann surface recipe also applies to plotting the square root function $w(z) = z^{1/2}$. Seeking a form like Equation (14.61), we easily write

$$z(w) = w^2 = \langle u, v \rangle \cdot \langle u, v \rangle = \langle u^2 - v^2, 2uv \rangle.$$

Turning this into a three-dimensional system, we write

$$X(u, v) = u^2 - v^2$$

$$Y(u, v) = 2uv$$

$$Z(u, v) = u,$$

while treating u, v as parameters. Plotting this system near $(u = 0, v = 0)$ leads to Fig. 14.3. The phase of $z^{1/2}$ abruptly jumps at the branch cut $\phi = \pm\pi$.

Choosing the principal root and plotting the square root in the complex plane leads to Figure 14.4. The bold-shaded lines correspond to positive integer outputs of $z^{1/2}$. Below the line $\text{Re}(z) = 0$, the sign on $\text{Im}(z)$ flips, but $\text{Re}(z)$ remains positive.

This is quickly tested by letting let $n = 2$ to find $|z| = \sqrt{|a|}$, and $\phi_0 = \theta/2$, $\phi_1 = \theta/2 + \pi$, corresponding to

two complex solutions

$$z_0 = \sqrt{|a|} e^{i\theta/2}$$

$$z_1 = -\sqrt{|a|} e^{i\theta/2}.$$

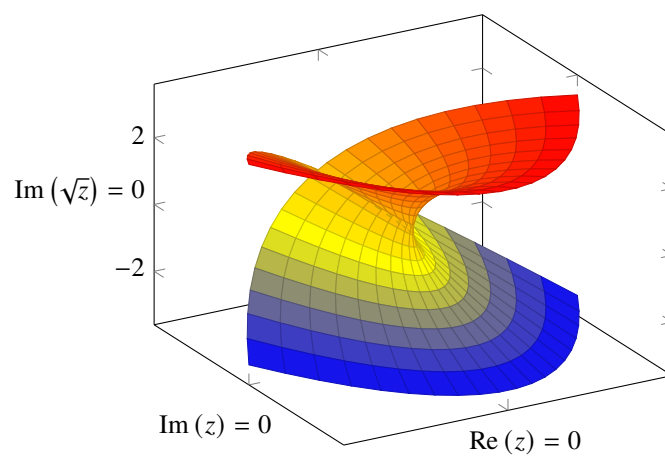


Figure 14.3. Complex square root $z^{1/2}$.

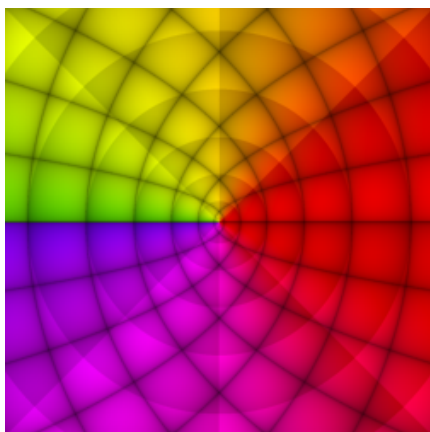


Figure 14.4. Complex square root on the branch $(-\pi; \pi]$.

Chapter 15

Linear Systems

1 Matrix Formalism

Formally, a *matrix* is a collection of numbers or variables arranged in a block with fixed rows M and columns N . Each element, i.e. *component* in the matrix requires two subscripts.

Matrix-Operator Equivalence

A primary use for a matrix is to ‘operate’ on a vector of dimension N , yielding a new vector of dimension M . (The term ‘matrix’ is often interchanged with the term ‘operator’.) Symbolically, this is written

$$\mathbf{A}\mathbf{x} = \mathbf{y} ,$$

and in full *block notation*, the same statement looks like

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2N} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{M1} & A_{M2} & A_{M3} & \cdots & A_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_M \end{bmatrix}$$

More compactly, we use *index notation* to express the same calculation:

$$\sum_{k=1}^N A_{jk} x_k = y_j \quad (15.1)$$

$$j = 1, 2, 3, \dots, M$$

Problem 1

Evaluate the product

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 4 & 0 & 5 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \\ 2 \end{bmatrix} .$$

Answer: $\mathbf{y} = \langle 3, 4, -2 \rangle$

Matrix Components

Consider two vectors $\mathbf{x}, \mathbf{y}$, each a linear combination in some N -dimensional basis such that

$$\mathbf{x} = \sum_{j=1}^N x_j \hat{\mathbf{e}}_j$$

$$\mathbf{y} = \sum_{j=1}^M y_j \hat{\mathbf{e}}_j .$$

While $\mathbf{y}$ is perfectly happy being expressed as a linear combination in the basis $\{\hat{\mathbf{e}}_j\}$, it’s instructive to re-express $\mathbf{y}$ in terms of its brother, $\mathbf{x}$. To do so, we propose an operator A such that

$$\mathbf{y} = \mathbf{A}\mathbf{x} .$$

To proceed, write the above as

$$\sum_{k=1}^M y_k \hat{\mathbf{e}}_k = \sum_{k=1}^N x_k A \hat{\mathbf{e}}_k ,$$

and multiply the basis vector $\hat{\mathbf{e}}_j$ (via dot product) into both sides:

$$\sum_{k=1}^M y_k \hat{\mathbf{e}}_j \cdot \hat{\mathbf{e}}_k = \sum_{k=1}^N x_k \hat{\mathbf{e}}_j \cdot A \hat{\mathbf{e}}_k$$

On the left, every term in the sum vanishes except that with $j = k$, and the above becomes

$$y_j = \sum_{k=1}^N (\hat{\mathbf{e}}_j \cdot A \hat{\mathbf{e}}_k) x_k$$

$$j = 1, 2, 3, \dots, M .$$

The parenthesized quantity is what we’re after:

$$A_{jk} = \hat{\mathbf{e}}_j \cdot A \hat{\mathbf{e}}_k \quad (15.2)$$

The term A_{jk} is the component of the matrix A corresponding to the j th row, k th column.

1.1 Projector

Consider the curious quantity

$$P_x = \mathbf{x} \mathbf{x} , \quad (15.3)$$

called the the *projector* of $\mathbf{x}$. By itself, P_x does nothing - there is no operation between the two copies of $\mathbf{x}$. What the projector *does* is ‘wait’ to be multiplied into another vector, resulting in a scaled version of $\mathbf{x}$. For example, applying the projector to a different vector $\mathbf{y}$ (of the same dimension as $\mathbf{x}$) goes like

$$P_x \mathbf{y} = \mathbf{x} (\mathbf{x} \cdot \mathbf{y}) .$$

1.2 Identity Operator

Consider a vector $\mathbf{x}$ as a linear combination in some N -dimensional basis:

$$\mathbf{x} = \sum_{j=1}^N x_j \hat{\mathbf{e}}_j$$

For any one of the basis vectors $\hat{\mathbf{e}}_k$, write the projector

$$P_{e_k} = \hat{\mathbf{e}}_k \hat{\mathbf{e}}_k ,$$

and then multiply $\mathbf{x}$ onto the right side to get

$$P_{e_k} \mathbf{x} = \hat{\mathbf{e}}_k \hat{\mathbf{e}}_k \cdot \mathbf{x} = x_k \hat{\mathbf{e}}_k$$

By summing over the index k , the right side is identically $\mathbf{x}$:

$$\left(\sum_{k=1}^N P_{e_k} \right) \mathbf{x} = \sum_{k=1}^N x_k \hat{\mathbf{e}}_k = \mathbf{x}$$

For the left side to also equal $\mathbf{x}$, the parenthesized quantity must be equivalent to ‘multiplying by one’, which we call the *identity* operator:

$$I = \sum_{k=1}^N P_{e_k} \quad (15.4)$$

The identity operator leaves a vector unchanged:

$$I\mathbf{x} = \mathbf{x}$$

The matrix-equivalence of I is square, has no mixed components, and has ones along the diagonal:

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad (15.5)$$

1.3 Unified Matrix Notation

Recall that a matrix A relates to its components A_{jk} in a way given by Equation (15.2), namely

$$A_{jk} = \hat{\mathbf{e}}_j \cdot A \hat{\mathbf{e}}_k.$$

To establish this directly using projectors, start with $A = IAI$ and watch what happens:

$$\begin{aligned} A &= IAI = \sum_{j=1}^M \sum_{k=1}^N P_{e_j} A P_{e_k} \\ &= \sum_{j=1}^M \sum_{k=1}^N \hat{\mathbf{e}}_j (\hat{\mathbf{e}}_j \cdot A \hat{\mathbf{e}}_k) \hat{\mathbf{e}}_k \end{aligned}$$

The parenthesized quantity is precisely A_{jk} . Evidently, the symbolic notation unifies with the index notation in the equation

$$A = \sum_{j=1}^M \sum_{k=1}^N \hat{\mathbf{e}}_j (A_{jk}) \hat{\mathbf{e}}_k \quad (15.6)$$

The presence of $\hat{\mathbf{e}}_j \hat{\mathbf{e}}_k$ is like a projector - it couples the component to the operator.

2 Matrix Operations

2.1 Matrix Addition

Two matrices A and B of identical dimensions, meaning M rows, N columns, can be combined to form a new matrix C such that

$$A + B = C,$$

or, to elaborate:

$$A_{jk} + B_{jk} = C_{jk} \quad (15.7)$$

$$\begin{cases} j = 1, 2, 3, \dots, M \\ k = 1, 2, 3, \dots, N \end{cases}$$

2.2 Scalar Multiplication

A scalar α can be multiplied into each component of a matrix A to form a new matrix B such that

$$\alpha A = B,$$

or:

$$\alpha A_{jk} = B_{jk} \quad (15.8)$$

$$\begin{cases} j = 1, 2, 3, \dots, M \\ k = 1, 2, 3, \dots, N \end{cases}$$

2.3 Matrix Multiplication

Two matrices A , B , of equal or different dimensions can be multiplied to form a new matrix C :

$$AB = C$$

The main ‘rule’ is that the number of *columns* in A must equal the number of *rows* in B :

$$A_{(M,K)} \times B_{(K,N)} = C_{(M,N)}$$

Matrix Non-Commutativity

If you’re paying attention, the commutated product BA may violate the above, and no product is defined. In any case, we should assume that the multiplication of two matrices is not commutative:

$$AB \neq BA \quad (15.9)$$

Multiplication Formula

To derive the formula for matrix multiplication, begin with the following ‘unified’ representation (15.6) of the respective matrices:

$$A = \sum_{m=1}^M \sum_{k=1}^K \hat{e}_m (A_{mk}) \hat{e}_k$$

$$B = \sum_{k'=1}^K \sum_{n=1}^N \hat{e}_{k'} (B_{k'n}) \hat{e}_n$$

Then, the product AB reads

$$AB = \sum_{m=1}^M \sum_{k=1}^K \hat{e}_m (A_{mk}) \hat{e}_k \sum_{k'=1}^K \sum_{n=1}^N \hat{e}_{k'} (B_{k'n}) \hat{e}_n$$

$$= \sum_{m=1}^M \sum_{k=1}^K \sum_{k'=1}^K \sum_{n=1}^N \hat{e}_m (A_{mk}) \hat{e}_k \hat{e}_{k'} (B_{k'n}) \hat{e}_n$$

Note that the quantity $\hat{e}_m \hat{e}_k \hat{e}_{k'} \hat{e}_n$ is the juxtaposition of two projectors, readily translating to $\hat{e}_m (\hat{e}_k \cdot \hat{e}_{k'}) \hat{e}_n$. Note further that the parenthesized product obeys

$$\hat{e}_k \cdot \hat{e}_{k'} = \begin{cases} 1 & k = k' \\ 0 & k \neq k' \end{cases},$$

which has the effect of equating $k = k'$ in the above, eliminating one of the sums. So far then, we have

$$AB = C = \sum_{m=1}^M \sum_{k=1}^K \sum_{n=1}^N (A_{mk} B_{kn}) \hat{e}_m \hat{e}_n$$

$$C = \sum_{m=1}^M \sum_{n=1}^N \left(\sum_{k=1}^K A_{mk} B_{kn} \right) \hat{e}_m \hat{e}_n.$$

The symbol C has replaced the quantity AB on the left. Comparing the right side to (15.6), we conclude that the component C_{mn} of matrix C is given by the famed *matrix multiplication* formula:

$$C_{mn} = \sum_{k=1}^K A_{mk} B_{kn} \quad (15.10)$$

$$\begin{cases} m = 1, 2, 3, \dots, M \\ n = 1, 2, 3, \dots, N \end{cases}$$

Equation (15.10) reminds that it’s only required that the number of columns in A match the number of rows in B . For

instance, the operation $A_{(2,4)} \times B_{(4,3)} = C_{(2,3)}$, explicitly written as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

is perfectly valid, whereas the commuted product $B_{(4,3)} \times A_{(2,4)}$ is undefined.

Matrix Associativity

A direct consequence of matrix multiplication is the associativity rule:

$$(AB)C = A(BC) \quad (15.11)$$

Problem 2

Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 4 & 0 & 5 \end{bmatrix}.$$

Compute A^2 . Answer:

$$A^2 = \begin{bmatrix} 7 & 4 & 11 \\ 28 & 11 & 42 \\ 20 & 4 & 29 \end{bmatrix}$$

2.4 Change of Basis

A square matrix A with components A_{jk} in the basis $\{\hat{e}_j\}$ can be represented by (15.6):

$$A = \sum_{j=1}^N \sum_{k=1}^N \hat{e}_j (A_{jk}) \hat{e}_k.$$

Under a change of basis $\{\hat{e}_j\} \rightarrow \{\hat{u}_j\}$, we can use

$$\hat{e}_j = \sum_{k=1}^N \tilde{U}_{jk} \hat{u}_k$$

to replace the unit vectors, leading to

$$A' = \sum_{m=1}^N \sum_{n=1}^N \hat{u}_m \left(\sum_{j=1}^M \sum_{k=1}^N U_{mj} A_{jk} \tilde{U}_{kn} \right) \hat{u}_n,$$

where the (first) term $\tilde{U}_{jm}$ has been replaced by U_{mj} . The parenthesized quantity is precisely the formula for the component A'_{mn} of the transformed matrix

$$A'_{mn} = \sum_{j=1}^N \sum_{k=1}^N U_{mj} A_{jk} \tilde{U}_{kn}, \quad (15.12)$$

or in symbolic form,

$$A' = U A \tilde{U}.$$

Note that the above verifies the associativity rule (15.11) for matrix multiplication. The order in which the sums are taken directly corresponds to which matrices are multiplied first. As a bonus, (15.12) tells us exactly how to take the product of three square matrices.

3 Linear Systems

3.1 Order-Two Formalism

Motivation

Consider the linear system of two equations containing two unknowns x and y ,

$$\begin{aligned} ax + by &= e \\ cx + dy &= f, \end{aligned}$$

where all coefficients are assumed nonzero. One way to solve the system begins by eliminating y , which means to multiply the top and bottom equations by d , b , respectively, and add the results:

$$x(ad - bc) = de - bf$$

Similarly, we can eliminate x to end up with

$$y(ad - bc) = af - ec,$$

and it is now trivial to solve for x and y .

If it just so happens that $ad - bc = 0$, the equations for x and y become indeterminate, meanwhile implying $de = bf$ and $af = ec$. To visualize this, treat each equation as a separate line

$$\begin{aligned} y_1 &= -\frac{a}{b}x + \frac{e}{b} \\ y_2 &= -\frac{c}{d}x + \frac{f}{d}, \end{aligned}$$

having respective slopes $m_1 = -a/b$, $m_2 = -c/d$. Take the difference of slopes to find

$$m_1 - m_2 = -\frac{a}{b} + \frac{c}{d} = \frac{1}{bd}(-ad + bc) = 0,$$

implying the lines are parallel. Moreover, $de = bf$ causes the lines to have the same y -intercept, thus the two lines y_1, y_2 are *identical*. This reduces the number of equations in the system back down to one equation that has an infinite number of solutions on the line $y_{1,2}$. In the special case that $e = 0$ or $f = 0$, the lines $y_{1,2}$ are parallel but non-overlapping, in which case no solutions exist at all.

Only when $ad - bc$ is *nonzero* does the line y_1 cross y_2 somewhere in the Cartesian plane at one point (x_0, y_0) . The intersection point is the ‘solution’ to the system of equations.

Matrix Formulation

Start with the same two-dimensional system and relabel all coefficients such that

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2, \end{aligned} \quad (15.13)$$

admitting a clean matrix representation

$$A\mathbf{x} = \mathbf{b},$$

or

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Determinant

The classification of solutions $\mathbf{x}$ depends on the quantity $A_{11}A_{22} - A_{12}A_{21}$, called the *determinant* of the matrix A :

$$\det A = \det \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A_{11}A_{22} - A_{12}A_{21} \quad (15.14)$$

If $\det A$ is nonzero, the vector $\mathbf{x}$ solves the system. On the other hand, any matrix with $\det A = 0$ is called *singular*, having a non-obvious number of solutions (zero or infinite, depending on $\mathbf{b}$). For the non-singular case $\det A \neq 0$, the solution to the system is given by:

$$\begin{aligned} x_1 &= \frac{1}{\det A} (A_{22}b_1 - A_{12}b_2) \\ x_2 &= \frac{1}{\det A} (A_{11}b_2 - A_{21}b_1) \end{aligned}$$

Cramer’s Rule

In the above solutions for x_1, x_2 , observe that the quantities

$$\begin{aligned} A_{22}b_1 - A_{12}b_2 \\ A_{11}b_2 - A_{21}b_1 \end{aligned}$$

are themselves the determinants of new matrices C_1, C_2 such that

$$\begin{aligned} x_1 &= \frac{1}{\det A} \det \begin{bmatrix} b_1 & A_{12} \\ b_2 & A_{22} \end{bmatrix} = \frac{\det C_1}{\det A} \\ x_2 &= \frac{1}{\det A} \det \begin{bmatrix} A_{11} & b_1 \\ A_{21} & b_2 \end{bmatrix} = \frac{\det C_2}{\det A} \end{aligned}$$

That is, the solution to the two-dimensional linear system (15.13) with nonzero determinant is

$$\begin{aligned} x_j &= \frac{\det C_j}{\det A} \\ j &= 1, 2, \end{aligned} \quad (15.15)$$

with

$$\begin{aligned} C_1 &= \begin{bmatrix} b_1 & A_{12} \\ b_2 & A_{22} \end{bmatrix} \\ C_2 &= \begin{bmatrix} A_{11} & b_1 \\ A_{21} & b_2 \end{bmatrix}. \end{aligned} \quad (15.16)$$

The ‘recipe’ that got us to this point is called *Cramer’s Rule*: if $\det A \neq 0$, there’s a solution to the system.

3.2 Order-N Formalism

Generalizing the 2×2 linear system to have M equations and N unknowns, we begin with

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1N}x_N &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \cdots + A_{2N}x_N &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \cdots + A_{3N}x_N &= b_3 \\ &\vdots \\ A_{M1}x_1 + A_{M2}x_2 + A_{M3}x_3 + \cdots + A_{MN}x_N &= b_M, \end{aligned} \quad (15.17)$$

admitting the matrix representation $A\mathbf{x} = \mathbf{b}$:

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2N} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{M1} & A_{M2} & A_{M3} & \cdots & A_{MN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{bmatrix} \quad (15.18)$$

At this stage, the relationship between M and N indicates the quality of solutions (if any) to the system. If $N > M$, the system is said to be *under-determined*, and there is not enough information to choose a solution. On the other hand, if $M > N$, the system is *over-determined*, and there may be zero, or perhaps infinite solutions.

In order to proceed, the matrix A is taken to be *square* with $M = N$. Then, the recipe for solving a two-dimensional linear system extrapolates to N dimensions. Although its not (yet) straightforward how to calculate the determinant of A , we can still use Cramer’s rule to write down the components of the solution vector $\mathbf{x}$, namely

$$\begin{aligned} x_j &= \frac{\det C_j}{\det A} \\ j &= 1, 2, 3, \dots, N \end{aligned} \quad (15.19)$$

The matrix C_j is constructed by starting with A and replacing the j th column with $\mathbf{b}$. That is:

$$C_j = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{1j} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & b_{2j} & \cdots & A_{2N} \\ A_{31} & A_{32} & \cdots & b_{3j} & \cdots & A_{3N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{N1} & A_{N2} & \cdots & b_{Nj} & \cdots & A_{NN} \end{bmatrix} \quad (15.20)$$

3.3 Row Operations

An important tool set called *row operations* can be applied to matrices. To illustrate these, consider again the N -dimensional linear system (15.17). The ‘game’ we play is to find ways to manipulate the system of equations without losing any information. Without much trouble, one finds the allowed operations to be:

- Exchange two rows.
- Multiply a row by a (nonzero) scalar.
- Replace a row by the sum of itself and another row.

For the sake of assigning symbols to the above row operations, let us denote row exchanges as E , scalar multiplication as M , and a row replacement as R . Using a four-dimensional square matrix as an example, row operations explicitly look like:

$$E^{23}A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix}$$

$$M_{\alpha}^2 A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ \alpha A_{31} & \alpha A_{32} & \alpha A_{33} & \alpha A_{34} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix}$$

$$R_3^2 A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{31} + A_{21} & A_{32} + A_{22} & A_{33} + A_{23} & A_{34} + A_{24} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix}$$

Note that the subscripts and superscripts on the symbols E , M , R are mere convenience of notation, often-omitted.

4 Determinant

The *determinant* is a scalar calculated from the components of a square ($N \times N$) matrix A . Of the many things the determinant can tell us, we've already seen that $\det A$ indicates the 'quality' of solutions to a linear system. Namely, if the determinant is nonzero, the linear system has a solution given by Cramer's rule. The determinant of a two-dimensional square matrix is given by Equation (15.14).

Three Dimensions

Consider the three-dimensional linear system

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 &= b_3, \end{aligned} \quad (15.21)$$

and we're interested in the determinant of the matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}.$$

Labeling each row R_1, R_2, R_3 , respectively, we deploy row operations to (i) multiply R_2 by a factor of A_{11}/A_{21} , and then (ii) replace R_2 with $R_2 - R_1$:

$$A \rightarrow \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & \frac{A_{11}A_{22}}{A_{21}} - A_{12} & \frac{A_{11}A_{23}}{A_{21}} - A_{13} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

Next, (iii) multiply R_3 by a factor of A_{11}/A_{31} , and (iv) replace R_3 with $R_3 - R_1$:

$$A \rightarrow \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & \frac{A_{11}A_{22}}{A_{21}} - A_{12} & \frac{A_{11}A_{23}}{A_{21}} - A_{13} \\ 0 & \frac{A_{11}A_{32}}{A_{31}} - A_{12} & \frac{A_{11}A_{33}}{A_{31}} - A_{13} \end{bmatrix}$$

Four Dimensions

Having witnessed the trick of performing row operations on an order-three square matrix A to condense all relevant information into a 2×2 square sub-matrix, this should also work for higher-order matrices. Indeed, the order-four matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix},$$

With the matrix configured as such, observe that the 'important' information is crammed into the lower 2×2 portion of the transformed matrix. Accordingly, we deploy the determinant formula (15.14) to write

$$\det A \propto \left(\frac{A_{11}A_{22}}{A_{21}} - A_{12} \right) \left(\frac{A_{11}A_{33}}{A_{31}} - A_{13} \right) - \left(\frac{A_{11}A_{23}}{A_{21}} - A_{13} \right) \left(\frac{A_{11}A_{32}}{A_{31}} - A_{12} \right),$$

which after simplifying, becomes

$$\begin{aligned} \det A \left(\frac{A_{21}A_{31}}{A_{11}} \right) &\propto A_{11} (A_{22}A_{33} - A_{32}A_{23}) \\ &\quad - A_{12} (A_{21}A_{33} - A_{31}A_{23}) \\ &\quad + A_{13} (A_{21}A_{32} - A_{31}A_{22}). \end{aligned}$$

Keeping in mind that the determinant of A is a single number that characterizes the solutions to the system, it follows that the right-side quantity in the above contains all of the required information. It's also easy (enough) to see that exchanging two rows in the original A will lead to the same final form of $\det A$, up to numerical factors and/or negative signs. In conclusion, we take the the order-three determinant to be

$$\begin{aligned} \det A &= A_{11} (A_{22}A_{33} - A_{32}A_{23}) \\ &\quad - A_{12} (A_{21}A_{33} - A_{31}A_{23}) \\ &\quad + A_{13} (A_{21}A_{32} - A_{31}A_{22}). \end{aligned} \quad (15.22)$$

permits a similar process to reduce the order of the problem. Doing so, the order-four determinant becomes the sum of four terms:

$$\begin{aligned} \det A = & A_{11} \det \begin{bmatrix} A_{22} & A_{23} & A_{24} \\ A_{32} & A_{33} & A_{34} \\ A_{42} & A_{43} & A_{44} \end{bmatrix} - A_{12} \det \begin{bmatrix} A_{21} & A_{23} & A_{24} \\ A_{31} & A_{33} & A_{34} \\ A_{41} & A_{43} & A_{44} \end{bmatrix} \\ & + A_{13} \det \begin{bmatrix} A_{21} & A_{22} & A_{24} \\ A_{31} & A_{32} & A_{34} \\ A_{41} & A_{42} & A_{44} \end{bmatrix} - A_{14} \det \begin{bmatrix} A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \\ A_{41} & A_{42} & A_{43} \end{bmatrix} \end{aligned} \quad (15.23)$$

4.1 Sub-Matrix and Matrix Minor

Taking another look at the three- and four-dimensional determinant formulas, we see the right side contains the sum of several ‘cross sections’ of the original matrix, each having dimension $N - 1$.

Sub-Matrix

The *sub-matrix* S_{jk} removes the j th row and the k th column from the original matrix A .

Matrix Minor

The *matrix minor*, denoted M_{jk} is the determinant of the sub-matrix S_{jk} . With matrix minor notation, equations (15.22), (15.23) can be written:

$$\begin{aligned} \det A_{(3)} &= A_{11}M_{11} - A_{12}M_{12} + A_{13}M_{13} \\ \det A_{(4)} &= A_{11}M_{11} - A_{12}M_{12} + A_{13}M_{13} - A_{14}M_{14} \end{aligned}$$

N Dimensions

Using matrix minor notation, the three and four-dimensional determinant formulas suggest of how to handle the N -dimensional case. Formally, the procedure is to use row operations to condense down-and-right all of the information on the matrix. After the dust settles, the general formula for the determinant of an order- N matrix is remarkably simple:

$$\begin{aligned} \text{if } N = 1 : \quad \det A &= \det [A_{11}] \\ \text{if } N > 1 : \quad \det A &= \sum_{j=1}^N \sum_{k=1}^N (-1)^{j+k} A_{jk} M_{jk} \end{aligned} \quad (15.24)$$

Note too that the summation can take place over rows *or* columns, meaning

$$\text{if } N > 1 : \quad \det A = \sum_{j=1}^N (-1)^{j+k} A_{kj} M_{kj}$$

also holds. Note that the variable j is fixed at some integer less than N . Only the k -variable is summed over.

Multiplication Rules

Readily shown from the properties of determinants are various multiplication rules. A scalar α multiplied by a matrix A of dimension N has the result

$$\det(\alpha A) = \alpha^N \det A.$$

Meanwhile, for the product of two matrices A, B :

$$\det(AB) = \det(A) \det(B)$$

Row Operations

The row operations E (row exchange), M (multiply by scalar), R (combine rows) have the following effect on the determinant:

$$\begin{aligned} \det(EA) &= -\det A \\ \det(MA) &= \alpha \det A \\ \det(RA) &= \det A \end{aligned}$$

Transpose Rule

Interestingly but not surprisingly, the matrix and its transpose have the same determinant:

$$\det(A^T) = \det A$$

5 Inverse Matrix

Definition

Given a square matrix A of dimension N , there *may* exist a special matrix A^{-1} that obeys the property

$$A^{-1}A = AA^{-1} = I, \quad (15.25)$$

where A^{-1} is called the *inverse matrix* to A , and I is the identity matrix of dimension N . The product of A and its inverse, or vice versa, results in the identity matrix.

Existence

For the notion of the inverse to make sense, the matrix A must perform a one-to-one mapping of a vector $\mathbf{x}$ to a vector $\mathbf{b}$ as in $A\mathbf{x} = \mathbf{b}$. By multiplying A^{-1} into both sides of $A\mathbf{x} = \mathbf{b}$, we end up with

$$A^{-1}A\mathbf{x} = A^{-1}\mathbf{b},$$

effectively ‘solving’ for the vector $\mathbf{x}$:

$$\mathbf{x} = A^{-1}\mathbf{b} \quad (15.26)$$

Formula

To come up with a formula for the inverse of A , consider another matrix B defined from the components A_{jk} such that

$$B_{jk} = (-1)^{j+k} M_{kj}, \quad (15.27)$$

where M_{kj} are the matrix minors of A . As innocent as it looks, (15.27) is quite ‘computationally expensive’, which means as N grows, it requires preposterous efforts to calculate the B -matrix by hand.

To proceed, calculate the product AB by matrix multiplication. Starting with the formula for matrix multiplication, and replacing the components of B using (15.27), we find

$$(AB)_{mn} = \sum_{k=1}^N A_{mk} B_{kn} = \sum_{k=1}^N A_{mk} (-1)^{k+n} M_{nk}.$$

In the case $m = n$, the above reduces to the formula for the determinant of A , namely (15.24). Any other case $m \neq n$ causes the right side to resolve to zero:

$$(AB)_{mn} = \begin{cases} \det A & m = n \\ 0 & m \neq n \end{cases}$$

In symbolic terms, the above reads

$$AB = (\det A) I,$$

where by comparison to (15.25), suggests the combination $B/\det A$ is equal to the inverse of A :

$$A^{-1} = \frac{1}{\det A} B \quad (15.28)$$

Products

The inverse of the product of two matrices is equal to the product of the individual inverses in reversed order:

$$(AB)^{-1} = B^{-1}A^{-1} \quad (15.29)$$

5.1 Cramer’s Rule Derived

For a linear system of N dimensions, we start with the dichotomy

$$\begin{aligned} A\mathbf{x} &= \mathbf{b} \\ \mathbf{x} &= A^{-1}\mathbf{b}, \end{aligned}$$

where accounting for (15.28), the ‘solution’ vector $\mathbf{x}$ is written

$$\mathbf{x} = \frac{1}{\det A} B\mathbf{b},$$

or using index notation,

$$x_j = \frac{1}{\det A} \sum_{k=1}^N B_{jk} b_k = \frac{1}{\det A} \sum_{k=1}^N (-1)^{j+k} M_{kj} b_k.$$

The product $M_{kj}b_k$ has the k, j indices in reverse order, indicating that the combination $M\mathbf{b}$ does *not* represent this situation. Instead, the sum constitutes the determinant of a matrix modified from A such that the k th column is replaced by $\mathbf{b}$. If this situation sounds familiar, it precisely describes the matrix introduced as Equation (15.20)

$$C_j = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{1j} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & b_{2j} & \cdots & A_{2N} \\ A_{31} & A_{32} & \cdots & b_{3j} & \cdots & A_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{N1} & A_{N2} & \cdots & b_{Nj} & \cdots & A_{NN} \end{bmatrix},$$

and the above reduces to Cramer’s rule (15.19) for the solution of the system:

$$\begin{aligned} x_j &= \frac{\det C_j}{\det A} \\ j &= 1, 2, 3, \dots, N \end{aligned}$$

Two Dimensions

Consider the two-dimensional matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

whose determinant is given by Equation (15.14). To calculate the inverse, begin with the B matrix given by Equation (15.27), coming out to

$$B = \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}.$$

Then, by the inverse formula (15.28), the inverse of the 2×2 matrix reads:

$$A^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix} \quad (15.30)$$

Three Dimensions

The three-dimensional matrix with

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

In matrix form, the above reads:

$$B = \begin{bmatrix} (A_{22}A_{33} - A_{32}A_{23}) & -(A_{12}A_{33} - A_{32}A_{13}) & (A_{12}A_{23} - A_{22}A_{13}) \\ -(A_{21}A_{33} - A_{31}A_{23}) & (A_{11}A_{33} - A_{31}A_{13}) & -(A_{11}A_{23} - A_{21}A_{13}) \\ (A_{21}A_{32} - A_{31}A_{22}) & -(A_{11}A_{32} - A_{31}A_{12}) & (A_{11}A_{22} - A_{21}A_{12}) \end{bmatrix}$$

With the matrix B fully specified in terms of A , the inverse A^{-1} is given by Equation (15.28), namely

$$A^{-1} = \frac{1}{\det A} B.$$

Note that $\det A$ was already calculated as Equation (15.22).

Problem 3

Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 4 & 0 & 5 \end{bmatrix}.$$

Compute A^{-1} . Answer:

$$A^{-1} = \frac{1}{13} \begin{bmatrix} 15 & -5 & 2 \\ 10 & 1 & -3 \\ -12 & 4 & 1 \end{bmatrix}$$

has a total of nine matrix minors M_{jk} , readily readable from A . Constructing the matrix B using

$$B_{jk} = (-1)^{j+k} M_{kj},$$

we find

$$B_{11} = (-1)^2 (A_{22}A_{33} - A_{32}A_{23})$$

$$B_{12} = (-1)^3 (A_{12}A_{33} - A_{32}A_{13})$$

$$B_{13} = (-1)^4 (A_{12}A_{23} - A_{22}A_{13})$$

$$B_{21} = (-1)^3 (A_{21}A_{33} - A_{31}A_{23})$$

$$B_{22} = (-1)^4 (A_{11}A_{33} - A_{31}A_{13})$$

$$B_{23} = (-1)^5 (A_{11}A_{23} - A_{21}A_{13})$$

$$B_{31} = (-1)^4 (A_{21}A_{32} - A_{31}A_{22})$$

$$B_{32} = (-1)^5 (A_{11}A_{32} - A_{31}A_{12})$$

$$B_{33} = (-1)^6 (A_{11}A_{22} - A_{21}A_{12})$$

Problem 4

Consider the system of equations

$$x + y + z = 5$$

$$2x + 3y + 5z = 8$$

$$4x + 5z = 2.$$

Tell the complete story of solving for x, y, z and do so.

Answer: Frame the problem in the form $A\mathbf{x} = \mathbf{b}$, namely

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 5 \\ 4 & 0 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ 8 \\ 2 \end{bmatrix},$$

implying $\mathbf{x} = A^{-1}\mathbf{b}$, and the problem becomes finding the inverse matrix A^{-1} . Proceed using the formula

$$A^{-1} = \frac{1}{\det A} B,$$

where components of matrix B are given by $B_{jk} =$ finally giving $x = 3, y = 4, z = -2$.

$(-1)^{j+k} M_{kj}$, and M_{kj} are the matrix minors of A . The matrix minor is the determinant of the kj th sub-matrix.

Given the information provided, we find $\det A = 13$ and

$$A^{-1} = \frac{1}{13} \begin{bmatrix} 15 & -5 & 2 \\ 10 & 1 & -3 \\ -12 & 4 & 1 \end{bmatrix},$$

6 Special Matrices

Transpose Matrix

Given a matrix A , there always exists the *transpose* of A , which swaps all rows for columns and vice versa. The transpose of a matrix A is denoted A^T , particularly

$$A_{jk}^T = A_{kj}. \quad (15.31)$$

Symmetric Matrix

A square matrix is said to be *symmetric* if the original matrix A is equal to the transposed matrix A^T :

$$\begin{aligned} A &= A^T \\ A_{jk} &= A_{kj} \end{aligned} \quad (15.32)$$

Anti-symmetric Matrix

A square matrix is said to be *anti-symmetric* if the original matrix A is equal to the negative transposed matrix A^T :

$$\begin{aligned} A &= -A^T \\ A_{jk} &= -A_{kj} \end{aligned} \quad (15.33)$$

This is sometimes known as *skew-symmetric*.

Orthogonal Matrix

A square matrix A whose transpose A^T is equal to the inverse A^{-1} is called an *orthogonal matrix*:

$$A^T = A^{-1} \quad (15.34)$$

Role of Row Operations

A general $M \times N$ matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} \end{bmatrix}$$

can be *reduced* by any operations E, M, R to produce a different matrix A' that contains the same information or similar information to A . This process can be applied sequentially to achieve various matrix forms cataloged below.

6.1 Triangular Forms

Square matrices with $M = N$ admit two special reduced forms called *triangular forms*.

Upper Triangular Form

If (by row operations or otherwise) a square matrix has $A_{jk} = 0$ when $j > k$, the form is called *upper triangular*:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \cdots & A_{2n} \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & A_{nn} \end{bmatrix} \quad (15.35)$$

Lower Triangular Form

If a square matrix has $A_{jk} = 0$ when $j < k$, the form is called *lower triangular*:

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 \\ \cdots & \cdots & \cdots & 0 \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \quad (15.36)$$

6.2 Diagonal Form

If a square matrix has $A_{jk} = 0$ when $j \neq k$, the form is called *diagonal*:

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & A_{nn} \end{bmatrix} \quad (15.37)$$

For any triangular or diagonal matrix A , the determinant is equal to the product of its diagonal entries:

$$\det A = A_{11}A_{22} \cdots A_{NN} = \prod_{j=1}^N A_{jj}$$

6.3 Augmented Matrix

For linear systems characterized by $A\mathbf{x} = \mathbf{b}$, where A is an $M \times N$ matrix, we can construct the *augmented* matrix by appending the components of $\mathbf{b}$ as an extra column:

$$A|\mathbf{b} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} & b_1 \\ A_{21} & A_{22} & \cdots & A_{2N} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} & b_M \end{bmatrix} \quad (15.38)$$

In the general case, $\mathbf{b}$ can be replaced with any matrix with M rows.

6.4 Row-Reduced Echelon Formthis is called the *row-reduced echelon form*.

If (by any means) a square matrix and a vector $\mathbf{x}$ can be written as

$$I|\mathbf{x} = \begin{bmatrix} 1 & 0 & \cdots & 0 & x_1 \\ 0 & 1 & \cdots & 0 & x_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & x_N \end{bmatrix}, \quad (15.39)$$

7 Elimination

Linear System

Consider an N -dimensional linear system $Ax = b$, represented by the $M = N$ -case of (15.17):

$$A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1N}x_N = b_1$$

$$A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \cdots + A_{2N}x_N = b_2$$

$$A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \cdots + A_{3N}x_N = b_3$$

$\cdots$

$$A_{N1}x_1 + A_{N2}x_2 + A_{N3}x_3 + \cdots + A_{NN}x_N = b_N$$

Equivalently, the above is represented by an augmented matrix $A|b$ of the form (15.38). Next, imagine having done all of the hard work to solve the system

$$x_1 + 0 + 0 + \cdots + 0 = x_1$$

$$0 + x_2 + 0 + \cdots + 0 = x_2$$

$$0 + 0 + x_3 + \cdots + 0 = x_3$$

$\cdots$

$$0 + 0 + 0 + \cdots + x_N = x_N ,$$

which appears like a tautological thing to write, but is in fact the row-reduced echelon form, $I|x$ cataloged as Equation (15.39). Written this way, the solutions x_j to the system are readily exportable as the right side of each equation.

The natural question is, how can we start with $A|b$ and somehow end up with $I|x$ using matrix trickery? The answer is called *elimination*, which is a sequence of row operations E, M, R that we carry out on the augmented matrix $A|b$ to bring it the form $I|x$. Representing the exact sequence of row operations as one ‘operator’ $\tilde{O}(E, M, R)$ or simply $\tilde{O}$, one writes

$$\tilde{O}(A|b) = I|x. \quad (15.40)$$

One may think of $\tilde{O}$ as a sequential list of procedures to carry out on $A|b$, much as a program receives input and returns output.

Example

Consider a linear system represented by the augmented matrix

$$A|b = \left[\begin{array}{cccc} 1 & 1 & 1 & 5 \\ 2 & 3 & 5 & 8 \\ 4 & 0 & 5 & 2 \end{array} \right].$$

Denoting the rows of $A|b$ as R_j , the first three operations may go as follows: (i) Subtract $2R_1$ from R_2 . (ii) Subtract $4R_1$ from R_3 . (iii) Add $4R_2$ to R_3 .

$$A|b \xrightarrow{(i)} \left[\begin{array}{cccc} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 4 & 0 & 5 & 2 \end{array} \right] \xrightarrow{(ii)} \left[\begin{array}{cccc} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 0 & -4 & 1 & -18 \end{array} \right] \xrightarrow{(iii)} \left[\begin{array}{cccc} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 13 & -26 \end{array} \right]$$

Note the new matrix has zeros down and left of the diagonal, i.e. upper triangular form. Don’t stop here though: (iv) Divide R_3 by 13 and subtract $3R_3$ from R_2 . (v) Subtract R_3 from R_1 . (vi) Subtract R_2 from R_1 .

$$\xrightarrow{(iv)} \left[\begin{array}{cccc} 1 & 1 & 1 & 5 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{array} \right] \xrightarrow{(v)} \left[\begin{array}{cccc} 1 & 1 & 0 & 7 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{array} \right] \xrightarrow{(vi)} \left[\begin{array}{cccc} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{array} \right] = I|x$$

Elimination halts when the ‘matrix part’ of the above reduces to the identity. Reading off the right-hand column, we see the solution to the system of equations is

$$\begin{aligned}x_1 &= 3 \\x_2 &= 4 \\x_3 &= -2.\end{aligned}$$

Reconciling what just happened with Equation (15.40), we see the operator $\tilde{O}$ is comprised of steps (i)-(vi), each being one particular E , M , R operation.

Matrix Inverse

Looking again at Equation (15.40), note that the sequence of row operations $\tilde{O}$ applies to A and $\mathbf{b}$ separately:

$$\begin{aligned}\tilde{O}A &= I \\ \tilde{O}\mathbf{b} &= \mathbf{x}\end{aligned}$$

While $\tilde{O}$ is not established as a matrix, it does precisely same job as A^{-1} , and must contain the same information as A^{-1} . As a point of comparison, note the similarity between the above versus familiar relations

$$\begin{aligned}A^{-1}A &= I \\ A^{-1}\mathbf{b} &= \mathbf{x}.\end{aligned}$$

Going with the hunch that $\tilde{O}$ can be treated as an operator that obeys the associativity rule of matrix multiplication, we would be able to do the following:

$$\begin{aligned}\tilde{O}A &= I \\ (\tilde{O}A)A^{-1} &= IA^{-1} \\ \tilde{O}(AA^{-1}) &= A^{-1} \\ \tilde{O}I &= A^{-1}\end{aligned}$$

Once again, we see the sequence $\tilde{O}$ is doing the same job as A^{-1} . Rounding up the circumstantial evidence, we see the set of steps $\tilde{O}$ that carries $A \rightarrow I$ is the *same* set of steps that carries $I \rightarrow A^{-1}$. In the language of augmented matrices, this is summarized by

$$\tilde{O}(A|I) = I|A^{-1}. \quad (15.41)$$

This conspiracy of mathematics is otherwise known as *Gauss-Jordan elimination*.

Two Dimensions

Demonstrating on a 2×2 matrix, begin with $A|I$ as

$$A|I = \left[\begin{array}{cc|cc} A_{11} & A_{12} & 1 & 0 \\ A_{21} & A_{22} & 0 & 1 \end{array} \right],$$

and perform row operations until form $I|A^{-1}$ is attained. In brief detail, the augmented matrix develops as:

$$\begin{aligned}A|I &\rightarrow \left[\begin{array}{cc|cc} A_{12}A_{21} - A_{22}A_{11} & 0 & -A_{22} & A_{21} \\ A_{21} & A_{22} & 0 & 1 \end{array} \right] \\ A|I &\rightarrow \frac{1}{\det A} \left[\begin{array}{cc|cc} 1 & 0 & -A_{22} & A_{21} \\ 0 & 1 & -A_{21} & A_{11} \end{array} \right]\end{aligned}$$

The final result is none other than Equation (15.30), the formula for the inverse of a 2×2 square matrix:

$$A^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \left[\begin{array}{cc} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{array} \right]$$

8 Eigenvectors and Eigenvalues

An important situation that arises in mathematics and physics is the so-called *eigenvalue problem*

$$A\mathbf{u} = \lambda\mathbf{u} . \quad (15.42)$$

The matrix A is taken to be square and N -dimensional. The vectors $\mathbf{u}^{(j)}$ that satisfy (15.42) are called *eigenvectors*, and the corresponding scalar $\lambda^{(j)}$ is called an *eigenvalue*.

8.1 Calculating Eigenvalues

The eigenvalue problem (15.42) can be equivalently framed as

$$(A - \lambda I)\mathbf{u} = 0 , \quad (15.43)$$

where I is the identity matrix to match the dimension of A .

Two Dimensions

Taking a two-dimensional case as an example, we have

$$A - \lambda I = \begin{bmatrix} A_{11} - \lambda & A_{12} \\ A_{21} & A_{22} - \lambda \end{bmatrix} ,$$

which, as a set of equations, looks like

$$\begin{aligned} (A_{11} - \lambda)x_1 &= -A_{12}x_2 \\ (A_{22} - \lambda)x_2 &= -A_{12}x_1 \end{aligned}$$

Multiply the pair of equations and cancel the product x_1x_2 to get

$$(A_{11} - \lambda)(A_{22} - \lambda) - A_{12}A_{21} = 0 . \quad (15.44)$$

The only unknown in the equation is λ , which can be isolated using the quadratic formula:

$$\lambda_{\pm} = \frac{A_{11} + A_{22}}{2} \pm \frac{1}{2}\sqrt{(A_{11} - A_{22})^2 + 4A_{12}A_{21}} \quad (15.45)$$

Note there are two solutions for λ , which we label λ_+ , and λ_- , respectively.

8.2 Characteristic Equation

When confronted with the eigenvalue problem (15.43), the first order of business, usually, is to calculate the eigenvalues λ . As we've seen for the two-dimensional case, this process boiled down to Equation (15.44). Pausing on this result for a moment, note that a quicker way to get there is to write

$$\det(A - \lambda I) = 0 , \quad (15.46)$$

which is in fact true in any number of dimensions. Equation (15.46) is called the *characteristic equation* of the system.

Characteristic Polynomial

The characteristic equation always 'simplifies' to the *characteristic polynomial*, a single equation embedding λ :

$$P_N(\lambda) = C_0 + C_1\lambda + C_2\lambda^2 + \cdots + C_N\lambda^N = 0 \quad (15.47)$$

The characteristic polynomial is suggestive of the *fundamental theorem of algebra*, stating that there are exactly N (complex) roots of a polynomial of degree N .

8.3 Calculating Eigenvectors

Once the eigenvalues λ are known, the components of each eigenvector $\mathbf{u}$ are readily calculated directly from

$$\begin{aligned} A\mathbf{u}^{(j)} &= \lambda_j\mathbf{u}^{(j)} \\ j &= 1, 2, 3, \dots, N . \end{aligned}$$

Two Dimensions

Developing the eigenvalue problem in two dimensions, there are two eigenvalues $\lambda_{\pm}$ given by Equation (15.45), and let us label the two corresponding eigenvectors $\mathbf{u}$, $\mathbf{v}$ such that

$$\begin{aligned} A\mathbf{u} &= \lambda_+\mathbf{u} \\ A\mathbf{v} &= \lambda_-\mathbf{v} . \end{aligned}$$

Working with the left equation first, it expands into two equations

$$\begin{aligned} A_{11}u_1 + A_{12}u_2 &= \lambda_+u_1 \\ A_{12}u_1 + A_{22}u_2 &= \lambda_+u_2 , \end{aligned}$$

which gives us two ways to solve for the ratio u_1/u_2 :

$$\begin{aligned} \frac{u_1}{u_2} &= \frac{-A_{12}}{A_{11} - \lambda_+} \\ \frac{u_1}{u_2} &= \frac{-(A_{22} - \lambda_+)}{A_{21}} \end{aligned} \quad (15.48)$$

As a sanity check, we may eliminate the ratio u_1/u_2 and recover the characteristic Equation (15.44). A similar set of steps isolates the ratio v_1/v_2 for the second eigenvalue/eigenvector

$$\frac{v_1}{v_2} = \frac{-A_{12}}{A_{11} - \lambda_-} \quad (15.49)$$

$$\frac{v_1}{v_2} = \frac{-(A_{22} - \lambda_-)}{A_{21}} , \quad (15.50)$$

which also combine to reproduce the characteristic equation, so we're on the right track.

Symmetric Matrix

Suppose the matrix A is given as

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}.$$

The eigenvalues of A are given by Equation (15.45), and simplify very nicely:

$$\lambda_{\pm} = a \pm b$$

Denoting the respective eigenvectors $\mathbf{u}$, $\mathbf{v}$, we apply (15.48) directly to find

$$\frac{u_1}{u_2} = \frac{-b}{-b} = 1.$$

Meanwhile, (15.49) similarly tells us

$$\frac{v_1}{v_2} = \frac{-b}{b} = -1,$$

and we're done. Evidently, the two eigenvectors are

$$\begin{aligned} \mathbf{u} &= \langle 1, 1 \rangle \\ \mathbf{v} &= \langle 1, -1 \rangle, \end{aligned}$$

or in normalized form,

$$\begin{aligned} \hat{\mathbf{u}} &= \frac{1}{\sqrt{2}} \langle 1, 1 \rangle \\ \hat{\mathbf{v}} &= \frac{1}{\sqrt{2}} \langle 1, -1 \rangle. \end{aligned}$$

Hermitian Matrix

For the Hermitian matrix

$$A = \begin{bmatrix} a & -ib \\ ib & a \end{bmatrix},$$

the characteristic equation is

$$(a - \lambda) + i^2 b^2 = 0,$$

or

$$\lambda_{\pm} = a \mp b.$$

Despite having complex components, the eigenvalues are real-valued.

Complex Eigenvalues

Modifying the above example, consider

$$A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}.$$

Following the same steps, we find the eigenvalues to be complex:

$$\lambda_{\pm} = a \pm ib$$

This is no hindrance, however. The complexity passes to the eigenvectors, which turn out to be:

$$\begin{aligned} \hat{\mathbf{u}} &= \frac{1}{\sqrt{2}} \langle 1, i \rangle \\ \hat{\mathbf{v}} &= \frac{1}{\sqrt{2}} \langle 1, -i \rangle \end{aligned}$$

8.4 Diagonalization

For the eigenvalue problem (15.42)

$$A\mathbf{u} = \lambda\mathbf{u}$$

of dimension N , suppose we already have the list of N eigenvalues λ and N eigenvectors $\mathbf{u}$.

Modal Matrix

It's instructive to condense all of the eigenvector information into a new object called the *modal matrix*, denoted C , whose j th column is comprised of the components of the j th eigenvector:

$$\begin{aligned} C &= \begin{bmatrix} u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(N)} \\ u_2^{(1)} & u_2^{(2)} & \cdots & u_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ u_N^{(1)} & u_N^{(2)} & \cdots & u_N^{(N)} \end{bmatrix} \\ &= [\mathbf{u}^{(1)} \quad \mathbf{u}^{(2)} \quad \cdots \quad \mathbf{u}^{(N)}] \end{aligned} \quad (15.51)$$

Then, the matrix product AC can be written

$$\begin{aligned} AC &= [\lambda_1 \mathbf{u}^{(1)} \quad \lambda_2 \mathbf{u}^{(2)} \quad \cdots \quad \lambda_N \mathbf{u}^{(N)}] \\ &= \begin{bmatrix} \lambda_1 u_1^{(1)} & \lambda_2 u_1^{(2)} & \cdots & \lambda_N u_1^{(N)} \\ \lambda_1 u_2^{(1)} & \lambda_2 u_2^{(2)} & \cdots & \lambda_N u_2^{(N)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 u_N^{(1)} & \lambda_2 u_N^{(2)} & \cdots & \lambda_N u_N^{(N)} \end{bmatrix}. \end{aligned}$$

Diagonal Matrix

The product AC , especially in matrix form, looks like the product of C with another, much simpler matrix. Consider a *diagonal* matrix Λ (Greek uppercase lambda) whose off-diagonal entries are all zero, and the eigenvalues occupy the diagonal:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \quad (15.52)$$

Indeed, the right result of AC is reproduced by the product $C\Lambda$, meaning the matrix products are equal:

$$AC = C\Lambda$$

Supposing the inverse of C can be attained, the diagonal matrix Λ can be isolated:

$$\Lambda = C^{-1}AC \quad (15.53)$$

The process of attaining Λ is called the *diagonalization* of the matrix A . If the columns of C happen to form an orthonormal basis, the inverse matrix C^{-1} may be replaced with its transpose C^T .

8.5 Hermitian Matrix

Eigenvectors as a Basis

It is no coincidence that a system of N dimensions has N eigenvectors. It makes sense to wonder if an arbitrary linear combination can be expressed via the change-of-basis formula for vectors.

$$\mathbf{V} = \sum_{j=1}^N V_j \hat{\mathbf{e}}_j \xrightarrow{?} (\mathbf{V})' = \sum_{j=1}^N V'_j \hat{\mathbf{u}}^{(j)}$$

In the above, the eigenvectors are assumed to be normalized (unit magnitude), which is always possible for nonzero vectors. However, we are *not* to assume that the eigenvectors $\{\hat{\mathbf{u}}^{(j)}\}$ form an orthogonal basis. That is, it's not always the case that any two eigenvectors are orthogonal.

Consider two solutions to the eigenvalue problem (15.42),

$$\begin{aligned} A\mathbf{u}^{(j)} &= \lambda_j \mathbf{u}^{(j)} \\ A\mathbf{u}^{(k)} &= \lambda_k \mathbf{u}^{(k)}, \end{aligned}$$

and multiply $\mathbf{u}^{(k)}$, $\mathbf{u}^{(j)}$, onto the left and right sides respectively into each:

$$\begin{aligned} \mathbf{u}^{(k)} \cdot A\mathbf{u}^{(j)} &= \lambda_j \mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} \\ A\mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} &= \lambda_k \mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} \end{aligned}$$

Looking at the left side of each equation, it appears as if

$$\mathbf{u}^{(k)} \cdot A\mathbf{u}^{(j)} = A\mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} \quad (15.54)$$

wants to be true, but simply isn't in the general case. The special case that satisfies (15.54) is called a *Hermitian* matrix.

Non-equal Eigenvalues

Pursuing the case where A is Hermitian, the above condenses to:

$$\lambda_j \mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} = \lambda_k \mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)}$$

Now, if we assume that $\lambda_j \neq \lambda_k$, the *only* way to reconcile this result is that *non-equal eigenvectors of a Hermitian matrix are orthogonal*:

$$\mathbf{u}^{(k)} \cdot \mathbf{u}^{(j)} = 0$$

Just as importantly, this reinforces that the eigenvectors of a non-Hermitian matrix may not be orthogonal.

Equal Eigenvalues

If m of the N eigenvalues are equal, one speaks of *m-fold degeneracy*. In this case, the corresponding eigenvectors form a *vector subspace* of the original vector space that might admit its own orthonormal basis.

9 Degenerate Systems

Concerning the eigenvalue problem (15.42)

$$A\mathbf{u} = \lambda\mathbf{u},$$

it could turn out that two eigenvalues λ_j, λ_k are equal, in which case we *may* be able to construct N unique eigenvectors $\mathbf{u}^{(j)}$, but not always. Specifically, for each repeated eigenvalue λ_j of multiplicity m_j , there must be m_j linearly independent eigenvectors. The ability to successfully do this depends on the system on hand.

Dead-end Case

Consider the matrix

$$A = \begin{bmatrix} -2 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix},$$

having a characteristic polynomial

$$(-2 - \lambda)(1 - \lambda)(-2 - \lambda) = 0.$$

Evidently we find three eigenvalues, with two identical:

$$\begin{aligned}\lambda_1 &= 1 \\ \lambda_2 &= -2 \\ \lambda_3 &= -2\end{aligned}$$

Handling the easy case first, the eigenvector corresponding to λ_1 is calculated from $A\mathbf{u} = 1\mathbf{u}$, resulting in

$$\mathbf{u}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Proceeding to the repeated eigenvalue case, we solve $A\mathbf{u} = -2\mathbf{u}$ to get a single eigenvector

$$\mathbf{u}^{(2,3)} = \frac{1}{\sqrt{10}} \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix}.$$

Note that $\mathbf{u}^{(1)}$ is linearly independent from $\mathbf{u}^{(2,3)}$, but not orthogonal. Since there is no obvious way to ‘peel apart’ the eigenvectors $\mathbf{u}^{(2,3)}$, the show stops here. The matrix A cannot be diagonalized.

Salvageable Case

Consider the matrix

$$A = \begin{bmatrix} 5 & -4 & 4 \\ 12 & -11 & 12 \\ 4 & -4 & 5 \end{bmatrix},$$

having a characteristic polynomial

$$0 = \lambda^3 + \lambda^2 - 5\lambda + 3,$$

which factors into

$$0 = (\lambda - 1)(\lambda - 1)(\lambda + 3).$$

We have three eigenvalues, with two identical:

$$\begin{aligned}\lambda_1 &= 1 \\ \lambda_2 &= 1 \\ \lambda_3 &= -3\end{aligned}$$

Handling the easy case first, the eigenvector corresponding to λ_3 is calculated from $A\mathbf{u} = -3\mathbf{u}$, leading to the relations

$$\begin{aligned}2u_1 - u_2 + u_3 &= 0 \\ u_1 - u_2 + 2u_3 &= 0 \\ 3u_1 - 2u_2 + 3u_3 &= 0,\end{aligned}$$

telling us the corresponding eigenvector is

$$\mathbf{u}^{(3)} = \frac{1}{\sqrt{11}} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix},$$

or any multiple.

Proceeding to the repeated eigenvalue case, we solve $A\mathbf{u} = \mathbf{u}$ to generate three copies of

$$u_1 - u_2 + u_3 = 0.$$

With one equation and three unknowns, we may choose any *two* values to be arbitrary. For instance, we may choose $u_1 = 1$ with $u_2 = 0$, causing $u_3 = -1$. We then construct an eigenvector from these numbers:

$$\mathbf{u}^{(1)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

On the other hand, we may choose $u_1 = 0, u_2 = 1$, causing $u_3 = 1$, to create another eigenvector, linearly independent from the others:

$$\mathbf{u}^{(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

With three eigenvectors in hand, a modal matrix can be defined such that

$$C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 3 \\ -1 & 1 & 1 \end{bmatrix},$$

allowing the matrix A to be diagonalized using $\Lambda = C^{-1}AC$.

Normalizable Case

Consider the matrix

$$A = \begin{bmatrix} 1 & 0 & \sqrt{2} \\ 0 & 2 & 0 \\ \sqrt{2} & 0 & 0 \end{bmatrix},$$

having a characteristic polynomial

$$0 = (2 - \lambda)(-\lambda^2 + \lambda + 2),$$

indicating three eigenvalues, with two identical:

$$\lambda_1 = 2$$

$$\lambda_2 = 2$$

$$\lambda_3 = -1$$

Handling the easy case first, the eigenvector corresponding to λ_3 is calculated from $A\mathbf{x} = -\mathbf{x}$, leading to

$$\mathbf{u}^{(3)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 0 \\ -\sqrt{2} \end{bmatrix}.$$

Proceeding to the repeated eigenvalue case, we solve $A\mathbf{u} = 2\mathbf{u}$ to get a single eigenvector

$$\mathbf{u}^{(1,2)} = \frac{1}{\sqrt{3\alpha^2/2 + \beta^2}} \begin{bmatrix} \alpha \\ \beta \\ \alpha/\sqrt{2} \end{bmatrix},$$

for two arbitrary constants α, β . The aim here is to tease two mutually orthogonal eigenvectors from the above, which means to require

$$\mathbf{u}^{(1)} \cdot \mathbf{u}^{(2)} = 0.$$

This amounts to finding pairs of α_j, β_j that satisfy

$$\frac{3}{2}\alpha_1\alpha_2 + \beta_1\beta_2 = 0.$$

Choosing $\alpha_1 = 0$ begins a fast avalanche that requires $\beta_1 = 1$, and also $\beta_2 = 0$, with α_2 remaining arbitrary. The remaining eigenvectors therefore read

$$\mathbf{u}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{u}^{(2)} = \frac{1}{\sqrt{3/2}} \begin{bmatrix} 1 \\ 0 \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \sqrt{2/3} \\ 0 \\ 1/\sqrt{3} \end{bmatrix}.$$

With three eigenvectors in hand, a modal matrix can be defined such that

$$C = \begin{bmatrix} 1 & 0 & \sqrt{2/3} \\ 0 & 1 & 0 \\ -\sqrt{2} & 0 & 1/\sqrt{3} \end{bmatrix},$$

allowing the matrix A to be diagonalized via $\Lambda = C^{-1}AC$. However, since the set of eigenvectors $\{\mathbf{u}^{(j)}\}$ form an orthonormal basis, we may further simplify the above using $C^{-1} = C^T$.

Part III

Calculus

Chapter 16

Differential Calculus

1 Slope at a Point

The notion of ‘rise over run’, which applies so naturally to straight lines, also applies to curves. Recall that for a line $y = mx + b$, the rise over run calculation $\Delta y / \Delta x$ always yields the same number m , the slope of the line, and the whole line has just one slope.

1.1 Derivative

On a curve, there is no single value m that characterizes the slope, but we can talk about the ‘local’ slope near a point. To illustrate, choose any point x_0 on a curve and begin ‘zooming in’ so the curve appears straighter and straighter until indistinguishable from a line. The slope of that line is the slope of the function in the place we’ve zoomed in.

This idea of *slope at a point* is also called the *derivative*. If the function is $f(x)$, the derivative is written $f'(x)$. A synonym for $f'(x)$ is written df/dx , called *Leibniz notation*. It’s slightly clearer than the f' notation, as df/dx is the ratio ‘differential f over differential x ’, which is the infinitesimal limit of $\Delta f / \Delta x$, or similarly $\Delta y / \Delta x$.

Yet another way to denote the derivative is to slip a parenthesized 1 next to f , i.e. $f^{(1)}(x)$. All of these expressions for ‘slope at a point’ are interchanged freely in greater literature:

$$\text{Slope at a point} = f'(x) = \frac{d}{dx} f(x) = f^{(1)}(x)$$

Definition

The formal definition of the derivative of $f(x)$ at any point x_0 is given as a limit:

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad (16.1)$$

Letting $h = x - x_0$, the definition can be written

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h},$$

which is in all ways the same as the above. This form is more common to your standard Calc 101 textbook.

Differentiable Functions

The definition of the derivative inevitably involves limits, thus all of the baggage pertaining to continuity, smoothness, etc. must become relevant.

When a curve is ‘well-behaved’, which is to say continuous and smooth, the function is *differentiable*, which means the definition of the derivative can be applied and returns useful information.

Things get woolly with the derivative at or across a discontinuity.

1.2 Elementary Derivatives

The definition of the derivative can be directly used on any differentiable function. While there are plenty of extra rules and shortcuts to make calculations easier, we’ll settle down a while and calculate a volley of derivatives the hard way.

Parabola

The easiest and most illustrative nontrivial derivative is the parabola $f(x) = x^2$. From the definition, we have

$$f'(x) = \lim_{h \rightarrow 0} \frac{(x+h)^2 - x^2}{h},$$

simplifying down to

$$f'(x) = \lim_{h \rightarrow 0} 2x + h = 2x.$$

That is, the slope on a parabola at point x is $2x$.

The same result comes from the formula that uses x_0 instead of h :

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{x^2 - x_0^2}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \frac{(x + x_0)(\cancel{x - x_0})}{\cancel{x - x_0}} = 2x_0 \end{aligned}$$

Whole Number Powers

Generalizing the parabolic case, consider the function with x raised to an arbitrary (whole) power n , $f(x) = x^n$. Using the definition, the derivative of $f(x)$ is

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{x^n - x_0^n}{x - x_0}.$$

To gain on this, recall an important identity attainable from polynomial division, namely

$$x^n - a^n = (x - a) \left(\sum_{k=1}^n a^{k-1} x^{n-k} \right).$$

Letting $a = x_0$ while letting $x \rightarrow x_0$ inside the sum, this reads

$$x^n - x_0^n = (x - x_0) \left(\sum_{k=1}^n x_0^{n-1} \right).$$

The final sum is n copies of the quantity x_0^{n-1} , and the term $x - x_0$ can be divided off to the left side:

$$\frac{x^n - x_0^n}{x - x_0} = nx_0^{n-1}$$

This is exactly what the definition of $f'(x_0)$ is asking for:

$$f'(x_0) = nx_0^{n-1}$$

For the sake of stating the function and the derivative on the same line, the result can be written:

$$\frac{d}{dx} (x^n) = nx^{n-1} \quad (16.2)$$

Reciprocal

For the reciprocal function $f(x) = 1/x$ we have

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{1/x - 1/x_0}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \frac{\cancel{-(x - x_0)}}{xx_0(\cancel{x - x_0})} = \frac{-1}{x_0^2}. \end{aligned}$$

That is, the slope of the reciprocal function at a point x_0 is $-1/x_0^2$. In summary:

$$\frac{d}{dx} \left(\frac{1}{x} \right) = \frac{-1}{x^2} \quad (16.3)$$

As an exercise, perhaps just a mental one, it's straightforward to show that a horizontally-shifted reciprocal function obeys:

$$\frac{d}{dx} \left(\frac{1}{x+a} \right) = \frac{-1}{(x+a)^2} \quad (16.4)$$

Inverse Square

The inverse square function $f(x)$ can be dealt with using Equation (16.2), but we'll suffer the brute force approach:

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{1/x^2 - 1/x_0^2}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \frac{-(x + x_0)(\cancel{x - x_0})}{x^2 x_0^2 (\cancel{x - x_0})} = \frac{-2}{x_0^3} \end{aligned}$$

Like the previous few cases, the denominator is eliminated by algebra and the derivative of the function becomes clear:

$$\frac{d}{dx} \left(\frac{1}{x^2} \right) = \frac{-2}{x^3} \quad (16.5)$$

To go with this is the shifted version

$$\frac{d}{dx} \left(\frac{1}{(x+a)^2} \right) = \frac{-2}{(x+a)^3}, \quad (16.6)$$

where a is a constant.

Square Root

The derivative of the square root $f(x) = \sqrt{x}$ is also straightforward, as:

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{\sqrt{x} - \sqrt{x_0}}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \frac{(\sqrt{x} - \sqrt{x_0})}{(\sqrt{x} - \sqrt{x_0})(\sqrt{x} + \sqrt{x_0})} = \frac{1}{2\sqrt{x_0}} \end{aligned}$$

In one line, this is:

$$\frac{d}{dx} (\sqrt{x}) = \frac{1}{2\sqrt{x}} \quad (16.7)$$

1.3 Exponential Derivatives

Standard Exponential

The exponential function $f(x) = n^x$ is a bit more tricky. Following traditional setup, we have

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{n^x - n^{x_0}}{x - x_0} = n^{x_0} \lim_{x \rightarrow x_0} \frac{n^{x-x_0} - 1}{x - x_0},$$

where the term n^{x_0} can be factored outside of the limit.

Substituting $h = x - x_0$, the remaining limit becomes

$$f'(x_0) = n^{x_0} \lim_{h \rightarrow 0} \frac{n^h - 1}{h},$$

is decidedly equivalent to the natural log of n . In summary:

$$\frac{d}{dx} (n^x) = n^x \ln(n) \quad (16.8)$$

Natural Exponential

A special case of Equation (16.8) has $n = e$, as in Euler's e , which gets rid of the $\ln$ -term because $\ln(e) = 1$. This tells us

$$\frac{d}{dx} (e^x) = e^x, \quad (16.9)$$

meaning $f(x) = e^x$ is its own derivative.

By making repeated use of Equation (16.2) for handling powers, one can show easily that another way to express e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \quad (16.10)$$

It's worth checking that the definition for e as an infinite limit, i.e. Equation

$$e = \lim_{h \rightarrow \infty} \left(1 + \frac{1}{h}\right)^h$$

can be derived from the definition of the derivative. Supposing we know nothing of e for a moment, consider a function $f(x) = E^x$ that is named to foreshadow the answer. The derivative of f reads

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{E^x - E^{x_0}}{x - x_0},$$

and E^{x_0} can be factored out:

$$f'(x_0) = E^{x_0} \left(\lim_{x \rightarrow x_0} \frac{E^{x-x_0} - 1}{x - x_0} \right)$$

If the function $f(x) = E^x$ is to be equal to its own derivative, then the parenthesized quantity in the above must resolve to one. Isolating this, we have

$$\lim_{x \rightarrow x_0} \frac{E^{x-x_0} - 1}{x - x_0} = 1,$$

and then solve for E :

$$E = \lim_{x \rightarrow x_0} (1 + (x - x_0))^{1/(x-x_0)}$$

Substitute $h = 1/(x - x_0)$, to write

$$E = \lim_{h \rightarrow \infty} \left(1 + \frac{1}{h}\right)^h,$$

thus $E = e$ and we're done.

Natural Exp with Squared Argument

Now comes a fun one. Consider the function $f(x) = e^{x^2}$. Starting off as usual, we have

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{e^{x^2} - e^{x_0^2}}{x - x_0}$$

Using Equation (16.10), the numerator expands out to

$$x^2 - x_0^2 + \frac{(x^2)^2 - (x_0^2)^2}{2!} + \frac{(x^2)^3 - (x_0^2)^3}{3!} + \cdots,$$

which is an algebraic mess, because we need to factor $x - x_0$ out of the whole expression. Going in chunks, we find:

$$\begin{aligned} x^2 - x_0^2 &= (x - x_0)(x + x_0) \\ (x^2)^2 - (x_0^2)^2 &= (x - x_0)(x + x_0)(x^2 + x_0^2) \\ (x^2)^3 - (x_0^2)^3 &= (x - x_0) \\ &\quad (x^2 + xx_0 + x_0^2)(x^3 + x_0^3) \\ (x^2)^4 - (x_0^2)^4 &= (x - x_0) \\ &\quad (x + x_0)(x^2 + x_0^2)(x^4 + x_0^4) \\ (x^2)^5 - (x_0^2)^5 &= (x - x_0) \\ &\quad (x^4 + x^3x_0 + x^2x_0^2 + xx_0^3 + x_0^4) \\ &\quad (x^5 + x_0^5) \end{aligned}$$

Assuming the pattern continues, we can say that $(x - x_0)$ can be factored out of each term in the expansion, and this resolves having to further deal with the denominator in the derivative.

What remains is to evaluate everything at $x = x_0$. Doing this carefully and spotting the pattern, we see

$$\begin{aligned} f'(x_0) &= 2x_0 + \frac{4x_0^3}{2!} + \frac{6x_0^5}{3!} + \frac{8x_0^7}{4!} \cdots \\ &= 2x_0 \left(1 + \frac{x_0^2}{1!} + \frac{x_0^4}{2!} + \frac{x_0^6}{3!} + \cdots \right). \end{aligned}$$

The parenthesized series is nothing more than $e^{x_0^2}$ according to Equation (16.10). Finally, the answer:

$$\frac{d}{dx} (e^{x^2}) = 2x e^{x^2} \quad (16.11)$$

Exponential with Squared Argument

The previous example can be done a different way by substituting $h = x - x_0$ before jumping into the algebra. To demonstrate on something more general, consider the function $f(x) = b^{x^2}$, where b is a real number. Setting up the derivative, we have

$$f'(x) = \lim_{x \rightarrow x_0} \frac{b^{x^2} - b^{x_0^2}}{x - x_0} = \lim_{h \rightarrow 0} \frac{b^{(x_0+h)^2} - b^{x_0^2}}{h}$$

Now we must spend a moment on the quantity $(x_0 + h)^2$. Expanding this out, we have

$$(x_0 + h)^2 = x_0^2 + 2x_0h + h^2.$$

In the limit that h is going to zero, the h^2 -term pales under the others and can be ignored. With this simplification, the derivative becomes

$$f'(x) = b^{x_0^2} \lim_{h \rightarrow 0} \frac{b^{2x_0h} - 1}{h}.$$

The remaining limit almost looks familiar as a natural logarithm. Make the substitution $h = 2x_0k$ to bring it into form:

$$f'(x) = 2x_0 b^{x_0^2} \lim_{k \rightarrow 0} \frac{b^k - 1}{k}.$$

The remaining limit now is unambiguously equivalent to the natural log of b . Finally, we find:

$$\frac{d}{dx} (b^{x^2}) = 2x b^{x^2} \ln(b) \quad (16.12)$$

The special case $b = e$ recovers Equation (16.11).

X to the X

A notorious derivative to figure out is that of $f(x) = x^x$. Setting up the calculation, we have

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{x^x - x_0^{x_0}}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \frac{x^{x_0}}{x - x_0} \left(x^{x-x_0} - \left(\frac{x_0}{x} \right)^{x_0} \right), \end{aligned}$$

and let $h = x - x_0$:

$$f(x) = \lim_{h \rightarrow 0} x^{x_0} \left(\frac{x^h - (1 - h/x)^{x_0}}{h} \right)$$

Recalling the limit-based expression for the natural logarithm, namely

$$\ln(x) = \lim_{h \rightarrow 0} \frac{x^h - 1}{h},$$

and the above becomes:

$$f(x) = \lim_{h \rightarrow 0} x^{x_0} \left(\ln(x) + \frac{1 - (1 - h/x)^{x_0}}{h} \right)$$

The right-hand limit is best handled in isolation. Letting

$$A = \lim_{h \rightarrow 0} \frac{1 - (1 - h/x)^{x_0}}{h},$$

rearrange to write

$$\lim_{h \rightarrow 0} (1 - Ah) = \lim_{h \rightarrow 0} \left(1 - \frac{h}{x} \right)^{x_0},$$

and raise each side to the $1/h$ power:

$$\lim_{h \rightarrow 0} (1 - Ah)^{1/h} = \lim_{h \rightarrow 0} \left(1 - \frac{h}{x} \right)^{x_0/h}$$

With one more substitution $q = 1/h$, this is

$$\lim_{q \rightarrow \infty} \left(1 - \frac{A}{q} \right)^q = \lim_{q \rightarrow \infty} \left(1 - \frac{1}{qx} \right)^{qx_0}$$

Keep in mind that $q \rightarrow \infty$ also means $x \rightarrow x_0$, and the above simplifies to:

$$e^{-A} = \left(e^{-1/x_0} \right)^{x_0} = e^{-1},$$

telling us finally that $A = 1$. Putting the answer together:

$$\frac{d}{dx} (x^x) = x^x (\ln(x) + 1) \quad (16.13)$$

1.4 Logarithmic Derivatives

Natural Logarithm

A keystone function is the natural logarithm, $f(x) = \ln(x)$. Setting up the derivative calculation, we have

$$\begin{aligned} f'(x_0) &= \lim_{x \rightarrow x_0} \frac{\ln(x) - \ln(x_0)}{x - x_0} \\ &= \frac{1}{x_0} \lim_{x \rightarrow x_0} \frac{\ln(x/x_0)}{x/x_0 - 1}, \end{aligned}$$

suggesting a substitution $x/x_0 = k$, and the limit becomes a matter of k approaching 1:

$$f'(x_0) = \frac{1}{x_0} \lim_{k \rightarrow 1} \frac{\ln(k)}{k - 1}$$

With another substitution $j = k - 1$, this is

$$f'(x_0) = \frac{1}{x_0} \lim_{j \rightarrow 0} \frac{\ln(1 + j)}{j},$$

equivalent to

$$f'(x_0) = \frac{1}{x_0} \lim_{j \rightarrow 0} \ln\left((1 + j)^{1/j}\right).$$

If it doesn't look familiar yet, let $h = 1/j$ to get

$$f'(x_0) = \frac{1}{x_0} \lim_{h \rightarrow \infty} \ln\left(\left(1 + \frac{1}{h}\right)^h\right).$$

The remaining limit patently resolves to e , and being enclosed as the argument to the $\ln$ function, resolves to just one, after all that. In summary, we find

$$\frac{d}{dx} (\ln(x)) = \frac{1}{x} \quad (16.14)$$

Shifted Natural Logarithm

The shifted natural logarithm $f(x) = \ln(1 + x)$ is handled much like the vanilla natural logarithm. Setting up the derivative calculation, we have

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{\ln(1 + x) - \ln(1 + x_0)}{x - x_0},$$

suggesting a substitution $z = 1 + x$. The above becomes

$$f'(x_0) = \lim_{z \rightarrow z_0} \frac{\ln(z) - \ln(z_0)}{z - z_0},$$

which now looks identical to the vanilla case in the variable z . Reversing the z -substitution gives the final answer:

$$\frac{d}{dx} (\ln(1 + x)) = \frac{1}{1 + x} \quad (16.15)$$

Nonlinear Natural Logarithm

For the function $f(x) = x \ln(x)$, the derivative calculation begins as

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{x \ln(x) - x_0 \ln(x_0)}{x - x_0}$$

This one is best attacked with polynomial division, which leads to

$$f'(x_0) = \lim_{x \rightarrow x_0} \left(\ln(x) + x_0 \left(\frac{\ln(x) - \ln(x_0)}{x - x_0} \right) \right),$$

and now the embedded limit should ring familiar as the derivative of the vanilla natural log, or simply $1/x_0$. Simplifying the rest, we find

$$\frac{d}{dx} (x \ln(x)) = \ln(x) + 1 \quad (16.16)$$

The same technique, namely polynomial division and then substitution from an easier derivative, is what is needed to find the derivative of the harder function, $f(x) = x^2 \ln(x)$. Going through the exercise (which you are encouraged to do), the result should be

$$\frac{d}{dx} (x^2 \ln(x)) = 2x \ln(x) + x. \quad (16.17)$$

Diminished Natural Logarithm

For another tough one, let us find the derivative of $f(x) = \ln(x)/x$. Going from the definition, we first write

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{\ln(x)/x - \ln(x_0)/x_0}{x - x_0}.$$

Using polynomial division and simplifying, we get to the intermediate step:

$$f'(x_0) = \frac{\ln(x_0)}{x_0^2} + \lim_{x \rightarrow x_0} \frac{1}{x_0^3} \left(\frac{x_0^2 \ln(x) - x^2 \ln(x_0)}{x - x_0} \right)$$

The remaining limit almost looks like Equation (16.16), i.e. the derivative of $x^2 \ln(x)$, but sadly isn't exact.

To proceed, write the derivative of the natural logarithm in the form

$$\lim_{x \rightarrow x_0} \frac{\ln(x) - \ln(x_0)}{x - x_0} = \frac{1}{x_0},$$

and then deduce the following:

$$\begin{aligned}\lim_{x \rightarrow x_0} \frac{x_0^2 \ln(x)}{x - x_0} &= \lim_{x \rightarrow x_0} \frac{x_0^2 \ln(x_0)}{x - x_0} + x_0 \\ \lim_{x \rightarrow x_0} \frac{x^2 \ln(x_0)}{x - x_0} &= \lim_{x \rightarrow x_0} \frac{x^2 \ln(x)}{x - x_0} - \frac{x^2}{x_0}\end{aligned}$$

Subtract the bottom equation from the top, and notice the left side can replace the parenthesized quantity in our $f'(x)$ equation. Doing so, we get:

$$\begin{aligned}f'(x_0) &= \frac{\ln(x_0)}{x_0^2} + \lim_{x \rightarrow x_0} \frac{1}{x_0^3} \left(x_0 + \frac{x^2}{x_0} \right) \\ &\quad + \lim_{x \rightarrow x_0} \frac{1}{x_0^3} \left(\frac{x_0^2 \ln(x_0) - x^2 \ln(x)}{x - x_0} \right)\end{aligned}$$

The latter term in the above contains the (negative) derivative of $x^2 \ln(x)$ and can be replaced by Equation (16.16). This gets rid of all singularities, and we can simplify to get the answer:

$$\frac{d}{dx} \left(\frac{\ln(x)}{x} \right) = \frac{1 - \ln(x)}{x^2} \quad (16.18)$$

Modified Natural Logarithm

For the function $f(x) = \ln(1 + x^2)$, the derivative calculation begins as

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{\ln(1 + x^2) - \ln(1 + x_0^2)}{x - x_0}$$

Letting $h = x - x_0$ and simplifying using the rules for manipulating logarithms, the above reduces way down to

$$f'(x_0) = \lim_{h \rightarrow 0} \ln \left(\left(1 + \frac{2x_0 h + h^2}{1 + x_0^2} \right)^h \right).$$

In the limit $h \rightarrow 0$, the h^2 -term is negligible, and the rest, after staring for long enough, contains the definition of the natural exponential:

$$f'(x_0) = \ln \left(e^{(2x_0 h)/(1+x_0^2)} \right)$$

Since the natural log and the natural exponential are mutually-annihilating, we get the result:

$$\frac{d}{dx} \left(\ln(1 + x^2) \right) = \frac{2x}{1 + x^2} \quad (16.19)$$

1.5 Trigonometric Derivatives

All of the elementary trigonometric functions are curves, so we're obligated now to find their derivatives.

Sine

For the sine function $f(x) = \sin(x)$, we first write

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\sin(x) - \sin(x_0)}{x - x_0}.$$

The difference of sines is handled by the trigonometric identity

$$\sin(a) - \sin(b) = 2 \sin\left(\frac{a-b}{2}\right) \cos\left(\frac{a+b}{2}\right),$$

and the derivative becomes

$$f'(x) = \cos(x_0) \lim_{x \rightarrow x_0} \frac{2}{x - x_0} \left(\sin\left(\frac{x - x_0}{2}\right) \right)$$

Let $h = (x - x_0)/2$ to uncover a sinc function lurking about:

$$f'(x) = \cos(x_0) \lim_{h \rightarrow 0} \left(\frac{\sin(h)}{h} \right)$$

Recall that we spent some effort deciding that the parenthesized quantity is identically one, and we're left with the answer:

$$\frac{d}{dx} (\sin(x)) = \cos(x) \quad (16.20)$$

Cosine

The steps for calculating the derivative of $f(x) = \cos(x)$ are about identical to that of the sine function, except the required trig identity is

$$\cos(a) - \cos(b) = -2 \sin\left(\frac{a-b}{2}\right) \sin\left(\frac{a+b}{2}\right).$$

This lands us at

$$f'(x) = -\sin(x_0) \lim_{h \rightarrow 0} \left(\frac{\sin(h)}{h} \right),$$

and the same sinc is also present. From this we conclude

$$\frac{d}{dx} (\cos(x)) = -\sin(x). \quad (16.21)$$

Tangent

For the tangent $f(x) = \tan(x)$, we start with

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\tan(x) - \tan(x_0)}{x - x_0},$$

and then the trick is to divide out $\cos^2(x)$ from the limit:

$$f'(x) = \frac{1}{(\cos(x_0))^2} \lim_{x \rightarrow x_0} \frac{\sin(x) \cos(x_0) - \sin(x_0) \cos(x)}{x - x_0}$$

The remaining limit contains a trig identity for the sum of two angles, namely $x - x_0$. This again resolves to $\sin(0)$, as was the case with the previous two trig functions. The result is

$$\frac{d}{dx}(\tan(x)) = \frac{1}{(\cos(x))^2}. \quad (16.22)$$

Cotangent

For the cotangent $f(x) = \cot(x)$, we start with

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\cot(x) - \cot(x_0)}{x - x_0},$$

and then the trick is to divide out $\sin^2(x)$ from the limit:

$$f'(x) = \frac{1}{(\sin(x_0))^2} \lim_{x \rightarrow x_0} \frac{\cos(x) \sin(x_0) - \cos(x_0) \sin(x)}{x - x_0}$$

The remaining limit contains a trig identity for the sum of two angles, namely $x_0 - x$. This resolves to $-\sin(0)$, and the final result is

$$\frac{d}{dx}(\cot(x)) = \frac{-1}{(\sin(x))^2}. \quad (16.23)$$

Secant

For the secant function $f(x) = 1/\cos(x)$, we start with

$$f'(x) = \lim_{x \rightarrow x_0} \frac{1/\cos(x) - 1/\cos(x_0)}{x - x_0},$$

and then divide out $-\cos^2(x)$ from the limit:

$$f'(x) = \frac{-1}{(\cos(x_0))^2} \lim_{x \rightarrow x_0} \frac{\cos(x) - \cos(x_0)}{x - x_0}$$

The limit now looks like the derivative of the cosine function and can be replaced by Equation (16.21), namely $-\sin(x_0)$. Reporting the result in standard form, we find

$$\frac{d}{dx}(\sec(x)) = \tan(x) \sec(x). \quad (16.24)$$

Cosecant

For the cosecant function $f(x) = 1/\sin(x)$, we start with

$$f'(x) = \lim_{x \rightarrow x_0} \frac{1/\sin(x) - 1/\sin(x_0)}{x - x_0},$$

and then divide out $-\sin^2(x)$ from the limit:

$$f'(x) = \frac{-1}{(\sin(x_0))^2} \lim_{x \rightarrow x_0} \frac{\sin(x) - \sin(x_0)}{x - x_0}$$

The limit now looks like the derivative of the sine function and can be replaced by Equation (16.20), namely $\cos(x_0)$. Reporting the result in standard form, we find

$$\frac{d}{dx}(\csc(x)) = -\cot(x) \csc(x). \quad (16.25)$$

Squared Argument

For the sine function with a squared argument $f(x) = \sin(x^2)$, we first write

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\sin(x^2) - \sin(x_0^2)}{x - x_0}.$$

Using the same trig identity that helped with the regular sine case, the above becomes

$$f'(x) = \lim_{x \rightarrow x_0} \left(\frac{2}{x - x_0} \right) \sin\left(\frac{(x - x_0)(x + x_0)}{2}\right) \cos\left(\frac{x^2 + x_0^2}{2}\right).$$

Let $h = (x - x_0)/2$ and simplify a little to get

$$f'(x) = \cos(x_0^2) \lim_{h \rightarrow 0} \frac{\sin(2x_0 h)}{h}.$$

There seems to be an extra term in the remaining sine function that cannot be ignored. To deal with this, make a new substitution $k = 2x_0 h$, which also limits to zero as h does so. With this, we now have

$$f'(x) = \cos(x_0^2) 2x_0 \lim_{k \rightarrow 0} \left(\frac{\sin(k)}{k} \right).$$

The final limit is identically one, and in conclusion,

$$\frac{d}{dx}(\sin(x^2)) = \cos(x^2) 2x. \quad (16.26)$$

The intermediate steps would be nearly the same had we started with cosine instead of sine, which would result in:

$$\frac{d}{dx}(\cos(x^2)) = -\sin(x^2) 2x. \quad (16.27)$$

X Times Sin(X)

One more before moving on. Consider the product of x and the sine of x , i.e. $f(x) = x \sin(x)$. Setting up this derivative, we write

$$f'(x) = \lim_{x \rightarrow x_0} \frac{x \sin(x) - x_0 \sin(x_0)}{x - x_0}$$

To crack this one, add and subtract the quantity $x_0 \sin(x)$ from the numerator, and then repack everything to get

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\sin(x_0)(x - x_0)}{x - x_0} + \lim_{x \rightarrow x_0} \frac{x_0(\sin(x) - \sin(x_0))}{x - x_0}.$$

Now, one hard limit is replaced by two easy limits. The former case has $x - x_0$ canceling, leaving just $\sin(x_0)$. The latter case has x_0 multiplied by the derivative of the sine function, which we know to be $\cos(x_0)$. In conclusion, we have

$$\frac{d}{dx}(x \sin(x)) = \sin(x) + x \cos(x). \quad (16.28)$$

The same recipe works for the $x \cos(x)$ case. Leaving the details as an exercise, the result is

$$\frac{d}{dx}(x \cos(x)) = \cos(x) - x \sin(x). \quad (16.29)$$

1.6 Small-Angle Approximation

In the limit of small angles, it's easy to show that the sine and cosine obey the aptly-named *small-angle approximation*:

$$\lim_{x \rightarrow 0} \sin(x) = x - \frac{x^3}{3!} + \dots \quad (16.30)$$

$$\lim_{x \rightarrow 0} \cos(x) = 1 - \frac{x^2}{2!} + \dots \quad (16.31)$$

Sine and Cosine Revisited

The derivatives for the sine and cosine can be derived in a nifty way using small angles with the angle-sum formulas from trigonometry. For the sum of two angles θ and ϕ , recall that:

$$\sin(\theta + \phi) = \sin(\theta) \cos(\phi) + \cos(\theta) \sin(\phi)$$

$$\cos(\theta + \phi) = \cos(\theta) \cos(\phi) - \sin(\theta) \sin(\phi)$$

Next, suppose that ϕ is a vanishingly small angle, i.e. a differential angle. (This makes the quantity $\theta + \phi$ analogous to $x + dx$ but we won't change letters.) In such a limit, the above identities become

$$\lim_{\phi \rightarrow 0} \sin(\theta + \phi) = \lim_{\phi \rightarrow 0} (\sin(\theta) + \phi \cos(\theta))$$

$$\lim_{\phi \rightarrow 0} \cos(\theta + \phi) = \lim_{\phi \rightarrow 0} (\cos(\theta) - \phi \sin(\theta)).$$

Solve the first equation for $\cos(\theta)$ and the second equation for $-\sin(\theta)$:

$$\cos(\theta) = \lim_{\phi \rightarrow 0} \frac{\sin(\theta + \phi) - \sin(\theta)}{\phi}$$

$$-\sin(\theta) = \lim_{\phi \rightarrow 0} \frac{\cos(\theta + \phi) - \cos(\theta)}{\phi}$$

This pair of results is none other than the derivative formulas for sine and cosine. By a change of variables these exactly reproduce Equations (16.20), (16.21).

2 Techniques of Differentiation

Fortunately, not every derivative needs to be calculated directly from the definition. To motivate a few tricks and shortcuts, suppose we have two functions of x , namely $f(x)$ and $g(x)$. We require that f and g be ‘well-behaved’ which is to say ‘differentiable’. If this is the case, each has a well-defined derivative, $f'(x)$ and $g'(x)$, respectively.

2.1 Product Rule

Suppose we define $p(x)$ as the product of $f(x)$ and $g(x)$, i.e.

$$p(x) = f(x)g(x).$$

The derivative of $p(x)$ is

$$p'(x) = \lim_{x \rightarrow x_0} \frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0},$$

and the job is to cook this down to something more useful.

Proceed by subtracting and adding the quantity $f(x_0)g(x)$ into the limit's numerator

$$\begin{aligned} p'(x_0) &= \lim_{x \rightarrow x_0} \left(\frac{f(x)g(x) - f(x_0)g(x)}{x - x_0} \right) \\ &\quad - \lim_{x \rightarrow x_0} \left(\frac{f(x_0)g(x_0) - f(x_0)g(x)}{x - x_0} \right), \end{aligned}$$

and simplify:

$$\begin{aligned} p'(x_0) &= g(x_0) \lim_{x \rightarrow x_0} \left(\frac{f(x) - f(x_0)}{x - x_0} \right) \\ &\quad + f(x_0) \lim_{x \rightarrow x_0} \left(\frac{g(x) - g(x_0)}{x - x_0} \right) \end{aligned}$$

Notice the two remaining limits are the individual derivatives of $f(x)$ and $g(x)$, so the above can be written might tighter

$$p'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0),$$

known as the *product rule* for derivatives.

Using abbreviated Leibniz notation, the product rule reads for $f(x)$ and $g(x)$:

$$\frac{d}{dx}(fg) = \frac{df}{dx}g + f\frac{dg}{dx} \quad (16.32)$$

Or, a more economical way to write the same thing:

$$(fg)' = f'g + fg'$$

All of these notations are mixed and matched in the greater literature, and the same liberties will be taken as we proceed.

Examples

The product rule makes quick work of a few cases explored previously.

Problem 1

Find the derivative of $p(x) = x^2 \ln(x)$. Answer: Identifying $f(x) = x^2$ and $g(x) = \ln(x)$, we have

$$\begin{aligned} p'(x) &= f'g + fg' \\ &= \ln(x) \frac{d}{dx} \ln(x^2) + x^2 \frac{d}{dx} (\ln(x)) \\ &= 2x \ln(x) + x, \end{aligned}$$

in agreement with Equation (16.16).

Problem 2

Find the derivative of $p(x) = \ln(x)/x$. Answer: Identifying $f(x) = \ln(x)$ and $g(x) = 1/x$, we have

$$\begin{aligned} p'(x) &= f'g + fg' \\ &= \left(\frac{1}{x}\right) \frac{d}{dx} \ln(x) + \ln(x) \frac{d}{dx} \left(\frac{1}{x}\right) \\ &= \frac{1 - \ln(x)}{x^2}, \end{aligned}$$

in agreement with Equation (16.18).

Problem 3

Find the derivative of $p(x) = x \sin(x)$. Answer: Identifying $f(x) = x$ and $g(x) = \sin(x)$, we have

$$\begin{aligned} p'(x) &= f'g + fg' \\ &= (\sin(x)) \frac{dx}{dx} + x \frac{d}{dx} (\sin(x)) \\ &= \sin(x) + x \cos(x), \end{aligned}$$

in agreement with Equation (16.28).

Problem 4

Find the derivative of $p(x) = x \cos(x)$. Answer: Identifying $f(x) = x$ and $g(x) = \cos(x)$, we have

$$\begin{aligned} p'(x) &= f'g + fg' \\ &= (\cos(x)) \frac{dx}{dx} + x \frac{d}{dx} (\cos(x)) \\ &= \cos(x) - x \sin(x) \end{aligned}$$

in agreement with Equation (16.29).

2.2 Quotient Rule

Suppose we define $r(x)$ as the ratio of $f(x)$ and $g(x)$, i.e.

$$r(x) = \frac{f(x)}{g(x)}.$$

The derivative of $r(x)$ is

$$r'(x) = \lim_{x \rightarrow x_0} \frac{f(x)/g(x) - f(x_0)/g(x_0)}{x - x_0},$$

and like before, we need to simplify.

Proceed by multiplying the numerator and denominator by $g(x)g(x_0)$, and then add and subtract the quantity $f(x_0)g(x_0)$ from the numerator. Carefully treating each limit, the result is

$$r'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2},$$

known as the *quotient rule* for derivatives.

In Leibniz notation, the quotient rule reads for $f(x)$ and $g(x)$:

$$\frac{d}{dx} \left(\frac{f}{g} \right) = \frac{1}{g^2} \left(\frac{df}{dx} g - f \frac{dg}{dx} \right) \quad (16.33)$$

Or, a more economical way to write the same thing:

$$\left(\frac{f}{g} \right)' = \frac{f'g - fg'}{g^2}$$

Examples

Like the product rule, the quotient makes quick work of certain problems.

Problem 5

Find the derivative of $r(x) = \ln(x)/x$. Answer: Identifying $f(x) = \ln(x)$ and $g(x) = x$, we have

$$\begin{aligned} r'(x) &= \frac{f'g - fg'}{g^2} \\ &= \frac{(1/x)x - \ln(x)(1)}{x^2} \\ &= \frac{1 - \ln(x)}{x^2}, \end{aligned}$$

in agreement with Equation (16.18).

Problem 6

Find the derivative of $r(x) = \tan(x)$. Answer: Identifying $f(x) = \sin(x)$ and $g(x) = \cos(x)$, we have

$$\begin{aligned} r'(x) &= \frac{f'g - fg'}{g^2} \\ &= \frac{(\cos(x))^2 + (\sin(x))^2}{(\cos(x))^2} \\ &= \frac{1}{(\cos(x))^2}, \end{aligned}$$

in agreement with Equation (16.22).

2.3 Chain Rule

Composite Functions

Consider the *composite* function

$$c(x) = f(g(x)).$$

To unpack this, we have a function $g(x)$ that is a typical function of x . The function f depends on g , so abbreviating the above by $c = f(g)$ is valid in the same way we would write $y = f(x)$.

Composite functions really aren't news to us. Things like $\cos(x^2)$ and e^{4x} or any nontrivial function can all be written as composite functions.

Derivation of Chain Rule

The issue of composite functions raises a subtle point, for if we have the generic scenario $y = f(x)$, it could have been all along that x itself is a function of some other variable, say t , as in $y(t) = f(x(t))$. This has curious implications for derivatives of the functions involved.

Recall the definition of the derivative of a function $f(x)$ in the generic case:

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Now, suppose we just found out that x is a function of a deeper variable t such that

$$\begin{aligned} x &= x(t) \\ x_0 &= x(t_0). \end{aligned}$$

In other words, $f(x)$ just became a composite function $f(x(t))$.

Naturally, $x(t)$ has a derivative of its own with respect to t :

$$x'(t_0) = \lim_{t \rightarrow t_0} \frac{x(t) - x(t_0)}{t - t_0}$$

The derivative of f , though, now looks like this:

$$f'(x(t_0)) = \lim_{x(t) \rightarrow x(t_0)} \frac{f(x(t)) - f(x(t_0))}{x(t) - x(t_0)}$$

To simplify the above, we first acknowledge that all functions are ultimately dependent on t , so the comment under the $\lim$ symbol can be replaced simply by $t \rightarrow t_0$. Next comes the key move, which is to multiply both sides by $x'(t_0)$ to get:

$$f'(x(t_0)) \cdot x'(t_0) = \lim_{t \rightarrow t_0} \frac{f(x(t)) - f(x(t_0))}{\cancel{x(t) - x(t_0)}} \cdot \frac{\cancel{x(t) - x(t_0)}}{t - t_0}$$

The quantity $x - x_0$ conveniently cancels, and we can tidy things up to write:

$$f'(x_0) \cdot x'(t_0) = \lim_{t \rightarrow t_0} \frac{f(x(t)) - f(x(t_0))}{t - t_0}$$

The right side of the above is the derivative of $f(x(t))$ with respect to t . It would be a misnomer to shorthand the right side with an f' -like symbol, as the 'prime' notation is reserved (typically) for derivatives with respect to x . It's much wiser at this instant to switch to Leibniz notation and rewrite the above as

$$\frac{df}{dt} = \frac{df}{dx} \cdot \frac{dx}{dt}, \quad (16.34)$$

known as the *chain rule* for derivatives.

The chain rule can be stacked indefinitely, i.e. if we found out that t itself depends on a deeper variable u such that $t = t(u)$, the above swiftly becomes:

$$\frac{df}{du} = \frac{df}{dx} \cdot \frac{dx}{dt} \cdot \frac{dt}{du}$$

Notice on the right that all terms 'cancel', at least visually, except for df in the numerator and du in the denominator to match the left side. This is what's nice about the chain rule - if it looks right, it *is* right.

Elementary Examples

Problem 7

Find the derivative of $f(x) = e^{x^2}$. Answer: Letting $u = x^2$, f becomes a composite function $f(x) = e^{u(x)}$. The

derivative of f with respect to x proceeds as

$$\begin{aligned} \frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} e^u \right) \left(\frac{d}{dx} x^2 \right) \\ &= e^{u(x)} 2x \\ &= 2x e^{x^2}, \end{aligned}$$

in agreement with Equation (16.11).

Problem 8

Find the derivative of $f(x) = b^{x^2}$. Answer: Letting $u = x^2$, f becomes a composite function $f(x) = b^{u(x)}$. The derivative of f with respect to x proceeds as

$$\begin{aligned} \frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} b^u \right) \left(\frac{d}{dx} x^2 \right) \\ &= b^{u(x)} \ln(b) 2x \\ &= 2x b^{x^2} \ln(b), \end{aligned}$$

in agreement with Equation (16.12).

Problem 9

Find the derivative of $f(x) = \ln(x^x)$. Answer: Letting $u = x^x$, f becomes a composite function $f(x) = \ln(u)$. The derivative of f with respect to x proceeds as:

$$\begin{aligned} \frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} \ln(u) \right) \left(\frac{d}{dx} x^x \right) \\ &= \frac{1}{u(x)} x^x (\ln(x) + 1) \\ &= \ln(x) + 1 \end{aligned}$$

Note that Equation (16.13) was quietly used in the above, making this example rather lengthy in its totality. However, this result is attained more easily by realizing $\ln(x^x)$ is equivalent to $x \ln(x)$ which is a job for the product rule. Either way, we conclude

$$\frac{d}{dx} (\ln(x^x)) = \ln(x) + 1. \quad (16.35)$$

Problem 10

Find the derivative of $f(x) = \ln(1+x)$. Answer: Letting $u = 1+x$, f becomes a composite function

$f(x) = \ln(u)$. The derivative of f with respect to x proceeds as:

$$\begin{aligned}\frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} \ln(u) \right) \left(\frac{d}{dx} (1+x) \right) \\ &= \frac{1}{u(x)} \\ &= \frac{1}{1+x},\end{aligned}$$

in agreement with Equation (16.15).

Problem 11

Find the derivative of $f(x) = \ln(1+x^2)$. Answer: Letting $u = 1+x^2$, f becomes a composite function $f(x) = \ln(u)$. The derivative of f with respect to x proceeds as:

$$\begin{aligned}\frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} \ln(u) \right) \left(\frac{d}{dx} (1+x^2) \right) \\ &= \frac{1}{u(x)} 2x \\ &= \frac{2x}{1+x^2},\end{aligned}$$

in agreement with Equation (16.19).

Problem 12

Find the derivative of $f(x) = \sin(x^2)$. Answer: Letting $u = x^2$, f becomes a composite function $f(x) = \sin(u)$. The derivative of f with respect to x proceeds as

$$\begin{aligned}\frac{df}{dx} &= \frac{df}{du} \cdot \frac{du}{dx} \\ &= \left(\frac{d}{du} \sin(u) \right) \left(\frac{d}{dx} x^2 \right) \\ &= \cos(u(x)) 2x \\ &= \cos(x^2) 2x,\end{aligned}$$

in agreement with Equation (16.26).

Logarithm Trick

The chain rule allows for some interesting cheats when calculating derivatives. Suppose we have a function $f(x)$ that seems to be tricky to differentiate, which is to say $f'(x)$ is

not straightforwardly calculated. It may help to send $f(x)$ to the natural logarithm before calculating the derivative, and then exploit the chain rule to weasel out an answer for $f'(x)$.

Applying the chain rule in, this scenario looks like

$$\frac{d}{dx} (\ln(f(x))) = \frac{1}{f(x)} f'(x),$$

or

$$\frac{df}{dx} = f(x) \frac{d}{dx} (\ln(f(x))), \quad (16.36)$$

which we'll call the *logarithm trick*.

Problem 13

Find the derivative of $f(x) = b^{x^2}$. Answer: Using the logarithm trick, we find

$$\begin{aligned}\frac{df}{dx} &= b^{x^2} \frac{d}{dx} (\ln(b^{x^2})) \\ &= b^{x^2} \ln(b) \frac{d}{dx} x^2 \\ &= 2x b^{x^2} \ln(b),\end{aligned}$$

in agreement with Equation (16.12).

Problem 14

Find the derivative of $f(x) = x^x$. Answer: Using the logarithm trick, we find

$$\begin{aligned}\frac{df}{dx} &= x^x \frac{d}{dx} (\ln(x^x)) \\ &= x^x \frac{d}{dx} (x \ln(x)) \\ &= x^x (\ln(x) + 1),\end{aligned}$$

in agreement with Equation (16.13).

Power Rule Revisited

Recall that the derivative of a function $f(x) = x^n$ is established by Equation (16.2), namely

$$\frac{d}{dx} (x^n) = nx^{n-1}.$$

The derivation of this is guaranteed for integer n , but we did not explicitly cover what happens for non-integer n .

Luckily, the result is the same, i.e. Equation (16.2) holds for any n . To prove this, start with

$$f(x) = x^n = e^{\ln(x^n)} = e^{n \ln(x)}.$$

By the chain rule, we then find

$$\frac{d}{dx} (x^n) = e^{n \ln(x)} \frac{n}{x} = nx^{n-1}$$

and we're done.

Problem 15

Consider the function

$$f(x) = 4^{x^2}.$$

Find the derivative of $f(x)$ using (i) the definition of the derivative, (ii) the product rule, (iii) the chain rule, and (iv) the logarithm trick.

Hints: (i) Use Equation (16.12). (ii) Let $f(x) = g(x) = 2^{x^2}$ and use Equation (16.32). (iii) Let $f(g) = 4^g$ and $g(x) = x^2$ and calculate df/dx . (iv) Use Equation (16.36).

2.4 Implicit Differentiation

Derivative Operator

A subtlety that has been present all along, but not explicitly mentioned, is that we may talk about the derivative d/dx as an operation on some completely unspecified object, in the same way that we can talk about the 'add' (+) or 'multiply' ($\times$) operations without mentioning what is being added or multiplied.

Just like we'd multiply by two or divide by π , we can take the derivative across an entire equation. For instance, starting with

$$a(x) = b(x) + c(x),$$

it's reasonable to write

$$\frac{d}{dx}a = \frac{d}{dx}(a + b).$$

We can go a little further by understanding the derivative to be a *linear operator*. This means the right side of the above can be broken into two separate derivatives on $b(x)$ and $c(x)$:

$$\frac{d}{dx}a = \frac{d}{dx}b + \frac{d}{dx}c$$

The act of applying the derivative operator across a whole equation has a name called *implicit differentiation*, a tool that works beautifully alongside the chain rule.

Tangent Line to the Ellipse

An ellipse characterized by semi-major axis a and semi-minor axis b centered in the Cartesian plane is described by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

Solving for y gives a pair of proper functions to describe the ellipse:

$$y_{\pm}(x) = \pm b \sqrt{1 - \frac{x^2}{a^2}}$$

If we want the slope of a tangent line to the circle, simply calculate $y'(x) = dy/dx$ (looking only at the top half of the circle for a moment):

$$y'(x) = \pm b \left(\frac{1}{2\sqrt{1 - x^2/a^2}} \right) (-2x/a^2)$$

$$y'(x) = -\frac{b^2 x}{a^2 y}$$

Implicit differentiation is a quicker way to calculate $y'(x)$, and it works directly on the equation of the ellipse by throwing d/dx around both sides:

$$\frac{d}{dx} \left(\frac{x^2}{a^2} \right) + \frac{d}{dx} \left(\frac{y^2}{b^2} \right) = \frac{d}{dx} (1)$$

Using the standard rules for differentiation, including the chain rule on the y -term, this gives

$$\frac{2x}{a^2} + \frac{2y}{b^2} \frac{dy}{dx} = 0,$$

and solving for dy/dx gives the same result as above, and there was no square root to fiddle with:

$$\frac{dy}{dx} = -\frac{b^2 x}{a^2 y}$$

Regardless of how we know the slope of the tangent line at a given point (x_0, y_0) on the ellipse, the tangent line itself is

$$y = \left(\frac{-b^2 x_0}{a^2 y_0} \right) x + b,$$

which is equivalent to

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1.$$

The proof is an exercise for the reader.

3 Mixed Techniques

Certain derivatives can require a mixture of tricks to figure out. In the following we pick and choose from the definition of the derivative, the product and quotient rules, along with the chain rule to produce results.

3.1 Inverse Trig Derivatives

Consider the set of inverse trigonometric functions, namely:

$$\arccos(x) = \cos^{-1}(x)$$

$$\arcsin(x) = \sin^{-1}(x)$$

$$\arctan(x) = \tan^{-1}(x)$$

$$\operatorname{arcsec}(x) = \sec^{-1}(x)$$

$$\operatorname{arccsc}(x) = \csc^{-1}(x)$$

$$\operatorname{arccot}(x) = \cot^{-1}(x)$$

Arccosine

Begin with the statement

$$\cos(\arccos(x)) = x,$$

and apply the d/dx operator to both sides while using the chain rule

$$\sin(\arccos(x)) \frac{d}{dx} \arccos(x) = -1,$$

and then solve for the quantity we are after:

$$\frac{d}{dx} \arccos(x) = \frac{-1}{\sin(\arccos(x))}$$

There's still a little more work to do. Imagine a right triangle of hypotenuse 1 such that the adjacent side is $x = \cos(\theta)$. Comparing this to what's already written, identify $\theta = \arccos(x)$, and from geometry we also have $\sin(\theta) = \sqrt{1-x^2}$. This is enough to get the final answer:

$$\frac{d}{dx} \arccos(x) = \frac{-1}{\sqrt{1-x^2}} \quad (16.37)$$

Arcsine

Begin with the statement

$$\sin(\arcsin(x)) = x,$$

and apply the d/dx operator to both sides while using the chain rule

$$\cos(\arcsin(x)) \frac{d}{dx} \arcsin(x) = 1,$$

and then solve for the quantity we are after:

$$\frac{d}{dx} \arcsin(x) = \frac{1}{\cos(\arcsin(x))}$$

As with the previous case, there's still a little more work to do with a right triangle of hypotenuse 1 such that the opposite side is $x = \sin(\theta)$. Comparing this to what's already written, identify $\theta = \arcsin(x)$, and from geometry we also have $\cos(\theta) = \sqrt{1-x^2}$. This is enough to get the final answer:

$$\frac{d}{dx} \arcsin(x) = \frac{1}{\sqrt{1-x^2}} \quad (16.38)$$

Arctangent

Begin with the statement

$$\tan(\arctan(x)) = x,$$

and apply the d/dx operator to both sides while using the chain rule

$$\frac{1}{(\cos(\arctan(x)))^2} \frac{d}{dx} \arctan(x) = 1,$$

and then solve for the quantity we are after:

$$\frac{d}{dx} \arctan(x) = \cos^2(\arctan(x))$$

Eliminate the $\cos^2$ -term using the trig identity $\cos^2(\theta) = 1/(1 + \tan^2(\theta))$ and the answer emerges:

$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2} \quad (16.39)$$

Arcsecant, Arccosecant, Arccotangent

The remaining three inverse functions are handled by a similar process to the first three. Without belaboring the details, which is left for an exercise, the results should be:

$$\frac{d}{dx} \operatorname{arcsec}(x) = \frac{1}{x\sqrt{x^2-1}} \quad (16.40)$$

$$\frac{d}{dx} \operatorname{arccsc}(x) = \frac{-1}{x\sqrt{x^2-1}} \quad (16.41)$$

$$\frac{d}{dx} \operatorname{arccot}(x) = \frac{-1}{1+x^2} \quad (16.42)$$

3.2 Hyperbolic Derivatives

Sinh, Cosh, Tanh

The hyperbolic trigonometric functions

$$\begin{aligned}\cosh(x) &= \frac{e^x + e^{-x}}{2} \\ \sinh(x) &= \frac{e^x - e^{-x}}{2}\end{aligned}$$

are in many ways analogous to the ordinary trigonometric functions. For instance one may take $f(x) = \sinh(x)$ and use the definition of the derivative to write

$$f'(x) = \lim_{x \rightarrow x_0} \frac{\sinh(x) - \sinh(x_0)}{x - x_0},$$

which, by a similar process the led to Equation (16.20), gives:

$$\frac{d}{dx}(\sinh(x)) = \cosh(x) \quad (16.43)$$

It's a bit easier to use implicit differentiation to chug through the derivative calculation. Demonstrating on $\cosh(x)$, we have

$$\frac{d}{dx}(\cosh(x)) = \frac{d}{dx} \left(\frac{e^x + e^{-x}}{2} \right),$$

readily simplifying to

$$\frac{d}{dx}(\cosh(x)) = \sinh(x) \quad (16.44)$$

Note that this result differs from its trigonometric cousin by lacking a negative sign.

The derivative of $\tanh(x)$ follows easily from the quotient rule:

$$\begin{aligned}\frac{d}{dx}(\tanh(x)) &= \frac{d}{dx} \left(\frac{\sinh(x)}{\cosh(x)} \right) \\ &= \frac{(\cosh(x))^2 - (\sinh(x))^2}{(\cosh(x))^2},\end{aligned}$$

and the numerator simplifies via the identity

$$(\cosh(x))^2 - (\sinh(x))^2 = 1.$$

In conclusion:

$$\frac{d}{dx}(\tanh(x)) = (\operatorname{sech}(x))^2 \quad (16.45)$$

Coth, Sech, Csch

The hyperbolic cotangent, hyperbolic secant, and hyperbolic cosecant are also straightforwardly handled:

$$\frac{d}{dx}(\coth(x)) = -(\operatorname{csch}(x))^2 \quad (16.46)$$

$$\frac{d}{dx}(\operatorname{sech}(x)) = -\operatorname{sech}(x) \tanh(x) \quad (16.47)$$

$$\frac{d}{dx}(\operatorname{csch}(x)) = -\operatorname{csch}(x) \coth(x) \quad (16.48)$$

Arccosh, Arcsinh, Arctanh

The inverse hyperbolic functions are also straightforward to handle using the ordinary trig case as an analogy:

$$\frac{d}{dx}(\operatorname{arccosh}(x)) = \frac{1}{\sqrt{x^2 - 1}} \quad (16.49)$$

$$\frac{d}{dx}(\operatorname{arsinh}(x)) = \frac{1}{\sqrt{x^2 + 1}} \quad (16.50)$$

$$\frac{d}{dx}(\operatorname{arctanh}(x)) = \frac{1}{1 - x^2} \quad (16.51)$$

Arcsech, Arcsch, Arccoth

$$\frac{d}{dx}(\operatorname{arcsech}(x)) = \frac{-1}{x\sqrt{1 - x^2}} \quad (16.52)$$

$$\frac{d}{dx}(\operatorname{arcsch}(x)) = \frac{-1}{x\sqrt{x^2 + 1}} \quad (16.53)$$

$$\frac{d}{dx}(\operatorname{arccoth}(x)) = \frac{1}{1 - x^2} \quad (16.54)$$

4 Applied Differentiation

4.1 Tangent Line

Given a differentiable function $y = f(x)$, the derivative $f'(x_0)$ gives the slope of the tangent line that, by definition, just touches the curve at the point (x_0, y_0) . The equation of the *tangent line* is simply:

$$y - y_0 = f'(x_0)(x - x_0) \quad (16.55)$$

Problem 16

Consider the equation

$$2x^2y + y = 2x + 13,$$

and note that the ordered pair $(3, 1)$ is a solution. Use implicit differentiation to find $y'(x_0)$ and then use Equation (16.55) to estimate the solution at $x = 2.8$.

4.2 Mean Value Theorem

The *mean value theorem* is a piece of mathematics that concretizes something that may be obvious by inspecting any function. The theorem states that, for a given arc between two points in the Cartesian plane, there is at least one point on the arc where the instantaneous slope is parallel to the secant line formed by the two points.

In detail, suppose we have two points x_a, x_b that we feed to a function $f(x)$. The mean value theorem dictates that somewhere between x_a, x_b is a third point x_m satisfying

$$f'(x_m) = \frac{f(b) - f(a)}{b - a}. \quad (16.56)$$

Notice this looks a bit like the definition of the derivative, but starkly absent from the right side is any notion of limit.

The mean value theorem is easily proved. Consider the secant line

$$y(x) = \frac{f(b) - f(a)}{b - a}(x - a) + f(a)$$

that connects two points on a curve. Next, write the vertical distance between $f(x)$ and $y(x)$ as a new function $h(x)$

$$h(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a) - f(a),$$

and notice that $h = 0$ at the endpoints a, b .

Proceed by applying the d/dx operator (i.e. take a derivative) across the whole equation:

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}$$

Now, if $h(x)$ is zero at the endpoints and is nonzero in between, it must be that the derivative of $h(x)$ toggles between positive and negative at one (or more) points x_m in the interval $a < x_m < b$. This can only mean $h'(x_m) = 0$ at the transition(s), and the proof is done.

4.3 L'Hopital's Rule

When deploying tools of mathematics, there are all-too-often situations where indeterminate forms, infinities, division by zero, etc., can occur. This is supposed to be a show-stopper, however the notions of 'limit' and 'derivative' grant a new cutting edge.

Motivation

Consider the ratio L of two functions $f(x)$ and $g(x)$ evaluated at a particular point x_0 such that

$$L(x_0) = \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)},$$

where L is known to 'blow up' at x_0 , which is to say the ratio resolves to $0/0$, ∞/∞ , $0 \times \infty$, or similar indeterminate form.

With the notion of derivatives there is somewhere new to go, so let's try looking at the ratio of the *slope* of each function at x_0 ,

$$R(x_0) = \frac{f'(x_0)}{g'(x_0)},$$

and expand the right side using the definition of the derivative:

$$R = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{g(x) - g(x_0)} \left(\frac{x - x_0}{x - x_0} \right)$$

The 0/0 Case

Now impose the condition $f(x_0) = 0$ and $g(x_0) = 0$, and the ratio becomes

$$R(x_0) = \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)},$$

which is in fact identical to $L(x_0)$. This can only mean, for points x_0 that cause L to blow up,

$$L(x_0) = \lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \frac{f'(x_0)}{g'(x_0)}, \quad (16.57)$$

known as *L'Hopital's rule*.

In words, L'Hopital's rule says an indeterminate ratio of functions can be calculated *anyway* by calculating the

ratio of their slopes. If *that* result is indeterminate, apply L'Hopital's rule until an answer comes out. While L'Hopital's rule was established using the 0/0 case, the result is in fact quite general.

The ∞/∞ Case

To explore another extreme, suppose we have instead that $f(x_0) \rightarrow \infty$ and $g(x_0) \rightarrow \infty$.

Flipping the problem on its head slightly, one may write

$$L(x_0) = \lim_{x \rightarrow x_0} \frac{1/g(x)}{1/f(x)},$$

and then attack this using the chain rule. Doing so, we get

$$\begin{aligned} L(x_0) &= \frac{g'(x_0)}{f'(x_0)} \lim_{x \rightarrow x_0} \left(\frac{f(x)}{g(x)} \right)^2 \\ L(x_0) &= \frac{g'(x_0)}{f'(x_0)} (L(x_0))^2, \end{aligned}$$

and ultimately,

$$L(x_0) = \frac{f'(x_0)}{g'(x_0)},$$

familiar already as Equation (16.57).

The Infinite- x Case

Equation (16.57) is reinforced again by investigating the case there L blows up for $x \rightarrow \infty$. To proceed, define a variable $t = 1/x$ such that

$$L(x_0) = \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)} = \lim_{t \rightarrow 0} \frac{f(1/t)}{g(1/t)},$$

which transforms the problem into a 0/0-like problem.

Running the chain rule on the right side, we further find:

$$\begin{aligned} L(x_0) &= \lim_{t \rightarrow 0} \frac{f'(1/t)}{g'(1/t)} \left(\frac{-1/t^2}{-1/t^2} \right) \\ L(x_0) &= \lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)} \end{aligned}$$

Examples

You are encouraged to work through each of the following. For a bonus, pick out the example that helps establish that $0^0 = 1$.

Problem 17

Prove:

$$\lim_{x \rightarrow 0} \frac{\tan(x)}{x} = 1$$

Problem 18

Prove:

$$\lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x^2} = \lim_{x \rightarrow 0} \frac{\sin(x)}{2x} = \frac{1}{2}$$

Problem 19

Prove:

$$\lim_{x \rightarrow 0} \frac{e^x - 1 - x}{\sin^2(x)} = \frac{1}{2}$$

Problem 20

Prove:

$$\lim_{x \rightarrow 0^+} x \ln(x) = \lim_{x \rightarrow 0^+} \frac{\ln(x)}{1/x} = 0$$

Problem 21

Prove:

$$\lim_{x \rightarrow 0} \frac{\ln(x)}{x^p} = 0$$

4.4 Critical Points

Many mathematical functions $y = f(x)$, apart from lines and constants, exhibit features akin to 'hills' and 'valleys' in the Cartesian plane. The peak of any given hill is called a *local maximum*, unless it's the tallest hill, earning the title *global maximum*. Similar notions of 'local' and 'global' apply to valleys, i.e. *minima*.

Definition

The very peak of a hill or very bottom of a valley is called an *extreme* point, also known as *critical* point. A function $f(x)$ having critical point x_c implies that the left-sided and right-sided limits near $f(x_c)$ are equal:

$$\lim_{w \rightarrow 0} f\left(x_c - \frac{w}{2}\right) = \lim_{w \rightarrow 0} f\left(x_c + \frac{w}{2}\right) \quad (16.58)$$

Critical points are locations where the derivative of the function is zero, i.e.

$$f'(x_c) = 0.$$

While intuitive, this notion can be established using the definition

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

and then using the shift of variables

$$\begin{aligned}x &\rightarrow x_c + w/2 \\ x_0 &\rightarrow x_c - w/2\end{aligned}$$

such that $x - x_0 = w$, we write the symmetric form

$$f'(x_c) = \lim_{w \rightarrow 0} \frac{f(x_c + w/2) - f(x_c - w/2)}{w}.$$

This is perhaps a more ‘natural’ representation of the derivative compared to what we’ve been working with. Enforcing Equation (16.58) nukes the right side, leaving the result $f'(x_c) = 0$ as expected.

4.5 Optimization Problems

A very handy application of the derivative applies to problems of *optimization*. This is the broad set of ‘real-world’ problems that can be modeled as functions $f(x)$, where finding critical points $f'(x_c) = 0$ could mean maximizing profits, minimizing fuel consumption, etc.

The recipe for optimization problems is almost the same every time. From the situation on hand:

1. Identify the working variable x and construct a well-behaved function $f(x)$ that characterizes the problem.
2. Calculate $f'(x_c) = 0$ to identify critical point(s).
3. Feed any x_c back into $f(x)$ to produce the optimized solution(s).

Cylindrical Can

Problem 22

A cylindrical can of variable radius r and variable height h has fixed volume V . Find the dimensions of the can that minimize the surface area.

Answer: From the information given we can write the volume and surface area of the can:

$$\begin{aligned}V &= \pi r^2 h \\ A &= 2\pi r h + 2\pi r^2\end{aligned}$$

While there are two variables in play, r and h , we can write the area entirely in terms of r :

$$A(r) = \frac{2V}{r} + 2\pi r^2$$

The idea now is to find the critical point in $A(r)$. Do so by calculating $dA/dr = 0$, i.e.

$$\frac{dA}{dr} = 0 = -\frac{2V}{r^2} + 4\pi r,$$

implying $h_c = 2r_c$. Evidently, the most efficient can has the height equal to the diameter.

Distance to a Curve

Consider a curve $y = f(x)$ in the Cartesian plane, and also a given point (x_0, y_0) not on the curve. Let us find the minimal distance L that connects (x_0, y_0) to a point $(x_\star, y_\star)$ on the curve.

This is done by constructing the Euclidean distance

$$L^2 = (x_\star - x_0)^2 + (f(x_\star) - y_0)^2,$$

where $y_\star$ has been replaced with $f(x_\star)$. Use implicit differentiation to minimize with respect to $x_\star$:

$$2L \frac{dL}{dx} \Big|_{x_\star} = 2(x_\star - x_0) + 2(f(x_\star) - y_0) \frac{df}{dx} \Big|_{x_\star}$$

For L to be minimized, the left side is identically zero, and we settle on an equation that can be solved for $x_\star$:

$$0 = (x_\star - x_0) + (f(x_\star) - y_0) \frac{df}{dx} \Big|_{x_\star}$$

Problem 23

Find the point on the parabola $y = x^2$ that is closest to the point $(0, 1)$.

Answer: Establishing the distance L and minimizing gives

$$0 = (x_\star - 0) + (x_\star^2 - 1) 2x_\star,$$

indicating three possible solutions:

$$x_\star = 0, \quad x_\star = \frac{\pm 1}{\sqrt{2}}$$

This means there are two simultaneous solutions in y :

$$y_\star = 0, \quad y_\star = \frac{1}{2}$$

Of course $(x_\star, y_\star)$ is on the curve and we toss this solution out. In conclusion, we find two solutions $(\pm 1/\sqrt{2}, 1/2)$, and the minimum distance is:

$$L = \sqrt{\frac{1}{2} + \frac{1}{4}} = \frac{\sqrt{3}}{2}$$

Inscribed Rectangle**Problem 24**

Find the largest rectangle that fits inside a 3-4-5 right triangle where one of the rectangle's edges lies on the hypotenuse.

Answer: Place the ninety-degree corner at the origin so the hypotenuse connects $(0, 3)$ to $(4, 0)$. Parallel to the hypotenuse is the base of the inscribed rectangle with two corners at $(0, y_\star)$, $(x_\star, 0)$, having length

$$b = \sqrt{x_\star^2 + y_\star^2},$$

and obeying the ratio

$$\frac{y_\star}{x_\star} = \frac{3}{4}.$$

The height of the inscribed rectangle is

$$h = (4 - x_\star) \sin(\theta),$$

where $\sin(\theta) = 3/5$ from geometry.

The area of the rectangle is $A = bh$, or, all in terms of one variable:

$$\begin{aligned} A(x_\star) &= \sqrt{x_\star^2 + y_\star^2} (4 - x_\star) \frac{3}{5} \\ &= x_\star (4 - x_\star) \end{aligned}$$

The critical point x_c is found by calculating $dA/dx_\star = 0$, namely

$$\frac{dA}{dx_\star} = 0 = 4 - 2x_c,$$

solved by $x_c = 2$, immediately meaning $y_c = 3/2$. Calculating b from these values yields $5/2$, which is half of the length of the hypotenuse. The height comes out to $h = 6/5$, and thus the area of the rectangle is $A = 3$.

Pi versus E**Problem 25**

Prove that

$$e^\pi > \pi^e.$$

Answer: First use the logarithm operator to get like symbols on their own sides:

$$\begin{aligned} \pi \ln(e) &= e \ln(\pi) \\ \frac{\ln(e)}{e} &> \frac{\ln(\pi)}{\pi} \end{aligned}$$

This is suggestive of the function $f(x) = \ln(x)/x$, and the question translates to whether $f(e)$ is larger or smaller than $f(\pi)$.

Proceed by calculating $df/dx = 0$:

$$0 = \frac{d}{dx} \left(\frac{\ln(x)}{x} \right) = \frac{1 - \ln(x_c)}{x_c^2}$$

From this, we have that $1 = \ln(x_c)$, satisfied by $x_c = e$, and the proof is done.

Hole in a Bucket**Problem 26**

A large cylindrical bucket of height H is full of water. You punch a hole in the side of the bucket at height $h < H$. Using Bernoulli's principle, it's possible to show that the velocity of water coming from the hole is

$$v = \sqrt{2g(H - h)},$$

where g is the local gravity constant. Show that the maximum horizontal distance of the water is achieved when $h = H/2$.

Answer: From kinematics, we write the condition for hitting the ground at time $t = t_\star$:

$$\begin{aligned} x_\star &= vt_\star \\ y_\star &= 0 = h - gt_\star^2/2 \end{aligned}$$

Eliminate $t_\star$ and v in the $x_\star$ -equation to find, after simplifying

$$x_\star = 2\sqrt{Hh - h^2}.$$

The condition for maximal distance occurs at $dx_\star/dh = 0$, or:

$$0 = \frac{dx_\star}{dh} = \frac{H - 2h}{\sqrt{Hh - h^2}} \rightarrow h_\star = H/2$$

4.6 Related Rate Problems

Implicit differentiation has some utility for analyzing 'real world' problems that aren't a matter of optimization. Instead, we may be concerned with the way one rate of change related to another, a class called *related rate* problems.

Melting Ice SheetProblem 27

A circular ice sheet of radius $r(t)$ in meters and area $A(t)$ is melting at a rate of $-\alpha \text{ m}^2/\text{s}$. How quickly is the radius decreasing?

Answer: Begin with

$$A(t) = \pi (r(t))^2,$$

and use implicit differentiation with respect to time:

$$\frac{d}{dt} A(t) = -\alpha = 2\pi r(t) \frac{d}{dt} (r(t))$$

The time derivative of A is given as $-\alpha$, whereas the time derivative of $r(t)$ is the quantity we're solving for.

So far then, we have

$$\frac{d}{dt} (r(t)) = r'(t) = \frac{-\alpha}{2\pi r(t)},$$

or in terms of A instead of r :

$$r' = \frac{-\sqrt{\pi}\alpha}{2\sqrt{A(t)}}$$

Distance from a RocketProblem 28

A person stands distance D away from a rocket that launches straight up with speed v_0 at $t = 0$. Write an equation for the distance r from the person to the rocket as a function of time, and then determine its derivative, $r'(t)$.

Answer: Use the Pythagorean theorem

$$r^2 = D^2 + v_0^2 t^2$$

to get started, and use implicit differentiation with respect to time:

$$2r(t) \frac{d}{dt} (r(t)) = 0 + 2v_0^2 t$$

Isolate $r'(t)$ to finish the job:

$$r'(r) = \frac{v_0^2 t}{\sqrt{D^2 + v_0^2 t^2}}$$

Tapered CandleProblem 29

A tapered candle approximated by a cone is burning at a rate of $\alpha = 1 \text{ cm}^3/\text{hr}$. Determine the rate at which the height h of the candle is decreasing in time. The equation of a cone is

$$r^2 = x^2 + y^2 = \beta z^2$$

and the volume of a cone is

$$V = \frac{1}{3}\pi r^2 h.$$

Answer:

$$\frac{dh}{dt} = \frac{\alpha h}{3V}$$

5 Second Derivative

Recall that, for a differentiable function $f(x)$, the notion of ‘slope at a point’, i.e. the derivative, can be expressed a few ways:

$$\text{Slope at a point} = f'(x) = \frac{d}{dx}f(x) = f^{(1)}(x)$$

5.1 Slope of Slope

If $f'(x)$ is itself a differentiable function, there must exist the notion of ‘slope of slope’, also known as the *second derivative* of $f(x)$:

$$\text{Second Derivative} = f''(x) = \frac{d^2}{dx^2}f(x) = f^{(2)}(x)$$

Starting with the definition of the derivative, a formula for the second derivative can be straightforwardly written:

$$f''(x_0) = \lim_{x \rightarrow x_0} \frac{f'(x) - f'(x_0)}{x - x_0}$$

Carefully replacing each f' -term with the definition again, we get

$$f''(x_0) = \lim_{x \rightarrow x_0} \frac{1}{x - x_0} \left(\lim_{w \rightarrow x} \frac{f(w) - f(x)}{w - x} - \frac{f(x) - f(x_0)}{x - x_0} \right),$$

or, after some algebra,

$$f''(x_0) = \lim_{w \rightarrow x} \lim_{x \rightarrow x_0} \left(\frac{1}{x - x_0} \right)^2 (\lambda f(w) - (\lambda + 1)f(x) + f(x_0)),$$

where

$$\lambda = \frac{x - x_0}{w - x}.$$

The above contains two simultaneous limits, namely $w \rightarrow x$ and $x \rightarrow x_0$. Applying each limit together, it should make sense that the ratio λ resolve to $\lambda = 1$ provided that $x_0 < x < w$. With this, we can set $x - x_0 = h$ and $w - x = h$, and the above becomes

$$f''(x_0) = \lim_{h \rightarrow 0} \lim_{x \rightarrow x_0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2},$$

simplifying once more to the standard formula for the second derivative:

$$f''(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0+h) + f(x_0-h) - 2f(x_0)}{h^2} \quad (16.59)$$

In practice, one does not need to directly deploy Equation (16.59) to calculate the second derivative. So long as the first derivative $f'(x)$ is on hand, simply calculate the derivative of *that* to get a hold of $f''(x)$.

5.2 Critical Point Stability

The second derivative $f''(x)$ carries important information about the function $f(x)$. To illustrate, consider the cubic curve

$$f(x) = \left(x + \frac{1}{2}\right)^3 - 3\left(x + \frac{1}{2}\right) - \frac{1}{2}$$

as shown in Figure 16.1.

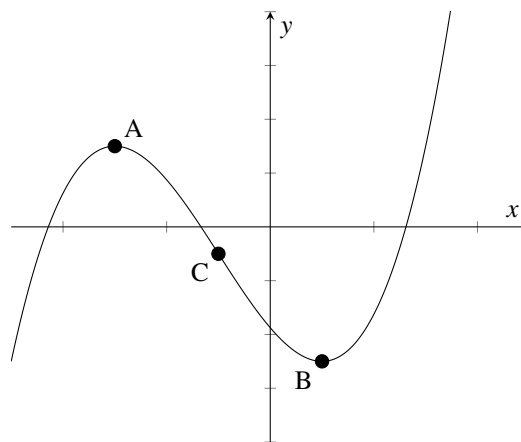


Figure 16.1. Cubic curve having two critical points and one inflection point.

Labeled in the Figure are three Cartesian points A , B , C . By a quick inspection, one sees that A and B correspond to critical points, and these can be found by the standard means of setting $f'(x) = 0$. Doing so, we first find

$$f'(x) = 3\left(x + \frac{1}{2}\right)^2 - 3$$

for the entire curve, and $f'(x) = 0$ is solved by:

$$\begin{aligned} x_A &= -3/2 \\ x_B &= 1/2 \end{aligned}$$

Concavity

While both A and B qualify as critical points, there is something clearly different about them in the sense that point A corresponds to a local maximum, and point B corresponds to a local minimum.

Introducing some new terminology, the curve $f(x)$ is *concave down* at and near point A , as if the only way to go is downhill. Conversely, in the ‘neighborhood’ of the local minimum at B , the curve is *concave up*.

In a mechanical analogy, a local maximum (such as A) is often called an *unstable equilibrium*, as if the curve $f(x)$ is embedded uniform gravity and there is a clear notion teetering on the top of a hill. A local minimum (such as B) is called a *stable equilibrium* for similar reasons.

The notion of concavity or stability begs a new question, namely how can we tell if part of a curve is concave up versus concave down without looking at the plot of the function? This is where the second derivative comes in. Starting from $f'(x)$ written above, the second derivative comes out to

$$f''(x) = 6x + 3$$

alone the whole curve.

Since we know the critical points occur at $x_A = -3/2$, $x_B = 1/2$, toss these into $f''(x)$ to learn

$$f''(-3/2) = 6(-3/2) + 3 = -6$$

$$f''(1/2) = 6(1/2) + 3 = 6.$$

Evidently, the *sign* on the second derivative tells the story of concavity. When negative, the curve is concave down. When positive, the curve is concave up.

Inflection

Since the derivative operation ‘knocks down’ one order of x from the function, it follows that the second derivative of a cubic curve is a straight line, $y = 6x + 3$ in this case. Furthermore, the second derivative has an x -intercept at $x_C = -1/2$, which is why point C is significant in Figure 16.1.

Point C is called an *inflection point*, which is where the second derivative $f''(x)$ is momentarily zero, and the concavity of the curve flips from downward to upward.

To summarize the role of the second derivative in general:

$$f''(x) = \begin{cases} < 0 & \text{Concave down} \\ = 0 & \text{Inflection} \\ > 0 & \text{Concave up} \end{cases}$$

5.3 Three Roads Problem

In the Cartesian plane, consider two parabola-shaped ‘roads’ described by

$$y_1(x) = -\frac{x^2}{4} - 1$$

$$y_2(x) = \frac{x^2}{4} + 1.$$

Also, suppose it’s our job to propose a new road $f(x)$ connecting the point $(-2, -2)$ on $y_1(x)$ to another point $(1, 5/4)$ on $y_2(x)$. The new road should not introduce any unwarranted forces on a vehicle.

Straight Line Approximation

The easiest solution to write down, which happens to also be the shortest road connecting the given points, is a straight line

$$f_1(x) = mx + b.$$

Using the information provided, it follows that the line connecting the points is specified via

$$f(-2) = y_1(-2)$$

$$f(1) = y_2(1),$$

and solved by:

$$m = \frac{5/4 - (-2)}{1 - (-2)} = \frac{13}{12}$$

$$b = \frac{1}{6}$$

However, the instantaneous derivative of each $y_{1,2}(x)$ tells us

$$\left(\frac{d}{dx} y_1(x) \right) \Big|_{-2} = 1$$

$$\left(\frac{d}{dx} y_2(x) \right) \Big|_1 = \frac{1}{2},$$

and neither is equal to m . This means that the straight line approximation induces a break in the smoothness of the ride at each transition.

Cubic Approximation

Trying a slightly more versatile candidate, consider the order-three approximation

$$f_2(x) = A_0 + A_1x + A_2x^2,$$

where each A_j is an unknown coefficient, three in total. However, we've already discerned that (at least) four equations govern the system. There are one too few unknowns on hand.

The next best thing to do is guess a cubic equation:

$$f_3(x) = A_0 + A_1x + A_2x^2 + A_3x^3$$

With the cubic approximation, we can find all unknown coefficients by requiring $f(p) = y_{1,2}(x)$ and $f'(x) = y'_{1,2}(x)$ where the roads intersect.

Note that whatever $f(x)$ is doing for $x < -2$ or $x > 1$ doesn't quite matter. This is why adding a cubic term, which undoubtedly changes the global shape of $f(x)$, happens to provide extra tuning in the domain $-2 \leq x \leq 1$.

Quintic Approximation

Two more equations can be extracted from the information provided, namely the second derivative of each road $y_{1,2}(x)$ at the points of transition:

$$\left(\frac{d^2}{dx^2} y_1(x) \right) \Big|_{-2} = \frac{-1}{2}$$

$$\left(\frac{d^2}{dx^2} y_2(x) \right) \Big|_1 = \frac{1}{2},$$

Including two more equations justifies introducing two more unknowns, so we may as well use an order-five polynomial

$$f_5(x) = \sum_{j=0}^5 A_j x^j$$

having six unknowns A_j . The first two derivatives of $f_5(x)$ are easy to jot down:

$$\frac{d}{dx} f_5(x) = \sum_{j=1}^5 j A_j x^{j-1}$$

$$\frac{d^2}{dx^2} f_5(x) = \sum_{j=2}^5 j(j-1) A_j x^{j-2}$$

Of course, the second derivative of $f_5(x)$ at the respective endpoints is $-1/2, 1/2$.

N Equations and Unknowns

To keep things general, let the given endpoints be represented by $(x_{1,2}, y_{1,2})$. Let the first and second derivatives of $y_{1,2}(x)$ at the endpoints be denoted $y'_{1,2}(x)$, $y''_{1,2}(x)$, respectively.

Note too that most of the curves $y_{1,2}(x)$ don't play into the solution. The relevant information is the location of each endpoints and the derivative(s). Working all of this out, the full information of the problem is specified in the augmented matrix

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 & y_1 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 & y_2 \\ 0 & 1 & 2x_1 & 3x_1^2 & 4x_1^3 & 5x_1^4 & y'_1 \\ 0 & 1 & 2x_2 & 3x_2^2 & 4x_2^3 & 5x_2^4 & y'_2 \\ 0 & 0 & 2 & 3 \cdot 2x_1^2 & 4 \cdot 3x_1^3 & 5 \cdot 4x_1^4 & y''_1 \\ 0 & 0 & 2 & 3 \cdot 2x_2^2 & 4 \cdot 3x_2^3 & 5 \cdot 4x_2^4 & y''_2 \end{bmatrix}.$$

Solution

To finish the example on hand with the data provided, the above becomes

$$\begin{bmatrix} 1 & -2 & 4 & -8 & 16 & -32 & -2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 5/4 \\ 0 & 1 & -4 & 12 & -32 & 80 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 1/2 \\ 0 & 0 & 2 & -12 & 48 & -160 & -1/2 \\ 0 & 0 & 2 & 6 & 12 & 20 & 1/2 \end{bmatrix},$$

with corresponding RREF:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 43/81 \\ 0 & 1 & 0 & 0 & 0 & 0 & 94/81 \\ 0 & 0 & 1 & 0 & 0 & 0 & -149/324 \\ 0 & 0 & 0 & 1 & 0 & 0 & -23/162 \\ 0 & 0 & 0 & 0 & 1 & 0 & 19/162 \\ 0 & 0 & 0 & 0 & 0 & 1 & 7/162 \end{bmatrix},$$

exposing the coefficients $\{A_j\}$. All three roads are plotted together in Figure 16.2.

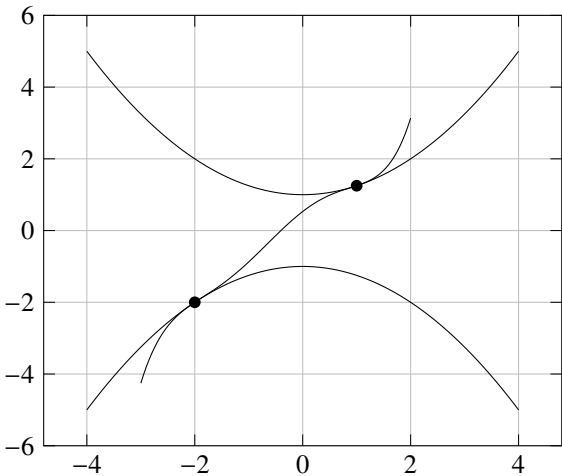


Figure 16.2. Three roads.

6 Taylor's Theorem

6.1 Kinematic Motivation

In freshman kinematics, one encounters the equations of motion under uniform acceleration

$$\begin{aligned}x(t) &= x_0 + v_0 t + \frac{1}{2} a t^2 \\v(t) &= v_0 + a t,\end{aligned}$$

where $x(t)$ and $v(t)$ are the position and velocity, respectively, with their initial values written as

$$\begin{aligned}x(0) &= x_0 \\v(0) &= v_0,\end{aligned}$$

all the while acceleration a is held constant in time t .

Uniform Jerk

Extending the picture of kinematics, we consider the acceleration being allowed to vary. The simplest regime has acceleration varying *linearly* in time such that the derivative of $a(t)$ is constant called *jerk*, denoted j . In such a case, two kinematic equations are readily evident:

$$\begin{aligned}v(t) &= v_0 + a_0 t + \frac{1}{2} j t^2 \\a(t) &= a_0 + j t\end{aligned}$$

The equation for $x(t)$ is a little more tricky though. Going from the pattern, there should be a new term proportional to $j t^3$, but the leading coefficient must be left as a variable

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3,$$

and the issue is deciding what A should be.

6.2 Time-Shift Analysis

Solving for A

To solve the riddle of the $1/A$ -coefficient, consider a shift in the time variable

$$t \rightarrow t + h,$$

where h is any constant. Inserting this into the above gives

$$\begin{aligned}x(t+h) &= x_0 + v_0(t+h) \\&\quad + \frac{1}{2} a_0 (t+h)^2 + \frac{1}{A} j (t+h)^3,\end{aligned}$$

and now the job is to expand all factors involving $(t+h)$. Doing so, and then combining like terms in powers of h , something interesting happens:

$$\begin{aligned}x(t+h) &= \left(x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3 \right) \\&\quad + h \left(v_0 + a_0 t + \frac{3}{A} j t^2 \right) \\&\quad + \frac{1}{2} h^2 \left(a_0 + \frac{6}{A} j t \right) + \frac{1}{6} h^3 (j)\end{aligned}$$

From this, we see the only way to correctly recover the identities already written is to have

$$A = 6$$

and no other choice suffices.

Time-Shifted Kinematics

Looking at the expanded $x(t+h)$ equation while knowing that $A = 6$, recall that the first parenthesized group of terms is just $x(t)$. Similarly the second group is $v(t)$, the third group, $a(t)$, and so on. Therefore we can write the same equation as

$$x(t+h) = x_t + v_t h + \frac{1}{2} a_t h^2 + \frac{1}{6} j h^3,$$

which is quite a beautiful result. In effect, we can pretend that t is constant, and h does the whole job of the time variable. Any point t along the path of motion can be considered as the 'initial' state.

6.3 Generalized Kinematics

Using the same procedures that led us to finding $A = 6$ in the kinematics-with-jerk analysis, it's straightforward to incorporate higher derivatives into the equations of motion. The derivative of jerk is called *snip*, denoted k . (Beyond this the derivatives aren't conventionally named.) Going through the exercise, one finds

$$x(t+h) = x_t + v_t h + \frac{a_t}{2!} h^2 + \frac{j_t}{3!} h^3 + \frac{k_t}{4!} h^4 + \dots$$

The factorial operator is used to tightly represent the kinematic coefficients.

To handle t being considered fixed while h is the varying quantity, let us relabel t to t_p , as in 'time at some special point', and write the *effective* time variable as

$$t = t_p + h.$$

The above transforms into

$$x(t) = x_{t_p} + v_{t_p}(t - t_p) + \frac{1}{2!}a_{t_p}(t - t_p)^2 + \frac{1}{3!}j_{t_p}(t - t_p)^3 + \dots$$

Using a generalized notation to represent the velocity, acceleration, jerk, and so on, let us make the associations

$$\begin{aligned} x_{t_p} &\rightarrow x_{t_p}^{(0)} \\ v_{t_p} &\rightarrow x_{t_p}^{(1)} \\ a_{t_p} &\rightarrow x_{t_p}^{(2)} \\ j_{t_p} &\rightarrow x_{t_p}^{(3)} \\ k_{t_p} &\rightarrow x_{t_p}^{(4)}, \end{aligned}$$

and so on. On the left we've run out of 'named' items after snap, thus the general symbol $x_{t_p}^{(q)}$ is utilized to denote the q th coefficient.

Taylor Polynomial

In condensed form, $x(t)$ can be written in a most general way using summation notation

$$x(t) = x_{t_p} + \sum_{q=1}^n \frac{1}{q!} x_{t_p}^{(q)} (t - t_p)^q + R_n(t), \quad (16.60)$$

known as the *Taylor polynomial*. The upper limit n can be any natural number, depending on the total number of motion coefficients in play.

The so-called 'remainder' term $R_n(t)$ contains the rest of the terms not included in the main sum. If $R_n(t)$ tends to zero for increasing n , the Taylor polynomial converges. The polynomial diverges if $R_n(t)$ fails to vanish for large n .

6.4 Taylor's Theorem

The Taylor polynomial for generalized kinematics is an extremely powerful and general result that is the center of *Taylor's theorem*. In a phrase, Taylor's theorem states that any n -times differentiable function can be approximated a Taylor polynomial or order n .

In terms of a function $f(x)$, near the point x_0 , Taylor's theorem reads

$$p(x) = f(x_0) + \sum_{q=1}^n \frac{1}{q!} f^{(q)}(x_0) (x - x_0)^q + R_n(x), \quad (16.61)$$

where $f^{(q)}(x_0)$ is the q th derivative of $f(x)$ evaluated at x_0 . The approximation $p(x)$ may or may not successfully approximate the entire function in its domain, but it does a great job in the neighborhood of x_0 in any case.

A less pedantic statement of Taylor's theorem omits the remainder unless it becomes necessary, and also acknowledges its approximate nature by replacing the equal sign:

$$f(x) \approx f(x_0) + \sum_{q=1}^n \frac{1}{q!} f^{(q)}(x_0) (x - x_0)^q$$

Proof of Taylor's Theorem

A formal proof of Taylor's theorem can begin with a new function $h_n(x)$ defined as:

$$h_n(x) = \begin{cases} (f(x) - p(x)) / ((x - x_0)^n) & x \neq x_0 \\ 0 & x = x_0 \end{cases}$$

It is required that $f(x)$ is an n -times differentiable function and $p(x)$ is the Taylor polynomial appearing in Equation (16.61).

The theorem is considered proven when we show that $h_n(x) = 0$ for all x near x_0 . Setting this up, begin with

$$H_n = \lim_{x \rightarrow x_0} h_n(x) = \lim_{x \rightarrow x_0} \frac{f(x) - p(x)}{(x - x_0)^n},$$

and recognize the right side as an indeterminate ratio.

Indeterminate ratios of this kind are handled by L'Hopital's rule, thus apply the d/dx operator to the numerator and denominator:

$$\lim_{x \rightarrow x_0} \frac{\frac{d}{dx}(f(x) - p(x))}{\frac{d}{dx}((x - x_0)^n)} = \lim_{x \rightarrow x_0} \frac{f^{(1)}(x) - p^{(1)}(x)}{n(x - x_0)^{n-1}}$$

The right side is again an indeterminate ratio, calling for another application of L'Hopital's rule:

$$H_n = \lim_{x \rightarrow x_0} \frac{f^{(2)}(x) - p^{(2)}(x)}{n(n-1)(x - x_0)^{n-2}}$$

In fact, this pattern continues $n-1$ times, with each application of L'Hopital's rules knocking down the exponent in the denominator. Exhausting this loop, one should find:

$$H_n = \frac{1}{n!} \left(\lim_{x \rightarrow x_0} \frac{f^{(n-1)}(x) - p^{(n-1)}(x)}{x - x_0} \right)$$

The parenthesized limit is equivalent to the definition of the derivative of $f^{(n)}(x)$ evaluated at x_0 . Or, use L'Hopital

once more to sap the denominator entirely, and the quantity H_n evaluates to

$$H_n = \frac{1}{n!} \left(f^{(n)}(x_0) - f^{(n)}(x_0) \right) = 0$$

and the proof is done.

Order of Approximation

In certain scenarios, especially when working near the point x_0 , it suffices to truncate the Taylor polynomial to a small, finite number of terms. This works only when the sum converges 'rapidly enough' so that higher powers of $x - x_0$ become negligible. To put a label to the first few approximations, we have, for a function $f(x)$:

Zeroth Order:

$$p_0(x) = f(x_0)$$

First Order:

$$p_1(x) = f(x_0) + f'(x_0)(x - x_0)$$

Second Order:

$$p_2(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2$$

6.5 Testing Taylor's Theorem

In a certain sense, Taylor's theorem contains the entire lesson of elementary calculus. Here we spend a moment recovering some already-known results.

Geometric Series

Consider the function

$$f(x) = \frac{1}{1-x}$$

in the domain $|x| < 1$. Near any point x_0 , the infinite Taylor polynomial approximation to $f(x)$ reads:

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \dots$$

The derivatives f' , f'' , etc., are straightforwardly attained from $f(x)$:

$$\begin{aligned} f'(x) &= 1/(x - x_0)^2 \\ f''(x) &= 2!/(x - x_0)^3 \\ f'''(x) &= 3!/(x - x_0)^4 \end{aligned}$$

Substituting these into $p(x)$ and performing the obvious cancellations gives:

$$p(x) = \frac{1}{1-x_0} \left(1 + \lambda + \lambda^2 + \dots \right),$$

where for brevity, λ contains the x -dependence via

$$\lambda = \frac{x - x_0}{1 - x_0}.$$

Note, of course, that the parenthesized sum containing powers of λ is a geometric series guaranteed to converge because $|\lambda| < 1$. Realizing this, replace the infinite sum with the ratio $1/(1 - \lambda)$ as

$$p(x) = \frac{1}{1-x_0} \frac{1}{1-\lambda},$$

simplifying to

$$p(x) = \frac{1}{1-x} = f(x).$$

Evidently, the infinite Taylor polynomial approximation of the geometric series is no approximation at all - the result is exact.

Trigonometric Functions

Another regime where the Taylor polynomial exactly approximates the function is in trigonometry, namely the sine and cosine. Choosing the sine function to play with, we have $f(x) = \sin(x)$, and then,

$$\begin{aligned} f'(x_0) &= \cos(x_0) \\ f''(x_0) &= -\sin(x_0) \\ f'''(x_0) &= -\cos(x_0) \\ f^{(4)}(x_0) &= \sin(x_0) \end{aligned}$$

and by the fourth derivative we're back to $\sin(x)$.

The simplest case has $x_0 = 0$, which nixes all of the sine terms in the list of evaluated derivatives. The cosines all resolve to either 1 or -1 , and we quickly find

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

Choosing a juicier example, let $x_0 = \pi/2$ and all the signs change:

$$\sin(x) = 1 - \frac{(x - \pi/2)^2}{2!} + \frac{(x - \pi/2)^4}{4!} + \dots$$

On the right is the polynomial expression for the cosine offset by $-\pi/2$, and the whole thing reduces to the trig identity

$$\sin(x) = \cos\left(x - \frac{\pi}{2}\right).$$

6.6 Recovering Differentiation Rules

Taylor's theorem also jives with the rules of differentiation.

Product and Quotient Rules

Consider two differentiable functions $f(x)$, $g(x)$. From these, construct the product $P(x) = f(x) \cdot g(x)$ along with the quotient $Q(x) = f(x)/g(x)$. A question that immediately arises from this is, what are the first-order approximations to $P(x)$, $Q(x)$?

To handle the product case, write each function $f(x)$, $g(x)$ to first-order approximation,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0)$$

with the understanding that x is near x_0 .

Denoting

$$\Delta x = x - x_0,$$

the product $P(x)$ reads

$$\begin{aligned} P(x) &\approx f(x_0)g(x_0) \\ &+ \Delta x (f'(x_0)g(x_0) + f(x_0)g'(x_0)) \\ &+ \cancel{(\Delta x)^2 f'(x_0)g'(x_0)} \end{aligned}$$

where the term $(\Delta x)^2$ is negligible compared to the others and is dropped.

The middle term in the above is Δx multiplied by the derivative of the product $f(x)g(x)$ per Equation (16.32), i.e. the product rule. After simplifying, we can summarize by writing the first-order approximation to $P(x)$:

$$P_1(x) = f(x_0)g(x_0) + \frac{d}{dx}(f(x)g(x)) \Big|_{x_0} \Delta x \quad (16.62)$$

The case for quotients is a little harder. To prepare, let us apply Taylor's theorem to $1/g(x)$ on its own. Begin using the first order approximation for $g(x)$ via

$$\frac{1}{g(x)} \approx \lim_{x \rightarrow x_0} \frac{1}{g(x_0) + g'(x_0)(x - x_0)},$$

and then factor out $1/g(x_0)$:

$$\frac{1}{g(x)} \approx \frac{1}{g(x_0)} \lim_{x \rightarrow x_0} \frac{1}{1 + \lambda},$$

where the x -dependence is wrapped up in λ :

$$\lambda = \frac{g'(x_0)}{g(x_0)}(x - x_0)$$

Like we've seen before, it suffices to proceed with $|\lambda| < 1$ for all x , and the fraction $1/(1 + \lambda)$ can be replaced with the geometric series:

$$\frac{1}{1 + \lambda} = 1 - \lambda + \lambda^2 - \lambda^3 + \dots$$

Of course, terms λ^2 and above are omitted in the first-order approximation, thus we have

$$\frac{1}{g(x)} \approx \frac{1}{g(x_0)} \lim_{x \rightarrow x_0} \left(1 - \frac{g'(x_0)}{g(x_0)}(x - x_0) \right).$$

The first-order approximation to $Q(x)$ can be taken as the product $f_1(x)$ and $1/g_1(x)$. Doing this out while dropping the inevitable $(\Delta x)^2$ term, we find:

$$Q(x) \approx \frac{f(x_0)}{g(x_0)} + \Delta x \left(\frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g(x_0))^2} \right)$$

The latter term in the above is Δx multiplied by the derivative of the quotient $f(x)/g(x)$ per Equation (16.33), i.e. the quotient rule. After simplifying, we can summarize by writing the first-order approximation to $Q(x)$:

$$Q_1(x) = \frac{f(x_0)}{g(x_0)} + \frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) \Big|_{x_0} \Delta x \quad (16.63)$$

Chain Rule

Consider the composite function

$$C(x) = f(g(x)).$$

To a first-order approximation the functions f and g obey

$$f(g) \approx f(g_0) + f'(g_0)(g - g_0)$$

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0),$$

where $g_0 = g(x_0)$.

With these, the composite function reads

$$f(g(x)) \approx f(g(x_0)) + f'(g(x_0))g'(x_0)(x - x_0),$$

or more succinctly:

$$C_1(x) = f(g(x_0)) + f'(g(x_0))g'(x_0)\Delta x \quad (16.64)$$

Second Derivative

The formula for the second derivative can also be wriggled from the Taylor polynomial. First write the standard approximation of $f(x)$:

$$\begin{aligned} f(x) &\approx f(x_0) + f'(x_0)(x - x_0) \\ &\quad + \frac{f''(x_0)}{2!}(x - x_0)^2 \\ &\quad + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \dots \end{aligned}$$

The same function can be approximated from a different base point, namely $x \rightarrow 2x_0 - x$. Writing this out, we have

$$\begin{aligned} f(2x_0 - x) &\approx f(x_0) + f'(x_0)(x_0 - x) \\ &\quad + \frac{f''(x_0)}{2!}(x_0 - x)^2 \\ &\quad + \frac{f'''(x_0)}{3!}(x_0 - x)^3 + \dots, \end{aligned}$$

which has the effect of reversing the sign on Δx on the odd-powered terms.

Next take the sum of the two equations to make all odd-powered terms cancel, and then reshuffle a little to write

$$\begin{aligned} \frac{f(x) + f(2x_0 - x) - 2f(x_0)}{(x - x_0)^2} &\approx f''(x_0) \\ &\quad + \frac{2}{2!} f'''(x_0) \Delta x^2, \end{aligned}$$

where terms containing powers of Δx^2 and above are negligible in a first-order approximation.

In the infinitesimal limit $x \rightarrow x_0$, the above reduces to the exact formula for the second derivative, Equation (16.59). Specifically, let $x - x_0 = h$ and let h go to zero.

6.7 Binomial Expansion

Consider the function

$$f(x) = (x + a)^r,$$

where a is a constant and r is an arbitrary exponent. In preparation for Taylor's theorem, crank out the first few derivatives of $f(x)$ and spot the pattern:

$$\begin{aligned} f^{(1)}(x_0) &= r(x_0 + a)^{r-1} \\ f^{(2)}(x_0) &= r(r-1)(x_0 + a)^{r-2} \\ f^{(3)}(x_0) &= r(r-1)(r-2)(x_0 + a)^{r-3} \\ f^{(q)}(x_0) &= \frac{r!}{(r-q)!}(x_0 + a)^{r-q} \end{aligned}$$

As a sum, the approximation for $f(x)$ then reads

$$\begin{aligned} f(x) &\approx (x_0 + a)^r \\ &\quad + \sum_{q=1}^n \frac{r!}{q!(r-q)!} (x_0 + a)^{r-q} (x - x_0)^q. \end{aligned}$$

Next, impose the condition

$$x \approx x_0 = 0,$$

which causes increasing powers of Δx^q tend to zero quickly. The above becomes

$$f(x) \approx a^r + a^r \sum_{q=1}^n \frac{r!}{q!(r-q)!} \left(\frac{x}{a}\right)^q,$$

and the condition $x \approx 0$ is represented by $(x/a)^q$ tending to zero for increasing q .

Binomial Coefficients

The pattern of factorials in the above are the familiar *binomial coefficients*, which follow a special notation:

$$\binom{r}{q} = \frac{r!}{q!(r-q)!}$$

In terms of binomial coefficients, the sum representing $f(x)$ is written

$$(x + a)^r \approx a^r \sum_{q=0}^n \binom{r}{q} \left(\frac{x}{a}\right)^q \quad (16.65)$$

valid for 'small' x . This is called the *binomial expansion* formula. The above can be written in open form for more practical use:

$$\begin{aligned} (x + a)^r &\approx a^r + r a^{r-1} x + \frac{r(r-1)}{2!} a^{r-2} x^2 \\ &\quad + \frac{r(r-1)(r-2)}{3!} a^{r-3} x^3 + \dots \end{aligned}$$

Examples

Problem 30

Expand $\sqrt{1+x}$ for small x . Answer:

$$\sqrt{1+x} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \dots \quad (16.66)$$

Problem 31

Expand $1/\sqrt{1+x}$ for small x . Answer:

$$\frac{1}{\sqrt{1+x}} \approx 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \dots \quad (16.67)$$

6.8 Generalized Taylor Expansion

Taylor's theorem can be used to approximate any differentiable function in addition to polynomials.

Shifted Natural Logarithm

Consider the shifted natural logarithm

$$f(x) = \ln(1+x) .$$

At a point x_0 , the derivatives of $f(x)$ are:

$$f^{(1)}(x_0) = 1/(1+x_0)$$

$$f^{(2)}(x_0) = -1/(1+x_0)^2$$

$$f^{(3)}(x_0) = 2/(1+x_0)^3$$

$$f^{(4)}(x_0) = -3 \cdot 2/(1+x_0)^4$$

$$f^{(q)}(x_0) = (-1)^{q-1} (q-1)!/(1+x_0)^q$$

Then, the approximation for $f(x)$ near x_0 reads

$$f(x) \approx \ln(1+x_0) + \sum_{q=1}^n \frac{(-1)^{q-1}}{q} \frac{(x-x_0)^q}{(1+x_0)^q} .$$

This result boils down to a quaint infinite series for x near zero:

$$\ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \quad (16.68)$$

Arctangent Near Zero

Using Taylor's theorem involves derivative calculations that can get increasingly messy without an obvious pattern showing.

To demonstrate, let's run through the exercise using $f(x) = \arctan(x)$, where we have:

$$f^{(1)}(x_0) = \frac{1}{1+x_0^2}$$

$$f^{(2)}(x_0) = \frac{-2x_0}{(1+x_0^2)^2}$$

$$f^{(3)}(x_0) = \frac{6x_0^2 - 2}{(1+x_0^2)^3}$$

$$f^{(4)}(x_0) = \frac{-24x_0(x_0^2 - 1)}{(1+x_0^2)^4}$$

$$f^{(5)}(x_0) = \frac{24(5x_0^4 - 10x_0^2 + 1)}{(1+x_0^2)^5}$$

Clearly the derivatives are not exhibiting a clear pattern. To reign in the work we're doing, let $x_0 = 0$ and simplify to end up with an infinite series approximation for the $\arctan(x)$ near $x = 0$:

$$\arctan(x) \approx x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad (16.69)$$

The last term $x^7/7$ wasn't directly calculated, but tacked on due to the prevailing pattern in the coefficients. This move isn't safe unless you're sure the pattern really is there.

Arctangent Near One

Taking another easy case, the arctangent near $x = 1$ can be found in the same way as above, resulting in:

$$\begin{aligned} \arctan(x)_{x \approx 1} &= \frac{\pi}{4} + \frac{x-1}{2} \\ &\quad - \frac{(x-1)^2}{4} + \frac{(x-1)^3}{12} \\ &\quad - \frac{(x-1)^5}{40} + \frac{(x-1)^6}{48} - \dots \end{aligned} \quad (16.70)$$

Arctangent Near Two

For the sake of completeness, the arctangent near $x = 2$ case works out to be:

$$\begin{aligned} \arctan(x)_{x \approx 2} &= \arctan(2) + \frac{x-2}{5} \\ &\quad - \frac{2(x-2)^2}{25} + \frac{11(x-2)^3}{375} - \dots \end{aligned} \quad (16.71)$$

Arctangent at Two

It's worth pausing a moment on the quantity $\arctan(2)$, which is required to evaluate Equation (16.71).

A situation with $\arctan(2)$ could arise from a right triangle with adjacent side 1, opposite side 2, and hypotenuse $\sqrt{5}$. No rational multiple of π radians or degrees corresponds to the interior angles of such a triangle. Moreover, Equation (16.70) cannot be used as $x = 2$ is outside the valid domain of approximation.

To crack this problem, consider some argument z and list the trig identity

$$\cot\left(\frac{\pi}{2} - z\right) = \tan(z) .$$

Then substitute $z = \arctan(x)$ to get

$$\cot\left(\frac{\pi}{2} - \arctan(x)\right) = \tan(\arctan(x)) = x ,$$

and simplifying further:

$$\frac{\pi}{2} - \arctan(x) = \operatorname{arccot}(x)$$

To deal with $\operatorname{arccot}(x)$, recall also from trigonometry that

$$\operatorname{arccot}(x) = \arctan\left(\frac{1}{x}\right),$$

thus we land at a powerful identity:

$$\arctan(x) = \frac{\pi}{2} - \arctan\left(\frac{1}{x}\right) \quad (16.72)$$

This puts us in position to finally calculate $\arctan(2)$ via

$$\arctan(2) = \frac{\pi}{2} - \arctan\left(\frac{1}{2}\right).$$

Either of Equations (16.69), (16.70) is sufficient to calculate $\arctan(1/2)$.

Tangent Near Zero

The tangent function is a bit ugly for having vertical asymptotes at integer multiples of $\pm\pi/2$, $\pm3\pi/2$, etc. The function is otherwise handled in typical fashion, first by listing off the first few derivatives of $f(x_0) = \tan(x_0)$:

$$\begin{aligned} f^{(1)}(x_0) &= \sec^2(x_0) \\ f^{(2)}(x_0) &= 2 \sec^2(x_0) \tan(x_0) \\ f^{(3)}(x_0) &= 2 \sec^2(x_0) \left(\sec^2(x_0) + 2 \tan^2(x_0) \right) \\ f^{(4)}(x_0) &= 16 \sec^4(x_0) \tan(x_0) \\ &\quad + 8 \sec^2(x_0) \tan^3(x_0) \\ f^{(5)}(x_0) &= 88 \sec^4(x_0) \tan^2(x_0) + 16 \sec^6(x_0) \\ &\quad + 16 \tan^4(x_0) \sec^2(x_0) \end{aligned}$$

The above simplifies differently depending on which x_0 is chosen. Going with $x_0 = 0$ first, acknowledge that

$$\begin{aligned} \sec(0) &= 1 \\ \tan(0) &= 0, \end{aligned}$$

and quickly find:

$$\begin{aligned} f^{(1)}(0) &= 1 & f^{(2)}(0) &= 0 \\ f^{(3)}(0) &= 2 & f^{(4)}(0) &= 0 \\ f^{(5)}(0) &= 16 \end{aligned}$$

Plugging this information into Taylor's theorem yields a useful approximation to the tangent function:

$$\tan(x) \approx x + \frac{x^3}{3} + \frac{2x^5}{15} + O(x^7) \quad (16.73)$$

The symbol $O(x^7)$ signifies that the next nonzero term in the approximation is of order 7, and then the terms get smaller after that. In this particular case, it happens that

$$O(x^7) = \frac{17x^7}{315},$$

which you are welcome to verify.

Tangent Near Pi/4

Shifting the base point to $x_0 = \pi/4$, we can recycle all of the work in calculating the derivatives of $\tan(x)$ and re-evaluate using

$$\begin{aligned} \sec(\pi/4) &= \sqrt{2} \\ \tan(\pi/4) &= 1, \end{aligned}$$

which gives:

$$\begin{aligned} f^{(1)}(\pi/4) &= 2 & f^{(2)}(\pi/4) &= 4 \\ f^{(3)}(\pi/4) &= 16 & f^{(4)}(\pi/4) &= 80 \\ f^{(5)}(\pi/4) &= 512 \end{aligned}$$

Not forgetting the shift by x_0 units, the approximation for the tangent near $x = \pi/4$ reads

$$\begin{aligned} \tan(x)_{x \approx \pi/4} &= 1 + 2 \left(x - \frac{\pi}{4} \right) \\ &\quad + 2 \left(x - \frac{\pi}{4} \right)^2 + \frac{8}{3} \left(x - \frac{\pi}{4} \right)^3 \\ &\quad + \frac{10}{3} \left(x - \frac{\pi}{4} \right)^4 + \frac{64}{15} \left(x - \frac{\pi}{4} \right)^5 \\ &\quad + O \left(x - \frac{\pi}{4} \right)^6 \end{aligned} \quad (16.74)$$

Cotangent Near Pi/2

The same routine can be applied to the cotangent. For $f(x) = \cot(x)$, find:

$$\begin{aligned} f^{(1)}(x_0) &= -\csc^2(x_0) \\ f^{(2)}(x_0) &= 2\csc^2(x_0)\cot(x_0) \\ f^{(3)}(x_0) &= -2\csc^2(x_0)\left(\csc^2(x_0) + 2\cot^2(x_0)\right) \\ f^{(4)}(x_0) &= 16\csc^4(x_0)\cot(x_0) \\ &\quad + 8\csc^2(x_0)\cot^3(x_0) \\ f^{(5)}(x_0) &= -88\csc^4(x_0)\cot^2(x_0) - 16\csc^6(x_0) \\ &\quad - 16\cot^4(x_0)\csc^2(x_0) \end{aligned}$$

Setting $x_0 = \pi/2$ first, note that

$$\begin{aligned} \csc(\pi/2) &= 1 \\ \cot(\pi/2) &= 0, \end{aligned}$$

and quickly find:

$$\begin{aligned} f^{(1)}(0) &= -1 & f^{(2)}(0) &= 0 \\ f^{(3)}(0) &= -2 & f^{(4)}(0) &= 0 \\ f^{(5)}(0) &= -16 \end{aligned}$$

Evidently, the expansion for the cotangent near $x = \pi/2$ is somewhat like the tangent near $x = 0$ with the signs reversed. For conciseness, let $z = x - \pi/2$ and find

$$\cot(z) \approx -z - \frac{z^3}{3} - \frac{2z^5}{15} - O(z^7). \quad (16.75)$$

6.9 Expansion Near Asymptotes

Tangent near Pi/2

Returning to the problem of the tangent function, we know $\tan(x)$ has a hopeless singularity at $x = \pi/2$ tending to $+\infty$ on the left and $-\infty$ on the right. With this, how can derivatives evaluated at $\pi/2$, which are surely divergent, mean anything?

It seems that Taylor would have nothing to say about expansion near an asymptote, but there is a trick. Since the tangent and cotangent are mutually reciprocal, then it should make sense to approximate the ratio $1/\cot(x)$ near $x = \pi/2$ and get the answer we want.

Letting $z = x - \pi/2$, this means we start with

$$\tan(z)_{z \approx 0} = \frac{1}{\cot(z)_{z \approx 0}} = \frac{-1}{z + z^3/3 + 2z^5/15},$$

where any terms of order 7 or higher are ignored as negligible. Carrying out the polynomial division leads to an infinite sum:

$$\tan(z) \approx -\frac{1}{z} + \frac{z}{3} + \frac{z^3}{45} + O(z^5) \quad (16.76)$$

Notice the result is two orders lower than the quantity we started with, thus any terms of order 5 or greater can't be trusted. More important are the low-order terms, and we see $-1/z$ being the dominant one. This captures the divergent behavior of the tangent near its first asymptote and the trailing terms improve accuracy.

Cotangent Near Zero

The cotangent function behaves asymptotically near $x = 0$, thus the same trick is needed to explore this case. That is, take the approximation for $\tan(x)$ near $x = 0$ and perform long division. Leaving the details for an exercise, the result is:

$$\cot(x) \approx \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - O(x^5) \quad (16.77)$$

6.10 Hyperbolic Expansions

Hyperbolic Arctangent Near Zero

Problem 32

For the hyperbolic arctangent, show that the Taylor expansion near $x = 0$ is:

$$\operatorname{arctanh}(x) \approx x + \frac{x^3}{3} + \frac{x^5}{5} + O(x^7)$$

6.11 Kinematics with Air Damping

When studying kinematics, one comes to understand that it all starts with a uniform gravitational field in vacuum, which on earth near sea level means

$$a = -g = -9.8 \text{ m/s}^2.$$

From this, we know the velocity will be a linear function in time, i.e.

$$v(x) = v_0 + at,$$

and the position is a quadratic:

$$x(t) = x_0 + v_0t + \frac{1}{2}at^2$$

Of course, this is the most baseline picture of kinematics in the sense that there are no jerk-like terms or higher derivatives. Starkly absent too are real-world effects that would

alter the idealized image of projectile motion, particularly the presence of the atmosphere as a resisting fluid.

Using a simplified model for air damping, we can imagine a new component to the acceleration that tries to slow down an object by an amount proportional to its speed. To capture this, we let the acceleration vary in time via

$$a(t) = -g - bv(t),$$

where b is a *linear damping* coefficient and $v(t)$ is the velocity.

Acknowledging that $a(t)$ is the second derivative of $x(t)$ and $v(t)$ is the first derivative, the above is more clearly stated in Leibniz notation as

$$\frac{d^2x}{dt^2} = -g - b \frac{dx}{dt},$$

which is an honest-to-goodness *differential equation*. The signature of a differential equation is that the function $x(t)$ is tied up in some kind of relationship with its own derivative(s), and solving for x can't be done algebraically. A less intimidating version of the same equation can be written strictly in terms of $v(t)$:

$$\frac{dv}{dt} = -g - bv \quad (16.78)$$

Frobenius Method

There is a brilliant trick attributed to Ferdinand Georg Frobenius (1849-1917) for solving equations like the one above. Suppose $v(t)$ takes the form of an infinite, Taylor-like polynomial with unknown coefficients:

$$v(t) = v_0 + A_1t + A_2t^2 + A_3t^3 + \dots$$

Without knowing much else about $v(t)$, we can still compute the derivative:

$$\frac{dv}{dt} = A_1 + 2A_2t + 3A_3t^2 + 4A_4t^3 + \dots$$

Now, plug both of these into $-g = bv + dv/dt$:

$$\begin{aligned} -g &= bv_0 + bA_1t + bA_2t^2 + bA_3t^3 + \dots \\ &\quad + A_1 + 2A_2t + 3A_3t^2 + 4A_4t^3 + \dots \end{aligned}$$

This seems to be a greater mess than we started with until *matching coefficients*, which means the coefficients on matching powers of t must balance. This means

$$\begin{aligned} -g &= bv_0 + A_1 \\ 0 &= bA_1 + 2A_2 \\ 0 &= bA_2 + 3A_3 \\ 0 &= bA_3 + 4A_4, \end{aligned}$$

and the pattern continues forever.

Velocity

Astonishingly, notice that all of the coefficients can all be related back to the first few, and $v(t)$ can now be written:

$$v(t) = v_0 + A_1t - \frac{bA_1}{2!}t^2 + \frac{b^2A_1}{3!}t^3 - \frac{b^3A_1}{4!}t^4 + \dots$$

The polynomial on the right looks tantalizingly close to an exponential, which it indeed is. Proceeding carefully, we next have

$$v(t) = v_0 + \frac{A_1}{b} \left(1 - e^{-bt}\right),$$

simplifying once more to

$$v(t) = \frac{-g}{b} + \left(v_0 + \frac{g}{b}\right) e^{-bt}. \quad (16.79)$$

The final unknown v_0 is the initial velocity $v(0)$.

If an object is left in freefall in atmosphere for a long time, it's likely to achieve a state called *terminal velocity* where the force of gravity balances the force of air damping. To see this, let t run to infinity in the above, and we find

$$v_{\text{terminal}} = \lim_{t \rightarrow \infty} v(t) = \frac{-g}{b}.$$

Position

It just happens that the Frobenius method works for attaining $x(t)$. Postulating

$$x(t) = x_0 + B_1t + B_2t^2 + B_3t^3 + \dots,$$

one finds, after plugging into

$$-g = \frac{d^2x}{dt^2} + b \frac{dx}{dt},$$

that

$$x(t) = x_0 - \frac{gt}{b} - \frac{2B_2}{b^2} \left(1 - e^{-bt}\right),$$

where x_0 is the initial position $x(0)$.

Since the derivative of $x(t)$ is identically $v(t)$, we can relate the coefficient B_2 to the initial velocity via

$$B_2 = -\frac{1}{2} (g + bv_0),$$

and the position equation takes the form:

$$x(t) = x_0 - \frac{gt}{b} + \frac{1}{b} \left(v_0 + \frac{g}{b}\right) \left(1 - e^{-bt}\right) \quad (16.80)$$

Small- b Limit

In the case that the damping constant b is small, the velocity and position equations ought to restore to their ideal form, or at least approximately. Doing the $v(t)$ -case first, Equation (16.79) in the small- b limit reads

$$v(t) \approx \frac{-g}{b} + \left(v_0 + \frac{g}{b}\right)(1 - bt) ,$$

reducing readily to

$$v(t) \approx v_0 - (g + v_0 b) t$$

with no factor of b in the denominator. The $v_0 b$ term plays essentially no role in the numerator and the form $v \approx v_0 - gt$ is recovered.

The position equation ought to churn out something similar. Starting from Equation (16.80) and expanding the exponential to second order gives:

$$x(t) \approx x_0 - \frac{gt}{b} + \frac{1}{b} \left(v_0 + \frac{g}{b}\right) \left(bt - \frac{1}{2}b^2 t^2\right) ,$$

which simplifies to

$$x(t) \approx x_0 + v_0 t - \frac{1}{2}(g + v_0 b) t^2$$

as expected.

7 Antiderivative

Reflecting on the notion of the derivative $f'(x)$ as it relates to the original function $f(x)$, there is a sense that the derivative operator is a one-way arrow from one ‘space’ of functions to another. For any given $f(x)$, we can more-or-less confidently calculate $f'(x)$ using the techniques gained above.

The derivative calculation begs an interesting question though, namely, can we start with $f'(x)$ and infer what the original $f(x)$ could have been? This idea is like running the derivative operator backwards, and is aptly named the *antiderivative*.

7.1 Motivation

A natural way to frame the antiderivative question starts with definition of the derivative

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

and then set the left side to a function that is given, call it $g(x)$:

$$g(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Next, consider any sequence of manipulations, symbolized by Q , that is applied to both sides of the above. By ‘manipulations’, we mean adding zero, multiplying by one, and so on. Symbolically, this would mean

$$Q(g(x_0)) = Q\left(\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}\right).$$

If the sequence Q is chosen properly, the quantity $Q(g(x_0))$ on the left is some new function of x_0 , which, and this is the key - should match the form of a known derivative. That is, we should be able to recognize $Q(g(x_0))$ as the derivative of some previously-cataloged function $r(x)$. This lets us replace the right side of the above:

$$Q(g(x_0)) = \frac{d}{dx}(r(x)) \Big|_{x_0}$$

Finally, isolate $g(x_0)$ algebraically via

$$g(x_0) = Q^{-1}\left(\frac{d}{dx}(r(x)) \Big|_{x_0}\right),$$

where Q^{-1} reverses the manipulations represented by Q .

7.2 Exemplary Cases

Natural Logarithm

Going for an interesting example, it turns out that the natural logarithm $\ln(x)$ didn’t turn up as the result of any derivative calculation previously done. Letting $f'(x) = \ln(x)$, we can puzzle out $f(x)$ starting with:

$$\begin{aligned} \ln(x_0) &= \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \\ \ln(x_0) + 1 &= \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} + 1 \end{aligned}$$

By adding 1 to each side, the left is suddenly recognizable from Equation (16.35), which reads

$$\frac{d}{dx}(\ln(x^x)) = \ln(x) + 1.$$

Knowing this, replace the right side of our working equation:

$$\begin{aligned} \ln(x_0) + 1 &= \frac{d}{dx}(\ln(x^x)) \Big|_{x_0} \\ \ln(x_0) + 1 &= \lim_{x \rightarrow x_0} \frac{\ln(x^x) - \ln(x_0^{x_0})}{x - x_0} \end{aligned}$$

Now subtract 1 from each side, thereby applying Q^{-1} , and simplify:

$$\ln(x_0) = \lim_{x \rightarrow x_0} \frac{(x \ln(x) - x) - (x_0 \ln(x_0) - x_0)}{x - x_0}$$

The right side is none other than the derivative of $x \ln(x) - x$, and we’re done:

$$\ln(x) = \frac{d}{dx}(x \ln(x) - x) \quad (16.81)$$

X Times Cos(X)

It’s a bit more practical to work in Leibniz notation if we have a good handle of how to isolate the desired derivative.

To illustrate, suppose we want to know which function has a slope of $f'(x) = x \cos(x)$. Reaching for a table of derivatives, recall Equation (16.28), namely

$$\frac{d}{dx}(x \sin(x)) = \sin(x) + x \cos(x),$$

which contains the answer as the rightmost term. To proceed, note that the sine term can be replaced by the negative derivative of the cosine:

$$\frac{d}{dx}(x \sin(x)) = \frac{d}{dx}(-\cos(x)) + x \cos(x)$$

Since the derivative operator is a linear one, we can cram all derivative terms on the same side to write the final answer:

$$x \cos(x) = \frac{d}{dx} (x \sin(x) + \cos(x)) \quad (16.82)$$

By identical reasoning, one can work out the case $f'(x) = x \sin(x)$ using Equation (16.29). Leaving the details as an exercise, the result is

$$x \sin(x) = \frac{d}{dx} (\sin(x) - x \cos(x)) . \quad (16.83)$$

7.3 Powers and Roots

The rule governing powers and roots is covered by Equation (16.2), namely

$$\frac{d}{dx} (x^n) = nx^{n-1} ,$$

which reads cleanly in both directions, i.e. it's ready for derivative and antiderivatives.

In light of the chain rule, we can replace x with a function $f(x)$ to have:

$$\frac{d}{dx} ((f(x))^n) = \frac{n}{(f(x))^{n-1}} \frac{d}{dx} (f(x))$$

Despite the above result being general, it's still a bit messy and not worth memorizing. One exception, though, is the special case $n = 1/2$:

$$\frac{d}{dx} \sqrt{f(x)} = \frac{f'(x)}{2\sqrt{f(x)}}$$

As you train your eye solve antiderivative problems, it helps to know that the ratio $f'/\sqrt{f}$ can be dealt with using the above.

To illustrate, consider the case

$$g'_\pm(x) = \frac{\pm x}{\sqrt{1 \pm x^2}} ,$$

which has $f(x) = 1 \pm x^2$ and $f'(x) = \pm 2x$. Immediately from this, we can write:

$$\frac{\pm x}{\sqrt{1 \pm x^2}} = \frac{d}{dx} \sqrt{1 \pm x^2} \quad (16.84)$$

Reciprocal

One exception to the usual pattern for powers and roots is the reciprocal function $f'(x) = 1/x$. This antiderivative is handled by Equation (16.14) going backwards, namely:

$$\frac{1}{x} = \frac{d}{dx} (\ln(x))$$

7.4 Logarithmic Antiderivatives

Diminished Natural Logarithm

Much of the struggle in calculating antiderivatives is deciding which functions to try. For instance, suppose we have the diminished natural logarithm represented by

$$f'(x) = \frac{\ln(x)}{x} .$$

After some fiddling with the chain rule, one eventually stumbles upon

$$\frac{d}{dx} ((\ln(x))^2) = \frac{2}{x} \ln(x) . \quad (16.85)$$

Recognizing the original problem embedded on the right, we have the answer:

$$\frac{\ln(x)}{x} = \frac{1}{2} \frac{d}{dx} ((\ln(x))^2) . \quad (16.86)$$

Shifted Natural Logarithm

Consider the case $f'(x) = \ln(x+1)$. For this, use the product rule to establish

$$\frac{d}{dx} ((x+1) \ln(x+1)) = \ln(x+1) + 1 .$$

In order to isolate $x \ln(x)$, everything else must be part of the same derivative, and this can be done by replacing the 1-term via

$$1 = \frac{dx}{dx} .$$

Substituting this into the above and simplifying gives the result we're after:

$$\ln(x+1) = \frac{d}{dx} ((x+1) \ln(x+1) - x) \quad (16.87)$$

Nonlinear Natural Logarithm

Consider the case $f'(x) = x \ln(x)$. For this, use the product rule to establish

$$\frac{d}{dx} (x^2 \ln(x)) = 2x \ln(x) + x .$$

Like the previous case, in order to isolate $x \ln(x)$, everything else must be part of the same derivative, and this can be done by replacing the x -term via

$$x = \frac{1}{2} \frac{d}{dx} (x^2) .$$

Substituting this into the above and simplifying gives the result we're after:

$$x \ln(x) = \frac{1}{2} \frac{d}{dx} \left(x^2 \left(\ln(x) - \frac{1}{2} \right) \right) \quad (16.88)$$

Modified Natural Logarithm

Finding the antiderivative of the modified natural logarithm $f'(x) = \ln(1+x^2)$ is a challenge. To begin we'll write something that contains $1+x^2$ and hope for the best, particularly:

$$\frac{d}{dx} \left(x \ln(1+x^2) \right) = \ln(1+x^2) + \frac{2x^2}{1+x^2}$$

The rightmost term can be split apart by algebra

$$\frac{2x^2}{1+x^2} = 2 \left(1 - \frac{1}{1+x^2} \right),$$

and now the problem reduces to writing the parenthesized quantity as a derivative. Luckily, we know exactly how to do this:

$$\frac{2x^2}{1+x^2} = 2 \left(\frac{dx}{dx} - \frac{d}{dx} \arctan(x) \right)$$

Putting all derivative terms on the same side yields the answer:

$$\begin{aligned} \ln(1+x^2) &= \frac{d}{dx} \left(x \left(\ln(1+x^2) - 2 \right) \right) \\ &\quad + 2 \frac{d}{dx} (\arctan(x)) \end{aligned} \quad (16.89)$$

7.5 Exponential Antiderivatives**Exponential Times X**

To handle the case $f'(x) = xe^x$, use the product rule on the same quantity

$$\frac{d}{dx} (xe^x) = e^x + xe^x,$$

and notice the right-most term contains the answer we want. Exploiting the fact that e^x is its own derivative, we can write everything else as a total derivative to have the answer:

$$xe^x = \frac{d}{dx} (e^x (x-1)) \quad (16.90)$$

Exponential Times X*X

To handle the case $f'(x) = x^2e^x$, use the product rule on the same quantity

$$\frac{d}{dx} (x^2e^x) = 2xe^x + x^2e^x,$$

and notice the right-most term contains the answer we want. The middle term would be show-stopper if it weren't for

Equation (16.90), which allows the rest to be written as a total derivative:

$$x^2e^x = \frac{d}{dx} \left(e^x (x^2 - 2x + 2) \right) \quad (16.91)$$

Exponential Times Cos(X)

The set of problems

$$f'_1(x) = e^x \cos(x)$$

$$f'_2(x) = e^x \sin(x)$$

can be handled simultaneously. First, write two results easily attainable by the product rule:

$$\frac{d}{dx} (e^x \sin(x)) = e^x \sin(x) + e^x \cos(x)$$

$$\frac{d}{dx} (e^x \cos(x)) = e^x \cos(x) - e^x \sin(x)$$

Next, take the sum and the difference of the two above equations and exploit the linearity of the derivative operator to get both results at once:

$$e^x \cos(x) = \frac{1}{2} \frac{d}{dx} (e^x \cos(x) + e^x \sin(x)) \quad (16.92)$$

$$e^x \sin(x) = \frac{1}{2} \frac{d}{dx} (e^x \sin(x) - e^x \cos(x)) \quad (16.93)$$

7.6 Trigonometric Antiderivatives**Tangent and Cotangent**

The case for $f'(x) = \tan(x)$ is a bit tricky. Hunting for any derivative calculation that has $\tan(x)$ as part of the answer, Equation (16.24) comes to mind, namely

$$\frac{d}{dx} (\sec(x)) = \tan(x) \sec(x).$$

Proceed by letting $u = \sec(x)$ and separate variables:

$$\frac{1}{u} \frac{du}{dx} = \tan(x)$$

By the chain rule, or equivalently by the 'logarithm trick' represented by Equation (16.36), the left side is equivalent to the derivative of the natural log of u :

$$\frac{d}{dx} (\ln(u)) = \tan(x)$$

Reversing the u -substitution, the final answer is

$$\tan(x) = \frac{d}{dx} (-\ln(\cos(x))) \quad (16.94)$$

By a similar line of reasoning, the cotangent version can also be done, with the details left as an exercise:

$$\cot(x) = \frac{d}{dx} (\ln(\sin(x))) . \quad (16.95)$$

Secant and Cosecant

The case of $f'(x) = \sec(x)$ can be attacked with partial fractions. Following the algebra, we find

$$\begin{aligned} \frac{1}{\cos(x)} &= \frac{\cos(x)}{\cos^2(x)} = \frac{\cos(x)}{1 - \sin^2(x)} \\ &= \frac{1}{2} \left(\frac{\cos(x)}{1 - \sin(x)} + \frac{\cos(x)}{1 + \sin(x)} \right) . \end{aligned}$$

Now, spotting this takes some getting used to, but the above can be rewritten using the logarithm trick

$$\begin{aligned} \frac{1}{\cos(x)} &= -\frac{1}{2} \frac{d}{dx} (\ln(1 - \sin(x))) \\ &\quad + \frac{1}{2} \frac{d}{dx} (\ln(1 + \sin(x))) , \end{aligned}$$

which can be simplified and the problem is solved:

$$\sec(x) = \frac{1}{2} \frac{d}{dx} \left(\ln \left(\frac{1 + \sin(x)}{1 - \sin(x)} \right) \right) \quad (16.96)$$

With a little more algebra, the above can be simplified to

$$\sec(x) = \frac{d}{dx} (\ln(\sec(x) + \tan(x))) .$$

Leaving the details for an exercise, a similar exercise leads to the cosecant version

$$\csc(x) = -\frac{1}{2} \frac{d}{dx} \left(\ln \left(\frac{1 + \cos(x)}{1 - \cos(x)} \right) \right) , \quad (16.97)$$

or

$$\csc(x) = -\frac{d}{dx} (\ln(\csc(x) + \cot(x))) .$$

Cos(X) Squared

The set of problems

$$\begin{aligned} f_1'(x) &= \cos^2(x) \\ f_2'(x) &= \sin^2(x) \end{aligned}$$

can be handled simultaneously. First, note from the product rule that

$$\frac{d}{dx} (\sin(x) \cos(x)) = \cos^2(x) - \sin^2(x) ,$$

which is equivalent to both of:

$$\frac{d}{dx} (\sin(x) \cos(x)) = 1 - 2\sin^2(x)$$

$$\frac{d}{dx} (\sin(x) \cos(x)) = 2\cos^2(x) - 1$$

Finally, note that the factor 1 is equivalent to dx/dx , which allows the left side (in each) to be written as a total derivative, leading to

$$\sin^2(x) = \frac{1}{2} \frac{d}{dx} (x - \sin(x) \cos(x)) \quad (16.98)$$

$$\cos^2(x) = \frac{1}{2} \frac{d}{dx} (x + \sin(x) \cos(x)) \quad (16.99)$$

Cos(X) Times Sin(X)

The case $f'(x) = \cos(x) \sin(x)$ has two unique answers. Handling both possibilities in one blow, use the chain rule to write

$$\frac{d}{dx} (\sin^2(x)) = -\frac{d}{dx} (\cos^2(x)) = 2\cos(x) \sin(x) ,$$

and the result is isolated:

$$\cos(x) \sin(x) = \frac{1}{2} \frac{d}{dx} ((\sin(x))^2) \quad (16.100)$$

$$\cos(x) \sin(x) = \frac{-1}{2} \frac{d}{dx} ((\cos(x))^2) \quad (16.101)$$

7.7 Inverse Trig Antiderivatives

Arccosine and Arcsine

The roulette of inverse trigonometric functions can also be tackled. Going for $f'(x) = \arccos(x)$ first, consider the following application of the product rule:

$$\frac{d}{dx} (x \arccos(x)) = \arccos(x) + x \frac{d}{dx} (\arccos(x))$$

The derivative of $\arccos(x)$ can be replaced by Equation (16.37), namely

$$\frac{d}{dx} \arccos(x) = \frac{-1}{\sqrt{1-x^2}} ,$$

and the above becomes

$$\frac{d}{dx} (x \arccos(x)) = \arccos(x) - \frac{x}{\sqrt{1-x^2}} .$$

The square root term is itself the derivative of a function obeying Equation (16.84), or

$$\frac{-x}{\sqrt{1 \pm x^2}} = \frac{d}{dx} \sqrt{1-x^2} .$$

Condensing derivatives on one side with $\arccos(x)$ on the other gives the answer:

$$\arccos(x) = \frac{d}{dx} \left(x \arccos(x) - \sqrt{1-x^2} \right) \quad (16.102)$$

By similar reasoning, the $\arcsin(x)$ case works out as

$$\arcsin(x) = \frac{d}{dx} \left(x \arcsin(x) + \sqrt{1-x^2} \right). \quad (16.103)$$

Arctangent and Arccotangent

To handle $f'(x) = \arctan(x)$, consider the following application of the product rule:

$$\frac{d}{dx} (x \arctan(x)) = \arctan(x) + x \frac{d}{dx} (\arctan(x))$$

The derivative of $\arctan(x)$ can be replaced by Equation (16.39), namely

$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2},$$

and the above becomes

$$\frac{d}{dx} (x \arctan(x)) = \arctan(x) + \frac{x}{1+x^2}.$$

The rightmost term needs to be replaced with the derivative of something. It turns out that Equation (16.15) does the job, namely

$$\frac{d}{dx} \left(\ln(1+x^2) \right) = \frac{2x}{1+x^2}.$$

Condensing derivatives on one side with $\arctan$ on the other gives the answer:

$$\arctan(x) = \frac{d}{dx} \left(x \arctan(x) - \frac{\ln(1+x^2)}{2} \right) \quad (16.104)$$

By similar reasoning, the $\operatorname{arccot}(x)$ case works out as

$$\operatorname{arccot}(x) = \frac{d}{dx} \left(x \operatorname{arccot}(x) + \frac{\ln(1+x^2)}{2} \right). \quad (16.105)$$

7.8 Trigonometric Substitution

Arcsecant and Arccosecant

The last two inverse trig functions, namely the arc-secant and the arc-cosecant, don't follow as easily as the others. To

handle $f'(x) = \operatorname{arcsec}(x)$, consider the application of the product rule

$$\frac{d}{dx} (x \operatorname{arcsec}(x)) = \operatorname{arcsec}(x) + x \frac{d}{dx} (\operatorname{arcsec}(x)),$$

which, by Equation (16.40), is equivalent to

$$\frac{d}{dx} (x \operatorname{arcsec}(x)) = \operatorname{arcsec}(x) + \frac{1}{\sqrt{x^2-1}}.$$

As usual, the rightmost term needs to be the derivative of something else. Innocent as this looks, a different technique called *trigonometric substitution* must be used. For the example on hand, introduce a new variable θ such that

$$x = \sec(\theta).$$

Then, standard trig identities tell us

$$\tan^2(\theta) = \sec^2(\theta) - 1 = x^2 - 1,$$

or

$$\tan(\theta(x)) = \sqrt{x^2-1}.$$

Meanwhile, we can take the θ -derivative of x to write

$$\frac{dx}{d\theta} = \frac{d}{d\theta} (\sec(\theta)) = \sec(\theta) \tan(\theta) = x \sqrt{x^2-1},$$

and by the chain rule, this means

$$\frac{d\theta}{dx} = \frac{1}{x \sqrt{x^2-1}}.$$

Now, if the term $1/\sqrt{x^2-1}$ is to be the derivative of some unknown function $q(x)$, we have, by the chain rule,

$$\frac{dq}{dx} = \frac{dq}{d\theta} \frac{d\theta}{dx} = \frac{dq}{d\theta} \frac{1}{x \sqrt{x^2-1}},$$

which can only mean

$$1 = \frac{dq}{d\theta} \frac{1}{x}.$$

Since x is already known as $\sec(\theta)$, the question has come to looking for a function whose derivative is $\sec(\theta)$. For this we may refer to Equation (16.96) using θ as the variable:

$$\sec(\theta) = \frac{d}{d\theta} (\ln(\sec(\theta) + \tan(\theta))),$$

or

$$q(\theta) = \ln(\sec(\theta) + \tan(\theta)).$$

Switching variables back to x , the above reads

$$q(x) = \ln(x + \sqrt{x^2-1}).$$

This result alone is worth noting,

$$\frac{1}{\sqrt{x^2 - 1}} = \frac{d}{dx} \left(\ln \left(x + \sqrt{x^2 - 1} \right) \right), \quad (16.106)$$

and more importantly, we can write the final answer to the $\operatorname{arcsec}(x)$ antiderivative:

$$\operatorname{arcsec}(x) = \frac{d}{dx} \left(x \operatorname{arcsec}(x) - \ln \left(x + \sqrt{x^2 - 1} \right) \right) \quad (16.107)$$

A similar exercise gives the $\operatorname{arccsc}(x)$ version:

$$\operatorname{arccsc}(x) = \frac{d}{dx} \left(x \operatorname{arccsc}(x) + \ln \left(x + \sqrt{x^2 - 1} \right) \right) \quad (16.108)$$

7.9 Hyperbolic Cases

The nontrivial hyperbolic trig antiderivatives are mostly analogous to their ordinary trig counterparts:

$$\tanh(x) = \frac{d}{dx} (\ln(\cosh(x))) \quad (16.109)$$

$$\coth(x) = \frac{d}{dx} (\ln(\sinh(x))) \quad (16.110)$$

The hyperbolic secant and cosecant are a little less obvious, but turn out to be:

$$\operatorname{sech}(x) = \frac{d}{dx} \left(2 \arctan \left(\tanh \left(\frac{x}{2} \right) \right) \right) \quad (16.111)$$

$$\operatorname{csch}(x) = \frac{d}{dx} \left(\ln \left(\tanh \left(\frac{x}{2} \right) \right) \right) \quad (16.112)$$

The details are left for an exercise.

Arccosh and Arcsinh

Luckily, some of the inverse hyperbolic trig derivatives are analogous to calculations previously done. For the cases $\operatorname{arccosh}$, arsinh , one straightforwardly finds

$$\operatorname{arccosh}(x) = \frac{d}{dx} \left(x \operatorname{arccosh}(x) - \sqrt{x^2 - 1} \right) \quad (16.113)$$

$$\operatorname{arsinh}(x) = \frac{d}{dx} \left(x \operatorname{arsinh}(x) - \sqrt{x^2 + 1} \right). \quad (16.114)$$

Arctanh and Arccoth

The $\operatorname{arccosh}$, arsinh functions have antiderivatives that are remarkably similar to their trig counterparts:

$$\operatorname{arctanh}(x) = \frac{d}{dx} \left(x \operatorname{arctanh}(x) + \frac{\ln(1 - x^2)}{2} \right) \quad (16.115)$$

$$\operatorname{arccoth}(x) = \frac{d}{dx} \left(x \operatorname{arccoth}(x) + \frac{\ln(1 - x^2)}{2} \right) \quad (16.116)$$

Arcsech and Arccsch

The $\operatorname{arcsech}$, $\operatorname{arccsch}$ functions can be aced with a trick. Handling both at the same time, introduce two unknown functions $f(x)$ and $g(x)$ to write the following total derivatives

$$\operatorname{arcsech}(x) = \frac{d}{dx} (f(x) + x \operatorname{arcsech}(x))$$

$$\operatorname{arccsch}(x) = \frac{d}{dx} (g(x) + x \operatorname{arccsch}(x)),$$

and now the whole problem is about finding out what $f(x)$ and $g(x)$ must be.

Applying the product rule across the right side, we find

$$f'(x) = \frac{1}{\sqrt{1 - x^2}}$$

$$g'(x) = \frac{1}{\sqrt{1 + x^2}},$$

telling us, according to Equations (16.38), (16.50), that

$$f(x) = \arcsin(x)$$

$$g(x) = \operatorname{arsinh}(x),$$

and thus:

$$\operatorname{arcsech}(x) = \frac{d}{dx} (x \operatorname{arcsech}(x) + \arcsin(x)) \quad (16.117)$$

$$\operatorname{arccsch}(x) = \frac{d}{dx} (x \operatorname{arccsch}(x) + \operatorname{arsinh}(x)) \quad (16.118)$$

8 Simple Harmonic Oscillator

The *simple harmonic oscillator* is a mathematical model used for approximating many real-world systems. Common simple harmonic oscillators are (i) a mass attached to a spring moving in a frictionless environment, (ii) a hanging pendulum making small deflections from equilibrium, or most generally, (iii) small displacements in any system featuring a local minimum in the potential energy.

Problem Setup

To get going, consider a body of mass m tethered to the point $x = 0$ subject to the linear restoring force

$$F = -kx$$

corresponding to a potential energy

$$U_{\text{spring}}(x) = \frac{1}{2}kx^2.$$

By Newton's second law

$$m \frac{d^2}{dt^2} x(t) = -\frac{d}{dx} U(x),$$

we can assemble an equation governing the so-called harmonic motion of the oscillator:

$$\frac{d^2}{dt^2} x(t) = -\frac{k}{m} x(t) \quad (16.119)$$

Finding the Solution

The task now is to 'solve' the above equation, which means to find the correct $x(t)$ that satisfies it. With $x(t)$ in hand, we will know the position of the body as a function of time.

We seek $x(t)$ as a function whose second derivative is equal to the negative of itself multiplied by a constant. Right away, two trigonometric functions come to mind:

$$\begin{aligned} \frac{d^2}{dt^2} \cos\left(\sqrt{\frac{k}{m}} t\right) &= \frac{-k}{m} \cos\left(\sqrt{\frac{k}{m}} t\right) \\ \frac{d^2}{dt^2} \sin\left(\sqrt{\frac{k}{m}} t\right) &= \frac{-k}{m} \sin\left(\sqrt{\frac{k}{m}} t\right) \end{aligned}$$

Angular Frequency

Both the cosine and the sine seem to satisfy Equation (16.119), so let's keep track of both for a moment. The

quantity $\sqrt{k/m}$ is called the *angular frequency* and is designated the symbol ω :

$$\omega = \sqrt{\frac{k}{m}}$$

So far we can sketch out two possible solutions:

$$x_1(t) \propto \cos(\omega t)$$

$$x_2(t) \propto \sin(\omega t)$$

Scaling Constants

Notice now that scaling each of these by an unknown constant to make $A \cos(\omega t)$, $B \sin(\omega t)$ would leave the oscillator equation unchanged, yet the presence of scaling constants would clearly affect the final solution. The updated solutions are now

$$x_1(t) = A \cos(\omega t)$$

$$x_2(t) = B \sin(\omega t)$$

for two undetermined coefficients A , B .

General Solution

With the preparatory work done, we can write a general solution for the problem:

$$x(t) = A \cos(\omega t) + B \sin(\omega t),$$

and from $x(t)$ we can take a time derivative to get the velocity:

$$v(t) = -A\omega \sin(\omega t) + B\omega \cos(\omega t)$$

For a sanity check, we should be able to take the time derivative of $v(t)$ to recover Equation (16.119). Doing so, we find

$$\begin{aligned} \frac{d^2}{dt^2} x(t) &= -A\omega^2 \cos(\omega t) - B\omega^2 \sin(\omega t) \\ &= -\omega^2 (A \cos(\omega t) + B \sin(\omega t)) \\ &= -\omega^2 x(t) \\ &= \frac{-k}{m} x(t) \end{aligned}$$

as expected.

Initial Conditions

To refine the solution to the harmonic oscillator equation, suppose at $t = 0$ the body is known to be at position x_0 with initial velocity v_0 . Such *initial conditions* can be worked into the solution by setting $t = 0$ in the x - and v -equations

$$x_0 = x(0) = A \cos(0) + 0$$

$$v_0 = v(0) = 0 + B\omega \cos(0)$$

to discern:

$$A = x_0$$

$$B = v_0/\omega$$

With this, the updated position and velocity read:

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t)$$

$$v(t) = -x_0\omega \sin(\omega t) + v_0 \cos(\omega t)$$

Magnitude and Phase

While the above is a workable solution to the simple harmonic oscillator, everything can be made tighter by introduc-

ing a *magnitude* coefficient R , along with a *phase* coefficient ϕ such that:

$$x_0 = R \cos(\phi)$$

$$\frac{v_0}{\omega} = -R \sin(\phi)$$

The magnitude and phase have a trigonometric relationship to the initial conditions, namely

$$R = \sqrt{x_0^2 + v_0^2/\omega^2}$$

$$\phi = \arctan\left(\frac{-v_0}{\omega x_0}\right).$$

In terms of R and ϕ , the solution $x(t)$ reads

$$x(t) = R \cos(\phi) \cos(\omega t) - R \sin(\phi) \sin(\omega t),$$

which, using a trigonometric angle-sum formula, simplifies to:

$$x(t) = R \cos(\omega t + \phi)$$

Chapter 17

Integral Calculus

1 Area Under a Curve

The workhorse equation of differential calculus is undoubtedly the definition of the derivative of a function $f(x)$

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which gives the instantaneous slope of the function at x_0 .

An extension of the derivative comes in the form of Taylor's theorem, which attempts to approximate the function $f(x)$ near a given point x_0 :

$$f(x) \approx f(x_0) + \sum_{q=1}^n \frac{1}{q!} f^{(q)}(x_0) (x - x_0)^q$$

Of course, Taylor's theorem embeds the first derivative as its first-order case.

As it turns out, the derivative formulas from differential calculus can be used backwards, so to speak, which is the basis of *integral calculus*.

Motivation

Consider a point x_0 in the domain of a 'well-behaved' function $f(x)$, and also consider a point x_1 that is arbitrarily close to x_0 . By the derivative formula, we can surely write

$$\lim_{x_1 \rightarrow x_0} f(x_1) - f(x_0) = f'(x_0) \lim_{x_1 \rightarrow x_0} (x_1 - x_0).$$

Also, consider another point x_2 that is arbitrarily close to x_1 , which means

$$\lim_{x_2 \rightarrow x_1} f(x_2) - f(x_1) = f'(x_1) \lim_{x_2 \rightarrow x_1} (x_2 - x_1),$$

and just to start a pattern, consider yet another point x_3 obeying

$$\lim_{x_3 \rightarrow x_2} f(x_3) - f(x_2) = f'(x_2) \lim_{x_3 \rightarrow x_2} (x_3 - x_2).$$

For definiteness, let's have the x -variables relate by

$$x_0 < x_1 < x_2 < x_3 < \cdots < x_n,$$

assuming the pattern keeps going. It also helps to label the interval between each:

$$\Delta x_j = x_{j+1} - x_j$$

With a list of n total equations, one sensible step is to take their sum. Tallying the right-hand sides first, we have

$$\begin{aligned} \text{RHS} &= f'(x_0) \Delta x_0 + f'(x_1) \Delta x_1 \\ &\quad + f'(x_2) \Delta x_2 + \cdots, \end{aligned}$$

which can be written concisely as a sum

$$\text{RHS} = \sum_{j=0}^{n-1} f'(x_j) \Delta x_j$$

that goes out to n total terms.

As for the left-hand side, we have

$$\begin{aligned} \text{LHS} &= \lim_{x_1 \rightarrow x_0} f(x_1) - f(x_0) \\ &\quad + \lim_{x_2 \rightarrow x_1} f(x_2) - f(x_1) \\ &\quad + \lim_{x_3 \rightarrow x_2} f(x_3) - f(x_2) \\ &\quad + \cdots + \lim_{x_n \rightarrow x_{n-1}} f(x_n) - f(x_{n-1}) \\ &= f(x_n) - f(x_0) \end{aligned}$$

Notice how any given $f(x_j)$ present in the above has a negative counterpart, thus most terms in the above cancel

out in pairs. This obliterates any notion of ‘limit’ on the left, as only the difference $f(x_n) - f(x_0)$ remains.

Putting the left side and the right side together, we seem to have discovered

$$f(x_n) - f(x_0) \approx \sum_{j=0}^{n-1} f'(x_j) \Delta x_j. \quad (17.1)$$

On the left is simply the difference of a function at two points in its domain. The right, however, seems to be the total *area* of n rectangles, with the j th rectangle having height $f'(x_j)$ and width Δx_j .

As a whole, Equation (17.1) suggests a way to approximate the area under the curve $f'(x)$ between the endpoints x_0, x_n . The tricky part, in general, is finding whatever function $f(x)$ corresponds to the slope $f'(x)$, i.e. the notorious *antiderivative*.

1.1 Riemann Sums

At face value, Equation (17.1) can be implemented ‘as-is’ to approximate the area under $f'(x)$. To clean up the notation, make the substitution $f'(x) = g(x)$, and write the above as

$$f(x_n) - f(x_0) \approx S = \sum_{j=0}^{n-1} g(x_j^*) \Delta x,$$

where the argument sent to $g(x)$ is denoted x_j^* . Furthermore, the subscript on Δx_j has been dropped with the understanding that each Δx_j is one and the same length given by

$$\Delta x = \frac{x_n - x_0}{n},$$

implying

$$x_j = x_0 + j \Delta x.$$

Left, Right, Midpoint Sum

The reason x_j^* gets special attention is there are no natural restrictions on where x_j^* occurs within the interval Δx_j . Right off the bat, there are three obvious options

$$x_j^* = \begin{cases} x_j & \text{Left sum} \\ x_{j+1} & \text{Right sum} \\ (x_j + x_{j+1})/2 & \text{Midpoint sum} \end{cases},$$

which sample from $g(x)$ differently. Explicitly, these mean:

$$\begin{aligned} \frac{S_{\text{Left}}}{\Delta x} &= g(x_0) + g(x_0 + \Delta x) \\ &\quad + g(x_0 + 2\Delta x) + \cdots + g(x_n - \Delta x) \\ \frac{S_{\text{Right}}}{\Delta x} &= g(x_0 + \Delta x) + g(x_0 + 2\Delta x) \\ &\quad + g(x_0 + 3\Delta x) + \cdots + g(x_n) \\ \frac{S_{\text{Mid}}}{\Delta x} &= g\left(x_0 + \frac{\Delta x}{2}\right) + g\left(x_0 + \frac{3\Delta x}{2}\right) \\ &\quad + g\left(x_0 + \frac{5\Delta x}{2}\right) + \cdots + g\left(x_n - \frac{\Delta x}{2}\right) \end{aligned}$$

Problem 1

Using the midpoint sum rule with $n = 10$ bins, approximate the area under the function

$$g(x) = 5x + 2$$

in the domain

$$-2 \leq x \leq 3.$$

Answer: Let $x_0 = -2$, let $x_n = 3$, and $n = 10$ so that

$$\Delta x = \frac{x_n - x_0}{n} = \frac{3 - (-2)}{10} = \frac{1}{2}.$$

At step j in the sum we further have

$$x_j = x_0 + j \Delta x = -2 + \frac{j}{2}.$$

To prepare for the midpoint sum, note that

$$x_j^* = \frac{x_j + x_{j+1}}{2} = \frac{-7}{4} + \frac{j}{2}.$$

The midpoint sum S_M is given by

$$S_M = \sum_{j=0}^{n-1} g(x_j^*) \Delta x = \sum_{j=0}^9 \left(5 \left(\frac{-7}{4} + \frac{j}{2} \right) + 2 \right) \frac{1}{2},$$

which simplifies nicely:

$$\begin{aligned} S_M &= \frac{1}{2} \sum_{j=0}^9 \left(2 - \frac{35}{4} \right) + \frac{5}{4} \sum_{j=0}^9 j \\ &= \frac{1}{2} (10) \left(2 - \frac{35}{4} \right) + \frac{5}{4} (45) \\ &= \frac{45}{2} = 22.5 \end{aligned}$$

Using the midpoint rule, 22.5 happens to be the exact solution to the stated problem, regardless of how many bins

we choose. This brings out a special relationship between the midpoint rule and straight lines: the approximation is perfect.

Problem 2

Using the right sum rule with any number n bins, approximate the area under the function

$$g(x) = 4x - x^2$$

in the domain

$$0 \leq x \leq 4.$$

Answer: Let $x_0 = 0$, let $x_n = 4$, and $n = 10$ so that

$$\Delta x = \frac{x_n - x_0}{n} = \frac{4 - 0}{n} = \frac{4}{n}.$$

At step j in the sum we further have

$$x_j = x_0 + j\Delta x = \frac{4j}{n}.$$

To prepare for the right sum rule, note that

$$x_{j+1} = \frac{4j}{n} + \frac{4}{n}.$$

Then, the right sum rule is

$$\begin{aligned} S_R &= \sum_{j=0}^{n-1} f'(x_{j+1}) \Delta x \\ &= \sum_{j=0}^{n-1} \left(4 \left(\frac{4j}{n} + \frac{4}{n} \right) - \left(\frac{4j}{n} + \frac{4}{n} \right)^2 \right) \frac{4}{n} \end{aligned}$$

Let $k = j + 1$ and simplify the right side to get

$$S_R = \frac{4^3}{n^3} \left(n \sum_{k=1}^n k - \sum_{k=1}^n k^2 \right).$$

By analyzing the remaining sums, it's straightforward to show that

$$\begin{aligned} \sum_{k=1}^n k &= \frac{n(n+1)}{2} \\ \sum_{k=1}^n k^2 &= \frac{n(n+1)(2n+1)}{6}, \end{aligned}$$

and the sum simplifies to

$$S_R = \frac{32}{3} \left(1 - \frac{1}{n^2} \right)$$

This result contains a factor of n , allowing the exactness of S_R to be tuned. Note that $1/n^2$ vanishes for sufficiently large n , telling us the exact area under the curve is $32/3$.

Trapezoid Rule

An improvement over rectangle-based methods is the average the left- and right rules, which effectively turns rectangles into trapezoids, giving (you guessed it) the *trapezoid rule*:

$$\begin{aligned} S_{\text{Trap}} &= \frac{1}{2} (S_{\text{Left}} + S_{\text{Right}}) \\ &= \frac{1}{2} \sum_{j=0}^{n-1} (g(x_j) + g(x_{j+1})) \Delta x \\ &= \frac{g(x_0) + g(x_n)}{2} \Delta x + \sum_{j=1}^{n-1} g(x_j) \Delta x \end{aligned}$$

Without summation notation, the above reads

$$\begin{aligned} \frac{S_{\text{Trap}}}{\Delta x} &= \frac{g(x_0)}{2} + g(x_0 + \Delta x) \\ &\quad + g(x_0 + 2\Delta x) + \cdots + \frac{g(x_n)}{2}. \end{aligned}$$

2 The Integral

There is a regime where all versions of the Riemann sum converge to the same answer, and that is when we impose the limit $\Delta x \rightarrow 0$ and simultaneously $n \rightarrow \infty$. In this limit, the entire picture gets squeezed together, and the area under a curve is approximated by an infinite number of vertical lines. In other words, the Riemann sum becomes an exact solution to the area under the curve $f'(x)$:

$$f(x_n) - f(x_0) = \lim_{n \rightarrow \infty} \sum_{j=0}^{n-1} f'(x_j) \Delta x_j.$$

Integral Notation

To update the above with cleaner notation, the summation is replaced by the ‘integral’, literally a giant ‘S’, via

$$\lim_{\Delta x \rightarrow 0} \sum \Delta x \rightarrow \int dx,$$

which also replaces Δx with dx . The limits on the sum turn into *integration limits*, one ‘lower’ limit and one ‘upper’ limit:

$$\lim_{\Delta x \rightarrow 0} \sum_{j=0}^{n-1} \Delta x \rightarrow \int_{x_0}^{x_n} dx$$

All j -subscripts have also been dropped, as x is now understood to be a continuous variable inside the integral.

2.1 Fundamental Theorem

Using integral notation, the above is written

$$f(x_n) - f(x_0) = \int_{x_0}^{x_n} f'(x) dx.$$

While this is workable, it’s customary to drop the n -subscript from $f(x_n)$, and this term becomes $f(x)$. To prevent a naming conflict on the right, swap the integration variable from x to t :

$$f(x) - f(x_0) = \int_{x_0}^x f'(t) dt. \quad (17.2)$$

This result is called the *fundamental theorem of calculus*, which is the full inversion of the definition of the derivative.

Role of the Antiderivative

A less tautological way to write Equation (17.2) is

$$f(x) - f(x_0) = \int_{x_0}^x g(t) dt,$$

where $f'(t)$ is renamed to some given or otherwise evident function $g(t)$. The left-side function $f(x)$ is considered unknown.

In order to ‘solve’ the integral, $g(t)$ must be expressed as the derivative of something else, which means to find the antiderivative of $g(t)$. The ‘something else’ in this case has already been named, particularly $f(t)$:

$$f(x) - f(x_0) = \int_{x_0}^x \frac{d}{dt} (f(t)) dt$$

The ability to evaluate an integral usually comes down to the ability to find the antiderivative of the function being integrated. This can be quite the chore, if not impossible.

With the proper antiderivative in place, the derivative and the integral on the right mutually annihilate, leaving $f(t)$ alone evaluated at the integration limits, i.e., the quantity $f(x) - f(x_0)$. One way to think of this is to cancel the factors of dt in a way inspired by the chain rule:

$$f(x) - f(x_0) = \int_{x_0}^x \frac{d}{dt} f(t) dt = \int_{x_0}^x df(t)$$

2.2 Definite Integral

indexintegral, definite

When the integration limits x_0, x are specified, either numerically or symbolically, the integral is called *definite*. In order to ‘fully’ solve a definite integral, the antiderivative $f(t)$ must be evaluated at each limit, and the answer is the difference between $f(x_0)$ and $f(x)$. For this, the ‘vertical bar’ notation is used:

$$\int_{x_0}^x df(t) = f(t) \Big|_{x_0}^x = f(x) - f(x_0)$$

Swapping the Limits

One can readily see that swapping the integration limits makes the integral ‘run backwards’, and gains an overall negative sign:

$$\int_x^{x_0} df(t) = f(x_0) - f(x) = - \int_{x_0}^x df(t)$$

Breaking the Interval

The integral remains intact if we split the interval into two or more parts. Introducing a variable a in the domain $x_0 \leq a \leq x$, we may write:

$$\int_{x_0}^x g(t) dt = \int_{x_0}^a g(t) dt + \int_a^x g(t) dt$$

Symmetric Domain

Consider the definite integral over a symmetric domain, meaning $-x_0$ is the lower limit and x_0 is the upper limit:

$$f(x_0) - f(-x_0) = \int_{-x_0}^{x_0} g(t) dt$$

From studying functions, recall that even functions obey

$$f_{\text{even}}(x) - f_{\text{even}}(-x) = 0,$$

and correspondingly for odd functions,

$$f_{\text{odd}}(x) + f_{\text{odd}}(-x) = 0,$$

meaning

$$f_{\text{odd}}(x) - f_{\text{odd}}(-x) = 2f_{\text{odd}}(x).$$

Since the integral of $g(x)$ effectively bumps up its order by one, it follows that the even-ness or oddness of function f is exactly the opposite of function g . We thus gain two cases:

$$f_{\text{even}}(x_0) - f_{\text{even}}(-x_0) = \int_{-x_0}^{x_0} g_{\text{odd}}(t) dt$$

$$f_{\text{odd}}(x_0) - f_{\text{odd}}(-x_0) = \int_{-x_0}^{x_0} g_{\text{even}}(t) dt$$

The first of these results is immediately zero from the properties of even functions. In fact, the integral of any odd function over any symmetric interval, as we've shown, is *always* zero:

$$0 = \int_{-x_0}^{x_0} g_{\text{odd}}(t) dt \quad (17.3)$$

For the other case, we correspondingly find

$$2f_{\text{odd}}(x_0) = \int_{-x_0}^{x_0} g_{\text{even}}(t) dt,$$

which means the integral of an even function over a symmetric interval effectively sums the same area twice. The above is also captured by

$$f_{\text{odd}}(x_0) = \int_0^{x_0} g_{\text{even}}(t) dt. \quad (17.4)$$

2.3 Integration Constant

When the lower integration limit x_0 is unspecified, the term $-f(x_0)$ is called the *integration constant*, denoted C . Setting $f(x_0) = -C$, this means:

$$\int^x f'(t) dt = f(x) + C$$

One way to justify the presence of the integration constant is to realize that any function $f(x) + C$ has the same derivative $f'(x)$, which is to say the absolute vertical offset of the curve has no bearing on its slope. To say this backwards, it follows that any antiderivative calculation without specific limits is only certain up to an arbitrary but non-ignorable constant C .

2.4 Indefinite Integral

indexintegral, indefinite

The integral still retains meaning if we ambiguate both integration limits by writing

$$\int f'(t) dt = f(x) + C,$$

where C is the integration constant.

In this abstraction, the upper integration limit is always understood to be x , which kills the naming conflict in the x -variable on the right. Thus we also have

$$\int f'(x) dx = f(x) + C, \quad (17.5)$$

which is called the *indefinite integral*.

2.5 Integral Operator

In the same sense that one can apply d/dx as an operator to both sides of an equation, we can do the opposite move, which is to apply $\int dx$ across both sides of an equation as well. If

$$g(x) = \frac{d}{dx} f(x),$$

then

$$\int g(x) dx = \int \frac{d}{dx} (f(x)) dx = f(x) + C.$$

On the right, the integral and the derivative are mutually-obliterating, leaving just the enclosed function up to a constant.

Interchangeability

As a sanity check, we should be able to apply d/dx across the whole equation and recover the starting point. Explicitly, this is

$$\frac{d}{dx} \left(\int g(x) dx \right) = \frac{d}{dx} f(x) + \frac{dC}{dx},$$

which readily reduces to $g(x) = f'(x)$, provided that:

$$\frac{d}{dx} \left(\int g(x) dx \right) = \int \frac{d}{dx} (g(x)) dx$$

That is, it's not harmful to move the derivative operation inside the enclosure of the integral.

3 Techniques of Integration

Integral calculations are trickier than anything else in introductory calculus. Here we go through the standard bag of tricks for solving integrals by hand. (Most integrals in the wild are not solvable by hand.)

3.1 Antiderivative Exploit

The most direct way to solve an integral is pick out (by experience or by luck) the antiderivative of the function being integrated. For instance, consider

$$I = \int_0^{\sqrt{\pi/2}} x \cos(x^2) dx.$$

Right away, note that the function being integrated can be written as a derivative

$$x \cos(x^2) = \frac{d}{dx} \left(\frac{1}{2} \sin(x^2) \right),$$

so then

$$I = \int_0^{\sqrt{\pi/2}} \frac{d}{dx} \left(\frac{1}{2} \sin(x^2) \right) dx,$$

and then the integral and derivative operators cancel, leaving only the evaluation:

$$I = \frac{1}{2} \sin(x^2) \Big|_0^{\sqrt{\pi/2}} = \frac{1}{2} (1 - 0) = \frac{1}{2}$$

3.2 Exponents and Roots

Powers

Starting with the power rule for differentiation

$$\frac{d}{dx} (x^n) = nx^{n-1},$$

replace $n \rightarrow n + 1$ for convenience and write the same rule:

$$(n + 1)x^n = \frac{d}{dx} (x^{n+1})$$

From this, we can apply the integral operator to derive the rule for integrating powers and roots:

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C \quad (17.6)$$

Going through a few exemplary cases, i.e. playing with common values of n , we generate some useful information. You are encouraged to work through each of these:

$$\int dx = x + C \quad (17.7)$$

$$\int x dx = \frac{1}{2}x^2 + C \quad (17.8)$$

$$\int x^2 dx = \frac{1}{3}x^3 + C \quad (17.9)$$

$$\int \sqrt{x} dx = \frac{2}{3}x^{3/2} + C \quad (17.10)$$

$$\int x^{3/2} dx = \frac{2}{5}x^{5/2} + C \quad (17.11)$$

$$\int x^{-2} dx = \frac{-1}{x} + C \quad (17.12)$$

$$\int x^{-3} dx = \frac{-1}{2x^2} + C \quad (17.13)$$

$$\int \frac{dx}{\sqrt{x}} = 2\sqrt{x} + C \quad (17.14)$$

$$\int x^{-3/2} dx = \frac{-2}{\sqrt{x}} + C \quad (17.15)$$

Reciprocal

One special case to power rule formula is the integral of $1/x$. Recalling that the derivative of the natural logarithm yields this result, i.e.

$$\frac{d}{dx} (\ln(x)) = \frac{1}{x},$$

the following must hold:

$$\int \frac{1}{x} dx = \ln(x) + C \quad (17.16)$$

Exponential

Starting with the derivative rule for exponents

$$\frac{d}{dx} (n^x) = n^x \ln(n),$$

it must follow that:

$$\int n^x dx = \frac{n^x}{\ln(n)} + C \quad (17.17)$$

Natural Exponential

In Equation (17.17), set $n = e$ to see that the area under the natural exponential is equal to the height of the curve itself (up to a constant):

$$\int e^x dx = e^x + C \quad (17.18)$$

Applied Chain Rule

Using the power rule and chain rule for derivatives, it's straightforward to derive

$$\frac{d}{dx} \sqrt{f(x)} = \frac{f'(x)}{2\sqrt{f(x)}}.$$

Applying the $\int dx$ operator across both sides and simplifying leads to a useful identity:

$$\int \frac{f'(x)}{2\sqrt{f(x)}} dx = \sqrt{f(x)} + C \quad (17.19)$$

It takes some effort to train the eye to make use of identities such as the above. Exploring one case, suppose we have

$$f(x) = 1 \pm x^2$$

with $f'(x) = \pm 2x$. Plugging all of this in and simplifying gives a two-channel result:

$$\int \frac{\pm x}{\sqrt{1 \pm x^2}} dx = \sqrt{1 \pm x^2} + C \quad (17.20)$$

Problem 3

Prove that the area under a parabolic segment of base b and height h is

$$A = \frac{2}{3}bh.$$

Problem 4

Prove that the area of the 'lens' formed between the curves

$$\begin{aligned} y_1 &= x^2 \\ y_2 &= ax + b \end{aligned}$$

is

$$A = \frac{1}{6} \left(a^2 + 4b \right)^{3/2}.$$

3.3 U-Substitution

The standard integral

$$\int_{x_0}^x f'(t) dt = f(x) - f(x_0)$$

can sometimes be made simpler by a technique called *u-substitution*, which entails choosing a function $u(x)$ and then recasting the integral in this variable.

The *u*-substitution can be established by multiplying $du/du = 1$ into the standard integral, i.e.

$$\int_{x_0}^x \frac{df}{dt} dt = \int \frac{df}{dt} \frac{du}{du} dt = \int_{u(x_0)}^{u(x)} \frac{df}{du} du,$$

where the factor dt/dt cancels out. Importantly, note that the limits on the integral are also modified to respect $u(x)$. Once the result is attained as $f(u)$, reverse-substitute to attain $f(x)$.

A pragmatic way to choose the *correct* *u*-substitution can be established. Consider an indefinite integral

$$I = \int f(x) g(x) dx$$

for two functions $f(x)$, $g(x)$. Under the substitution $u = u(x)$, the above still must come out to

$$I = \int f(u) du,$$

which can only mean

$$g(x) = \frac{du}{dx}.$$

That is, the function $g(x)$ must be (at least) proportional to the derivative of the substitution $u(x)$.

Exemplary Case

Consider again the definite integral

$$I = \int_0^{\sqrt{\pi/2}} x \cos(x^2) dx.$$

To solve this with *u*-substitution, let

$$u(x) = x^2$$

such that

$$du = 2x dx.$$

The limits of the integral must change to reflect the u -substitution as well. With this, the integral becomes

$$I = \int_0^{\pi/2} \frac{1}{2} \cos(u) du ,$$

which has a straightforward solution:

$$I = \frac{1}{2} \sin(u) \Big|_0^{\pi/2}$$

From here, one may stay in the u -domain to get the final answer, or switch back to the x -variable to recover

$$I = \frac{1}{2} \sin(x^2) \Big|_0^{\sqrt{\pi/2}} = \frac{1}{2} (1 - 0) = \frac{1}{2} .$$

Constant Shift

If the x -dependence in the integrand is shifted by a constant λ , i.e.

$$u(x) = x + \lambda ,$$

then

$$du = dx$$

always holds.

For instance, in

$$I = \int (x + 3)^n dx ,$$

we can let $u = x + 3$ so the above becomes

$$I = \int u^n du ,$$

which is easy to solve using Equation (17.6) as

$$I = \frac{u^{n+1}}{n+1} + C .$$

Reverse-substitute to get the answer in terms of x :

$$\int (x + 3)^n dx = \frac{(x + 3)^{n+1}}{n + 1} + C$$

Problem 5

Use u -substitution to prove Equation (17.20).

Problem 6

Use u -substitution to prove:

$$\int \frac{dx}{1+x} = \ln(1+x) + C \quad (17.21)$$

Problem 7

For constant a , prove:

$$\int (x - a)^{n-1} dx = \frac{1}{n} (x - a)^n$$

The $d(\sin)$ Shortcut

Integrals of the form

$$I = \int f(\sin(x)) \cos(x) dx$$

are transformed by standard u -substitution. Letting

$$u(x) = \sin(x)$$

such that

$$\frac{du}{dx} = \cos(x) ,$$

the above readily takes a more standard form:

$$I = \int f(\sin(x)) \cos(x) dx = \int f(u) du$$

The combination $\cos(x) dx$ is written $d(\sin(x))$ as a shortcut, which embeds the notions $u = \sin(x)$, $du = \cos(x) dx$ simultaneously:

$$\cos(x) dx = d(\sin(x))$$

For example, consider the indefinite integral

$$J = \int \sin^2(x) \cos(x) dx ,$$

which looks like a rather messy antiderivative to wrestle with. Applying the so-called $d \sin()$ shortcut, the integral reads

$$J = \int \sin^2(x) d(\sin(x)) = \int u^2 du ,$$

and the problem is now simpler in the u -variable. To finish the job, we have

$$J = \frac{1}{3} u^3 + C = \frac{1}{3} \sin^3(x) + C .$$

3.4 Integrands with Roots

Not every integral involving a square root (or worse) can be solved by simple u -substitution. In these cases, it's worth including the exponent of the embedded root in the u -substitution.

To illustrate, consider the indefinite integral

$$I = \int \frac{x}{(x-4)^{1/3}} dx,$$

which begs trying $u = x - 4$, but this makes an absolute mess. Instead, let us take

$$u = (x - 4)^{1/3}$$

such that

$$\begin{aligned} x &= u^3 + 4 \\ dx &= 3u^2 du. \end{aligned}$$

Then, the integral looks much easier in the u -domain:

$$I = 3 \int (u^4 + 4u) du$$

Problem 8

Use the above as a starting point to prove:

$$\int \frac{x}{(x-4)^{1/3}} = \frac{3}{5} (x-4)^{2/3} (x+6) + C$$

3.5 Partial Fractions

...

3.6 Integration by Parts

Consider the product $H(x)$ of two functions $U(x)$, $V(x)$,

$$H(x) = U(x) V(x),$$

and take the derivative of H , minding the product rule:

$$\frac{d}{dx} H(x) = V(x) \frac{d}{dx} U(x) + U(x) \frac{d}{dx} V(x)$$

Next, apply the integral operator $\int dx$ across the whole equation:

$$\begin{aligned} \int \frac{d}{dx} H(x) dx &= \int V(x) \frac{d}{dx} U(x) dx \\ &\quad + \int U(x) \frac{d}{dx} V(x) dx \end{aligned}$$

Since the integral and derivative operators are mutually annihilating, the left side is simply $H(x)$ evaluated at the integration limits. It suffices to leave the vertical bar empty while working in indefinite form:

$$\int \frac{d}{dx} H(x) dx = H(x) \Big| = U(x) V(x) \Big|$$

Introducing the shorthand notation

$$\frac{d}{dx} U(x) = dU$$

and similar for dV , the above is written

$$UV \Big| = \int V dU + \int U dV,$$

where all quantities are understood to be functions of x .

The reason for doing this is, suppose you are handed an integral of the form $\int U dV$ that is difficult to solve. If we can somehow manage to identify $V(x)$, then perhaps the integral $\int V dU$ is easier than its counterpart. All of this inspires the *integration by parts* formula:

$$\int U dV = UV \Big| - \int V dU \quad (17.22)$$

Exemplary Case

To demonstrate integration by parts, consider the definite integral

$$I = \int_0^{\pi/2} x \cos(x) dx,$$

which we immediately rewrite as

$$\int_0^{\pi/2} x \cos(x) dx = \int_0^{\pi/2} U dV.$$

Then identify

$$\begin{aligned} U &= x \\ dV &= \cos(x) dx, \end{aligned}$$

and we now have two 'mini problems' of determining $dU(x)$ and $V(x)$.

For this example, dU is simply equal to dx . (It's always easy to calculate dU .) As for V , we have $dV/dx = \cos(x)$, which can only mean $V(x) = \sin(x)$.

The integration by parts formula then tells us:

$$\int_0^{\pi/2} x \cos(x) dx = x \sin(x) \Big|_0^{\pi/2} - \int_0^{\pi/2} \sin(x) dx$$

Notice how the ‘hard’ integral on the left is replaced by an ‘easy’ integral on the right. The answer is now straightforward:

$$I = \int_0^{\pi/2} x \cos(x) dx = \frac{\pi}{2} - 1$$

Natural Logarithm

The integration by parts recipe also works when there is one function in the integrand, and this is how to find the integral of the natural logarithm. Starting with

$$I = \int \ln(x) dx,$$

let

$$\begin{aligned} U &= \ln(x) \\ dV &= dx \end{aligned}$$

such that

$$\begin{aligned} dU &= dx/x \\ V &= x. \end{aligned}$$

Then, we have

$$I = x \ln(x) \Big| - \int dx,$$

simplifying to:

$$\int \ln(x) dx = x \ln(x) - x + C \quad (17.23)$$

Problem 9

Use u -substitution to find the integral of the shifted natural logarithm:

$$\int \ln(1+x) dx = (1+x) \ln(1+x) + x + C \quad (17.24)$$

Problem 10

Use integration by parts to prove:

$$\int_0^{\infty} x^2 e^{-x} dx = 2$$

3.7 Label Trick

Area of the Unit Circle

Consider the definite integral that attempts to calculate the area of one quarter of the unit circle:

$$A = \int_0^{\pi/2} \sin^2(\theta) d\theta$$

This can be attacked with integration by parts by letting

$$\begin{aligned} U &= \sin(\theta) \\ dV &= \sin(\theta) d\theta \end{aligned}$$

such that

$$\begin{aligned} dU &= \cos(\theta) d\theta \\ V &= -\cos(\theta), \end{aligned}$$

and then

$$A = \cancel{-\sin(\theta)\cos(\theta)} \Big|_0^{\pi/2} + \int_0^{\pi/2} \cos^2(\theta) d\theta.$$

All we’ve managed to show is that the function $\sin^2(\theta)$ can be replaced by $\cos^2(\theta)$ and the integral remains the same.

Now make use of the fundamental trigonometric identity

$$\sin^2(\theta) + \cos^2(\theta) = 1$$

to write

$$\int_0^{\pi/2} \cos^2(\theta) d\theta = \int_0^{\pi/2} d\theta - \int_0^{\pi/2} \sin^2(\theta) d\theta.$$

The left-most and right-most integrals are both equal to A , and all of the hard work suddenly vanishes with the so-called *label trick*:

$$A = \int_0^{\pi/2} d\theta - A$$

Solving for A is a matter of algebra, and the remaining integral is trivial:

$$A = \frac{1}{2} \int_0^{\pi/2} d\theta = \frac{\pi}{4}$$

Tricky Logarithmic Integral

A tricky problem that you’re welcome to stop reading and try on your own is the following definite integral:

$$I = \int_0^{\infty} \frac{\ln(x)}{1+x+x^2} dx$$

The key to this problem is the substitution $u = 1/x$. From this, we have $du/dx = -1/x^2$, and furthermore $\ln(1/u) = -\ln(u)$. The integration limits also end up swapping, and the integral becomes

$$I = \int_{\infty}^0 \frac{-\ln(u)}{1+1/u+1/u^2} \frac{-du}{u^2}.$$

Simplifying further, we find

$$I = \int_{\infty}^0 \frac{\ln(u)}{1+u+u^2} du,$$

and swap the integration limits by paying with a negative sign:

$$I = - \int_0^{\infty} \frac{\ln(u)}{1+u+u^2} du$$

This result is exactly opposite to the problem we started with, up to a trivial change of letters. In effect, we have found

$$I = -I,$$

which can *only* mean $I = 0$:

$$0 = \int_0^{\infty} \frac{\ln(x)}{1+x+x^2} dx$$

3.8 Trigonometric Integrals

Standard Functions

The integral of each trigonometric function is straightforwardly calculated using antiderivatives or other integration techniques. In indefinite form, these are:

$$\int \sin(x) dx = -\cos(x) + C \quad (17.25)$$

$$\int \cos(x) dx = \sin(x) + C \quad (17.26)$$

$$\begin{aligned} \int \tan(x) dx &= - \int \frac{d(\cos(x))}{\cos(x)} \\ &= -\ln(\cos(x)) + C \end{aligned} \quad (17.27)$$

$$\begin{aligned} \int \cot(x) dx &= \int \frac{d(\sin(x))}{\sin(x)} \\ &= \ln(\sin(x)) + C \end{aligned} \quad (17.28)$$

$$\int \sec(x) dx = \ln(\sec(x) + \tan(x)) + C \quad (17.29)$$

$$\int \csc(x) dx = -\ln(\csc(x) + \cot(x)) + C \quad (17.30)$$

We can also recall the derivative of each trigonometric function and make use of the $\int dx$ operator to come up with

a few more:

$$\int \sec^2(x) dx = \tan(x) + C \quad (17.31)$$

$$\int \csc^2(x) dx = -\cot(x) + C \quad (17.32)$$

$$\int \tan(x) \sec(x) dx = \sec(x) + C \quad (17.33)$$

$$\int \cot(x) \csc(x) dx = -\csc(x) + C \quad (17.34)$$

Squared Integrand

The pair of indefinite integrals

$$I_1 = \int \sin^2(x) dx$$

$$I_2 = \int \cos^2(x) dx$$

can be solved simultaneously. Using the fundamental trig identity, we see

$$I_1 + I_2 = \int (\sin^2(x) + \cos^2(x)) dx = \int dx = x \Big|,$$

or equivalently

$$I_1 + I_2 = x + C.$$

Now integrate I_1 by parts via

$$U = \sin(x)$$

$$dV = \sin(x) dx$$

such that

$$dU = \cos(x) dx$$

$$V = -\cos(x),$$

and I_1 is written

$$I_1 = -\sin(x) \cos(x) \Big| + \int \cos^2(x) dx,$$

simplifying to

$$I_1 - I_2 = -\sin(x) \cos(x) + C.$$

With two equations and two unknowns, I_1 and I_2 can be isolated independently, resulting in

$$\int \sin^2(x) dx = \frac{-\sin(x) \cos(x)}{2} + \frac{x}{2} + C \quad (17.35)$$

$$\int \cos^2(x) dx = \frac{\sin(x) \cos(x)}{2} + \frac{x}{2} + C \quad (17.36)$$

The pair of indefinite integrals

$$I_3 = \int \tan^2(x) dx$$

$$I_4 = \int \sec^2(x) dx$$

can also be solved together. Using another fundamental trig identity, find

$$I_4 - I_3 = x + C,$$

which means only I_3 or I_4 need be calculated and we get the other for free.

Choosing I_4 , recall that the derivative of the tangent is the square of the secant, so

$$I_4 = \int \frac{d}{dx} \tan(x) dx = \tan(x) + C,$$

and conclude:

$$\int \tan^2(x) dx = \tan(x) - x + C \quad (17.37)$$

$$\int \sec^2(x) dx = \tan(x) + C$$

Finally, the pair of indefinite integrals

$$I_5 = \int \cot^2(x) dx$$

$$I_6 = \int \csc^2(x) dx$$

can also be solved together. Using another fundamental trig identity, find

$$I_6 - I_5 = x + C.$$

The easiest way to proceed is to remember that the derivative of the cotangent is the negative of the square of the cosecant. Just kidding, that's not terribly easy to remember, but nonetheless the integral I_6 becomes

$$I_6 = \int \frac{d}{dx} (-\cot(x)) dx = -\cot(x) + C.$$

From the above we get the pair of answers:

$$\int \cot^2(x) dx = -\cot(x) - x + C \quad (17.38)$$

$$\int \csc^2(x) dx = -\cot(x) + C$$

Inverse Functions

Integrals of the inverse trigonometric functions can be tricky to find. Integration by parts works well on a few of them, such as the arctangent. For

$$I = \int \arctan(x) dx,$$

let

$$U = \arctan(x)$$

$$dV = dx$$

such that:

$$dU = \frac{dx}{1+x^2}$$

$$V = x$$

With this, the integral reads

$$I = x \arctan(x) \Big| - \int \frac{x}{1+x^2} dx$$

The remaining integral is solved by standard u -substitution, namely $u = 1 + x^2$ such that $du = 2x dx$. After simplifying, we get the answer:

$$\int \arctan(x) dx = x \arctan(x)$$

$$- \frac{1}{2} \ln(1+x^2) + C \quad (17.39)$$

The same recipe works for several other inverse trigonometric functions, namely the arccosine, arcsine, and arccotangent:

$$\int \arccos(x) dx = x \cos(x)$$

$$- \frac{1}{2} \ln(1-x^2) + C \quad (17.40)$$

$$\int \arcsin(x) dx = x \sin(x)$$

$$+ \frac{1}{2} \ln(1-x^2) + C \quad (17.41)$$

$$\int \operatorname{arccot}(x) dx = x \operatorname{arccot}(x)$$

$$+ \frac{1}{2} \ln(1+x^2) + C \quad (17.42)$$

Conspicuously absent from our stack of results are the integrals of the arcsecant and arccosecant. These require more than a simple u -substitution that we haven't hit yet, so stay tuned.

Reduction Formulas

For positive integer m , consider the indefinite integral

$$I = \int \sin^m(x) dx .$$

Integrating by parts, we first write

$$\begin{aligned} U &= \sin^{m-1}(x) \\ dV &= \sin(x) dx \end{aligned}$$

and also

$$\begin{aligned} dU &= (m-1) \sin^{m-2}(x) \cos(x) dx \\ V &= -\cos(x) . \end{aligned}$$

From this, we have

$$\begin{aligned} I &= -\sin^{m-1}(x) \cos(x) \Big| \\ &\quad + (m-1) \int \sin^{m-2}(x) \cos^2(x) dx . \end{aligned}$$

Next, replace $\cos^2(x)$ with $1 - \sin^2(x)$ and use the label trick, giving

$$\begin{aligned} I &= -\sin^{m-1}(x) \cos(x) \Big| \\ &\quad + (m-1) \int \sin^{m-2}(x) dx - (m-1) I , \end{aligned}$$

and solving for I we arrive at a *trigonometric reduction formula*:

$$\begin{aligned} \int \sin^m(x) dx &= \frac{-1}{m} \sin^{m-1}(x) \cos(x) \Big| \\ &\quad + \frac{m-1}{m} \int \sin^{m-2}(x) dx \quad (17.43) \end{aligned}$$

Similar reduction formulas exist for each of the elementary trig functions. Each of the following is attained by integration by parts and the label trick:

$$\begin{aligned} \int \cos^m(x) dx &= \frac{1}{m} \cos^{m-1}(x) \sin(x) \Big| \\ &\quad + \frac{m-1}{m} \int \cos^{m-2}(x) dx \quad (17.44) \end{aligned}$$

$$\begin{aligned} \int \tan^m(x) dx &= \frac{1}{m-1} \tan^{m-1}(x) \Big| \\ &\quad - \int \tan^{m-2}(x) dx \quad (17.45) \end{aligned}$$

$$\begin{aligned} \int \csc^m(x) dx &= \frac{-1}{m-1} \csc^{m-2}(x) \cot(x) \Big| \\ &\quad + \frac{m-2}{m-1} \int \csc^{m-2}(x) dx \quad (17.46) \end{aligned}$$

$$\begin{aligned} \int \sec^m(x) dx &= \frac{1}{m-1} \sec^{m-2}(x) \tan(x) \Big| \\ &\quad + \frac{m-2}{m-1} \int \sec^{m-2}(x) dx \quad (17.47) \end{aligned}$$

$$\begin{aligned} \int \cot^m(x) dx &= \frac{-1}{m-1} \cot^{m-1}(x) \Big| \\ &\quad - \int \cot^{m-2}(x) dx \quad (17.48) \end{aligned}$$

Another reduction formula that mixes the sine and cosine can be established. Consider the case

$$I = \int \sin^m(x) \cos^n(x) dx .$$

By letting $u = \sin^{m-1}(x)$ and following the consequences, one finds

$$\begin{aligned} \int \sin^m(x) \cos^n(x) dx &= \\ &= -\frac{1}{m+n} \sin^{m-1}(x) \cos^{n+1}(x) \Big| \\ &\quad + \frac{m-1}{m+n} \int \sin^{m-2}(x) \cos^n(x) dx . \quad (17.49) \end{aligned}$$

Note that this result reproduces Equation (17.43) for $n = 0$. A different result is attained by letting $u = \cos^{n-1}(x)$:

$$\begin{aligned} \int \sin^m(x) \cos^n(x) dx &= \\ &= \frac{1}{m+n} \sin^{m+1}(x) \cos^{n-1}(x) \Big| \\ &\quad + \frac{n-1}{m+n} \int \sin^m(x) \cos^{n-2}(x) dx . \quad (17.50) \end{aligned}$$

Note that this result reproduces Equation (17.44) for $m = 0$.

Mixed Wavelengths

Starting with the product formula

$$2 \sin(\alpha) \cos(\beta) = \sin(\alpha + \beta) + \sin(\alpha - \beta),$$

suppose that α, β are multiples of an angle θ

$$\begin{aligned}\alpha &= m\theta \\ \beta &= n\theta\end{aligned}$$

for non-equal integers m, n .

Next apply the integral operator $\int d\theta$ across the whole equation

$$\begin{aligned}2 \int \sin(m\theta) \cos(n\theta) d\theta &= \int \sin(m\theta + n\theta) d\theta \\ &+ \int \sin(m\theta - n\theta) d\theta,\end{aligned}$$

and simplify:

$$\begin{aligned}\int \sin(m\theta) \cos(n\theta) d\theta &= \\ &\frac{-\cos((m+n)\theta)}{2(m+n)} - \frac{\cos((m-n)\theta)}{2(m-n)} + C\end{aligned}\quad (17.51)$$

More product formula exploits lead to additional mixed-wavelength integral identities:

$$\begin{aligned}\int \cos(m\theta) \cos(n\theta) d\theta &= \\ &\frac{\sin((m+n)\theta)}{2(m+n)} + \frac{\sin((m-n)\theta)}{2(m-n)} + C\end{aligned}\quad (17.52)$$

$$\begin{aligned}\int \sin(m\theta) \sin(n\theta) d\theta &= \\ &\frac{-\sin((m+n)\theta)}{2(m+n)} + \frac{\sin((m-n)\theta)}{2(m-n)} + C\end{aligned}\quad (17.53)$$

Orthogonality

Evaluating the mixed-wavelength integral identities (17.51)-(17.53), in various domains of length 2π leads to some additional information called *orthogonality relations*. (Keep in mind that m, n are different integers.)

Choosing $[-\pi : \pi]$ first, we find

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(m\theta) \cos(n\theta) d\theta &= 0 \\ \int_{-\pi}^{\pi} \cos(m\theta) \cos(n\theta) d\theta &= 0 \\ \int_{-\pi}^{\pi} \sin(m\theta) \sin(n\theta) d\theta &= 0\end{aligned}$$

The same results hold when the domain is changed to $[0 : 2\pi]$:

$$\begin{aligned}\int_0^{2\pi} \sin(m\theta) \cos(n\theta) d\theta &= 0 \\ \int_0^{2\pi} \cos(m\theta) \cos(n\theta) d\theta &= 0 \\ \int_0^{2\pi} \sin(m\theta) \sin(n\theta) d\theta &= 0\end{aligned}$$

When the wavelengths m, n are one and the same integer m , two results switch to nonzero

$$\begin{aligned}\int_{-\pi}^{\pi} \cos^2(m\theta) d\theta &= \int_0^{2\pi} \cos^2(m\theta) d\theta = \pi \\ \int_{-\pi}^{\pi} \sin^2(m\theta) d\theta &= \int_0^{2\pi} \sin^2(m\theta) d\theta = \pi,\end{aligned}$$

and the case that mixes sine and cosine remains zero:

$$\begin{aligned}\int_{-\pi}^{\pi} \sin(m\theta) \cos(m\theta) d\theta &= \\ &= \int_0^{2\pi} \sin(m\theta) \cos(m\theta) d\theta = 0\end{aligned}$$

3.9 Hyperbolic Integrals

...

3.10 Trigonometric Substitution

Each of the following integrals

$$\begin{aligned}I_1 &= \int \frac{dx}{x^2 \sqrt{x^2 + a^2}} \\ I_2 &= \int \frac{\sqrt{a^2 - x^2}}{x^2} dx \\ I_3 &= \int \frac{dx}{(x^2 - a^2)^{3/2}}\end{aligned}$$

for nonzero constant a are difficult to solve by standard u -substitution or integration by parts. In fact, each requires a different trick called *trigonometric substitution*.

Tangent Substitution

When the integrand contains $x^2 + a^2$, let

$$x = a \tan (\theta) ,$$

so then

$$dx = a \sec^2 (\theta) d\theta .$$

By standard trig identities, the quantity $x^2 + a^2$ becomes

$$x^2 + a^2 = a^2 \sec^2 (\theta) .$$

With this, the integral I_1 transforms into something we can solve:

$$I_1 = \int \frac{a \sec^2 (\theta)}{a^3 \tan^2 (\theta) \sec (\theta)} d\theta = \frac{1}{a^2} \int \frac{d(\sin (\theta))}{\sin^2 (\theta)}$$

Problem 11

Use the above as a starting point to prove:

$$\int \frac{dx}{x^2 \sqrt{x^2 + a^2}} = -\frac{\sqrt{x^2 + a^2}}{a^2 x} + C$$

Sine Substitution

When the integrand contains $a^2 - x^2$, let

$$x = a \sin (\theta) ,$$

so then

$$dx = a \cos (\theta) d\theta .$$

By standard trig identities, the quantity $a^2 - x^2$ becomes

$$a^2 - x^2 = a^2 \cos^2 (\theta) .$$

With the sine substitution, the integral I_2 reduces to a simpler problem:

$$I_2 = \int \frac{a^2 \cos^2 (\theta)}{a^2 \sin^2 (\theta)} d\theta = \int \cot^2 (\theta) d\theta$$

Problem 12

Use the above as a starting point to prove:

$$\int \frac{\sqrt{a^2 - x^2}}{x^2} dx = -\arcsin \left(\frac{x}{a} \right) - \frac{\sqrt{a^2 - x^2}}{x} + C$$

Secant Substitution

When the integrand contains $x^2 - a^2$, let

$$x = a \sec (\theta) ,$$

so then

$$dx = a \sec (\theta) \tan (\theta) d\theta .$$

By standard trig identities, the quantity $x^2 - a^2$ becomes

$$x^2 - a^2 = a^2 \tan^2 (\theta) .$$

With the sine substitution, the integral I_3 reduces to a simpler problem:

$$I_3 = \frac{1}{a^2} \int \frac{d(\sin (\theta))}{\sin^2 (\theta)}$$

Problem 13

Use the above as a starting point to prove:

$$\int \frac{dx}{(x^2 - a^2)^{3/2}} = \frac{-x}{a^2 \sqrt{x^2 - a^2}} + C$$

Trigonometric Ratios

Rational functions of sine and cosine land to a particular u -substitution:

$$u = \tan (\theta/2) .$$

From the trigonometric half-angle formulas, we can next write

$$\begin{aligned} \cos (\theta) &= \frac{1 - u^2}{1 + u^2} \\ \sin (\theta) &= \frac{2u}{1 + u^2} , \end{aligned}$$

and

$$\begin{aligned} u &= \frac{\sin (\theta)}{1 + \cos (\theta)} \\ du &= \frac{1}{2} (1 + u^2) d\theta . \end{aligned}$$

With this substitution, integrals of the form

$$I = \int f(\sin (\theta), \cos (\theta)) d\theta$$

can be written:

$$I = \int f\left(\frac{2u}{1 + u^2}, \frac{1 - u^2}{1 + u^2}\right) \frac{du}{1 + u^2}$$

In the general case, this substitution works when the function being integrated is a polynomial of two variables or a ratio of two polynomials.

To illustrate, consider the indefinite integral

$$J = \int \frac{d\theta}{3 + \cos(\theta)} .$$

Using the above substitutions, the integral becomes

$$J = \int \frac{du}{2 + u^2} .$$

Problem 14

Use the above as a starting point to prove:

$$\int \frac{d\theta}{3 + \cos(\theta)} = \frac{1}{\sqrt{2}} \arctan \left(\frac{1}{\sqrt{2}} \tan \left(\frac{\theta}{2} \right) \right) + C$$

Arcsecant and Arccosecant

The integrals of the arcsecant and the arccosecant have to be cracked with trigonometric substitution. For

$$I = \int \operatorname{arcsec}(x) dx ,$$

proceed with integration by parts to write

$$\begin{aligned} U &= \operatorname{arcsec}(x) \\ dV &= dx \end{aligned}$$

such that

$$\begin{aligned} dU &= \frac{dx}{x\sqrt{x^2-1}} \\ V &= x . \end{aligned}$$

The integral becomes

$$\int \operatorname{arcsec}(x) dx = x \operatorname{arcsec}(x) \Big| - \int \frac{dx}{\sqrt{x^2-1}} .$$

The new integral on the right is handled by a secant substitution. Let

$$x = \sec(\theta)$$

such that

$$dx = \sec(\theta) \tan(\theta) d\theta ,$$

and

$$\sqrt{x^2-1} = \tan(\theta) ,$$

so we have

$$\int \frac{dx}{\sqrt{x^2-1}} = \int \sec(\theta) d\theta .$$

The integral of the secant has a known solution, namely

$$\int \sec(\theta) d\theta = \ln(\sec(\theta) + \tan(\theta)) + C ,$$

or, in terms of the x -variable,

$$\int \sec(\theta) d\theta = \ln(x + \sqrt{x^2-1}) + C .$$

Finally, we have the answer:

$$\begin{aligned} \int \operatorname{arcsec}(x) dx &= x \operatorname{arcsec}(x) \Big| \\ &\quad - \ln(x + \sqrt{x^2-1}) + C \end{aligned} \quad (17.54)$$

Problem 15

Do a similar exercise for the arccosecant:

$$\begin{aligned} \int \operatorname{arccsc}(x) dx &= x \operatorname{arccsc}(x) \Big| \\ &\quad + \ln(x + \sqrt{x^2-1}) + C \end{aligned} \quad (17.55)$$

Area of the Ellipse

For the ellipse defined by

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 ,$$

the area contained in the first quadrant (a quarter of the ellipse) is given by

$$A = \int_0^a y(x) dx .$$

It also happens that the same ellipse can be described using a pair of parametric equations, particularly

$$\begin{aligned} x &= a \cos(\theta) \\ y &= b \sin(\theta) , \end{aligned}$$

easily shown to reproduce the Cartesian formula. Substituting the above equations for x , y into the area integral changes the integration variable to θ :

$$A = -ab \int_{\pi/2}^0 \sin^2(\theta) d\theta$$

The integral of the square of the sine is well known by known, particularly by equation (17.35). Evaluating the definite integral gives the final answer:

$$A = -ab \left(\frac{-\pi}{4} \right) = \frac{1}{4} \pi ab$$

The area of the complete ellipse is πab .

Problem 16

Show that the area of the ellipse

$$ax^2 + bxy + cy^2 = 1$$

is equal to

$$A = \frac{2\pi}{\sqrt{4ac - b^2}}.$$

Hint: Rotate the coordinates and write the area of the same ellipse in the rotated system.

3.11 Mirror Trick

A lesser-known technique we'll call the *mirror trick* can help with integrals such as

$$J = \int_0^{\pi/2} \frac{\sqrt{\sin(\theta)}}{\sqrt{\sin(\theta)} + \sqrt{\cos(\theta)}} d\theta.$$

For practice, consider the definite integral of a well-behaved function $g(x)$:

$$I = \int_a^b g(x) dx$$

By making the substitution

$$\begin{aligned} u &= b + a - x \\ du &= -dx, \end{aligned}$$

we find

$$I = \int_b^a g(b + a - u) (-du).$$

Of course, the integration variable itself can be swapped with any other letter, so we come up with a second equation for J involving the integral in the x -domain:

$$I = \int_a^b g(b + a - x) dx$$

The same idea can be applied to a different integral

$$K = \int_a^b \frac{g(x)}{g(b + a - x) + g(x)} dx,$$

which, using the same u -substitution $u = b + a - x$, becomes

$$K = \int_b^a \frac{g(b + a - u)}{g(u) + g(b + a - u)} (-du),$$

or equivalently

$$K = \int_a^b \frac{g(b + a - x)}{g(x) + g(b + a - x)} dx.$$

Take the two expressions for K and take their sum,

$$2K = \int_a^b \frac{g(x) + g(b + a - x)}{g(x) + g(b + a - x)} dx,$$

and notice the entire integrand cancels, leaving

$$K = \frac{b - a}{2}.$$

Evidently, the result of integral K has nothing to do with the function being integrated, only the limits matter:

$$\int_a^b \frac{g(x)}{g(b + a - x) + g(x)} dx = \frac{b - a}{2} \quad (17.56)$$

Returning to the problem on hand, the integral J can be written

$$J = \int_0^{\pi/2} \frac{\sqrt{\sin(\theta)}}{\sqrt{\sin(\theta)} + \sqrt{\sin(\pi/2 - \theta)}} d\theta.$$

Comparing this to Equation (17.56), let $a = 0$ and $b = \pi/2$ and the result is half their difference:

$$J = \frac{\pi}{4}$$

3.12 Series Expansion

Integration and series expansion play nicely together and are used often to approximate the solution to otherwise insoluble integrals.

Physical Pendulum

It's possible to show using energy conservation that a frictionless pendulum of length L and mass m in uniform gravity is governed by

$$\frac{d\theta}{dt} = \sqrt{\frac{2g}{L} \sqrt{\cos(\theta) - \cos(\theta_0)}},$$

where θ is the deflection of the pendulum from vertical and θ_0 represents the highest position attainable where motion momentarily stops. This is a 'separable' equation, and can be reshuffled as an indefinite integral:

$$\int \frac{d\theta}{\sqrt{\cos(\theta) - \cos(\theta_0)}} = \sqrt{\frac{2g}{L}} \int dt$$

The left side needs some preparation before proceeding. The cosine terms are replaced using the half-angle formula

$$1 - \cos(\theta) = 2 \sin^2\left(\frac{\theta}{2}\right).$$

Also define

$$\sin(\phi) = \frac{1}{a} \sin\left(\frac{\theta}{2}\right),$$

where

$$a = \sin\left(\frac{\theta_0}{2}\right),$$

implying

$$d\theta = 2a \frac{\sqrt{1 - \sin^2(\phi)}}{\sqrt{1 - a^2 \sin^2(\phi)}} d\phi.$$

With these substitutions, the integral on hand becomes

$$\int \frac{d\phi}{\sqrt{1 - a^2 \sin^2(\phi)}} = \sqrt{\frac{g}{L}} \int dt.$$

The left side is called an *elliptic integral*, and has no simple closed-form solution in general.

Despite the above elliptic integral, we can still use it to crank out an answer. Let $t = 0$ correspond to $\theta = 0$ and $\phi = 0$, which is the lowest position available to the pendulum. After one period of motion at $t = T$, i.e. once the angle θ has returned to zero again, and the value 2π has accumulated in ϕ . We then have a formula for the period of the motion:

$$\sqrt{\frac{g}{L}} \int_0^T dt = \int_0^{2\pi} \frac{d\phi}{\sqrt{1 - a^2 \sin^2(\phi)}}.$$

On the right, use the Taylor expansion of the radical to write

$$\begin{aligned} \frac{1}{\sqrt{1 - a^2 \sin^2(\phi)}} &\approx 1 + \frac{1}{2} a^2 \sin^2(\phi) \\ &+ \frac{3}{8} a^4 \sin^4(\phi) + \dots, \end{aligned}$$

$$T = 2\pi \sqrt{\frac{L}{g}} \sqrt{1 + \left(\frac{1}{2}\right)^2 a^2 + \left(\frac{1 \cdot 3}{2 \cdot 4}\right)^2 a^4 + \left(\frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}\right)^2 a^6 + \dots}$$

Shifted Natural Logarithm

Starting with Equation (17.21), namely

$$\ln(1+x) + C = \int \frac{dx}{1+x},$$

which only works when $a \sin(\phi)$ is a relatively ‘small’ number.

While we have paid with some accuracy and generality, the thing we gain is that the right side can be integrated. Going term by term it helps to know

$$\begin{aligned} \int_0^{2\pi} \sin^2(\phi) d\phi &= \pi \\ \int_0^{2\pi} \sin^4(\phi) d\phi &= \frac{3\pi}{4}, \end{aligned}$$

attainable by elementary means or using a trigonometric reduction formula.

The integral for the period reduces to

$$T \approx 2\pi \sqrt{\frac{L}{g}} \left(1 + \frac{a^2}{4} + \frac{9a^4}{64} + \dots\right).$$

If the initial angle θ_0 is much less than one, we further have

$$a^2 \approx \frac{\theta_0^2}{4},$$

or

$$T \approx 2\pi \sqrt{\frac{L}{g}} \left(1 + \frac{\theta_0^2}{16}\right).$$

From this we get the familiar period of the simple pendulum, along with a correction that accounts for more extreme initial conditions.

Problem 17

Re-work the physical pendulum calculation without oversimplifying the series coefficients to derive:

consider the scenario $|x| < 1$.

In this case, the fraction $1/(1+x)$ can be replaced via

the geometric series:

$$\ln(1+x) + C = \int (1 - x + x^2 - x^3 + \dots) dx$$

The whole right side can be integrated quite easily:

$$\ln(1+x) + C = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

The integration constant is zero by construction, and we end up with an infinite series for the shifted natural logarithm:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

This result is in fact the same thing we'd get by Taylor expanding $\ln(1+x)$ near $x=0$. Unlike the Taylor expansion however, we can now say for certain that the series approximation of $\ln(1+x)$ converges for $|x| < 1$. We can be a little naughty and try $x=1$ exactly to come up with an infinite approximation for $\ln(2)$:

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots \quad (17.57)$$

Arctangent

Recall that the derivative of the arctangent function

$$\frac{d}{dx} \arctan(x) = \frac{1}{1+x^2},$$

and consider the case $|x| < 1$. The right side expands as a geometric series:

$$\frac{d}{dx} \arctan(x) = 1 - x^2 + x^4 - x^6 + \dots$$

Apply the $\int dx$ operator to both sides and simplify, to get, for $|x| < 1$,

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

The integration constant is easily ruled to be zero and is omitted. Nor surprisingly, this is what emerges when Taylor expanding $\arctan(x)$ near $x=0$.

The infinite expression for the arctangent can be used to come up with an expression for $\pi/4$ by setting $x=1$,

$$\frac{\pi}{4} \approx 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots, \quad (17.58)$$

called the *Leibniz formula*.

It's important to note that Equations (17.57), (17.58) each send $x=1$ to the geometric series, which may seem illegal, as this is where the geometric series is supposed to lose jurisdiction. Technically, each result is attained by letting $x \rightarrow 1$ in a formal limit, and making sure divergence does not occur.

Sine of X Squared

Innocent as it appears, the indefinite integral

$$I = \int \sin(x^2) dx$$

has no elementary solution. To make headway, replace the sine function with its exact polynomial representation, namely

$$\sin(x^2) = x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \dots$$

Suddenly, we see a path forward. By trading any possibility of a closed solution, we can at least deal with the right side. Integrate each term and a strange answer emerges:

$$\int \sin(x^2) dx = \frac{x^3}{3} - \frac{x^7}{42} + \frac{x^{11}}{1320} - \dots \quad (17.59)$$

3.13 Stirling's Approximation

There is an important relationship governing very large whole numbers called *Stirling's approximation*, given by

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

While a full derivation is beyond the scope of this section, we can establish a slightly weaker version, namely

$$n! \approx \left(\frac{n}{e}\right)^n. \quad (17.60)$$

Derivation

To begin, write $n!$ in open form, namely

$$n! = n(n-1)(n-2)(n-3)\dots(2)(1),$$

then take the natural log of both sides to write

$$\ln(n!) = \ln(n) + \ln(n-1) + \ln(n-2) + \dots,$$

and condense the right using summation notation:

$$\ln(n!) = \sum_{j=1}^n \ln(j)$$

Now, this almost looks like a Riemann sum if it weren't for the conspicuous absence of a Δx -like term. However,

since the sum runs over whole numbers only, there is an effective $\Delta x_j = 1$ at play:

$$\ln(n!) = \sum_{j=1}^n \ln(j) \Delta x_j$$

Even though Δx_j cannot be pushed to zero, the above sum can be approximated as continuous *anyway*, but only for very large n . Working in this regime, we can replace the above with

$$\ln(n!) \approx \int_1^n \ln(x) dx,$$

solved by

$$\ln(n!) \approx (\ln(x) - x) \Big|_1^n,$$

having approximate solution

$$\ln(n!) \approx \ln(n) - n.$$

Apply the $\exp()$ operator to isolate the factorial term, and Equation (17.60) emerges.

Strange Product

Let us simplify the quantity

$$A = \lim_{n \rightarrow \infty} \left(\frac{(n+1)(n+2) \cdots (3n)}{n^{2n}} \right)^{1/n}$$

as far as possible.

One way to proceed is to take the natural log of both sides and simplify:

$$\ln(A) = \lim_{n \rightarrow \infty} \frac{\ln(n+1) + \cdots + \ln(3n) - 2n \ln(n)}{n}$$

There are $2n$ total positive terms in the sum above, so we can break apart the negative term into $2n$ parts and subtract $\ln(n)$ from each positive term to get:

$$\ln(A) = \lim_{n \rightarrow \infty} \frac{1}{n} \cdot \left(\ln\left(1 + \frac{1}{n}\right) + \ln\left(1 + \frac{2}{n}\right) + \cdots + \ln\left(1 + \frac{2n}{n}\right) \right)$$

Simplifying, this is

$$\ln(A) = \lim_{n \rightarrow \infty} \sum_{j=1}^{2n} \ln\left(1 + \frac{j}{n}\right) \frac{1}{n}.$$

Using the same trick that led to Stirling's approximation, argue that because the largeness of j will dominate anything

to do with small j , the sum can be considered continuous with

$$x_j = j/n \\ \Delta x = 1/n.$$

In this regime, we have, approximately:

$$\ln(A) \approx \int_0^2 \ln(1+x) dx,$$

equivalent to

$$\ln(A) \approx \int_1^3 \ln(u) du,$$

having solution

$$\ln(A) \approx (u \ln(u) - u) \Big|_1^3,$$

or

$$\ln(A) \approx 3 \ln(3) - 2,$$

and, finally,

$$A \approx \frac{3^3}{e^2}.$$

Let us now do the same calculation using Stirling's approximation. First notice A can be written

$$A = \lim_{n \rightarrow \infty} \left(\frac{1}{n^{2n}} \frac{(3n)!}{n!} \right)^{1/n},$$

and then Equation (17.60) tells us

$$A \approx \lim_{n \rightarrow \infty} \left(\frac{1}{n^{2n}} \frac{(3n)^{3n}}{e^{3n}} \frac{e^n}{n^n} \right)^{1/n},$$

reducing to $A \approx 3^3/e^2$, as expected. All n -dependence cancels out.

Strange Function

Consider the function

$$A(x) = \lim_{n \rightarrow \infty} \left(\frac{1}{n^{xn}} \frac{((x+1)n)!}{n!} \right)^{1/n},$$

where $x = 2$ reproduces the previous product.

Following similar steps, it's straightforward to show that

$$\ln(A(x)) = \lim_{n \rightarrow \infty} \sum_{j=1}^{xn} \ln\left(1 + \frac{j}{n}\right) \frac{1}{n},$$

or

$$\ln(A(x)) \approx \int_0^x \ln(1+t) dt,$$

but let's resist solving the integral.

Attack the problem a second way using Stirling's approximation to get

$$A(x) \approx \frac{(x+1)^{x+1}}{e^x},$$

or

$$\ln(A(x)) = (x+1) \ln(x+1) - x.$$

With two ways to express $\ln(A)$, eliminate it to conclude

$$\int \ln(1+x) dx = (x+1) \ln(x+1) - x + C,$$

which happens to be correct.

4 Integrals and Geometry

4.1 Arc Length

Integration is the tool for calculating the arc length of a differentiable curve $y = f(x)$. At a given point (x, y) on such a curve, there is a neighboring point $(x + dx, y + dy)$ connected by a straight line of length

$$dS = \sqrt{dx^2 + dy^2}.$$

The term dx can be pulled out of the radical to get

$$dS = dx \sqrt{1 + \left(\frac{dy}{dx}\right)^2},$$

and notice the ratio dy/dx is none other than the slope $f'(x)$ of the curve being measured.

The integral over dS is the total length of the curve between a set of endpoints x_0, x_1 :

$$S = \int dS = \int_{x_0}^{x_1} \sqrt{1 + (f'(x))^2} dx \quad (17.61)$$

Note that a similar formula can be derived by removing dy from the radical and ending up with an integral in the y -domain.

Problem 18

Show that the arc length of a symmetric parabolic segment of base $2a$ and height h is:

$$\begin{aligned} L &= \frac{a^2}{h} \int_0^{2h/a} \sqrt{1 + x^2} dx \\ &= \sqrt{a^2 + 4h^2} + \frac{a^2}{2h} \ln \left(\frac{2h + \sqrt{a^2 + 4h^2}}{a} \right) \end{aligned}$$

Hint: You may need the secant reduction formula.

Problem 19

Show that the arc length of an ellipse with eccentricity e is given by the *complete elliptic integral of the second kind*:

$$L = 4a \int_0^{\pi/2} \sqrt{1 - e^2 \sin^2(\theta)} d\theta$$

Problem 20

Show that the arc length of a hyperbola with eccentricity e is given by another elliptic integral:

$$L = 4a \int \sqrt{e^2 \cosh^2(\theta) - 1} d\theta$$

4.2 Volume of Revolution

A sneaky way to calculate certain three-dimensional volumes using one-dimensional integrals can be established. For this we require differentiable functions $y = f(x)$ that are greater than zero in the domain $x_0 \leq x \leq x_1$.

Circular Disk Method

A three-dimensional volume with axial symmetry can be produced by rotating the curve $y(x)$ about the x -axis. Each height y on the curve is swung around one full revolution to trace out a disk of area πy^2 , and the total volume enclosed is the sum across the grain of many infinitely-thin disks. As an integral, such a *volume of revolution* is given by:

$$V = \int_{x_0}^{x_1} \pi (f(x))^2 dx \quad (17.62)$$

Problem 21

Show that a cone of height H and base radius R has volume

$$V = \frac{1}{3} \pi R^2 H.$$

Problem 22

Use elementary methods to show that a *cone frustum* of height H with end radii R_1, R_2 has volume

$$V = \frac{1}{3} \pi (R_1^2 + R_1 R_2 + R_2^2) H.$$

Use the disk method with the line

$$y = \left(\frac{R_2 - R_1}{H} \right) x + R_1$$

to get the same answer.

Problem 23

A *paraboloid* is the volume formed by a parabola rotated about its axis of symmetry. Show that the volume of a paraboloid of height H and base radius R is given by

$$V = \frac{1}{2} \pi R^2 H.$$

Hint: Rotate the parabola $y = Hx^2/R^2$ about the y -axis and x becomes the disk radius.

Square Disk Method

Modifying the circular disk method, one can imagine summing across square disks instead. To illustrate, suppose a pyramid with square cross section has height h , length l , and width w .

We'll take the square cross section as parallel to the xy -plane, and we will integrate vertically along z . For a given height $z \leq h$, the dimensions of a 'square disk' are

$$\begin{aligned}x(z) &= z l / h \\y(z) &= z w / h.\end{aligned}$$

The total volume the pyramid is

$$V = \frac{lw}{h^2} \int_0^h z^2 dz = \frac{lwh}{3}.$$

Washer Method

Introducing a second function $g(x)$ that is less than $f(x)$ but greater than zero in the domain, we can calculate the volume of revolution trapped between the two curves. In this case, simply subtract the area of one disk from the other to form a 'washer'. The corresponding volume integral becomes:

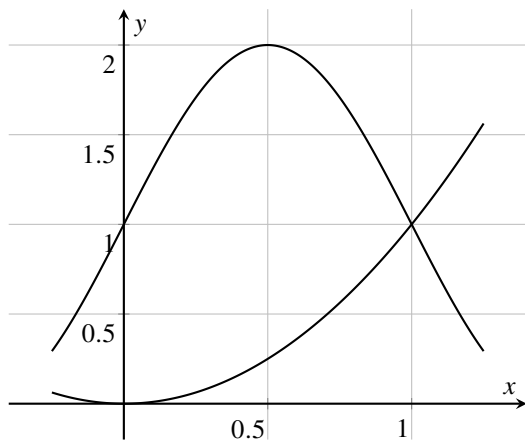
$$V = \int_{x_0}^{x_1} \pi \left((f(x))^2 - (g(x))^2 \right) dx \quad (17.63)$$

Problem 24

In the domain $0 \leq x \leq 1$, consider the two curves

$$\begin{aligned}y_1 &= 1 + \sin(\pi x) \\y_2 &= x^2\end{aligned}$$

as shown. Write an expression for the volume of revolution about the x -axis and also the y -axis.



Hint: For the x -axis rotation, you should find:

$$V_x = \pi \int_0^1 \left((1 + \sin(\pi x))^2 - x^4 \right) dx$$

Then, with

$$\begin{aligned}x_1 &= \sqrt{y} \\x_2 &= \frac{1}{\pi} \arcsin(y - 1),\end{aligned}$$

find

$$V_y = \pi \int_0^1 y dy + \pi \int_1^2 \left((1 - x_2)^2 - x_2^2 \right) dy.$$

Cylindrical Shell Method

A different volume of revolution is attained by rotating the function $y = f(x)$ about the y -axis. In this case, a three-dimensional volume is made of many concentric cylindrical shells.

For a point x in the domain, along with a neighboring point $x + dx$, rotating about the y -axis traces a pair of circles whose radii differ by dx . The height of each circle is $f(x)$, $f(x + dx)$ respectively. This defines a cylindrical 'shell' having volume

$$\begin{aligned}dV_{\text{shell}} &= \pi (x + dx)^2 f(x + dx) \\&\quad - \pi (x)^2 f(x),\end{aligned}$$

or, in the first-order limit,

$$dV_{\text{shell}} = 2\pi x f(x) dx.$$

In essence, we see that the volume of a thin cylindrical shell is the same as that of a rectangle of thickness dx , height $f(x)$, and width $2\pi x$. The total volume is the integral of thin shells:

$$V = \int dV_{\text{shell}} = \int_{x_0}^{x_1} 2\pi x f(x) dx \quad (17.64)$$

Problem 25

Show that the volume of the upper half of a sphere of radius R is given by

$$V = \int_0^R 2\pi x \sqrt{R^2 - x^2} dx = \frac{2}{3} \pi R^3.$$

Problem 26

Use the offset circle

$$(x - R)^2 + y^2 = a^2$$

to find the volume of a torus:

$$V = 2 \int_{R-a}^{R+a} 2\pi x \sqrt{a^2 - (x - R)^2} dx = 2\pi^2 R a^2$$

4.3 Surface of Revolution

A technique similar to the volume of revolution can tell us the *surface area of revolution* of a solid generated by a function $y = f(x)$.

For a point x in the domain, along with a neighboring point $x + dx$, rotating about the x -axis traces a pair of circles parallel to the yz plane. The circumference of each circle is $2\pi y$, $2\pi(y + dy)$, respectively. We can take each circumference as the edges of a skinny trapezoid whose width is the arc length

$$dw = \sqrt{dx^2 + dy^2},$$

and the area of such a trapezoid is

$$dA = \pi(2y + dy)dw.$$

In the first-order limit, we can write the differential area:

$$dA = 2\pi y \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx$$

Summing across the grain of many thin strips will cover the surface and reveal the total area of revolution for $y = f(x)$:

$$A = \int_{x_0}^{x_1} 2\pi y \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx \quad (17.65)$$

Gabriel's Horn

Consider the hyperbola

$$y = \frac{1}{x}$$

in the domain

$$1 \leq x < \infty.$$

The volume of revolution of this particular shape is called *Gabriel's horn*, and contains an interesting 'paradox'. Computing the volume of Gabriel's horn is straightforward:

$$V = \int_1^\infty \pi \left(\frac{1}{x}\right)^2 dx = 1$$

Watch what happens if we try to compute the surface area:

$$A = \int_1^\infty 2\pi \left(\frac{1}{x}\right) \sqrt{1 + \frac{1}{x^4}} dx$$

The square root term makes the integral rather ugly, but notice how its presence always scales the integrand higher. This means we can also write

$$A > \int_1^\infty 2\pi \left(\frac{1}{x}\right) dx,$$

which means

$$A > 2\pi (\ln(\infty) - \ln(1)).$$

What? The area is somehow infinite - the math was done correctly. But this shouldn't be, because the volume is a finite number. Some argued that filling the horn with a finite volume of paint is equivalent to painting the inside, which ought to make the area finite. Others pointed out that an infinite horn cannot be physically constructed, and that paint flows at a finite speed and would take forever to flow into the horn.

This 'paradox', originally brought to public attention by Torricelli, was known to seventeenth-century mathematicians, not excluding Hobbes, Wallis, and Galileo.

There really is no paradox on hand, and paint is a bad analogy. Keep in mind that paint is a three-dimensional fluid. Filling Gabriel's horn with fluid returns to the original problem - what's the surface area of the paint (excluding the end disc)?

Another way to illustrate the point is to compare the rates of change of the volume and surface with respect to x . Using

$$\begin{aligned} \frac{dV}{dx} &= \pi \left(\frac{1}{x}\right)^2 \\ \frac{dA}{dx} &= 2\pi \left(\frac{1}{x}\right) \sqrt{1 + \frac{1}{x^4}}, \end{aligned}$$

define the rate

$$R = \frac{dV/dx}{dA/dx},$$

simplifying to

$$R = \frac{1}{2x\sqrt{1 + 1/x^4}}.$$

This rate vanishes in the limit $x \rightarrow \infty$, which means the area outpaces the volume in the long run.

4.4 Centroid

...

Problem 27

Show that the centroid of a parabolic segment of height h is $\bar{y} = 2h/5$.

Problem 28

Show that the centroid of a half-ellipse of base $2a$ and vertex height b is $\bar{y} = 4b/3\pi$.

5 Series Analysis

Integration is a powerful addition to the toolkit for analyzing infinite sums, particularly on the issues of convergence and divergence.

5.1 Taylor Series

The most versatile series is surely the Taylor series, which tells that a function $f(x)$ at a point x_0 is approximated by a polynomial $p(x)$ involving derivatives $f^{(q)}(x_0)$:

$$p(x) = f(x_0) + \sum_{q=1}^n \frac{1}{q!} f^{(q)}(x_0) (x - x_0)^q + R_n(x)$$

For large n approaching infinity, the remainder term $R_n(x)$ vanishes if the series is to converge.

Derivation

To derive Taylor's theorem, begin with the fundamental theorem of calculus, i.e. Equation (17.2), and isolate $f'(x)$:

$$f(x) = f(x_0) + \int_{x_0}^x f'(t) dt$$

Of course, the function $f'(t)$ could itself be approximated to first order using the fundamental theorem

$$f'(t) = f'(x_0) + \int_{x_0}^t f''(u) du,$$

which begs substitution into the above, giving:

$$f(x) = f(x_0) + \int_{x_0}^x \left(f'(x_0) + \int_{x_0}^t f''(u) du \right) dt$$

After simplifying, we see the familiar first-order Taylor series term trailed by a messy integral:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \int_{x_0}^x \left(\int_{x_0}^t f''(u) du \right) dt$$

Trudging forward, take $f''(u)$ to first order

$$f''(u) = f''(x_0) + \int_{x_0}^u f'''(w) dw,$$

and substitute into the preceding integral. This first means having to solve

$$I = \int_{x_0}^x \left(\int_{x_0}^t f''(x_0) du \right) dt.$$

Knowing $f''(x_0)$ is constant, proceed using brute force to find

$$\begin{aligned} I &= f''(x_0) \int_{x_0}^x \left(\int_{x_0}^t du \right) dt \\ &= f''(x_0) \int_{x_0}^x (t - x_0) dt \\ &= f''(x_0) \left(\frac{t^2}{2} - x_0 t \right) \Big|_{x_0}^x \\ &= \frac{1}{2} f''(x_0) (x - x_0)^2. \end{aligned}$$

Interestingly, this is the second-order term in the Taylor series of $f(x)$. To summarize:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) \\ &\quad + \frac{1}{2} f''(x_0)(x - x_0)^2 \\ &\quad + \int_{x_0}^x \left(\int_{x_0}^t \left(\int_{x_0}^u f'''(w) dw \right) du \right) dt \end{aligned}$$

Repeating the steps that got us this far, use the first-order approximation of $f'''(w)$. The obligatory integral to solve is

$$J = f'''(x_0) \int_{x_0}^x \left(\int_{x_0}^t \left(\int_{x_0}^u dw \right) du \right) dt,$$

which after a bit of grinding, comes out to

$$J = \frac{1}{3!} f'''(x_0) (x - x_0)^3.$$

By now we're seeing a pattern, particularly:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) \\ &\quad + \frac{1}{2} f''(x_0)(x - x_0)^2 \\ &\quad + \frac{1}{3!} f'''(x_0)(x - x_0)^3 \\ &\quad + R_3(x), \end{aligned}$$

where $R_3(x)$ is given by

$$\int_{x_0}^x \left(\int_{x_0}^t \left(\int_{x_0}^u \left(\int_{x_0}^v f^{(4)}(v) dv \right) dw \right) du \right) dt.$$

Remainder

In the general case, the remainder term $R_n(x)$ always contains a polynomial term plus an integral. Since the integral part ends up being higher order than n , we can always push the hard work to the next step, so to speak, and take as the remainder term:

$$R_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(x_0) (x - x_0)^{n+1}$$

6 Applied Integration

6.1 Kinematics

The equations of constant-acceleration kinematics in one dimension fall out trivially from integration. Starting with the definition

$$\frac{d}{dt}v(t) = a = \text{constant},$$

integrate with respect to time to find

$$v(t) = v_0 + at,$$

where v_0 is the integration constant. Integrate with respect time once more to write

$$\begin{aligned} \int v(t) dt &= \int (v_0 + at) dt \\ x(t) &= x_0 + v_0 t + \frac{1}{2}at^2. \end{aligned}$$

The integration constant has been labeled x_0 .

Similarly, one can develop the regime of constant jerk by starting with

$$\frac{d}{dt}a(t) = j = \text{constant}$$

such that

$$a(t) = a_0 + jt.$$

Turning the same crank, one eventually finds

$$x(t) = x_0 + v_0 t + \frac{1}{2}a_0 t^2 + \frac{1}{6}j t^3.$$

6.2 The Cycloid

Let a ‘generating’ circle of radius R roll on the x axis. As the circle moves, the point on the rim originally at $(0, 0)$ traces the shape of a *cycloid* as shown in Figure 17.1.

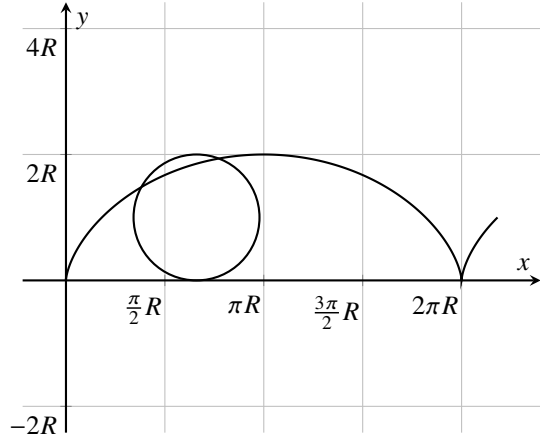


Figure 17.1. The cycloid with generating circle.

Parameterization

There is no simple expression $y(x)$ for the cycloid. Instead we introduce a parameter θ that tracks the evolution of the generating circle. In terms of θ , the shape of the cycloid is given by

$$x(\theta) = R\theta - R \sin(\theta) \quad (17.66)$$

$$y(\theta) = R - R \cos(\theta). \quad (17.67)$$

The cycloid is clearly periodic in the variable θ . While θ can take on any real value and still represent a cycloid, we'll stay interested in the domain $[0 : 2\pi]$.

Velocity Envelope

Supposing θ evolves in a smooth and differentiable manner, we can take derivatives with respect to θ . For brevity, define

$$\omega(t) = \frac{d}{d\theta}\theta(t),$$

and calculate the time derivative of $x(\theta)$, $y(\theta)$ to get:

$$\frac{dx}{dt} = R\omega - R\omega \cos(\theta)$$

$$\frac{dy}{dt} = R\omega \sin(\theta)$$

If we isolate the trig terms and square each equation, the fundamental trig identity can be used to derive

$$\left(\frac{dx}{dt} - R\omega\right)^2 + \left(\frac{dy}{dt}\right)^2 = (R\omega)^2,$$

which is called the *envelope* of velocities of the cycloid. Plotted in velocity space, the above depicts a circle of radius $R\omega$ centered at $(R\omega, 0)$.

Tangent Line

At a point (x_0, y_0) on the cycloid, the slope is still given by dy/dx at that point, despite using a parameterized representation of the curve. Calculating the slope is a matter of the chain rule:

$$\frac{dy}{dx} = \frac{dy}{d\theta} \frac{d\theta}{dx} = \frac{dy}{d\theta} \left(\frac{dx}{d\theta} \right)^{-1}$$

Carrying this out, we find

$$\frac{dy}{dx} = \frac{\sin(\theta)}{1 - \cos(\theta)} = \cot\left(\frac{\theta}{2}\right).$$

This is enough to write down an equation for the tangent line to the cycloid:

$$y_{\text{tan}} = y_0 + \cot\left(\frac{\theta}{2}\right)(x - x_0)$$

Replacing x_0, y_0 with their representations in θ gives a neater formula, after some simplifying:

$$y_{\text{tan}} = 2R + \cot\left(\frac{\theta}{2}\right)(x - R\theta)$$

Interestingly, we see that the tangent line always passes through the point $(R\theta, 2R)$, which is the top of the generating circle as it goes along.

Problem 29

A stone lodged on the rim of a bicycle tire of radius R dislodges at the height of its cycloidal path. Determine its trajectory after leaving the tire. Answer:

$$x(t) = R\pi + 2R\omega t$$

$$y(t) = 2R - gt^2/2$$

Normal Line

Knowing the slope at any point (x_0, y_0) on the cycloid, we can write an expression for the normal line at the same point:

$$y_{\text{norm}} = y_0 - \tan\left(\frac{\theta}{2}\right)(x - x_0)$$

Like the case for the tangent line, replacing x_0, y_0 with their representations in θ gives a neater formula, after some simplifying:

$$y_{\text{norm}} = -\tan\left(\frac{\theta}{2}\right)(x - R\theta)$$

From this, we see that the normal line always hits the point of contact between the generating circle and the line on which it rolls.

Arc Length

The arc length of the cycloid is straightforwardly calculated from Equation (17.61). For this, we start with

$$S = \int_0^{2\pi R} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx,$$

which, after substituting $x(\theta), y(\theta)$ becomes

$$S = \sqrt{2}R \int_0^{2\pi} \sqrt{1 - \cos(\theta)} d\theta,$$

readily simplifying to

$$S = 2R \int_0^{2\pi} \sin\left(\frac{\theta}{2}\right) d\theta = 8R.$$

Area Enclosed

The area enclosed by the cycloid and the x -axis is given by the standard setup:

$$A = \int_0^{2\pi R} y dx = R^2 \int_0^{2\pi} (1 - \cos(\theta))^2 d\theta$$

The remaining integral is straightforwardly solved, and we find the enclosed area to be three times that of the generating circle:

$$A = 3\pi R^2$$

Volume Enclosed

A cycloid revolved about the x -axis encloses a volume that we can calculate with the circular disk method, i.e. Equation (17.62). For this case, we have, after simplifying

$$V = \int_0^{2\pi} \pi R^2 (1 - \cos(\theta))^3 d\theta.$$

The remaining integral is a bit tedious but isn't difficult, ending with

$$V = 5\pi^2 R^3.$$

Surface Area

The surface of revolution made by revolving a cycloid about the x -axis is straightforwardly given by Equation (17.65). Here, we have

$$A = 2\sqrt{2}\pi R^2 \int_0^{2\pi} (1 - \cos(\theta))^{3/2} d\theta.$$

The remaining integral is tricky to evaluate but not impossible. Leaving the details for an exercise, we ultimately find

$$A = \left(2\sqrt{2}\pi R^2\right) \left(\frac{16\sqrt{2}}{3}\right) = \frac{64}{3}\pi R^2.$$

Tautochrone

Consider a cycloid flipped upside-down, described by

$$x(\theta) = R\theta - R\sin(\theta)$$

$$y(\theta) = -R + R\cos(\theta).$$

Pretending we have constructed a ramp in such a shape, let us analyze the sliding (not rolling) motion of a body of mass m placed at rest on the ramp.

In uniform gravity, the system respects an energy constant

$$E = \frac{1}{2}mv^2 + mgy,$$

where v is the velocity of the body in motion, g is the local gravitational acceleration, and y is the height above $y = 0$. Assuming the object begins at rest, we also have

$$E = mgy_0,$$

where y_0 is the initial height of the body.

With this setup, it's useful to know the total time T required for the body to slide to the bottom of the inverted cycloid. As an integral, we have, at least provisionally,

$$T = \int_{y_0}^{-2R} dt,$$

and the job is recast the integral in variables we know.

Proceed by replacing dt with something akin to arc length, namely

$$dS = v(t) dt.$$

Meanwhile, we know from geometry that

$$dS^2 = dx^2 + dy^2.$$

This is enough to wrestle the time integral into something manageable:

$$T = \int_{x_0}^{R\pi} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \frac{dx}{\sqrt{2g(y_0 - y)}}$$

We haven't used the equations of the cycloid yet, so proceed by using

$$y_0 - y = R(\cos(\theta_0) - \cos(\theta))$$

$$dx = R(1 - \cos(\theta)) d\theta$$

$$dy = -R\sin(\theta) d\theta,$$

and the above simplifies to

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sqrt{1 - \cos(\theta)} d\theta}{\sqrt{\cos(\theta_0) - \cos(\theta)}}.$$

Note that θ_0 corresponds to the initial position (x_0, y_0) , and $\theta = \pi$ occurs when the sliding body reaches the bottom of the curve.

Ugly as it is, the time integral can be solved after making a few substitutions that are left for an exercise to the reader to find. As a hint, you should first have

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sin(\theta/2) d\theta}{\sqrt{\cos^2(\theta_0/2) - \cos^2(\theta/2)}},$$

and then

$$T = \sqrt{\frac{R}{g}} \int_1^0 \frac{-2du}{\sqrt{1 - u^2}}.$$

Keep on solving with yet another u -substitution, and the final answer comes out to

$$T = \pi\sqrt{\frac{R}{g}}.$$

Remarkably, the final answer $T = \pi\sqrt{R/g}$ makes no mention of the initial position (x_0, y_0) of the sliding body. This is to say that the time to slide to the bottom of a cycloid is always the same. No other known curve has this feature. The Ancient Greeks called this the *tautochrone*.

6.3 Horizontal Tank

Consider a cylindrical tank of volume $V = \pi R^2 L$, where R is the radius and L is the length. The tank is resting on its side and is filled partially with water to a depth h above the lowest point in the tank. For a given ratio h/R , what fraction of the tank's volume is filled with water?

For convenience, redefine 'down' as pointing along the positive x -axis so the empty tank fills horizontally from the right. Placing the origin at the center of the circular cross section, the line defining the water's surface has length

$$2y = 2R \sin(\theta) ,$$

where θ is placed conventionally. In this frame, we can forget about the length of the tank entirely.

Letting x represent the distance from the center of the tank to the water's surface, presuming $h < R$, then the x -coordinate obeys

$$x = R \cos(\theta) ,$$

along with

$$x + h = R .$$

We'll need a formula for h/R , which we take from the above:

$$\frac{h}{R} = 1 - \frac{x}{R} = 1 - \cos(\theta)$$

Per our setup, the ratio h/R picks out a unique value for θ_* , so long as $0 \leq \theta_* \leq \pi/2$.

Setting up the area integral for the cross section of water occupying the tank, we begin with

$$A = -2 \int_R^{x_*} y \, dx ,$$

which after using the trigonometric substitutions above becomes

$$A = -2R^2 \int_0^{\theta_*} \sin^2(\theta) \, d\theta ,$$

simplifying further to

$$A = R^2 \left(\theta_* - \frac{1}{2} \sin(2\theta_*) \right) .$$

Keep in mind that θ_* is nontrivially related to the ratio h/R . Using a machine equipped to run Newton's method in one dimension, we generate the following results:

Ratio $h/2R$	Angle θ_*	Vol. Rat.
0	0	0
0.05	0.45103	0.018693
0.1	0.64350	0.052044
0.15	0.79540	0.094060
0.2	0.92730	0.14238
0.25	1.0472	0.19550
0.3	1.1593	0.25232
0.35	1.2661	0.31192
0.4	1.3694	0.37353
0.45	1.4706	0.43644
0.5	$\pi/2$	0.5

The handle the case $\pi/2 \leq \theta_* \leq \pi$, write

$$\frac{h}{R} = 1 + \cos(\theta)$$

and regenerate the list of θ_* -values to find:

Ratio $h/2R$	Angle θ_*	Vol. Rat.
0.55	1.6710	0.56356
0.6	1.7722	0.62647
0.65	1.8755	0.68808
0.7	1.9823	0.74768
0.75	2.0944	0.80450
0.8	2.2144	0.85762
0.85	2.3462	0.90594
0.9	2.4981	0.94796
0.95	2.6906	0.98131
1	π	1

6.4 Mass Between Springs

Consider a point mass m in the center of two springs pulled tight and mounted distance L apart, ignoring gravity. The left spring has constant k_a , and the right has spring constant k_b , and both springs have rest length $L_0 < L/2$.

Rest Condition

When the system is not in motion, the mass will rest somewhere between the endpoints toward the stiffer spring, not necessarily at $x = L/2$. To work this out, balance all relevant forces in the x - and y -directions:

$$m \frac{d^2 y}{dt^2} = F_{\text{net}}^y = 0$$

$$m \frac{d^2 x}{dt^2} = F_{\text{net}}^x = F_a + F_b ,$$

and each left size is zero for the rest condition. Each force F_a, F_b obeys Hooke's law:

$$F_{\text{spring}} = -kx$$

Letting a constant q denote the position of the mass away from $x = L/2$, the above tells us:

$$0 = -k_a \left(-L_0 + \frac{L}{2} - q \right) + k_b \left(-L_0 + \frac{L}{2} + q \right)$$

Solving for q tells us where the system rests:

$$q = \left(\frac{k_a - k_b}{k_a + k_b} \right) \left(\frac{L}{2} - L_0 \right)$$

Looking at a few special cases, note first that q vanishes if $k_a = k_b$, giving the symmetric result. Note also that if $L/2 = L_0$, the system is under no tension at all, and q vanishes again. More curiously, if it happens that $L/2 < L_0$, this corresponds to the system being compressed rather than stretched, and the sign on q flips. That is, the offset would be away from the stiffer spring. (This situation is unstable.)

Longitudinal Vibrations

If the mass-between-springs system is perturbed in a direction that is purely longitudinal, i.e. parallel to the springs, then resulting motion is confined to one dimension. To prepare for this, define two constants

$$\begin{aligned} x_a &= -L_0 + L/2 - q \\ x_b &= -L_0 + L/2 + q, \end{aligned}$$

so the rest condition is written

$$0 = -k_a x_a + k_b x_b.$$

For the non-rest case, use Newton's second law and Hooke's law combine to write

$$m \frac{d^2}{dt^2} x(t) = -k_a (x_a + x(t)) + k_b (x_b - x(t)),$$

readily simplifying to

$$m \frac{d^2}{dt^2} x(t) = -x(t) (k_a + k_b)$$

This is a simple harmonic oscillator with effective angular frequency:

$$\omega = \sqrt{\frac{k_a + k_b}{m}}$$

Transverse Vibrations

Things get more interesting when we examine vibrations in the direction perpendicular to the springs. Taking the

two spring constants as the same, i.e. $k_a = k_b = k$, an initial displacement of the mass in the y -direction results in one-dimensional motion.

In this case, we have $F_{\text{net}}^x = 0$ for the x -direction, and for the y -direction,

$$F_{\text{net}}^y = 2F_{\text{spring}} \sin(\theta),$$

where θ is the angle formed between a spring and the horizontal, and from geometry we pick out

$$\sin(\theta) = \frac{y}{\sqrt{(L/2)^2 + y^2}}.$$

The magnitude of the spring force is given by

$$F_{\text{spring}} = -k \left(\sqrt{\left(\frac{L}{2} \right)^2 + y^2} - L_0 \right),$$

which, as long as $L/2 \neq L_0$, has a nonzero value for $y = 0$, affirming the springs are always under tension. All together, transverse vibrations are summarized by

$$F_{\text{net}}^y = m \frac{d^2}{dt^2} y(t) = -2ky \left(1 - \frac{L_0}{\sqrt{(L/2)^2 + y^2}} \right).$$

Small Vibrations

In the special case that the displacement $|y|$ is always much less than $L/2$, the above becomes

$$\begin{aligned} F_{\text{net}}^y &\approx -2ky \left(1 - \frac{2L_0}{L} \left(1 - \frac{1}{2} \frac{4y^2}{L^2} \right) \right) \\ &\approx -2ky \left(1 - \frac{2L_0}{L} \right), \end{aligned}$$

where the square root has been eliminated by Taylor expansion.

Defining a new quantity

$$p = \frac{L}{2} - L_0,$$

the above simplifies to, of course, the equation of a harmonic oscillator

$$m \frac{d^2}{dt^2} y(t) \approx - \left(\frac{2k}{1 + L_0/p} \right) y(t).$$

The angular frequency is given by

$$\omega = \sqrt{\frac{2k}{m} \left(\frac{1}{1 + L_0/p} \right)},$$

which is scaled by the tension in the springs. This is in fact a crude model for a plucked guitar string - the greater the tension, the greater the frequency of vibration.

Critical Vibrations

The problem becomes a different beast when we consider $L_0 = L/2$, meaning there is no resting tension in the system. Staying in the regime of transverse small oscillations, i.e. $|y| \ll L/2$, let us jot down a previous result without canceling the y^2 -term:

$$F_{\text{net}}^y \approx -2ky \left(1 - \frac{2l_0}{L} \left(1 - \frac{1}{2} \frac{4y^2}{L^2} \right) \right)$$

Setting $2L_0 = L$, the above simplifies to

$$m \frac{d^2}{dt^2} y(t) \approx -k \left(\frac{2}{L} \right)^2 (y(t))^3,$$

which is classified as a nonlinear second-order differential equation.

Energy Constraint

Despite the scary name, we can wrestle with the above equation anyway. Letting

$$\lambda = \frac{4k}{mL^2}$$

and using the ‘dot’ operator as a shorthand for the time derivative, we must solve

$$\ddot{y} = -\lambda y^3.$$

Proceed by multiplying both sides by $\dot{y}$, and condense the left using the product rule:

$$\frac{1}{2} \frac{d}{dt} (\dot{y}^2) = \ddot{y} \dot{y} = -\lambda \frac{dy}{dt} y^3$$

Multiply dt onto each side to attain a so-called ‘differential form’

$$\frac{1}{2} \frac{d}{dt} (\dot{y}^2) dt = -\lambda y^3 dy,$$

which can be cleanly integrated with respect to t on the left, y on the right:

$$\frac{1}{2} \dot{y}^2 = -\frac{\lambda}{4} y^4 + C$$

This result looks very much like a conservation of energy statement. If we multiply through by a mass constant m , the left side is the kinetic energy then Cm is the total energy E . The potential energy term is proportional to y^4 , not y^2 , which is not a simple harmonic oscillator potential.

Initial Condition

One typical scenario for this system would have the mass released from rest at some initial value A above $y = 0$. In this case, the above equation reads

$$0 = -\frac{\lambda}{4} A^4 + C$$

at $t = 0$, and the integration constant C can be eliminated. Doing so, we get

$$\frac{1}{2} \dot{y}^2 = \frac{\lambda}{4} (A^4 - y^4) = \frac{\lambda A^4}{4} \left(1 - \left(\frac{y}{A} \right)^4 \right),$$

or

$$\frac{dy}{dt} = \sqrt{\dot{y}^2} = \pm \sqrt{\lambda} \frac{A^2}{2} \sqrt{1 - \left(\frac{y}{A} \right)^4},$$

which can be separated with all y ’s on one side, t ’s on the other:

$$\frac{dy}{\sqrt{1 - (y/A)^4}} = \pm \left(\sqrt{\lambda} \frac{A^2}{2} \right) dt$$

Proceed with the substitution

$$y = A \cos(\phi)$$

$$dy = -A \sin(\phi) d\phi,$$

and the above becomes

$$\frac{-A \sin(\phi) d\phi}{\sqrt{(1 - \cos^2(\phi)) (1 + \cos^2(\phi))}} = \pm \left(\sqrt{\lambda} \frac{A^2}{2} \right) dt,$$

allowing each side to be integrated:

$$\int \frac{d\phi}{\sqrt{1 + \cos^2(\phi)}} = \mp \left(\sqrt{\lambda} \frac{A}{2} \right) \int dt$$

There are many choices for integration limits. One simple correspondence emerges by letting ϕ run from 0 to $\pi/2$, in which case the t -variable elapses a quarter-period:

$$\begin{aligned} \int_0^{\pi/4} \frac{d\phi}{\sqrt{1 + \cos^2(\phi)}} &= \mp \left(\sqrt{\lambda} \frac{A}{2} \right) \int_0^{T/4} dt \\ &= \mp \left(\sqrt{\lambda} \frac{A}{2} \right) \frac{1}{4} T \end{aligned}$$

The integral on the left is strictly numerical. Condensing constants and sifting out the period's relation to the amplitude gives the result:

$$AT = \text{constant}$$

Chapter 18

Analytic Geometry

1 Parametric Equations

Throughout the study of algebra, trigonometry, and calculus, one inevitably runs into *parametric equations*. This is the generalization of the typical $y = f(x)$ construction, where instead of using one function, the same information is contained in two equations

$$\begin{aligned}x &= g(t) \\ y &= h(t) ,\end{aligned}$$

where t is the parameter. Parametric equations are helpful because the form $y = f(x)$ may not be straightforwardly attained.

1.1 Parametric Systems

Kinematics

Let v_x, v_y be the components of the initial velocity of a moving body in the presence of uniform gravity without resistance. Starting from the position (x_0, y_0) and evolving with the time parameter t , the kinematic equations of motion read:

$$\begin{aligned}x(t) &= x_0 + v_x t \\ y(t) &= y_0 + v_y t - \frac{1}{2} g t^2\end{aligned}$$

Ellipse

Consider the ellipse centered at the origin:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

Defining a parameter ϕ in the domain $[0 : 2\pi]$, the same ellipse has the following parametric representation:

$$\begin{aligned}x &= a \cos(\phi) \\ y &= b \sin(\phi)\end{aligned}$$

Hyperbola

Consider the hyperbola centered at the origin:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

Defining a parameter ϕ in the domain $[0 : 2\pi]$, the same hyperbola has the following parametric representation:

$$\begin{aligned}x &= a \cosh(\phi) \\ y &= b \sinh(\phi)\end{aligned}$$

Cycloid

The shape that solves the tautochrone problem, i.e. the cycloid, can *only* be represented by a pair of parametric equations. For a generating circle of radius R , we have

$$\begin{aligned}x &= R\phi - R \sin(\phi) \\ y &= R - R \cos(\phi) ,\end{aligned}$$

where the parameter ϕ can be any real number. For one period of the cycloid, confine $0 \leq \phi \leq 2\pi$.

Involute

Unwind a string from a circle of radius R while maintaining tension. The endpoint of the string traces a shape called

an involute. In terms of an ‘unwinding parameter’ ϕ , the involute is given by

$$\begin{aligned}x(\phi) &= R \cos(\phi) + R\phi \sin(\phi) \\y(\phi) &= R \sin(\phi) - R\phi \cos(\phi) .\end{aligned}$$

Lissajous Figures

...

Scalar Multiplication

For a vector $\mathbf{m}$ that represents the slope of a straight line, along with a vector $\mathbf{b} = \langle 0, b \rangle$ as the y -intercept, the whole line can be represented by the vector

$$\mathbf{r}(\alpha) = \alpha \mathbf{m} + \mathbf{b} ,$$

where the parameter α is any real number.

1.2 Polar Coordinate System

The polar coordinate system is a two-parameter apparatus. A point (x, y) in the plane requires two pieces information to specify:

$$\begin{aligned}x &= x(r, \theta) = r \cos(\theta) \\y &= y(r, \theta) = r \sin(\theta)\end{aligned}$$

It’s also possible to frame r and θ in terms of each other, which is to say $r(\theta)$ and $\theta(r)$ are legal residents.

Of course, each can also be framed in terms of a general parameter t to make $r = r(t)$, $\theta = \theta(t)$. In this case, a point in the plane is represented by:

$$\begin{aligned}x &= x(t) = r(t) \cos(\theta(t)) \\y &= y(t) = r(t) \sin(\theta(t))\end{aligned}$$

2 Parametric Derivatives

In a typical parametric system, we have separate entities $x(t)$ and $y(t)$ representing a single curve. Despite not having a $y = f(x)$ representation, we're still allowed to ask about the slope dy/dx of such a curve in the Cartesian plane. To this end, we simply employ the chain rule. For a parametric system, we have

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt}.$$

As per usual with the chain rule, you can see the dt -factor canceling on the right.

Of course, the above presumes that dx/dt is not zero at the point(s) of interest. The above can be easily inverted to get

$$\frac{dx}{dy} = \frac{dx/dt}{dy/dt}.$$

2.1 Slope in Polar Coordinates

Consider the polar coordinate system where r is known to be a function of θ :

$$x = r(\theta) \cos(\theta)$$

$$y = r(\theta) \sin(\theta)$$

With respect to the parameter θ , we find:

$$dx/d\theta = r'(\theta) \cos(\theta) - r(\theta) \sin(\theta)$$

$$dy/d\theta = r'(\theta) \sin(\theta) + r(\theta) \cos(\theta)$$

Then, by the chain rule, we have:

$$\frac{dy}{dx} = \frac{dy/d\theta}{dx/d\theta} = \frac{r'(\theta) \sin(\theta) + r(\theta) \cos(\theta)}{r'(\theta) \cos(\theta) - r(\theta) \sin(\theta)}$$

Divide by the cosine term to get a slightly neater formula:

$$\frac{dy}{dx} = \frac{r'(\theta) \tan(\theta) + r(\theta)}{r'(\theta) - r(\theta) \tan(\theta)}$$

Of course, the slope dy/dx is the instantaneous rise over run, which means the ratio of these lengths is the tangent of another angle ϕ measured from the positive x -axis:

$$\frac{dy}{dx} = \tan(\phi)$$

Psi Parameter

Staying in this picture, consider yet another angle ψ (Greek 'psi') measured from the position vector (rather than the

x -axis) that terminates at the tangent line, which brings forth the identity

$$\theta + \psi - \phi = 0.$$

Using the trig identity

$$\tan(\psi) = \frac{\tan(\phi) - \tan(\theta)}{1 + \tan(\phi) \tan(\theta)},$$

we find

$$\tan(\psi) = \frac{(dy/dx) - \tan(\theta)}{1 + (dy/dx) \tan(\theta)},$$

where dy/dx can be replaced by the slope formula. Simplifying like crazy, the above boils down to:

$$\tan(\psi) = \frac{r(\theta)}{r'(\theta)}$$

We can keep going with this. Take the reciprocal of the the equation and then notice the right side looks like a derivative

$$\cot(\psi) = \frac{r'(\theta)}{r(\theta)} = \frac{d}{d\theta} (\ln(r(\theta))),$$

and then this can be integrated to simplify the right side:

$$\int \cot(\phi - \theta) d\theta = \ln(r(\theta))$$

Logarithmic Spiral

While the above is a general statement, it's a bit unworkable in the general case. However, consider a regime where

$$\cot(\phi - \theta) = k,$$

where k is constant. Then, the integral immediately becomes the equation of a *logarithmic spiral*:

$$r(\theta) = r_0 e^{k\theta}$$

2.2 Parametric Second Derivative

The second derivative d^2y/dx^2 is straightforwardly stated:

$$\frac{d^2y}{dx^2} = \frac{d}{dx} \left(\frac{dy}{dx} \right)$$

Then, using the chain rule, we find:

$$\frac{d^2y}{dx^2} = \left(\frac{1}{dx/dt} \right) \frac{d}{dt} \left(\frac{dy}{dx} \right)$$

3 Parametric Integrals

In standard integral calculus, we know the area A under the curve $y(x)$ between the limits x_0, x_1 is given by

$$A = \int_{x_0}^{x_1} y(x) dx .$$

In the parametric regime, begin instead with

$$\begin{aligned} x &= g(t) \\ y &= h(t) , \end{aligned}$$

where $y(x)$ may be difficult or perhaps impossible to attain. Proceed by noting

$$dx = \frac{dg}{dt} dt = g'(t) dt ,$$

and also

$$\begin{aligned} x_0 &= g(t_0) \\ x_1 &= g(t_1) , \end{aligned}$$

i.e., the integration limits are indicated by t_0, t_1 . In this set of variables, the area integral takes the form:

$$A = \int_{t_0}^{t_1} h(t) g'(t) dt$$

3.1 Parametric Arc Length

The typical starting point for the arc length calculation is given by the indefinite integral

$$S = \int dS = \int \sqrt{dx^2 + dy^2} ,$$

where the ‘standard’ move is to then factor out dx from the radical.

To handle the parametric regime, multiply dt/dt (a factor of one) into the integrand and simplify:

$$S = \int \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt$$

In terms of the parameterization $x = g(t)$, $y = h(t)$, this is

$$S = \int \sqrt{(g'(t))^2 + (h'(t))^2} dt .$$

Integral of Speed

Of course, the above integrand can be condensed back down via

$$\frac{dS}{dt} = \sqrt{(g'(t))^2 + (h'(t))^2} ,$$

in which case dS/dt is interpreted as the speed:

$$S = \int \frac{dS}{dt} dt = \int v dt$$

3.2 Parametric Surface of Revolution

The surface area of a curve $y(x)$ rotated about the x axis is straightforwardly given by

$$A = \int_{x_0}^{x_1} 2\pi y \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx .$$

This can be reverse-engineered by multiplying dt/dt into the integrand and simplifying:

$$A = \int_{t_0}^{t_1} 2\pi h(t) \sqrt{(g'(t))^2 + (h'(t))^2} dt$$

3.3 Polar Area Integral

The area integral in polar coordinates is a different beast than its Cartesian counterpart.

In the plane, consider a line that connects the origin $(0, 0)$ to some position (x, y) , corresponding to polar coordinates (r, θ) . Next, imagine a neighboring point located at $(x + \Delta x, y + \Delta y)$, or $(r + \Delta r, \theta + \Delta \theta)$, connected to the origin by a second line.

This setup constitutes two sides of a triangle with one vertex at $(0, 0)$ and two sides $r, r + \Delta r$. The third side of the triangle has length $\sqrt{\Delta x^2 + \Delta y^2}$. Further, it's straightforward to show that the area of such a triangle is

$$\Delta A = \frac{1}{2} (r + \Delta r) \Delta r \sin(\Delta \theta) .$$

In the differential limit, we may approximate

$$\begin{aligned} r + \Delta r &\approx r \\ \sin(\Delta \theta) &\approx \Delta \theta , \end{aligned}$$

and the differential area of our very skinny triangle becomes

$$dA = \frac{1}{2} r^2 d\theta .$$

This is enough to write the formula for the area of a polar function $r(\theta)$ by integrating in the θ -variable:

$$A = \frac{1}{2} \int_{\theta_0}^{\theta_1} (r(\theta))^2 d\theta$$

3.4 Polar Arc Length

Consider a point (x, y) in the Cartesian plane, having polar representation

$$x = r \cos (\theta)$$

$$y = r \sin (\theta) .$$

Also consider a neighboring point $(x + dx, y + dy)$, with

$$x + dx = (r + dr) \cos (\theta + d\theta)$$

$$y + dy = (r + dr) \sin (\theta + d\theta) .$$

Starting from the usual construction for arc length, we're interested in summing many consecutive lengths given by

$$dS = \sqrt{dx^2 + dy^2} .$$

With this in mind, use the angle-sum formulas on the pair of equations above to establish:

$$x + dx = (r + dr) (\cos (\theta) \cos (d\theta) - \sin (\theta) \sin (d\theta))$$

$$y + dy = (r + dr) (\sin (\theta) \cos (d\theta) + \cos (\theta) \sin (d\theta))$$

Of course, we can take

$$\cos (d\theta) \approx 1$$

$$\sin (d\theta) \approx d\theta$$

$$dr \cdot d\theta \approx 0$$

because differential quantities are vanishing, and the above simplifies to

$$dx = -y d\theta + \cos (\theta) dr$$

$$dy = x d\theta + \sin (\theta) dr .$$

Note that the ratio dy/dx returns the proper slope of a curve in polar coordinates.

Taking the sum $dx^2 + dy^2$ will eliminate most of the ugliness:

$$dx^2 + dy^2 = r^2 d\theta^2 + dr^2 ,$$

and now we have an equation for dS in polar coordinates:

$$dS = \sqrt{r^2 d\theta^2 + dr^2}$$

Assuming that r occurs as a function $r(\theta)$, the term $d\theta$ can be pulled out of the root

$$dS = \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2} d\theta ,$$

or more succinctly,

$$dS = \sqrt{r^2 + (r'(\theta))^2} d\theta .$$

This is enough to write the formula for the arc length of a polar function $r(\theta)$ by integrating in the θ -variable:

$$S = \int_{\theta_0}^{\theta_1} \sqrt{r^2 + (r'(\theta))^2} d\theta$$

4 Vector Derivatives

Rules of Differentiation

The rules for differentiating vector quantities are exactly analogous to those for scalars. For instance, if we have a position vector $\mathbf{r}(t)$, then the velocity vector is the time derivative by definition:

$$\mathbf{v}(t) = \frac{d}{dt}(\mathbf{r}(t)) = \lim_{h \rightarrow 0} \frac{\mathbf{r}(t+h) - \mathbf{r}(t)}{h}$$

In the following, suppose c is a constant, $\lambda(t)$ is a function of time, and so too are the vectors $\mathbf{r}$, $\mathbf{s}$. Then, we always have:

$$\begin{aligned} \frac{d}{dt}(c\mathbf{r}) &= c \frac{d\mathbf{r}}{dt} \\ \frac{d}{dt}(\lambda\mathbf{r}) &= \frac{d\lambda}{dt}\mathbf{r} + \lambda \frac{d\mathbf{r}}{dt} \\ \frac{d}{dt}(\mathbf{r} + \mathbf{s}) &= \frac{d\mathbf{r}}{dt} + \frac{d\mathbf{s}}{dt} \\ \frac{d}{dt}(\mathbf{r} \cdot \mathbf{s}) &= \frac{d\mathbf{r}}{dt} \cdot \mathbf{s} + \mathbf{r} \cdot \frac{d\mathbf{s}}{dt} \\ \frac{d}{dt}(\mathbf{r} \times \mathbf{s}) &= \frac{d\mathbf{r}}{dt} \times \mathbf{s} + \mathbf{r} \times \frac{d\mathbf{s}}{dt} \end{aligned}$$

4.1 Basis Vector Derivatives

Derivatives of the Cartesian basis vectors are all zero. As for polar coordinates, since $\hat{\mathbf{r}}$, $\hat{\boldsymbol{\theta}}$ are allowed to swivel about as θ changes, it makes sense to ask about the derivatives of these. Proceeding carefully, we find

$$\frac{d}{d\theta}\hat{\mathbf{r}} = -\sin(\theta)\hat{\mathbf{x}} + \cos(\theta)\hat{\mathbf{y}} = \hat{\boldsymbol{\theta}}$$

and

$$\frac{d}{d\theta}\hat{\boldsymbol{\theta}} = -\cos(\theta)\hat{\mathbf{x}} - \sin(\theta)\hat{\mathbf{y}} = -\hat{\mathbf{r}}.$$

Velocity

Supposing x , y are each functions of time, it follows that r , θ are also functions of time. For the Cartesian case, we easily differentiate the position with respect to time to get the velocity vector

$$\mathbf{V} = \frac{d}{dt}\mathbf{R} = \left(\frac{d}{dt}x(t)\right)\hat{\mathbf{x}} + \left(\frac{d}{dt}y(t)\right)\hat{\mathbf{y}}.$$

Velocity in Polar Coordinates

The story in polar coordinates is a little different. For shorthand, express the *angular velocity* as the time derivative of the θ coordinate as ω (Greek ‘omega’):

$$\omega = \omega(t) = \frac{d}{dt}\theta(t)$$

Next, we’ll need the time derivative of the *entire* position vector

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d}{dt}(r(t)\hat{\mathbf{r}}(t)),$$

which calls for the product rule:

$$\mathbf{v} = \left(\frac{d}{dt}r(t)\right)\hat{\mathbf{r}}(t) + r(t)\frac{d}{dt}\hat{\mathbf{r}}(t)$$

Expand the right-most term with the with the chain rule

$$\mathbf{v} = \frac{dr}{dt}\hat{\mathbf{r}} + r\frac{d\theta}{dt}\frac{d\hat{\mathbf{r}}}{d\theta},$$

and simplify to get the velocity in polar coordinates:

$$\mathbf{v} = \frac{dr}{dt}\hat{\mathbf{r}} + r\omega\hat{\boldsymbol{\theta}}$$

Problem 1

If the magnitude $|\mathbf{r}| = r$ is constant in time, show that $\mathbf{r}$ and $\mathbf{v}$ are perpendicular. Hint: differentiate $\mathbf{r} \cdot \mathbf{r}$

4.2 Acceleration

The acceleration vector is the time derivative of the velocity vector. This is straightforward in Cartesian coordinates:

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2}{dt^2}\mathbf{R} = \left(\frac{d^2}{dt^2}x(t)\right)\hat{\mathbf{x}} + \left(\frac{d^2}{dt^2}y(t)\right)\hat{\mathbf{y}}$$

Acceleration in Polar Coordinates

As expected, the acceleration in polar coordinates is messy. Starting the calculation, we have

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d}{dt}\left(\frac{dr}{dt}\hat{\mathbf{r}}\right) + \frac{d}{dt}(r\omega\hat{\boldsymbol{\theta}}).$$

Leaving the details for an exercise, the result comes out to

$$\mathbf{a} = \left(\frac{d^2r}{dt^2} - r\omega^2\right)\hat{\mathbf{r}} + \left(r\frac{d\omega}{dt} + 2\frac{dr}{dt}\omega\right)\hat{\boldsymbol{\theta}}.$$

Problem 2

Derive the acceleration vector in polar coordinates by taking two derivatives of $\mathbf{r} = r \hat{\mathbf{r}}$.

Problem 3

A particle has a known trajectory

$$r(t) = \frac{r_0}{\cos(\omega t)},$$

where ω is constant. Graph the motion and determine the velocity vector and the acceleration vector.

Uniform Circular Motion

The case of uniform circular motion has $r(t) = r_0$ and $\omega(t) = \omega_0$. In this case, the acceleration vector still changes in time and is given by

$$\mathbf{a} = -r_0 \omega_0^2 \hat{\mathbf{r}},$$

which affirms that the acceleration points to the center of the circle.

Complex Number Analogy

Interestingly, the velocity and acceleration equations for polar coordinates can arise from taking derivatives of the complex number

$$z = r e^{i\theta}.$$

Assuming r, θ are functions of time, take a time-derivative of z to get something like the velocity:

$$\frac{d}{dt} z = \frac{dr}{dt} e^{i\theta} + r \omega (i e^{i\theta})$$

Comparing this result to the velocity in polar coordinates, a direct analogy emerges:

$$\begin{aligned} e^{i\theta} &\leftrightarrow \hat{\mathbf{r}} \\ i e^{i\theta} &\leftrightarrow \hat{\boldsymbol{\theta}} \end{aligned}$$

Evidently, the terms $e^{i\theta}, \hat{\mathbf{r}}$ play similar roles in complex numbers and polar coordinates, which is no surprise since each is of the format $\langle \cos(\theta), \sin(\theta) \rangle$. Similar comments apply to the pair $i e^{i\theta}, \hat{\boldsymbol{\theta}}$.

4.3 Differential Line Element

Starting from the velocity vector, whether it be the Cartesian representation or polar representation

$$\begin{aligned} \mathbf{V} &= \frac{dx}{dt} \hat{\mathbf{x}} + \frac{dy}{dt} \hat{\mathbf{y}} \\ \mathbf{v} &= \frac{dr}{dt} \hat{\mathbf{r}} + r \frac{d\theta}{dt} \hat{\boldsymbol{\theta}}, \end{aligned}$$

and multiply through by dt :

$$\begin{aligned} d\mathbf{R} &= \mathbf{V} dt = dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}} \\ d\mathbf{r} &= \mathbf{v} dt = dr \hat{\mathbf{r}} + r d\theta \hat{\boldsymbol{\theta}} \end{aligned}$$

The terms $d\mathbf{R}$ (Cartesian), $d\mathbf{r}$ (polar) are each called the *differential line element*, also called $d\mathbf{S}$. For a point (x_0, y_0) in the plane, the differential line element provides a local coordinate system for moving to a nearby point $(x_0 + dx, y_0 + dy)$.

Notice that the differential line element always has units of length, hence the factor of r attached to the $d\theta$ term, which it itself dimensionless.

Differential Interval

The square of the differential line element is called the *differential interval*, always denoted dS^2 , regardless of coordinate system. The reason for this is that the differential element is the same in all coordinate systems.

To check this, first write the differential interval in Cartesian coordinates, namely

$$dS^2 = (dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}}) \cdot (dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}}) = dx^2 + dy^2.$$

Then, for polar coordinates:

$$dS^2 = d\mathbf{r} \cdot d\mathbf{r} = dr^2 + r^2 d\theta^2$$

In other words, we have recovered

$$dx^2 + dy^2 = dr^2 + r^2 d\theta^2,$$

which is a familiar identity relating Cartesian to polar coordinates.

The square root of the differential interval is the differential arc length. That is,

$$\sqrt{dx^2 + dy^2} = \sqrt{dr^2 + r^2 d\theta^2}$$

are each equal to dS as it appears in arc length calculations

$$S = \int dS = \int \sqrt{d\mathbf{S} \cdot d\mathbf{S}}.$$

Misguided Missile

Problem 4

A missile traveling at constant speed is homing in on a target at the origin. Do to an error in its circuitry, it is consistently misdirected by a constant angle α . Find its path. Show that if $|\alpha| < 90^\circ$, then it will eventually hit its target, taking $1/\cos(\alpha)$ as long as if it were correctly aimed.

Answer: Using polar coordinates, the velocity is

$$\mathbf{v} = -v \cos(\alpha) \hat{\mathbf{r}} + v \sin(\alpha) \hat{\boldsymbol{\theta}},$$

which means

$$\begin{aligned} \frac{dr}{dt} &= -v \cos(\alpha) \\ r \frac{d\theta}{dt} &= v \sin(\alpha). \end{aligned}$$

Eliminating dt between the above and canceling v leads to

$$\frac{dr}{r} = -\cot(\alpha) d\theta,$$

which can be integrated and simplified to

$$r(\theta) = r_0 e^{-\cot(\alpha)\theta},$$

where the integration constant is recast as r_0 . The path happens to be a logarithmic spiral.

The time taken for the correctly-aimed missile to reach the target is

$$T = \int dt = \frac{-1}{v} \int_{r_0}^0 dr = \frac{r_0}{v}.$$

When the motion is not on a straight line, the time is instead

$$T' = \frac{-1}{v} \int_{r_0}^0 dS.$$

Going from the above, it also follows that

$$dS = \frac{dr}{\cos(\alpha)},$$

and thus

$$T' = T/\cos(\alpha),$$

as we wanted to show.

4.4 Differential Area Element

The existence of the differential line element hints at the *differential area element*, which is a small 'patch' that covers part of the plane.

Cartesian Area Element

In Cartesian coordinates, it makes sense to propose

$$dA = dx dy$$

as the differential area element, which happens to be correct.

Let us derive this more carefully, though. Start with the differential line element

$$d\mathbf{R} = dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}},$$

and then write two versions of this - one with $dy = 0$, the other with $dx = 0$:

$$d\mathbf{R}_1 = dx \hat{\mathbf{x}} + 0 \hat{\mathbf{y}}$$

$$d\mathbf{R}_2 = 0 \hat{\mathbf{x}} + dy \hat{\mathbf{y}}$$

This defines two vectors that form the sides of a parallelogram (just a rectangle in this case). The magnitude of the cross product $d\mathbf{R}_1 \times d\mathbf{R}_2$ yields the area of said rectangle:

$$dA = |d\mathbf{R}_1 \times d\mathbf{R}_2| = (dx \hat{\mathbf{x}}) \times (dy \hat{\mathbf{y}}) = dx dy$$

Polar Area Element

The reason for deriving the Cartesian area element in such a belabored way is to make easy the polar area element. For the polar case, we have

$$d\mathbf{r} = dr \hat{\mathbf{r}} + r d\theta \hat{\boldsymbol{\theta}},$$

and then

$$d\mathbf{r}_1 = dr \hat{\mathbf{r}} + 0 \hat{\boldsymbol{\theta}}$$

$$d\mathbf{r}_2 = 0 \hat{\mathbf{r}} + r d\theta \hat{\boldsymbol{\theta}}.$$

In this case, $d\mathbf{r}_1, d\mathbf{r}_2$ form the sides of a parallelogram (not rectangular), whose area is the the magnitude of their cross product:

$$dA = |d\mathbf{r}_1 \times d\mathbf{r}_2| = r dr d\theta$$

5 Plane Curve Analysis

Consider a curve described by a position vector $\mathbf{r}(t)$ in Cartesian coordinates parameterized by a real number t :

$$\mathbf{r}(t) = x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}}$$

5.1 Tangent Vector

The derivative of $\mathbf{r}$ with respect to the parameter t yields the *tangent vector* to the curve.

Velocity Vector

If t is taken as an accumulation of time, the tangent vector is the velocity:

$$\mathbf{v}(t) = \frac{d}{dt}\mathbf{r}(t) = \frac{dx}{dt}\hat{\mathbf{x}} + \frac{dy}{dt}\hat{\mathbf{y}} = v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}}$$

The magnitude of the velocity is the speed,

$$v = \sqrt{v_x^2 + v_y^2} = \frac{dS}{dt},$$

and dividing the velocity by the speed returns the unit tangent vector:

$$\hat{\mathbf{T}} = \frac{\mathbf{v}}{v} = \frac{v_x\hat{\mathbf{x}} + v_y\hat{\mathbf{y}}}{\sqrt{v_x^2 + v_y^2}}$$

Going further, we can divide out a factor of dt from the numerator and denominator to write

$$\hat{\mathbf{T}} = \frac{dx\hat{\mathbf{x}} + dy\hat{\mathbf{y}}}{\sqrt{dx^2 + dy^2}}.$$

In this form, we see the right side is the ratio of the differential line element to the differential arc length. From this we have

$$\hat{\mathbf{T}} = \frac{d\mathbf{S}}{dS},$$

which is the tightest definition for the tangent vector to a curve. The derivative of the position vector with respect to the arc length is the direction of ‘motion’.

In terms of a standard parameter ϕ , the normalized tangent vector can be expressed as

$$\hat{\mathbf{T}} = \cos(\phi)\hat{\mathbf{x}} + \sin(\phi)\hat{\mathbf{y}},$$

in which case the slope of the curve dy/dx is

$$\frac{dy}{dx} = \frac{\sin(\phi)}{\cos(\phi)} = \tan(\phi) = \frac{v_y}{v_x}.$$

Note that the tangent vector also applies to non-parametric curves, i.e. the classic function $y = f(x)$. For this, the general equation for $\hat{\mathbf{T}}$ can be configured as:

$$\hat{\mathbf{T}} = \frac{\hat{\mathbf{x}} + (dy/dx)\hat{\mathbf{y}}}{\sqrt{1 + (dy/dx)^2}} = \frac{\hat{\mathbf{x}} + y'\hat{\mathbf{y}}}{\sqrt{1 + (y')^2}}$$

Elliptic Motion

Problem 5

Suppose the position of an object moving in two dimensions is given by

$$\mathbf{r}(t) = a \cos(\omega t)\hat{\mathbf{x}} + b \sin(\omega t)\hat{\mathbf{y}}.$$

Show that the object moves on an elliptical path and find the distance from the origin as a function of time. Also write the velocity and acceleration vectors.

Answer:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

$$r = \sqrt{\mathbf{r} \cdot \mathbf{r}} = \sqrt{a^2 \cos^2(\omega t) + b^2 \sin^2(\omega t)}$$

$$\mathbf{v}(t) = -a\omega \sin(\omega t)\hat{\mathbf{x}} + b\omega \cos(\omega t)\hat{\mathbf{y}}$$

$$\mathbf{a}(t) = -\omega^2 \mathbf{r}(t)$$

5.2 Normal Vector

Given the tangent vector $\hat{\mathbf{T}}$ to a curve, one can imagine that differential changes in $\hat{\mathbf{T}}$ are always perpendicular to the tangent, and are thus perpendicular to the curve:

$$\frac{d}{d\phi}\hat{\mathbf{T}} = -\sin(\phi)\hat{\mathbf{x}} + \cos(\phi)\hat{\mathbf{y}}$$

One can explicitly check that

$$\hat{\mathbf{T}} \cdot \frac{d}{d\phi}\hat{\mathbf{T}} = 0.$$

This implies the existence of the *normal vector*. In the general case, the normalized vector is crudely defined as

$$\mathbf{N} = \frac{d}{dt}\mathbf{T},$$

where the parameter t is considered generic. The unit normal vector divides out its own magnitude

$$\hat{\mathbf{N}} = \frac{1}{|d\mathbf{T}/dt|} \frac{d\mathbf{T}}{dt},$$

and by the chain rule, we can essentially swap the t -parameter for the arc length:

$$\hat{N} = \frac{1}{|d\mathbf{T}/dS|} \frac{d\mathbf{T}}{dS}$$

Problem 6

Show that the tangent vector and normal vector always obey the orthogonality relation

$$\mathbf{T} \cdot \mathbf{N} = 0.$$

Hint: differentiate $\hat{\mathbf{T}} \cdot \hat{\mathbf{T}}$ with respect to arc length.

5.3 Acceleration Vector

The position vector

$$\mathbf{r}(t) = x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}}$$

and the velocity vector

$$\mathbf{v}(t) = \frac{d}{dt}\mathbf{r}(t) = \frac{dx}{dt}\hat{\mathbf{x}} + \frac{dy}{dt}\hat{\mathbf{y}}$$

imply the existence of the acceleration vector

$$\mathbf{a}(t) = \frac{d}{dt}\mathbf{v}(t) = \frac{d^2x}{dt^2}\hat{\mathbf{x}} + \frac{d^2y}{dt^2}\hat{\mathbf{y}}.$$

While none of the above is news, we can frame these items in terms of tangent and normal vectors. Most easy is the velocity, which is

$$\mathbf{v}(t) = v(t)\hat{\mathbf{T}}(t).$$

Now recalculate the acceleration vector using the above definition, which gives

$$\mathbf{a}(t) = \frac{d}{dt}\mathbf{v}(t) = \frac{dv}{dt}\hat{\mathbf{T}} + v \frac{d\hat{\mathbf{T}}}{dt},$$

or

$$\mathbf{a}(t) = \left(\frac{d^2S}{dt^2}\right)\hat{\mathbf{T}} + \left|\frac{d\hat{\mathbf{T}}}{dS}\right|\left(\frac{dS}{dt}\right)^2\hat{\mathbf{N}},$$

where $v = dS/dt$ and the definition of the normal vector have been used.

Problem 7

Derive the following:

$$\hat{N} = \frac{1}{v|d\hat{\mathbf{T}}/dt|}\mathbf{a} - \frac{dv/dt}{v^2|d\hat{\mathbf{T}}/dt|}\mathbf{v}$$

Instantaneous Basis Vectors

Since the tangent vector and normal vector are always perpendicular, these form a local set of basis vectors to the curve:

$$\begin{aligned}\mathbf{v} &= v\hat{\mathbf{T}} \\ \mathbf{a} &= a_T\hat{\mathbf{T}} + a_N\hat{\mathbf{N}}\end{aligned}$$

Problem 8

A particle has a known trajectory

$$\mathbf{r}(t) = \frac{r_0}{\cos(\omega t)},$$

where ω is constant. Find the velocity and acceleration vectors.

5.4 Curvature

The magnitude of the derivative of the normalized tangent vector with respect to arc length, i.e. $|d\hat{\mathbf{T}}/dS|$, is called the curvature, denoted κ (Greek 'kappa'):

$$\kappa = \left|\frac{d\hat{\mathbf{T}}}{dS}\right|$$

In terms of the curvature, the normal vector is

$$\hat{N} = \frac{1}{\kappa} \frac{d\hat{\mathbf{T}}}{dS},$$

and the acceleration vector is

$$\mathbf{a}(t) = \left(\frac{d^2S}{dt^2}\right)\hat{\mathbf{T}} + \kappa\left(\frac{dS}{dt}\right)^2\hat{\mathbf{N}}.$$

Interpreting Curvature

The curvature κ , having units of inverse length, can be regarded as one divided by the radius ρ of the circle that instantaneously approximates the curve. To see this, suppose, much like we do with a straight-line approximation, that the curve is approximated by a circle of radius $\rho(t)$ at some instant t .

To go around such a circle, the tangent vector

$$\hat{\mathbf{T}} = \cos(\phi)\hat{\mathbf{x}} + \sin(\phi)\hat{\mathbf{y}}$$

runs the parameter ϕ from 0 to 2π . Meanwhile, the total arc length traveled throughout the trip is $2\pi\rho(t)$, and we establish

$$\frac{d\phi}{dS} = \frac{2\pi}{2\pi\rho(t)} = \frac{1}{\rho(t)},$$

which is harmlessly inverted:

$$\rho(t) = \frac{dS}{d\phi} = \frac{dS/dt}{d\phi/dt}$$

Proceed by calculating the time derivative of $\hat{T}$

$$\frac{d\hat{T}}{dt} = \frac{d\phi}{dt} (-\sin(\phi)\hat{x} + \cos(\phi)\hat{y}),$$

where the parenthesized portion is a unit normal vector. Taking the magnitude of the above allows $d\phi/dt$ to be isolated:

$$\left| \frac{d\hat{T}}{dt} \right| = \frac{d\phi}{dt}$$

Now rewrite the equation for $\rho(t)$:

$$\rho(t) = \frac{dS/dt}{|d\hat{T}/dt|} = \frac{1}{|d\hat{T}/dS|} = \frac{1}{\kappa(t)}$$

Calculating Curvature

Another way to isolate the curvature κ comes from the cross product between the velocity vector and the acceleration vector. For this, begin with

$$\mathbf{v} \times \mathbf{a} = (v\hat{T}) \times \left(\left(\frac{d^2S}{dt^2} \right) \hat{T} + \kappa \left(\frac{dS}{dt} \right)^2 \hat{N} \right).$$

The cross product distributes into both acceleration components, but $\hat{T} \times \hat{T}$ is automatically zero:

$$\hat{v} \times \hat{a} = v\kappa \left(\frac{dS}{dt} \right)^2 \hat{T} \times \hat{N}$$

The terms v and dS/dt are the same, and the remaining cross product yields a unit vector oriented perpendicular to the plane of the curve:

$$|\hat{T} \times \hat{N}| = 1$$

This is enough to isolate the curvature κ in terms of the vectors of motion:

$$\kappa = \frac{|\mathbf{v} \times \mathbf{a}|}{v^3}$$

Curvature in Polar Coordinates

For a polar curve

$$\mathbf{r}(t) = r(t)\hat{r},$$

the velocity and acceleration are

$$\begin{aligned} \mathbf{v} &= \frac{dr}{dt} \hat{r} + r \frac{d\theta}{dt} \hat{\theta} \\ \mathbf{a} &= \left(\frac{d^2r}{dt^2} - r \left(\frac{d\theta}{dt} \right)^2 \right) \hat{r} + \left(r \frac{d^2\theta}{dt^2} + 2 \frac{dr}{dt} \frac{d\theta}{dt} \right) \hat{\theta}. \end{aligned}$$

The magnitude of the velocity is

$$v = \sqrt{\left(\frac{dr}{dt} \right)^2 + r^2 \left(\frac{d\theta}{dt} \right)^2}.$$

Anticipating the cross product $\mathbf{v} \times \mathbf{a}$, note that

$$\hat{r} \times \hat{r} = \hat{\theta} \times \hat{\theta} = 0,$$

and also

$$\hat{r} \times \hat{\theta} = -\hat{\theta} \times \hat{r}.$$

With this, we find

$$\begin{aligned} \mathbf{v} \times \mathbf{a} &= (v_r \hat{r} + v_\theta \hat{\theta}) \times (a_r \hat{r} + a_\theta \hat{\theta}) \\ &= (v_r a_\theta - v_\theta a_r) (\hat{r} \times \hat{\theta}), \end{aligned}$$

or, noting $\omega = d\theta/dt$,

$$|\mathbf{v} \times \mathbf{a}| = \left| r \frac{dr}{dt} \frac{d\omega}{dt} + 2\omega \left(\frac{dr}{dt} \right)^2 - r\omega \frac{d^2r}{dt^2} + r^2\omega^3 \right|.$$

In the special case $t = \theta$, we have $dt = d\theta$, meaning $\omega = 1$ and $d\omega/dt = 0$. This is the setup for plotting polar curves $r = r(\theta)$ in the plane, and the corresponding curvature is:

$$\kappa = \frac{|r^2 + 2(dr/d\theta)^2 - r(d^2r/d\theta^2)|}{\left(r^2 + (dr/d\theta)^2 \right)^{3/2}}$$

6 Curves in Three Dimensions

6.1 Generalizing Plane Curves

In three dimensions, the position vector representing a parametric curve can be written

$$\mathbf{r}(t) = x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}} + z(t)\hat{\mathbf{z}}.$$

This is just a generalization of the two-dimensional case, where of course, we've used the Cartesian representation for simplicity, but any three-dimensional system works just as well.

Derivatives

The position vector admits time derivatives to yield the velocity vector and acceleration vector, so long as the parametric equations $x(t)$, $y(t)$, $z(t)$ are differentiable:

$$\begin{aligned}\mathbf{v}(t) &= \left(\frac{d}{dt}x(t)\right)\hat{\mathbf{x}} + \left(\frac{d}{dt}y(t)\right)\hat{\mathbf{y}} + \left(\frac{d}{dt}z(t)\right)\hat{\mathbf{z}} \\ \mathbf{a}(t) &= \left(\frac{d^2}{dt^2}x(t)\right)\hat{\mathbf{x}} + \left(\frac{d^2}{dt^2}y(t)\right)\hat{\mathbf{y}} + \left(\frac{d^2}{dt^2}z(t)\right)\hat{\mathbf{z}}\end{aligned}$$

Tangent Vector

The (unit) tangent vector $\hat{\mathbf{T}}$ is given by the velocity vector divided by the speed

$$\hat{\mathbf{T}} = \frac{\mathbf{v}}{v},$$

and the velocity vector can also be written in terms of the speed and the tangent vector:

$$\mathbf{v}(t) = v\hat{\mathbf{T}}$$

Another equation for the tangent vector reads

$$\hat{\mathbf{T}} = \frac{d\mathbf{S}}{dS},$$

where dS is the differential line element, and dS is the differential arc length.

Curvature

The notion of curvature straightforwardly applies in three dimensions from the definition, particularly

$$\kappa = \left| \frac{d\hat{\mathbf{T}}}{dS} \right|.$$

The tight formula for curvature also survives migration to three dimensions:

$$\kappa = \frac{|\mathbf{v} \times \mathbf{a}|}{v^3}$$

Normal Vector

The (unit) normal vector $\hat{\mathbf{N}}$ also needs no modification from the two-dimensional case:

$$\hat{\mathbf{N}} = \frac{1}{\kappa} \frac{d\hat{\mathbf{T}}}{dS}$$

Acceleration Vector

In terms of the tangent vector, the formula for the acceleration vector remains in tact as well:

$$\mathbf{a}(t) = \left(\frac{d^2S}{dt^2}\right)\hat{\mathbf{T}} + \kappa \left(\frac{dS}{dt}\right)^2 \hat{\mathbf{N}}$$

6.2 Binormal Vector

For a curve in three dimensions, the tangent vector $\hat{\mathbf{T}}$ and normal vector $\hat{\mathbf{N}}$ form the basis vectors for a plane, which itself has a normal vector called the *binormal*. The binormal vector is a unit vector $\hat{\mathbf{B}}$, and is defined by

$$\hat{\mathbf{B}} = \hat{\mathbf{T}} \times \hat{\mathbf{N}}.$$

There are two cyclic permutations of the definition that contain the same information:

$$\begin{aligned}\hat{\mathbf{T}} &= \hat{\mathbf{N}} \times \hat{\mathbf{B}} \\ \hat{\mathbf{N}} &= \hat{\mathbf{B}} \times \hat{\mathbf{T}}\end{aligned}$$

Torsion

We can learn more about $\hat{\mathbf{B}}$ by writing two straightforward consequences of its definition, namely

$$\begin{aligned}\hat{\mathbf{B}} \cdot \hat{\mathbf{T}} &= 0 \\ \hat{\mathbf{B}} \cdot \hat{\mathbf{B}} &= 1,\end{aligned}$$

and apply an arc length derivative to each:

$$\begin{aligned}\frac{d\hat{\mathbf{B}}}{dS} \cdot \hat{\mathbf{T}} + \hat{\mathbf{B}} \cdot \frac{d\hat{\mathbf{T}}}{dS} &= 0 \\ 2\hat{\mathbf{B}} \cdot \frac{d\hat{\mathbf{B}}}{dS} &= 0\end{aligned}$$

Evidently, the derivative $d\hat{\mathbf{B}}/dS$ is perpendicular to both $\hat{\mathbf{T}}$ and $\hat{\mathbf{B}}$. This can only mean $d\hat{\mathbf{B}}/dS$ is parallel to the normal vector:

$$\frac{d\hat{\mathbf{B}}}{dS} = -\tau \hat{\mathbf{N}}$$

The proportionality constant τ (Greek 'tau'), defined with a minus sign, is called the *torsion* along the curve.

6.3 Serret-Frenet Formulas

The equations for $d\hat{T}/dS$ and $d\hat{B}/dS$ constitute two of the three *Serret-Frenet* formulas. To complete the set we need to find $d\hat{N}/dS$. By brute force, we have:

$$\begin{aligned}\frac{d\hat{N}}{dS} &= \frac{d}{dS} (\hat{B} \times \hat{T}) \\ &= -\tau \hat{N} \times \hat{T} + \kappa \hat{B} \times \hat{N} \\ &= \tau \hat{B} - \kappa \hat{T}\end{aligned}$$

TNB Frame

The tangent, normal, and binormal vectors constitute three coordinate axes that trace along the curve, sometimes called

the TNB frame.

Since the derivative of each vector is a linear combination in the TNB frame, the Serret-Frenet formulas lend nicely to matrix notation:

$$\frac{d}{dS} \begin{bmatrix} \hat{T} \\ \hat{N} \\ \hat{B} \end{bmatrix} = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix} \begin{bmatrix} \hat{T} \\ \hat{N} \\ \hat{B} \end{bmatrix}$$

The matrix containing the κ , τ coefficients is skew-symmetric.

7 Cartesian 3D Coordinates

7.1 Cartesian Coordinates

The extension of the Cartesian coordinate system to three dimensions is straightforward. The xy plane is upgraded to the xyz volume, which requires three coordinates to specify any point in the system as shown in Figure 18.1.

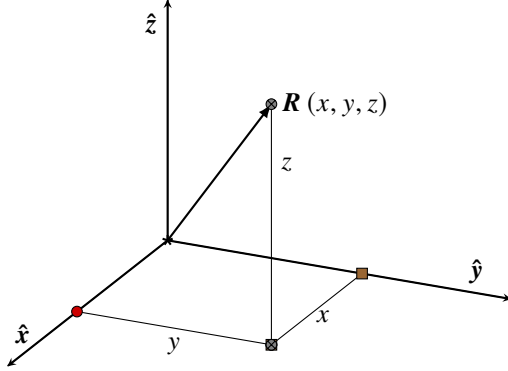


Figure 18.1. Cartesian coordinate system.

The position vector becomes

$$\mathbf{R} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}},$$

which requires three mutually-perpendicular basis vectors:

$$\hat{\mathbf{x}} = \hat{\mathbf{i}} = \hat{\mathbf{e}}_x = \hat{\mathbf{e}}_i = \langle 1, 0, 0 \rangle$$

$$\hat{\mathbf{y}} = \hat{\mathbf{j}} = \hat{\mathbf{e}}_y = \hat{\mathbf{e}}_j = \langle 0, 1, 0 \rangle$$

$$\hat{\mathbf{z}} = \hat{\mathbf{k}} = \hat{\mathbf{e}}_z = \hat{\mathbf{e}}_k = \langle 0, 0, 1 \rangle$$

Quantities like velocity and acceleration simply take on a new z -component:

$$\mathbf{v} = \frac{dx}{dt} \hat{\mathbf{x}} + \frac{dy}{dt} \hat{\mathbf{y}} + \frac{dz}{dt} \hat{\mathbf{z}}$$

$$\mathbf{a} = \frac{d^2x}{dt^2} \hat{\mathbf{x}} + \frac{d^2y}{dt^2} \hat{\mathbf{y}} + \frac{d^2z}{dt^2} \hat{\mathbf{z}}$$

The same goes for the differential line element

$$d\mathbf{S} = dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}} + dz \hat{\mathbf{z}}$$

and the differential interval

$$dS^2 = dx^2 + dy^2 + dz^2.$$

Cartesian Area Element

In three dimensions, the notion of ‘area element’ takes on three meanings. An area can be ‘facing’ along the x , y ,

or z direction. A differential patch of area parallel to the xy -plane is written

$$dA_z = dx dy.$$

Similarly, a differential patch of area parallel to the yz -plane is

$$dA_x = dy dz,$$

and finally, differential patch of area parallel to the zx -plane is

$$dA_y = dz dx.$$

Cartesian Volume Element

Each of the differential area elements can be turned into a differential volume element by multiplying by dz , dx , dy , respectively:

$$dz dA_z = dx dy dz$$

$$dx dA_x = dy dz dx$$

$$dy dA_y = dz dx dy$$

Each of these describes the same *Cartesian volume element*:

$$dV = dx dy dz$$

7.2 Planes

A line is given by $y = mx + b$ in Cartesian coordinates. This construction is convenient and ergonomic, but is more versatile if written

$$Ax + By + C = 0.$$

This is still the equation of a line, but the y -variable is treated on equal footing to the x -variable. The line passes through the origin if $C = 0$.

Equation of a Plane

The notion of the straight line can be extended by one dimension to *planes* in three-dimensional space. The equation of a plane in Cartesian coordinates is

$$ax + by + cz + d = 0.$$

The variables a , b , c together govern the overall slant of the plane. The plane passes through the origin if $d = 0$.

Vector Representation

Interestingly, observe that the quantity $ax + by + cz$ looks like the dot product of two vectors. Define a vector $\mathbf{P}$ to hold the coefficients

$$\mathbf{P} = a \hat{x} + b \hat{y} + c \hat{z},$$

and project this into the position vector

$$\mathbf{R} = x \hat{x} + y \hat{y} + z \hat{z}$$

to find

$$\mathbf{R} \cdot \mathbf{P} = -d.$$

7.3 Normal Vector to a Plane

Consider any given point $\mathbf{R}_0$ that is known to be in the plane. For any point $\mathbf{R}$ that is also in the plane, the difference

$$\Delta \mathbf{R} = \mathbf{R} - \mathbf{R}_0$$

is a tangent vector that stays embedded in the plane.

This is enough to define the notion of the normal vector to the plane $\mathbf{N}$ such that

$$\Delta \mathbf{R} \cdot \mathbf{N} = 0.$$

Peel apart the $\Delta \mathbf{R}$ term to write:

$$\mathbf{R} \cdot \mathbf{N} = \mathbf{R}_0 \cdot \mathbf{N}$$

The above looks very much like the equation of the plane when written

$$\mathbf{R} \cdot \mathbf{P} = -d.$$

Subtract the two equations to get

$$\mathbf{R} \cdot (\mathbf{N} - \mathbf{P}) = \mathbf{R}_0 \cdot \mathbf{N} + d,$$

and then notice d is also equivalent to $-\mathbf{R}_0 \cdot \mathbf{P}$. The above becomes:

$$\mathbf{R} \cdot (\mathbf{N} - \mathbf{P}) = \mathbf{R}_0 \cdot (\mathbf{N} - \mathbf{P})$$

This result needs to be true for all $\mathbf{R}$, which can only mean the parenthesized quantities are zero, which means $\mathbf{N} = \mathbf{P}$. The normal vector is the list of coefficients on x , y , and z :

$$\mathbf{N} = a \hat{x} + b \hat{y} + c \hat{z}$$

Problem 9

Show that the unit normal vector to the plane is given by:

$$\hat{\mathbf{N}} = \frac{a \hat{x} + b \hat{y} + c \hat{z}}{\sqrt{a^2 + b^2 + c^2}}$$

Two Planes Intersecting

Consider two non-parallel planes in three-dimensional space

$$ax + by + cz + d = 0$$

$$Ax + By + Cz + D = 0.$$

Somewhere out in the Cartesian volume is a line of intersection between the planes. To find such a line, take the cross product of the normal vector of each:

$$\mathbf{L} = \mathbf{N}_1 \times \mathbf{N}_2 = \begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ a & b & c \\ A & B & C \end{vmatrix} = \langle L_x, L_y, L_z \rangle$$

The vector $\mathbf{L}$ indicates the intersection of each plane. To establish this, note that the following pair of equations must be true:

$$\mathbf{N}_1 \cdot \mathbf{L} = 0$$

$$\mathbf{N}_2 \cdot \mathbf{L} = 0$$

Substitute the proposed form for $\mathbf{L}$,

$$\mathbf{N}_1 \cdot (\mathbf{N}_1 \times \mathbf{N}_2) = 0$$

$$\mathbf{N}_2 \cdot (\mathbf{N}_1 \times \mathbf{N}_2) = 0,$$

and notice each as a triple product to finish the job:

$$\mathbf{N}_2 \cdot (\mathbf{N}_1 \times \mathbf{N}_1) = 0$$

$$\mathbf{N}_1 \cdot (\mathbf{N}_2 \times \mathbf{N}_2) = 0$$

8 Non-Cartesian Coordinates

8.1 Cylindrical Coordinates

Another way of mapping the three-dimensional Cartesian space is with *cylindrical coordinates*. In this system, illustrated in Figure 18.2, the xy plane is replaced by the $\rho\phi$ plane, with radius ρ and angle ϕ playing their ‘usual’ roles in polar coordinates. The z -direction is handled much as in the ordinary Cartesian system.

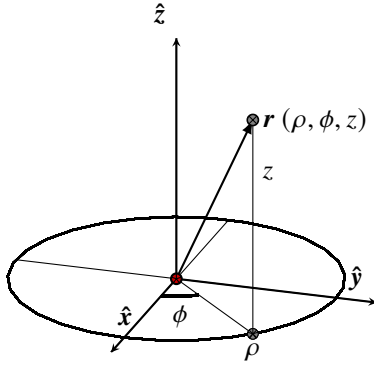


Figure 18.2. Cylindrical coordinate system.

Cylindrical Position Vector

In cylindrical coordinates, the position vector is:

$$\mathbf{r}(\rho, \phi, z) = \rho \hat{\rho} + z \hat{z},$$

or in terms of Cartesian directions,

$$\mathbf{r} = \rho \cos(\phi) \hat{x} + \rho \sin(\phi) \hat{y} + z \hat{z}.$$

As an extension of two already well-studied systems, it's worth mentioning but not writing that time derivatives of the position vector yield the velocity and acceleration. The $\hat{\rho}$ and $\hat{\phi}$ unit vectors are exactly analogous to $\hat{r}$, $\hat{\theta}$, and $\hat{z}$ has no derivative.

Cylindrical Basis Vectors

The basis vectors $\hat{\rho}$, $\hat{\phi}$ are analogous to those in plane polar coordinates. In particular:

$$\begin{aligned}\hat{\rho} &= \cos(\phi) \hat{x} + \sin(\phi) \hat{y} \\ \hat{\phi} &= -\sin(\phi) \hat{x} + \cos(\phi) \hat{y}\end{aligned}$$

The derivatives follow the familiar pattern:

$$\begin{aligned}\frac{d\hat{\rho}}{d\phi} &= \hat{\phi} \\ \frac{d\hat{\phi}}{d\phi} &= -\hat{\rho}\end{aligned}$$

Cylindrical Line Element

The differential version of the cylindrical position vector $d\mathbf{S} = d\mathbf{r}$ gives the *cylindrical line element*:

$$d\mathbf{S} = d\rho \hat{\rho} + \rho d\phi \hat{\phi} + dz \hat{z}$$

Differential Interval

The differential interval in cylindrical coordinates is straightforwardly calculated from the line interval:

$$dS^2 = d\mathbf{S} \cdot d\mathbf{S} = d\rho^2 + \rho^2 d\phi^2 + dz^2$$

Differential Arc Length

The differential arc length is the positive square root of the differential interval:

$$dS = \sqrt{d\rho^2 + \rho^2 d\phi^2 + dz^2}$$

Cylindrical Area Element

Like the Cartesian case, the notion of area element takes three meanings. For a patch of area swept by constant ρ with small changes in z and ϕ , the area is

$$dA_\rho = \rho d\phi dz.$$

A patch of area that keeps ϕ constant is

$$dA_\phi = d\rho dz.$$

A patch of area that has z constant has area

$$dA_z = \rho d\rho d\phi.$$

Note that each of these can be constructed from various cross products of special cases of the cylindrical line element:

$$\begin{aligned}dA_\rho &= |\rho d\phi \hat{\phi} \times dz \hat{z}| \\ dA_\phi &= |d\rho \hat{\rho} \times dz \hat{z}| \\ dA_z &= |d\rho \hat{\rho} \times \rho d\phi \hat{\phi}|\end{aligned}$$

Cylindrical Volume Element

The volume element in cylindrical coordinates is the product of components of the line element:

$$dV = \rho \, d\rho \, d\phi \, dz$$

More rigorously, dV can be written as the triple product

$$dV = d\rho \, \hat{\rho} \cdot (\rho \, d\phi \, \hat{\phi} \times dz \, \hat{z})$$

or any of its permutations.

8.2 Spherical Coordinates

The *spherical coordinate system* is another way to locate any point in three-dimensional space. There are three parameters: (i) a distance r from the origin, (ii) an angle θ measured from the positive z -axis, (iii) a polar parameter ϕ .

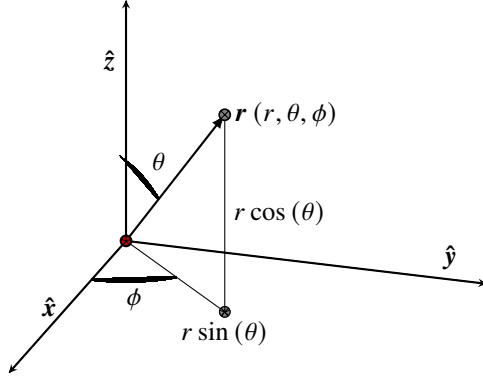


Figure 18.3. Spherical coordinate system.

Spherical Position Vector

In spherical coordinates, the position vector is

$$\begin{aligned} \mathbf{r}(r, \theta, \phi) &= r \sin(\theta) \cos(\phi) \hat{x} \\ &\quad + r \sin(\theta) \sin(\phi) \hat{y} \\ &\quad + r \cos(\theta) \hat{z}, \end{aligned}$$

which can be discerned by analyzing Figure 18.3. Explicitly, this is stating

$$\begin{aligned} x &= r \sin(\theta) \cos(\phi) \\ y &= r \sin(\theta) \sin(\phi) \\ z &= r \cos(\theta). \end{aligned}$$

It's a worthwhile exercise to check that

$$\sqrt{\mathbf{r} \cdot \mathbf{r}} = \sqrt{r^2} = r$$

still holds, which means to make sure all of the trigonometric terms cancel inside the square root.

Spherical Basis Vectors

The position vector in spherical coordinates can be abbreviated

$$\mathbf{r} = r \hat{\mathbf{r}}(r, \theta, \phi),$$

which tells us the radial basis vector:

$$\hat{\mathbf{r}} = \frac{\mathbf{r}}{r} = \langle \sin(\theta) \cos(\phi), \sin(\theta) \sin(\phi), \cos(\theta) \rangle$$

We need to find two more basis vectors, namely $\hat{\phi}$, $\hat{\theta}$ for spherical coordinates. To figure out $\hat{\phi}$, notice that $r \sin(\theta)$ is the projected radius in the xy -plane, and ϕ is free to vary without affecting the projected radius. This means $\hat{\phi}$ has no radial component and no z -component, and is therefore analogous to the polar basis vector in the two-dimensional case. We conclude

$$\hat{\phi} = -\sin(\phi) \hat{x} + \cos(\phi) \hat{y},$$

or in tighter notation:

$$\hat{\phi} = \langle -\sin(\phi), \cos(\phi), 0 \rangle$$

The $\hat{\theta}$ unit vector chops downward in the direction of increasing θ , and is confined to the plane that hangs under the position vector. As a ninety-degree rotation from $\hat{\mathbf{r}}$, we can quickly write

$$\begin{aligned} \hat{\theta}(\theta, \phi) &= \sin\left(\theta + \frac{\pi}{2}\right) \cos(\phi) \hat{x} \\ &\quad + \sin\left(\theta + \frac{\pi}{2}\right) \sin(\phi) \hat{y} \\ &\quad + \cos\left(\theta + \frac{\pi}{2}\right) \hat{z}, \end{aligned}$$

simplifying to

$$\hat{\theta} = \langle \cos(\theta) \cos(\phi), \cos(\theta) \sin(\phi), -\sin(\theta) \rangle.$$

Basis Vector Orthogonality

Despite their messiness, the unit vectors $\hat{\mathbf{r}}$, $\hat{\theta}$, $\hat{\phi}$ are easily shown to have unit length

$$\hat{\mathbf{r}} \cdot \hat{\mathbf{r}} = \hat{\phi} \cdot \hat{\phi} = \hat{\theta} \cdot \hat{\theta} = 1,$$

and to be mutually perpendicular

$$\hat{\mathbf{r}} \cdot \hat{\theta} = \hat{\mathbf{r}} \cdot \hat{\phi} = 0$$

$$\hat{\theta} \cdot \hat{\phi} = \hat{\theta} \cdot \hat{\mathbf{r}} = 0$$

$$\hat{\phi} \cdot \hat{\mathbf{r}} = \hat{\phi} \cdot \hat{\theta} = 0,$$

and subordinate to the cross product:

$$\hat{\mathbf{r}} \times \hat{\theta} = \hat{\phi}$$

$$\hat{\theta} \times \hat{\phi} = \hat{\mathbf{r}}$$

$$\hat{\phi} \times \hat{\mathbf{r}} = \hat{\theta}$$

Basis Vector Derivatives

Various derivatives of the basis vectors establish relationships between them. In particular:

$$\begin{aligned} d\hat{r}/d\theta &= \hat{\theta} \\ d\hat{r}/d\phi &= \sin(\theta) \hat{\phi} \\ d\hat{\theta}/d\theta &= -\hat{r} \\ d\hat{\theta}/d\phi &= \cos(\theta) \hat{\phi} \\ d\hat{\phi}/d\theta &= 0 \\ d\hat{\phi}/d\phi &= -\sin(\theta) \hat{r} - \cos(\theta) \hat{\theta} \end{aligned}$$

Matrix Representation

The basis vectors in spherical coordinates are nicely organized in matrix notation

$$\begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix} = R_{r\theta\phi} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix},$$

where:

$$R_{r\theta\phi} = \begin{bmatrix} \sin(\theta) \cos(\phi) & \sin(\theta) \sin(\phi) & \cos(\theta) \\ \cos(\theta) \cos(\phi) & \cos(\theta) \sin(\phi) & -\sin(\theta) \\ -\sin(\phi) & \cos(\phi) & 0 \end{bmatrix}$$

Resolving Cartesian Basis

The process for isolating $\hat{x}$, $\hat{y}$, $\hat{z}$ in terms of $\hat{r}$, $\hat{\theta}$, $\hat{\phi}$ is equivalent to inverting the matrix $R_{r\theta\phi}$:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = R_{r\theta\phi}^{-1} \begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix}$$

Leaving the details to the reader, the required matrix is:

$$R_{r\theta\phi}^{-1} = \begin{bmatrix} \sin(\theta) \cos(\phi) & \cos(\theta) \cos(\phi) & -\sin(\phi) \\ \sin(\theta) \sin(\phi) & \cos(\theta) \sin(\phi) & \cos(\phi) \\ \cos(\theta) & -\sin(\theta) & 0 \end{bmatrix}$$

Spherical Velocity Vector

The time derivative of $\mathbf{r} = r \hat{r}(r, \theta, \phi)$ gives the velocity vector in spherical coordinates. Doing this carefully, one should find:

$$\mathbf{v}(t) = \frac{dr}{dt} \hat{r} + r \frac{d\theta}{dt} \hat{\theta} + r \sin(\theta) \frac{d\phi}{dt} \hat{\phi}$$

Eliminating the dt -term by the chain rule gives the differential line element

$$dS = dr \hat{r} + r d\theta \hat{\theta} + r \sin(\theta) d\phi \hat{\phi},$$

and the square of the differential line element is the differential interval:

$$dS^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2(\theta) d\phi^2$$

The positive root of the differential interval is the differential arc length:

$$dS = \sqrt{dr^2 + r^2 d\theta^2 + r^2 \sin^2(\theta) d\phi^2}$$

Spherical Area Element

There are three ways to write an area element in spherical coordinates - one for each coordinate being fixed. Fixing r , the patch of spherical shell swept out is

$$dA_r = |r d\theta \hat{\theta} \times r \sin(\theta) d\phi \hat{\phi}| = r^2 \sin(\theta) d\theta d\phi.$$

The other two area elements come from fixing θ , ϕ respectively:

$$dA_\theta = |dr \hat{r} \times r \sin(\theta) d\phi \hat{\phi}| = r \sin(\theta) dr d\phi$$

$$dA_\phi = |dr \hat{r} \times r d\theta \hat{\theta}| = r dr d\theta$$

Spherical Volume Element

The volume element in spherical coordinates is the product of components of the line element:

$$dV = r^2 \sin(\theta) dr d\theta d\phi$$

More rigorously, dV can be written as the triple product

$$dV = dr \hat{r} \cdot (r d\theta \hat{\theta} \times r \sin(\theta) d\phi \hat{\phi})$$

or any of its permutations.

9 Intersections

9.1 Line Segments Intersecting

In the Cartesian plane, consider a line segment with endpoints located at $\mathbf{p}_A, \mathbf{p}_B$. Also consider a second line segment with endpoints located at $\mathbf{q}_A, \mathbf{q}_B$. The task is to determine whether line segments p, q are intersecting, or if parallel, whether the segments overlap.

In general, each line can be parameterized by a vector equation

$$\mathbf{y}_j = \mathbf{b}_j + \alpha_j \hat{\mathbf{t}}_j,$$

where $j = p, q$ toggles between segments. The vector $\mathbf{b}_j$ is any fixed point on the j th line segment (not necessarily the y -intercept). The unit vector $\hat{\mathbf{t}}_j$ registers the slope of the line segment. The parameter α_j is a real number confined to the domain

$$\alpha_j^{\min} \leq \alpha \leq \alpha_j^{\max}$$

to assure the finite extent of each segment.

Intersection

The condition for intersection is given by $\mathbf{y}_p = \mathbf{y}_q$, or

$$\mathbf{b}_p + \alpha_p \hat{\mathbf{t}}_p = \mathbf{b}_q + \alpha_q \hat{\mathbf{t}}_q.$$

Letting

$$\Delta \mathbf{b} = \mathbf{b}_p - \mathbf{b}_q$$

$$f = \hat{\mathbf{t}}_p \cdot \hat{\mathbf{t}}_q$$

$$w_p = \Delta \mathbf{b} \cdot \hat{\mathbf{t}}_p$$

$$w_q = \Delta \mathbf{b} \cdot \hat{\mathbf{t}}_q,$$

the intersection condition yields a pair of equations

$$\alpha_q - \alpha_p f = w_q$$

$$\alpha_q f - \alpha_p = w_p,$$

allowing α_p, α_q to be isolated:

$$\alpha_p = \frac{f w_q - w_p}{1 - f^2}$$

$$\alpha_q = \frac{w_q - f w_p}{1 - f^2}$$

For $f \neq 1$, the above gives the solution to the intersection of two infinite lines. For line segments, make sure

$$\alpha_{p,q}^{\min} \leq \alpha \leq \alpha_{p,q}^{\max}$$

is satisfied.

Overlap

If the two lines being compared are parallel, then $f = \hat{\mathbf{t}} \cdot \hat{\mathbf{t}} = \pm 1$, and the solutions for α become invalid. We must instead go back to:

$$\alpha_q = \alpha_p f + w_q$$

$$\alpha_p = \alpha_q f - w_p$$

Taking only the first equation, we can set α_p to be either of $\alpha_p^{\min}$ or $\alpha_p^{\max}$, and α_q on the left side of the equation must be between each of these. A similar story applies to the bottom equation.

9.2 Line Intersecting Ellipse

Consider an ellipse located somewhere in the plane parameterized by ϕ such that

$$\mathbf{r}_e = \mathbf{r}_0 + \mathbf{a} \cos(\phi) + \mathbf{b} \sin(\phi),$$

where $\mathbf{r}_0$ is the center, and the mutually perpendicular vectors $\mathbf{a}, \mathbf{b}$ orient the major and minor axes.

Also consider a straight line in the plane parameterized by α :

$$\mathbf{y} = \mathbf{y}_0 + \alpha \hat{\mathbf{v}},$$

where $\mathbf{y}_0$ is a constant vector, and $\hat{\mathbf{v}}$ is the unit tangent vector that tells us the slope of the line.

The case for intersections is given by:

$$\mathbf{r}_0 + \mathbf{a} \cos(\phi) + \mathbf{b} \sin(\phi) = \mathbf{y}_0 + \alpha \hat{\mathbf{v}}$$

From prior knowledge of lines and ellipses, we can anticipate at most two solutions to the above when the line hits the ellipse in the belly, one solution for when the line is tangent to the ellipse, and zero solutions for when the line misses entirely.

To gain on this, multiply through the equation above by $\mathbf{a}$, and again by $\mathbf{b}$, and then isolate the trig terms:

$$\cos(\phi) = \frac{\mathbf{a} \cdot (\mathbf{y}_0 - \mathbf{r}_0) + \alpha \mathbf{a} \cdot \hat{\mathbf{v}}}{a^2}$$

$$\sin(\phi) = \frac{\mathbf{b} \cdot (\mathbf{y}_0 - \mathbf{r}_0) + \alpha \mathbf{b} \cdot \hat{\mathbf{v}}}{b^2}$$

Let

$$\mathbf{q}_0 = \mathbf{y}_0 - \mathbf{r}_0,$$

and simplify further:

$$\cos(\phi) = \frac{\hat{\mathbf{a}} \cdot \mathbf{q}_0 + \alpha \hat{\mathbf{a}} \cdot \hat{\mathbf{v}}}{a}$$

$$\sin(\phi) = \frac{\hat{\mathbf{b}} \cdot \mathbf{q}_0 + \alpha \hat{\mathbf{b}} \cdot \hat{\mathbf{v}}}{b}$$

To proceed, use the fundamental trig identity to write

$$1 = \left(\frac{\hat{\mathbf{a}} \cdot \mathbf{q}_0 + \alpha \hat{\mathbf{a}} \cdot \hat{\mathbf{v}}}{a} \right)^2 + \left(\frac{\hat{\mathbf{b}} \cdot \mathbf{q}_0 + \alpha \hat{\mathbf{b}} \cdot \hat{\mathbf{v}}}{b} \right)^2.$$

Blowing out the algebra, the above takes the form

$$0 = A\alpha^2 + B\alpha + C,$$

where:

$$\begin{aligned} A &= \left(\frac{\hat{\mathbf{a}} \cdot \hat{\mathbf{v}}}{a} \right)^2 + \left(\frac{\hat{\mathbf{b}} \cdot \hat{\mathbf{v}}}{b} \right)^2 \\ B &= \frac{2(\hat{\mathbf{a}} \cdot \mathbf{q}_0)(\hat{\mathbf{a}} \cdot \hat{\mathbf{v}})}{a^2} + \frac{2(\hat{\mathbf{b}} \cdot \mathbf{q}_0)(\hat{\mathbf{b}} \cdot \hat{\mathbf{v}})}{b^2} \\ C &= \left(\frac{\hat{\mathbf{a}} \cdot \mathbf{q}_0}{a} \right)^2 + \left(\frac{\hat{\mathbf{b}} \cdot \mathbf{q}_0}{b} \right)^2 - 1 \end{aligned}$$

Finally, use the quadratic formula to establish

$$\alpha = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

In the special case $B^2 = 4AC$, the line is tangent to the ellipse. If the solutions become imaginary, the line misses the ellipse.

9.3 Circle Intersecting Circle

Let the position $\mathbf{r}_j$ on a circle of radius R_j centered at $\mathbf{C}_j$ be parameterized by ϕ_j :

$$\mathbf{r}_j = \mathbf{C}_j + R_j \langle \cos(\phi_j), \sin(\phi_j) \rangle$$

For the intersection of two such circles of different radii, we can write

$$\mathbf{D} = R_2 \langle \cos(\phi_2), \sin(\phi_2) \rangle - R_1 \langle \cos(\phi_1), \sin(\phi_1) \rangle,$$

where

$$\mathbf{D} = \mathbf{C}_1 - \mathbf{C}_2.$$

In component form, vector $\mathbf{D}$ reads

$$\begin{aligned} D_x &= R_2 \cos(\phi_2) - R_1 \cos(\phi_1) \\ D_y &= R_2 \sin(\phi_2) - R_1 \sin(\phi_1). \end{aligned}$$

Next, solve for $\cos^2(\phi_1) + \sin^2(\phi_1)$ among the two equations to get

$$D_x \cos(\phi_2) + D_y \sin(\phi_2) = \frac{R_2^2 - R_1^2 + D^2}{2R_2} = E,$$

where E is a constant. Similarly, isolate $\cos^2(\phi_2) + \sin^2(\phi_2)$ to find

$$D_x \cos(\phi_1) + D_y \sin(\phi_1) = \frac{R_2^2 - R_1^2 - D^2}{2R_1} = F,$$

where F is also constant.

To summarize, the problem reduces to solving either of:

$$\begin{aligned} \frac{D_x}{E} \cos(\phi_2) + \frac{D_y}{E} \sin(\phi_2) &= 1 \\ \frac{D_x}{F} \cos(\phi_1) + \frac{D_y}{F} \sin(\phi_1) &= 1 \end{aligned}$$

Choosing the first equation, write this as

$$\frac{D_x}{E} \cos(\phi_2) \pm \frac{D_y}{E} \sqrt{1 - \cos^2(\phi_2)} = 1,$$

equivalent to:

$$\left(\frac{D_x}{E} \right)^2 \cos^2(\phi_2) - \frac{2D_x}{E} \cos(\phi_2) + 1 - \left(\frac{D_y}{E} \right)^2 = 0$$

The term $\cos(\phi_2)$ can be isolated with the quadratic formula:

$$\cos(\phi_2) = \frac{D_x E \pm D_y E \sqrt{D^2/E^2 - 1}}{D^2}$$

An identical exercise in solving for $\sin(\phi_2)$ gives:

$$\sin(\phi_2) = \frac{D_y E \pm D_x E \sqrt{D^2/E^2 - 1}}{D^2}$$

Of course, the same two exercises can be done to isolate the terms involving ϕ_1 . The result would have all terms E replaced with F .

To continue, we need to decide how to handle the multi-channel aspect of the solution, i.e. what to do with the $\pm$ symbols. Multiply the pair of equations by D_x/E , D_y/E respectively. The sum must come to one, so

$$\begin{aligned} 1 &= \frac{D_x^2 \pm D_y D_x \sqrt{D^2/E^2 - 1}}{D^2} \\ &\quad + \frac{D_y^2 \pm D_x D_y \sqrt{D^2/E^2 - 1}}{D^2} \end{aligned}$$

tells us

$$0 = (\pm 1 \pm 1) \frac{D_x D_y \sqrt{D^2/E^2 - 1}}{D^2}.$$

In other words, we can have the combinations $+-$, $-+$, but the pure cases $++$, $--$ are invalid.

Explicitly, one pair of solutions to the system reads

$$\begin{aligned}\cos(\phi_2^+) &= \frac{D_x E + D_y E \sqrt{D^2/E^2 - 1}}{D^2} \\ \sin(\phi_2^-) &= \frac{D_y E - D_x E \sqrt{D^2/E^2 - 1}}{D^2},\end{aligned}$$

and the other pair of solutions has the signs outside the roots swapped. The intersection points of the two circles are finally:

$$\begin{aligned}X_{\text{int}}^1 &= C_2 + R_2 \langle \cos(\phi_2^+), \sin(\phi_2^-) \rangle \\ X_{\text{int}}^2 &= C_2 + R_2 \langle \cos(\phi_2^-), \sin(\phi_2^+) \rangle\end{aligned}$$

One Intersection

The condition for both intersection points overlapping at a single intersection point, i.e., when the two circles are tangent, occurs when $D = R_1 + R_2$, which is equivalent to $D = E$.

Zero Intersections

The circles clearly don't intersect in two cases: (i) the circles are sufficiently separated, or (ii) the smaller circle is inside the larger circle with no contact.

9.4 Circle Intersecting Points

Find the circle

$$(x - h)^2 + (y - k)^2 = R^2$$

that passes through three points in the plane $\mathbf{q}_j = \langle x_j, y_j \rangle$ with $j = 1, 2, 3$.

Proceed by blooming out the equation for the circle and substitute each data point:

$$\begin{aligned}q_1^2 - 2x_1 h - 2y_1 k &= R^2 - h^2 - k^2 \\ q_2^2 - 2x_2 h - 2y_2 k &= R^2 - h^2 - k^2 \\ q_3^2 - 2x_3 h - 2y_3 k &= R^2 - h^2 - k^2\end{aligned}$$

Take the difference between the second and first equations

$$q_2^2 - q_1^2 + h(2x_1 - 2x_2) + k(2y_1 - 2y_2) = 0,$$

and also take the difference between the third and first equations:

$$q_3^2 - q_1^2 + h(2x_1 - 2x_3) + k(2y_1 - 2y_3) = 0$$

This is merely a linear system of two equations and two unknowns. Packing the above coefficients on h, k into new variables $a_{1,2}, b_{1,2}, c_{1,2}$, the above reads

$$\begin{aligned}a_1 + b_1 h + c_1 k &= 0 \\ a_2 + b_2 h + c_2 k &= 0.\end{aligned}$$

Solving for h, k is a matter of elementary (linear) algebra. For results, we finally have:

$$\begin{aligned}h &= \frac{a_1/b_1 - a_2/b_2}{c_2/b_2 - c_1/b_1} \\ k &= \frac{a_1/c_1 - a_2/c_2}{b_2/c_2 - b_1/c_1}\end{aligned}$$

To solve for the radius R , consider any vector $\mathbf{P}_j$ that extends from the center of the circle to any given point $\mathbf{q}_j$:

$$\mathbf{P}_j = \mathbf{q}_j - \langle h, k \rangle$$

The radius is the magnitude P .

9.5 Ellipse Intersecting Ellipse

Consider two ellipses in the same plane given by:

$$\mathbf{r}_j = \mathbf{k}_j + \mathbf{a}_j \cos(\phi_j) + \mathbf{b}_j \sin(\phi_j),$$

where $j = 1, 2$, and, as usual for an ellipse $\mathbf{a}_j \cdot \mathbf{b}_j = 0$. The condition for intersection of the two ellipses is

$$\begin{aligned}\mathbf{k}_1 + \mathbf{a}_1 \cos(\phi_1) + \mathbf{b}_1 \sin(\phi_1) \\ = \mathbf{k}_2 + \mathbf{a}_2 \cos(\phi_2) + \mathbf{b}_2 \sin(\phi_2).\end{aligned}$$

Then, multiply $\mathbf{a}_1, \mathbf{b}_1$, separately into the above to generate two results:

$$\begin{aligned}\mathbf{a}_1 \cdot \mathbf{k}_1 + a_1^2 \cos(\phi_1) &= \\ \mathbf{a}_1 \cdot \mathbf{k}_2 + \mathbf{a}_1 \cdot \mathbf{a}_2 \cos(\phi_2) + \mathbf{a}_1 \cdot \mathbf{b}_2 \sin(\phi_2) \\ \mathbf{b}_1 \cdot \mathbf{k}_1 + b_1^2 \sin(\phi_1) &= \\ \mathbf{b}_1 \cdot \mathbf{k}_2 + \mathbf{b}_1 \cdot \mathbf{a}_2 \cos(\phi_2) + \mathbf{b}_1 \cdot \mathbf{b}_2 \sin(\phi_2)\end{aligned}$$

Denoting

$$\begin{aligned}\Delta \mathbf{k} &= \mathbf{k}_2 - \mathbf{k}_1 \\ A_{jk} &= \mathbf{a}_j \cdot \mathbf{a}_k \\ B_{jk} &= \mathbf{b}_j \cdot \mathbf{b}_k \\ C_{jk} &= \mathbf{a}_j \cdot \mathbf{b}_k\end{aligned}$$

to keep the algebra tame, the above rearrange to

$$\begin{aligned}\cos(\phi_1) &= \frac{\mathbf{a}_1 \cdot \Delta \mathbf{k} + A_{12} \cos(\phi_2) + C_{12} \sin(\phi_2)}{a_1^2} \\ \sin(\phi_1) &= \frac{\mathbf{b}_1 \cdot \Delta \mathbf{k} + C_{21} \cos(\phi_2) + B_{12} \sin(\phi_2)}{b_1^2}.\end{aligned}$$

Use the fundamental trig identity $\sin^2(\phi_1) + \cos^2(\phi_1) = 1$ to condense the two equations back into one, with the only unknown being ϕ_2 . This is still a mess to solve, but can be done numerically in the general case, or analytically in certain special cases.

Problem 10

Consider the two ellipses

$$\begin{aligned}\mathbf{r}_1 &= a \cos(\phi_1) \hat{\mathbf{x}} + b \sin(\phi_1) \hat{\mathbf{y}} \\ \mathbf{r}_2 &= b \cos(\phi_2) \hat{\mathbf{x}} + a \sin(\phi_2) \hat{\mathbf{y}}.\end{aligned}$$

Calculate the distance D from any intersection point to the origin.

Answer: For this we have $A_{12} = B_{12} = ab$, with everything else zero. The situation is then governed by

$$1 = \left(\frac{b}{a} \cos(\phi_2)\right)^2 + \left(\frac{a}{b} \sin(\phi_2)\right)^2.$$

Solutions to the problem are then

$$\mathbf{r}_1 = \mathbf{r}_2 = \frac{ab}{\sqrt{a^2 + b^2}} (\pm \hat{\mathbf{x}} \pm \hat{\mathbf{y}}),$$

or

$$D = \frac{\sqrt{2}ab}{\sqrt{a^2 + b^2}}$$

9.6 Ellipse Intersecting Parabola

A parabolic curve

$$y(x) = Ax^2 + Bx + C$$

can be represented in vector notation via

$$\mathbf{y} = x \hat{\mathbf{x}} + y(x) \hat{\mathbf{y}}.$$

Let us find the intersection between such a parabola and the ellipse:

$$\mathbf{r} = \mathbf{r}_0 + \mathbf{a} \cos(\phi) + \mathbf{b} \sin(\phi)$$

The condition for intersection is $\mathbf{y} = \mathbf{r}$, or

$$x \hat{\mathbf{x}} + y(x) \hat{\mathbf{y}} = \mathbf{r}_0 + \mathbf{a} \cos(\phi) + \mathbf{b} \sin(\phi).$$

Next, multiply $\mathbf{a}$, $\mathbf{b}$, separately into the above to generate two results,

$$\begin{aligned}x \mathbf{a} \cdot \hat{\mathbf{x}} + y(x) \mathbf{a} \cdot \hat{\mathbf{y}} &= \mathbf{a} \cdot \mathbf{r}_0 + a^2 \cos(\phi) \\ x \mathbf{b} \cdot \hat{\mathbf{x}} + y(x) \mathbf{b} \cdot \hat{\mathbf{y}} &= \mathbf{b} \cdot \mathbf{r}_0 + b^2 \sin(\phi),\end{aligned}$$

or:

$$\begin{aligned}\cos(\phi) &= \frac{x a_x + y(x) a_y - \mathbf{a} \cdot \mathbf{r}_0}{a^2} \\ \sin(\phi) &= \frac{x b_x + y(x) b_y - \mathbf{b} \cdot \mathbf{r}_0}{b^2}\end{aligned}$$

As expected, we're left with another mess, but the ϕ -parameter could be eliminated in a way similar to the cases above to proceed with the general case.

For a special case, suppose $\mathbf{r}_0 = 0$, and let $\mathbf{a} = a \hat{\mathbf{x}}$, $\mathbf{b} = b \hat{\mathbf{y}}$. Then, we have

$$1 = \left(\frac{x}{a}\right)^2 + \left(\frac{y(x)}{b}\right)^2,$$

or all in terms of one variable,

$$1 = \left(\frac{x}{a}\right)^2 + \left(\frac{Ax^2 + Bx + C}{b}\right)^2.$$

9.7 Point Intersecting Triangle

Consider three points in Cartesian space that mark the vertices of a triangle

$$\mathbf{P}_j = x_j \hat{\mathbf{x}} + y_j \hat{\mathbf{y}} + z_j \hat{\mathbf{z}},$$

where $j = 1, 2, 3$. Also consider a fourth point in space

$$\mathbf{Q} = Q_x \hat{\mathbf{x}} + Q_y \hat{\mathbf{y}} + Q_z \hat{\mathbf{z}}$$

that may or may not be embedded in the triangle. The job is to find out whether this is so.

Improvising the Normal Vector

From the coordinates given, define a pair of tangent vectors $\mathbf{U}$, $\mathbf{V}$ such that

$$\begin{aligned}\mathbf{U} &= \mathbf{P}_2 - \mathbf{P}_1 \\ \mathbf{V} &= \mathbf{P}_3 - \mathbf{P}_1.\end{aligned}$$

Since both vectors are in the same plane, their cross product will by definition point perpendicular to the plane, which is a normal vector:

$$\mathbf{N} = \mathbf{U} \times \mathbf{V}$$

The details of the cross product are straightforward and aren't necessary to spell out here.

Improvising a Coordinate System

The pair of vectors $\mathbf{U}$, $\mathbf{V}$, despite not being mutually perpendicular, can *still* be used as the basis for a coordinate system embedded in the plane of the triangle. Indeed, any point $\mathbf{p}$ in the plane can be expressed as a linear combination of the basis vectors

$$\mathbf{p} = \alpha \mathbf{U} + \beta \mathbf{V},$$

where α (Greek 'alpha') and β (Greek 'beta') are two real-valued parameters.

Since the normal vector adds a third dimension to the picture, it follows that any point in the three-dimensional Cartesian space can be located by the vector

$$\mathbf{R} = \alpha \mathbf{U} + \beta \mathbf{V} + \gamma \mathbf{N}$$

using a third parameter γ .

In this improvised coordinate system, the point $\mathbf{P}_1$ is considered to be the origin. That is, if all parameters are zero, points $\mathbf{p}$ and $\mathbf{R}$ land back at $\mathbf{P}_1$.

The basis vectors need not be normalized, although it is good practice to do so. Either way, the parameters take care of all scaling, and the magnitudes U , V , N are simple to calculate when needed.

Intersection Condition

The point $\mathbf{Q}$, despite being handed to us in Cartesian coordinates, can also be represented by the vector $\mathbf{R}$ for some choice of α , β , γ . If $\mathbf{Q}$ is in the same plane as the three points $\mathbf{P}_j$ provided, then the γ -parameter should be zero.

Let us propose that point $\mathbf{Q}$ is located in the *normalized* $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$, $\hat{\mathbf{N}}$ system with new coefficients α , β , γ (that are different than those that occur in the non-normalized version). Tracing from the origin, we have

$$\mathbf{Q} = \mathbf{P}_1 + \alpha \hat{\mathbf{U}} + \beta \hat{\mathbf{V}} + \gamma \hat{\mathbf{N}},$$

and project $\hat{\mathbf{N}}$ into both sides to get

$$\mathbf{Q} \cdot \hat{\mathbf{N}} = \mathbf{P}_1 \cdot \hat{\mathbf{N}} + \alpha \hat{\mathbf{U}} \cdot \hat{\mathbf{N}} + \beta \hat{\mathbf{V}} \cdot \hat{\mathbf{N}} + \gamma \hat{\mathbf{N}} \cdot \hat{\mathbf{N}},$$

and this lets us solve for gamma, which ought to be zero if $\mathbf{Q}$ is within the triangle:

$$\gamma = \hat{\mathbf{N}} \cdot (\mathbf{Q} - \mathbf{P}_1)$$

Containment Condition

Supposing point $\mathbf{Q}$ satisfies $\gamma = 0$, the job now is to figure out whether $\mathbf{Q}$ is inside or outside the boundaries of the triangle defined by $\mathbf{P}_j$. Define a third tangent vector $\mathbf{W}$ that satisfies

$$\mathbf{U} + \mathbf{V} + \mathbf{W} = \mathbf{0},$$

which is like a trip around the triangle. In accordance with the right hand rule, this trip around the triangle should go in the counterclockwise direction.

Now comes the important observation. In order for $\mathbf{Q}$ to be inside the triangle, $\mathbf{Q}$ must occur to the 'left' of the line formed by each vector $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$ simultaneously. (We say 'left' and not 'right' due to the counterclockwise flow of the tangent vectors.)

To see this, write out

$$\mathbf{U} \times (\mathbf{Q} - \mathbf{P}_1) = \delta_U \hat{\mathbf{N}}$$

for some parameter δ_U . Since $\mathbf{Q} - \mathbf{P}_1$ also lies in the plane, the above simplifies

$$U |\mathbf{Q} - \mathbf{P}_1| \sin(\phi) = \delta_U,$$

where ϕ is the angle formed between $\mathbf{U}$ and $\mathbf{Q} - \mathbf{P}_1$.

If $\mathbf{Q}$ is to the 'left' of the line made by $\mathbf{U}$ as described, then $\sin(\phi)$ is a positive quantity, and δ_U is also positive. Repeat for the $\mathbf{V}$ - and $\mathbf{W}$ -cases to decide if the point is inside the triangle. All δ -parameters must be positive.

9.8 Plane from Three Points

Three points not on the same line define an infinite plane. From the points

$$\mathbf{P}_1 = 1 \hat{\mathbf{x}} + 0 \hat{\mathbf{y}} + 1 \hat{\mathbf{z}}$$

$$\mathbf{P}_2 = 0 \hat{\mathbf{x}} + 1 \hat{\mathbf{y}} + 1 \hat{\mathbf{z}}$$

$$\mathbf{P}_3 = 1 \hat{\mathbf{x}} + 1 \hat{\mathbf{y}} + 0 \hat{\mathbf{z}},$$

find the equation of the implied plane and report the result as

$$ax + by + cz + d = 0.$$

Determining the plane implied by three points is identical to finding all points $\mathbf{Q}$ coplanar to the triangle. Proceed by defining a pair of vectors $\mathbf{U}$, $\mathbf{V}$:

$$\mathbf{U} = \mathbf{P}_2 - \mathbf{P}_1 = -1 \hat{\mathbf{x}} + 1 \hat{\mathbf{y}} + 0 \hat{\mathbf{z}}$$

$$\mathbf{V} = \mathbf{P}_3 - \mathbf{P}_1 = 0 \hat{\mathbf{x}} + 1 \hat{\mathbf{y}} - 1 \hat{\mathbf{z}}$$

For a change of taste, calculate the normal vector using $\mathbf{P}_1$, and propose the vector $\mathbf{Q}$ is always in the plane if determinant/bracket notation:

$$\mathbf{N} = \mathbf{U} \times \mathbf{V} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ -1 & 1 & 0 \\ 0 & 1 & -1 \end{vmatrix} = \langle -1, -1, -1 \rangle$$

In other words, the normal vector is

$$\mathbf{N} = -1 \hat{\mathbf{x}} - 1 \hat{\mathbf{y}} - 1 \hat{\mathbf{z}} .$$

Next take any point known to be on the plane, such as

$$(\mathbf{Q} - \mathbf{P}_1) \cdot \mathbf{N} = 0$$

is satisfied. In particular

$$((x-1)\hat{\mathbf{x}} + (y-0)\hat{\mathbf{y}} + (z-1)) \cdot \hat{\mathbf{N}} = 0 ,$$

or, after simplifying:

$$x + y + z - 2 = 0$$

Chapter 19

Multivariate Calculus

1 Surfaces and Solids

Before getting into heavy jargon, we'll do a brief tour of the extension of the curve $y = f(x)$ into more dimensions.

1.1 Surfaces

The natural extension of a single-input function $f(x)$ is one that takes two arguments. In analogy to $y = f(x)$, we may also write

$$z = f(x, y),$$

where $f(x, y)$ requires two independent inputs x and y . The domain of f is part (or all) of the Cartesian plane on which x and y occur. The range variable z may be regarded as the 'height above' the plane at (x, y) .

If $f(x, y)$ is continuous in both variables, the set of all z -points constitutes a surface. In the same sense that a curve is a continuous arrangement of points in the Cartesian plane, surfaces may be like 'sheets' in a Cartesian volume.

Level Curves

Fixing z constant on a surface restricts the freedom in the xy -plane at height z to a *level curve*. A level curve is the same as a contour line seen on a topographical (not topological) map, or on various weather maps. A small ring surrounds a local minimum or a local maximum. Level curves intersect at a saddle point.

Critical Points

As two-dimensional creatures, surfaces have three kinds of critical points. A maxima in the surface is when both the x - and the y -variables reach a high point simultaneously. The same comment applies to a minima on the surface.

Another kind of critical point is called a *saddle point*, which has one of the variables x , y is a maxima, and the other a minima.

1.2 Solids

Adding another variable into the mix, we can have functions of three variables

$$F = f(x, y, z),$$

where F is most generally called a *scalar field*. The temperature or air density in a room qualifies as a scalar field. Holding any variable F constant produces a *level surface*, the generalization of the level curve.

We'll stave off the discussion of critical points within scalar fields, or if you see it coming, vector fields, until after a few developments are made.

Topological Remarks

If the field F describes a finite solid, then the notion of minima, maxima, and saddle points on the surface of the solid are given the respective labels 'pits', 'peaks', and 'passes'. It's possible to show using arguments from topology that the following is always true:

$$\text{peaks} - \text{passes} + \text{pits} = 2$$

To understand this, consider an ice cream cone with one scoop of ice cream, supposed spherical. There are two peaks in this situation - the top of the ice cream, and the bottom of the cone, so $2 = 2$ checks out. Press your thumb into the ice cream to introduce a pit and a pass simultaneously. Then you get $2 - 1 + 1 = 2$.

The relationship between critical points has an analogous formula with respect volumes with flat faces and sharp edges. If V is the number of vertices, E is the number of

edges, and F is the number of faces, it's possible to prove, much like the above:

$$V - E + F = 2$$

2 Multiple Integration

The workhorse of integral calculus is the fundamental theorem

$$f(x_1) - f(x_0) = \int_{x_0}^{x_1} f'(t) dt ,$$

where $f'(x)$ is the derivative df/dx .

The integral calculates the area under the curve $y(x) = f'(x)$ between the endpoints x_0, x_1 and hands us the answer in the form $A = f(x_1) - f(x_0)$. A concise way to express the area under such a curve is

$$A = \int_{x_0}^{x_1} y(x) dx .$$

2.1 Double Integral

The standard area integral calculates the sum of an infinite number of heights $y(x)$ above the x -axis. It stands to reason that $y(x)$ itself could be the result of an integral:

$$y(x) = \int_0^{y(x)} dy$$

The lower limit need not be zero if we take $y(x)$ as the vertical length trapped between two curves $y_0(x), y_1(x)$, or:

$$y(x) = \int_{y_0(x)}^{y_1(x)} dy$$

Inserting this form for $y(x)$ into the one dimensional area integral yields a *double integral*:

$$A = \int_{x_0}^{x_1} \int_{y_0(x)}^{y_1(x)} dy dx$$

Notice that the *inner* integration limits are functions of the *outer* integration variable.

Supposing $y_0(x), y_1(x)$ are easily inverted, the same area can be expressed with the integration variables reversed:

$$A = \int_{y_0}^{y_1} \int_{x_0(y)}^{x_1(y)} dx dy$$

Order of Integration

Note that in any multiple integral, it's always implied that the order of integration goes from the inner-most to the outer-most, not unlike like simplifying expressions with parentheses.

Integration Region

The curves $y_0(x), y_1(x)$ are called *bounding functions*. The same comment applies to their inverted counterparts $x_0(y), x_1(y)$.

The total information contained in the integration limits, including bounding functions, is called the *integration region*, denoted $\mathcal{D}$. With this we can express the double integral in a less definite form:

$$A = \iint_{\mathcal{D}} dx dy .$$

By obscuring the product $dx dy$ into an area element dA , the above can be written free of the Cartesian coordinate system:

$$A = \iint_{\mathcal{D}} dA$$

Volume Integral

The double integral apparatus can be used to calculate the volume trapped between the xy -plane and a given surface $z = f(x, y)$. For this, we simply write

$$V = \iint_{\mathcal{D}} f(x, y) dx dy .$$

Integrating Functions

Supposing we need to calculate something more abstract than an area, the double integral apparatus takes any reasonable function in its integrand

$$B = \iint_{\mathcal{D}} f(x, y) dx dy ,$$

where $f(x, y)$ is an arbitrary function (that can be integrated).

2.2 Polar Integral

To work a specific case, the differential area element in polar coordinates reads

$$dA = r dr d\theta ,$$

which means

$$A = \iint_{\mathcal{D}} r dr d\theta$$

Note that the r -variable is usually a function of θ , which means the r -integral should be solved first. In detail,

$$A = \int \left(\int r dr \right) d\theta ,$$

where the integration limits are left in indefinite form.

Now, even if there were an additional function $f(\theta)$ in the integrand, thus changing the integral to

$$B = \int \left(\int r \, dr \right) f(\theta) \, d\theta,$$

notice how $f(\theta)$ has no r -dependence, and is situated outside the r -integral.

We're thus free to evaluate the inner r -integral, and the above becomes

$$B = \frac{1}{2} \int r^2 f(\theta) \, d\theta.$$

This is a nifty result, as if $f(\theta)$ weren't there, then the integral is simply the area under $r(\theta)$ in polar coordinates:

$$A = \frac{1}{2} \int r^2 \, d\theta$$

Area of a Circular Arc

Calculating the area of the circular arc in polar coordinates is as easy as

$$A = \int_{\theta_0}^{\theta_1} \int_0^R r \, dr \, d\theta = \frac{1}{2} (\theta_1 - \theta_0) R^2,$$

which is the entire story.

For a bit of self-torture, let us do the same calculation in Cartesian coordinates using double integration. To set up, consider a pair of points (x_0, y_0) , (x_1, y_1) with $x_0 > x_1$ that define the endpoints of the arc:

$$\begin{aligned} R \cos(\theta_0) &= x_0 \\ R \sin(\theta_0) &= y_0 \\ R \cos(\theta_1) &= x_1 \\ R \sin(\theta_1) &= y_1 \end{aligned}$$

Note also that the points (x_0, y_0) , (x_1, y_1) define straight lines through the origin (no y -intercept), with respective slopes:

$$\begin{aligned} m_0 &= \frac{y_0}{x_0} \\ m_1 &= \frac{y_1}{x_1} \end{aligned}$$

The area of the circular arc will be calculated in two parts, one that resolves to a triangle with all straight edges,

and another to handle the curved region. Working this out carefully, find:

$$A = \int_0^{x_1} \int_{m_0 x}^{m_1 x} dy \, dx + \int_{x_1}^{x_0} \int_{m_0 x}^{\sqrt{R^2 - x^2}} dy \, dx$$

The first integral can be evaluated fully with ease, but the second needs to be chipped away at. Doing a round of simplifying, reach the intermediate step

$$\begin{aligned} A &= \frac{x_1^2}{2} (m_1 - m_0) \\ &\quad + \int_{x_0}^{x_1} \sqrt{R^2 - x^2} \, dx \\ &\quad - \frac{m_0}{2} (x_0^2 - x_1^2). \end{aligned}$$

The remaining integral can be solved with a sine substitution, which has the general solution:

$$\begin{aligned} \int \sqrt{R^2 - x^2} \, dx &= \frac{x \sqrt{R^2 - x^2}}{2} \\ &\quad + \frac{R^2}{2} \arctan \left(\frac{x}{\sqrt{R^2 - x^2}} \right) + C \end{aligned}$$

For the problem on hand, this means

$$\begin{aligned} \int \sqrt{R^2 - x^2} \, dx &= \frac{m_0 x_0^2 - m_1 x_1^2}{2} \\ &\quad + \frac{R^2}{2} \left(\arctan \left(\frac{x_0}{y_0} \right) - \arctan \left(\frac{x_1}{y_1} \right) \right). \end{aligned}$$

The x - and m -terms all cancel, and the area reads

$$A = \frac{R^2}{2} \left(\arctan \left(\frac{x_0}{y_0} \right) - \arctan \left(\frac{x_1}{y_1} \right) \right).$$

From trigonometry, note in general that

$$\arctan(\cot(u)) = \frac{\pi}{2} - u,$$

and the area becomes

$$A = \frac{R^2}{2} \left(\frac{\pi}{2} - \theta_0 - \frac{\pi}{2} + \theta_1 \right),$$

and finally,

$$A = \frac{1}{2} (\theta_1 - \theta_0) R^2,$$

in agreement with the previous answer.

Gaussian Integral

Consider the definite integral

$$I_1 = \int_{-\infty}^{\infty} e^{-ax^2} dx,$$

which has no elementary solution. Instead of turning to a numerical approximation, which would ordinarily be the case for such an integral, consider the same exact integral with a swap of variables:

$$I_1 = \int_{-\infty}^{\infty} e^{-ay^2} dy$$

If this analysis doesn't seem insane yet, multiply each copy of the integral together to get to what seems like a dead end,

$$I_1^2 = \left(\int_{-\infty}^{\infty} e^{-ax^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-ay^2} dy \right),$$

and melt the notation down:

$$I_1^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-a(x^2+y^2)} dx dy$$

Now, one must be very careful when doing this, but it just happens that the x, y variables can be regarded as locations on the Cartesian plane, which lends to polar coordinates. Switching to polar, the above integral is

$$I_1^2 = \int_0^{2\pi} \int_0^{\infty} e^{-a\rho^2} \rho d\rho d\phi.$$

Observe how the region of integration (the infinite plane) makes the limits on each integral easy to write.

The ϕ -integral is trivial and resolves to 2π . The remaining ρ -integral can be solved straightforwardly by u -substitution. Chugging through each, we find $I_1^2 = \pi/a$, or

$$I_1 = \int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}.$$

This cheat works on several I_1 -like problems called *Gaussian integrals*. Let us work through

$$I_2 = \int_{-\infty}^{\infty} e^{-ax^2+bx} dx.$$

Completing the square within the exponential leads to:

$$I_2 = e^{b^2/4a} \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\frac{\pi}{a}} e^{b^2/4a}$$

Yet another Gaussian integral

$$I_3 = \int_{-\infty}^{\infty} x^2 e^{-ax^2} dx$$

can be solved by taking the derivative of $-I_1$ with respect to a (not x). In detail:

$$\frac{d}{da} (-I_1) = I_3 = -\frac{d}{da} \left(\sqrt{\frac{\pi}{a}} \right) = \sqrt{\frac{\pi}{4a^3}}$$

2.3 Triple Integral

The apparatus for multiple integration readily generalizes for three and more dimensions. For three dimensions, the idea of a bounding function becomes a *bounding surface*, and we have a *triple integral*:

$$V = \int_{x_0}^{x_1} \int_{y_0(x)}^{y_1(x)} \int_{z_0(x,y)}^{z_1(x,y)} dz dy dx$$

In the special case $z_0(x, y) = 0$, the above reduces to a standard volume integral.

The order of integration has greater significance as the dimension of the integral increases. Supposing the required bounding functions and surfaces are easily attained, the triple integral can be written several more ways, for instance:

$$V = \int_{y_0}^{y_1} \int_{z_0(y)}^{z_1(y)} \int_{x_0(y,z)}^{x_1(y,z)} dx dz dy$$

$$V = \int_{z_0}^{z_1} \int_{x_0(z)}^{x_1(z)} \int_{y_0(z,x)}^{y_1(z,x)} dy dx dz$$

Or, in terms of an integration region:

$$V = \iiint_{\mathcal{D}} dx dy dz$$

Of course, the triple integral can involve functions in the integrand. For a three-variable function $f(x, y, z)$, sometimes called a *scalar field*, we can calculate things like

$$B = \iiint_{\mathcal{D}} f(x, y, z) dx dy dz.$$

Non-Cartesian Volume Elements

To go from three-dimensional Cartesian coordinates to a different system, the volume element and specification of the integration need to be changed. For cylindrical coordinates, we may have

$$V = \int_{z_0}^{z_1} \int_{\phi_0(z)}^{\phi_1(z)} \int_{\rho_0(\phi,z)}^{\rho_1(\phi,z)} \rho d\rho d\phi dz,$$

and for spherical coordinates:

$$V = \int_{\phi_0}^{\phi_1} \int_{\theta_0(\phi)}^{\theta_1(\phi)} \int_{r_0(\theta, \phi)}^{r_1(\theta, \phi)} r^2 \sin(\theta) dr d\theta d\phi$$

Like the two-dimensional case, the volume element and integration region can be generalized via:

$$V = \iiint_{\mathcal{D}} dV$$

Problem 1

In cylindrical coordinates given by

$$\begin{aligned} x(\rho, \phi, z) &= \rho \cos(\phi) \\ y(\rho, \phi, z) &= \rho \sin(\phi) \\ z(\rho, \phi, z) &= z, \end{aligned}$$

show that the volume element dV is given by:

$$dV = dx dy dz = \rho d\rho d\phi dz$$

Problem 2

In spherical coordinates given by

$$\begin{aligned} x &= r \sin(\theta) \cos(\phi) \\ y &= r \sin(\theta) \sin(\phi) \\ z &= r \cos(\theta), \end{aligned}$$

show that the volume element dV is given by:

$$dV = dx dy dz = r^2 \sin(\theta) dr d\theta d\phi$$

Hurricane Problem

Problem 3

In a simplified model of a hurricane, the velocity of the wind is taken to be purely in the circumferential direction and of magnitude

$$v(\rho, z) = \Omega \rho e^{-z/h - \rho/a},$$

where ρ and z are cylindrical coordinates measured from the eye of the hurricane at sea level, and Ω, h, a are positive constants. The density of the atmosphere is approximated by

$$d(z) = d_0 e^{-z/h}.$$

Find the total kinetic energy of the motion.

Answer: As an integral, the kinetic energy is given by

$$T = \int \frac{1}{2} v^2 dm.$$

This can be converted to a volume integral via the relation

$$\frac{dm}{dV} = d(z),$$

where dV is the volume element in cylindrical coordinates. The kinetic energy integral becomes

$$T = \int_0^\infty \int_0^{2\pi} \int_0^\infty \frac{1}{2} d(z) (v(\rho, z))^2 dV,$$

or

$$T = \frac{d_0 \Omega^2}{2} \int_0^\infty \int_0^{2\pi} \int_0^\infty e^{-z/h} \rho^2 e^{-2z/h - 2\rho/a} \rho d\rho d\phi dz.$$

The integral can be broken apart into three separate integrals

$$T = \frac{d_0 \Omega^2}{2} \left(\int_0^\infty e^{-3z/h} dz \right) \left(\int_0^{2\pi} d\phi \right) \left(\int_0^\infty \rho^3 e^{-2\rho/a} d\rho \right),$$

each straightforwardly evaluated:

$$T = \frac{d_0 \Omega^2}{2} \left(\frac{h}{3} \right) (2\pi) \left(\frac{3a^4}{8} \right) = \frac{\pi}{8} \Omega^2 d_0 h a^4$$

3 Partial Derivative

Returning to the definition of a function, recall that a function f depends on an input variable x in the function's domain. Given any input value, the output of the function is written $y = f(x)$, and there is only one y for a given x . The set of all y -values constitute the function's range. A 'curve' given by function $y = f(x)$ may exhibit a myriad of features: asymptotic behavior, periodicity, singularities, critical points, inflection points, etc.

Derivative

One star result from the analysis of curves gives the slope of the function at a point x_0 , namely

$$\frac{d}{dx}f(x_0) = \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Taylor's Theorem

The so-called derivative turns out to be just the first-order version of something more general we known as Taylor's theorem. Near the point x_0 , we have:

$$f(x) \approx f(x_0) + \sum_{q=1}^{\infty} \frac{1}{q!} f^{(q)}(x_0) (x - x_0)^q$$

3.1 Slope on a Surface

The technical definition of the derivative generalizes to surfaces. For this, we require the surface $z = f(x, y)$ to be differentiable both in the x -direction and the y -direction, meaning that the slope of the surface at a point (x_0, y_0) has two answers: a slope along x , and a slope along y .

To express the slope on a surface at the point (x_0, y) , we write the usual slope formula treating x as the 'active' variable with y as a constant:

$$\frac{\partial}{\partial x}f(x_0, y) = \lim_{x \rightarrow x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} \quad (19.1)$$

Meanwhile, the slope at the point (x, y_0) allows y to vary while x is constant:

$$\frac{\partial}{\partial y}f(x, y_0) = \lim_{y \rightarrow y_0} \frac{f(x, y) - f(x, y_0)}{y - y_0} \quad (19.2)$$

The familiar d/dx -notation is replaced by $\partial/\partial x$. The symbol ∂ denotes the *partial derivative*.

3.2 Mixed Partial Derivatives

One issue that needs to be settled right away is the idea of *mixed* partial derivatives. For the surface $z = f(x, y)$, let us find out whether

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \frac{\partial}{\partial x} \left(\frac{\partial z}{\partial y} \right) \quad (19.3)$$

is true. Using brute force, start with

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \frac{\partial}{\partial y} \left(\lim_{x \rightarrow x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} \right),$$

which becomes

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \lim_{y \rightarrow y_0} \lim_{x \rightarrow x_0} \frac{f(x, y) - f(x_0, y) - f(x, y_0) + f(x_0, y_0)}{(y - y_0)(x - x_0)}.$$

Now, it takes little to imagine doing a similar calculation with the y -partial derivative first to have

$$\frac{\partial}{\partial x} \left(\frac{\partial z}{\partial y} \right) = \frac{\partial}{\partial x} \left(\lim_{y \rightarrow y_0} \frac{f(x, y) - f(x, y_0)}{y - y_0} \right),$$

which then simplifies to something nearly identical to the above, save one difference, which that the *order* of the limits is swapped. The task boils down to showing in this context that

$$\lim_{x \rightarrow x_0} \lim_{y \rightarrow y_0} \leftrightarrow \lim_{y \rightarrow y_0} \lim_{x \rightarrow x_0}$$

can be assumed.

To prove this, define two new functions

$$\begin{aligned} X(x, y) &= f(x, y) - f(x_0, y) \\ Y(x, y) &= f(x, y) - f(x, y_0), \end{aligned}$$

and notice the following equality:

$$X(x, y) - X(x, y_0) = Y(x, y) - Y(x_0, y)$$

The left- and right-hand sides of the above each represent the endpoints of a secant line on the surface. By the mean value theorem, each can be replaced by partial derivatives as

$$(y - y_0) \frac{\partial}{\partial y} X(x, b) = (x - x_0) \frac{\partial}{\partial x} Y(a, y),$$

where

$$\begin{aligned} x_0 &< a < x \\ y_0 &< b < y. \end{aligned}$$

Keep simplifying to write

$$(y - y_0) \left(\frac{\partial}{\partial y} f(x, b) - \frac{\partial}{\partial y} f(x_0, b) \right) = \\ (x - x_0) \left(\frac{\partial}{\partial x} f(a, y) - \frac{\partial}{\partial x} f(a, y_0) \right),$$

and use the mean value theorem a second time on each side to write

$$(y - y_0) (x - x_0) \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} f(\alpha, b) \right) = \\ (x - x_0) (y - y_0) \frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} f(a, \beta) \right),$$

where

$$x_0 < \alpha < x \\ y_0 < \beta < y.$$

Closing the limits tighter, we see that the pair a, α tend to x_0 , and also the pair b, β tend to y_0 . In the differential limit, the left and right sides are equal and the proof is done.

3.3 Partial Derivative Operator

Like the ordinary derivative operator d/dx , the partial derivative operator is written $\partial/\partial x$. For shorthand, the same operator is often written with one ‘partial’ symbol and a subscript:

$$\frac{\partial}{\partial x} = \partial_x$$

In this notation, the mixed partial derivative Equation (19.3) is simply written

$$\partial_{yx} f(x, y) = \partial_{xy} f(x, y),$$

or with just the operators,

$$\partial_{yx} = \partial_{xy}.$$

Yet another nomenclature for partial derivatives involves placing a subscript with the function itself:

$$\frac{\partial}{\partial x} f(x, y) = f_x$$

Second Derivative

With the notion of partial derivatives, the idea of the second derivative of a function can go three ways. Each of the following is a second derivative operator

$$\partial_{xx} \quad \partial_{xy} \quad \partial_{yy}$$

and each produces, in the general case, a different result.

The partial derivative operator obeys the same algebraic rules as the ordinary derivative operator. Without abusing the notation, we can establish things like:

$$(\partial_x + \partial_y)(\partial_x - \partial_y) = \partial_{xx} - \partial_{xy} + \partial_{yx} - \partial_{yy} \\ = \partial_{xx} - \partial_{yy}$$

Third Derivative

The equivalency of the mixed partial derivative extends to any depth. From Equation (19.3), we reason that

$$\partial_{yxx} \quad \partial_{xyx} \quad \partial_{xxy}$$

yield the same result. For this reason, it turns out there four *unique* third derivative operations:

$$\partial_{xxx} \quad \partial_{xxy} \quad \partial_{yyx} \quad \partial_{yyy}$$

The notation can be condensed once more by using exponent notation on repeated derivatives:

$$\partial_{xxx} = \partial_x^3 \\ \partial_{xxy} = \partial_x^2 \partial_y \\ \partial_{yyx} = \partial_{xy}^2 \\ \partial_{yyy} = \partial_y^3$$

3.4 Total Derivative

The notion of ‘regular’ derivative still survives the jump to more dimensions, and is given the name *total* derivative.

For a function $f(x, y, z)$ the total derivative with respect to a variable t sums across each partial derivative:

$$\frac{d}{dt} f(x, y, z) = \frac{\partial f}{\partial x} \frac{dx}{dt} + \frac{\partial f}{\partial y} \frac{dy}{dt} + \frac{\partial f}{\partial z} \frac{dz}{dt}$$

The Differential

Stripping away the dt -variable by the chain rule yields the so-called ‘differential of’ $f(x, y, z)$:

$$df(x, y, z) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz$$

3.5 Jacobian Matrix

Consider the generic volume integral

$$V = \iiint_{\mathcal{D}} f(x, y, z) dx dy dz$$

and introduce the coordinates u, v, w which relate to x, y, z in the most general way possible:

$$x = x(u, v, w)$$

$$y = y(u, v, w)$$

$$z = z(u, v, w)$$

(You may, for instance, think of u, v, w as r, θ, ϕ in spherical coordinates.) The goal is to cast the whole integral in the variables u, v, w .

Surely we know the domain will change via $\mathcal{D} \rightarrow \mathcal{D}'$, and the embedded function $f(x, y, z)$, after substitution, becomes a function $f(u, v, w)$. The volume element $dx dy dz$ doesn't simply change to $du dv dw$, rather an extra factor is required:

$$dx dy dz = |\det J| du dv dw,$$

where $|\det J|$ is the absolute value of the determinant of the *Jacobian matrix* J given by:

$$J = \begin{bmatrix} \frac{\partial(x,y,z)}{\partial(u,v,w)} \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{bmatrix}$$

Putting it all together, the volume integral changes to

$$V = \iiint_{\mathcal{D}'} f(u, v, w) |\det J| du dv dw.$$

The reason *why* we need the determinant of the Jacobian matrix is beyond the scope of this chapter. A study of calculus on manifolds via Riemannian geometry avails a simple proof of the matter.

3.6 Variable Integration Limits

It's possible for an integral to have variable limits, which can make a mess of things like integration by parts. Consider a function $y(x)$ and a two-variable function $f(x, y(x))$. By the fundamental theorem, this setup implies integrals of the form

$$F(x) = \int_{a(x)}^{b(x)} f(x, t) dt.$$

To attack this, define a helper function

$$G(x, y) = \int_{t_0}^y f(x, t) dt,$$

so then $F(x)$ reads

$$\begin{aligned} F(x) &= \int_{t_0}^{b(x)} f(x, t) dt - \int_{t_0}^{a(x)} f(x, t) dt \\ &= G(x, b(x)) - G(x, a(x)). \end{aligned}$$

Take the total derivative of $F(x)$:

$$\begin{aligned} \frac{d}{dx} F(x) &= \frac{\partial}{\partial x} G(x, b(x)) - \frac{\partial}{\partial x} G(x, a(x)) \\ &\quad + \frac{\partial}{\partial y} G(x, b(x)) \frac{db}{dx} \\ &\quad - \frac{\partial}{\partial y} G(x, a(x)) \frac{da}{dx} \end{aligned}$$

The first two terms combine to make $\partial F / \partial x$. To handle the $\partial G / \partial y$ factors, define Δy such that

$$G(x, y + \Delta y) = \int_{t_0}^{y+\Delta y} f(x, t) dt$$

Unpack the right side and divide through by Δy to find

$$\frac{G(x, y + \Delta y) - G(x, y)}{\Delta y} = \frac{1}{\Delta y} \int_y^{y+\Delta y} f(x, t) dt.$$

In the limit $\Delta y \rightarrow 0$, the left side is $\partial G / \partial y$. The right simplifies to $f(x, y)$. In other words:

$$\frac{\partial}{\partial y} G(x, y) = f(x, y)$$

Putting the whole answer together, we have found

$$\begin{aligned} \frac{d}{dx} \int_{a(x)}^{b(x)} f(x, t) dt &= \\ &= f(x, b(x)) \frac{db}{dx} - f(x, a(x)) \frac{da}{dx} \\ &\quad + \frac{\partial}{\partial x} \int_{a(x)}^{b(x)} f(x, t) dt \end{aligned} \quad (19.4)$$

as the chief result, called the *Leibniz integral rule*.

Finally, since the operator ∂_x doesn't touch the y -variable, the last integral obeys:

$$\frac{\partial}{\partial x} \int_{a(x)}^{b(x)} f(x, t) dt = \int_{a(x)}^{b(x)} \frac{\partial}{\partial x} f(x, t) dt$$

Problem 4

Show the following:

$$\frac{d}{dx} \int_0^x e^{xt^2} dt = e^{x^3} + \int_0^x \frac{\partial}{\partial x} e^{xt^2} dt$$

3.7 Two-Variable Taylor's Theorem

Taylor's theorem generalizes readily to surfaces. To get started, consider a fixed point (x_0, y_0) in the domain of a surface $z = f(x, y)$. Deviations from the fixed point are tracked by two quantities

$$\begin{aligned}\Delta x &= x - x_0 \\ \Delta y &= y - y_0.\end{aligned}$$

Like the one-dimensional Taylor's theorem, we're allowed to frame the final answer as an infinite sum. To do this, first notice that the one-dimensional case contains every whole number power of the quantity $x - x_0 = \Delta x$. Then, for two dimensions, we ought to need every whole number power of $\Delta x \Delta y$.

Of course, the factor $f^{(q)}(x_0)/q!$ that appears in the one-dimensional case can't work for two dimensions. Without knowing what to replace this with, let a set of unknown coefficients $\{C_{jk}\}$ stand in for now. With all this, the two-dimensional Taylor's theorem looks like:

$$\begin{aligned}f(x, y) &\approx f(x_0, y_0) \\ &+ C_{10}\Delta x + C_{01}\Delta y \\ &+ C_{20}\Delta x^2 + C_{11}\Delta x\Delta y + C_{02}\Delta y^2 \\ &+ C_{30}\Delta x^3 + C_{21}\Delta x^2\Delta y + C_{12}\Delta x\Delta y^2 + C_{03}\Delta y^3 \\ &+ \dots\end{aligned}$$

Now we have the problem of determining each unknown coefficient C_{jk} . Begin by applying the ∂_x operator across the whole equation, and then evaluate the equation at (x_0, y_0) . With almost no effort, we can see that any terms containing Δx^2 or higher power will zero out, and the whole result is

$$\partial_x f(x_0, y_0) = C_{10}.$$

Applying the ∂_y instead and doing the exercise again leads to a similar result

$$\partial_y f(x_0, y_0) = C_{01}.$$

For the next 'row' of coefficients, apply the ∂_{xx} , ∂_{xy} , ∂_{yy} operators respectively, and evaluate at (x_0, y_0) . This saps all but the order-two terms in the equation, from which we find:

$$\begin{aligned}\partial_{xx} f(x_0, y_0) &= 2 \cdot C_{20} \\ \partial_{xy} f(x_0, y_0) &= C_{11} \\ \partial_{yy} f(x_0, y_0) &= 2 \cdot C_{02}\end{aligned}$$

Solving for the order-three coefficients means using the four operators ∂_{x^3} , ∂_{x^2y} , ∂_{xy^2} , ∂_{y^3} to $f(x, y)$ and evaluate at (x_0, y_0) . This gives four new equations:

$$\begin{aligned}\partial_{x^3} f(x_0, y_0) &= 3 \cdot 2 \cdot C_{30} \\ \partial_{x^2y} f(x_0, y_0) &= 2 \cdot C_{21} \\ \partial_{xy^2} f(x_0, y_0) &= 2 \cdot C_{12} \\ \partial_{y^3} f(x_0, y_0) &= 3 \cdot 2 \cdot C_{03}\end{aligned}$$

To summarize and condense notation once more, use using the general shorthand

$$z_{x^j y^k} = \partial_{x^j y^k} f(x_0, y_0),$$

and we have found

$$\begin{aligned}C_{10} &= z_x \\ C_{01} &= z_y \\ C_{20} &= z_{x^2}/2 \\ C_{11} &= z_{xy} \\ C_{02} &= z_{y^2}/2 \\ C_{30} &= z_{x^3}/3! \\ C_{21} &= z_{x^2y}/2 \\ C_{12} &= z_{xy^2}/2 \\ C_{03} &= z_{y^3}/3!\end{aligned}$$

The two-dimensional Taylor's theorem now looks like:

$$\begin{aligned}f(x, y) &\approx f(x_0, y_0) \\ &+ z_x \Delta x + z_y \Delta y \\ &+ \frac{z_{x^2} \Delta x^2 + 2z_{xy} \Delta x \Delta y + z_{y^2} \Delta y^2}{2} \\ &+ \frac{z_{x^3} \Delta x^3 + 3z_{x^2y} \Delta x^2 \Delta y + 3z_{xy^2} \Delta x \Delta y^2 + z_{y^3} \Delta y^3}{3!} \\ &+ \dots\end{aligned}$$

Look for a moment at the pattern in the numerical coefficients in the numerator of each term written so far. Jotting these down:

$$\begin{array}{cccc} & & 1 & \\ & 1 & & 1 \\ & 1 & 2 & 1 \\ 1 & 3 & 3 & 1\end{array}$$

The pattern is clearly that of the binomial coefficients, i.e. the entries of Pascal's triangle. This means that the terms

in the infinite sum can be regrouped as binomials with the help of the partial derivative operator. For instance, the order-two terms are written

$$\begin{aligned} z_{xx^2}\Delta x^2 + 2z_{xy}\Delta x\Delta y + z_{yy^2}\Delta y^2 \\ = (\Delta x \partial_x + \Delta y \partial_y)^2 f(x_0, y_0) , \end{aligned}$$

and similarly for all orders.

Switching to summation notation, we finally have the

two-dimensional Taylor's theorem

$$\begin{aligned} f(x, y) \approx f(x_0, y_0) \\ + \sum_{q=1}^{\infty} \frac{1}{q!} (\Delta x \partial_x + \Delta y \partial_y)^q f(x_0, y_0) \end{aligned} \tag{19.5}$$

For a sanity check, you can see that if all $y = 0$ then the above reduces to the familiar one-dimensional form.

4 Vectors and Surfaces

In three-dimensional space, there are always three basis vectors from which everything is oriented. In Cartesian coordinates, these are just $\hat{x}$, $\hat{y}$, $\hat{z}$, and are fixed in space.

4.1 Basis Vectors as Derivatives

In non-Cartesian coordinate systems, such as cylindrical coordinates $\hat{\rho}$, $\hat{\phi}$, $\hat{z}$, and spherical coordinates $\hat{r}$, $\hat{\theta}$, $\hat{\phi}$, each basis vector depends on the coordinates themselves.

In each system mentioned, the respective position vector is:

$$\begin{aligned}\mathbf{r} &= x \hat{x} + y \hat{y} + z \hat{z} \\ \mathbf{r} &= \rho \cos(\phi) \hat{x} + \rho \sin(\phi) \hat{y} + z \hat{z} \\ \mathbf{r} &= r \sin(\theta) (\cos(\phi) \hat{x} + \sin(\phi) \hat{y}) + r \cos(\theta) \hat{z}\end{aligned}$$

Of course, one could manually invert the above to attain the basis vectors for each system, namely $\hat{\rho}$, $\hat{\phi}$, $\hat{z}$, and also $\hat{r}$, $\hat{\theta}$, $\hat{\phi}$.

Having suffered the tedious derivations once, you're entitled to a secret from the math department. Let q represent any parameter whatsoever - it could be x , or z , or ϕ , etc. It turns out that the basis vector $\hat{q}$ is the normalized q -derivative of the position vector. That is:

$$\hat{q} = \frac{1}{|\partial \mathbf{r} / \partial q|} \frac{\partial \mathbf{r}}{\partial q} \quad (19.6)$$

For example, if we want $\hat{\theta}$ from spherical coordinates, write

$$\frac{\partial \mathbf{r}}{\partial \theta} = r \cos(\theta) (\cos(\phi) \hat{x} + \sin(\phi) \hat{y}) - r \sin(\theta) \hat{z},$$

whose magnitude is r . Dividing this out delivers the result promised:

$$\frac{1}{r} \frac{\partial \mathbf{r}}{\partial \theta} = \hat{\theta}$$

4.2 Surface Tangent Vectors

In the same way that curves $y = f(x)$ can be represented with vectors and parameters, the story is similar for surfaces $z = f(x, y)$. In a generic case, a surface requires two parameters u, v such that

$$\mathbf{r}(u, v) = x(u, v) \hat{x} + y(u, v) \hat{y} + z(u, v) \hat{z},$$

which doesn't necessarily need to be framed in the Cartesian system.

Parametric Surface Tangents

Choosing any fixed point (u_0, v_0) on a parameterized surface, there exist a pair of embedded tangent vectors we'll call $\mathbf{u}, \mathbf{v}$ straightforwardly calculated directly from $\mathbf{r}(u, v)$:

$$\begin{aligned}\mathbf{u}(u_0, v_0) &= \left(\frac{\partial}{\partial u} \mathbf{r}(u, v_0) \right) \Big|_{u_0} \\ \mathbf{v}(u_0, v_0) &= \left(\frac{\partial}{\partial v} \mathbf{r}(u_0, v) \right) \Big|_{v_0}\end{aligned}$$

Like all vectors, the tangents $\mathbf{u}, \mathbf{v}$ can be converted to normal vectors by dividing out the magnitude:

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{u} / |\mathbf{u}| \\ \hat{\mathbf{v}} &= \mathbf{v} / |\mathbf{v}|\end{aligned}$$

Level Curve Tangents

The tangent vectors to a level curve of $z = f(x, y)$ are trickier to determine. To begin, propose choose a point (x_0, y_0) and write the pair of vectors

$$\begin{aligned}\mathbf{u}(x_0, y_0) &= u_x \hat{x} + u_z \hat{z} \\ \mathbf{v}(x_0, y_0) &= v_y \hat{y} + v_z \hat{z},\end{aligned}$$

where without loss of generality, $\mathbf{u}$ lacks a y -component and $\mathbf{v}$ lacks an x -component.

The ratios u_z/u_x , v_z/v_y , receptively, are the partial derivatives in disguise, as

$$\begin{aligned}\frac{u_z}{u_x} &= \left(\frac{\partial}{\partial x} f(x, y_0) \right) \Big|_{x_0} \\ \frac{v_z}{v_y} &= \left(\frac{\partial}{\partial y} f(x_0, y) \right) \Big|_{y_0},\end{aligned}$$

which allows the vectors $\mathbf{u}, \mathbf{v}$ to be written in terms of partial derivatives:

$$\begin{aligned}\mathbf{u}(x_0, y_0) &= u_x \left(\hat{x} + \left(\frac{\partial}{\partial x} f(x, y_0) \right) \Big|_{x_0} \hat{z} \right) \\ \mathbf{v}(x_0, y_0) &= v_y \left(\hat{y} + \left(\frac{\partial}{\partial y} f(x_0, y) \right) \Big|_{y_0} \hat{z} \right)\end{aligned}$$

For shorthand, denote the fully-evaluated partial derivatives as $f_x(x_0, y_0)$, $f_y(x_0, y_0)$, respectively. Dividing each vector by its own magnitude gives the normalized version

of each:

$$\hat{\mathbf{u}}_{z_0} = \frac{\hat{\mathbf{x}} + f_x \hat{\mathbf{z}}}{\sqrt{1 + f_x^2}}$$

$$\hat{\mathbf{v}}_{z_0} = \frac{\hat{\mathbf{y}} + f_y \hat{\mathbf{z}}}{\sqrt{1 + f_y^2}}$$

Problem 5

The notion of level curve tangent vectors was developed assuming a surface $z = f(x, y)$. Repeat the derivation for the other two cases, i.e. $y = g(x, z)$ and $x = h(y, z)$.

Answer:

$$\hat{\mathbf{u}}_{y_0} = \frac{\hat{\mathbf{x}} + g_x \hat{\mathbf{y}}}{\sqrt{1 + g_x^2}}$$

$$\hat{\mathbf{v}}_{y_0} = \frac{\hat{\mathbf{z}} + g_z \hat{\mathbf{y}}}{\sqrt{1 + g_z^2}}$$

$$\hat{\mathbf{u}}_{x_0} = \frac{\hat{\mathbf{y}} + h_y \hat{\mathbf{x}}}{\sqrt{1 + h_y^2}}$$

$$\hat{\mathbf{v}}_{x_0} = \frac{\hat{\mathbf{z}} + h_z \hat{\mathbf{x}}}{\sqrt{1 + h_z^2}}$$

4.3 Surface Normal Vector

With a pair of surface tangent vectors $\mathbf{u}, \mathbf{v}$ in hand for a given point, the cross product of the two yields the vector $\mathbf{n}$ that is normal to the surface:

$$\mathbf{n} = \mathbf{u} \times \mathbf{v}$$

Parametric Surface Normal

For the parametric surface $\mathbf{r}(u, v)$, the surface normal is straightforwardly calculated from

$$\mathbf{n}(u_0, v_0) = \mathbf{u}(u_0, v_0) \times \mathbf{v}(u_0, v_0) ,$$

which suggests a normalized version

$$\hat{\mathbf{n}} = \frac{\mathbf{u}(u_0, v_0) \times \mathbf{v}(u_0, v_0)}{|\mathbf{u}(u_0, v_0) \times \mathbf{v}(u_0, v_0)|} .$$

Of course, there is no need to normalize if we use unit vectors only:

$$\hat{\mathbf{n}} = \hat{\mathbf{u}} \times \hat{\mathbf{v}}$$

Cartesian Surface Normal

The normal vector to the surface $z = f(x, y)$ at a point (x_0, y_0) is the cross product of the tangent vectors $\mathbf{u}(x_0, y_0)$, $\mathbf{v}(x_0, y_0)$. Explicitly, this is:

$$\mathbf{n}(x_0, y_0) = \mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ u_x & 0 & u_x f_x \\ 0 & v_y & v_y f_y \end{vmatrix} ,$$

or

$$\mathbf{n} = u_x v_y (-f_x \hat{\mathbf{x}} - f_y \hat{\mathbf{y}} + \hat{\mathbf{z}}) .$$

Eliminate the stray coefficients by normalizing:

$$\hat{\mathbf{n}}_{z_0} = \frac{-f_x \hat{\mathbf{x}} - f_y \hat{\mathbf{y}} + \hat{\mathbf{z}}}{\sqrt{1 + f_x^2 + f_y^2}}$$

Problem 6

Attain the unit normal vector to the surfaces $y = g(x, z)$ and $x = h(y, z)$.

Answer:

$$\hat{\mathbf{n}}_{y_0} = \frac{-g_x \hat{\mathbf{x}} + \hat{\mathbf{y}} - g_z \hat{\mathbf{z}}}{\sqrt{1 + g_x^2 + g_z^2}}$$

$$\hat{\mathbf{n}}_{x_0} = \frac{\hat{\mathbf{x}} - h_y \hat{\mathbf{y}} - h_z \hat{\mathbf{z}}}{\sqrt{1 + h_y^2 + h_z^2}}$$

Problem 7

A unit sphere centered at the origin is described by

$$z = f(x, y) = \sqrt{1 - x^2 - y^2} .$$

Show that the normal vector to the sphere is given by

$$\hat{\mathbf{n}} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}} .$$

4.4 Tangent Plane

In either picture, whether it be parametric or Cartesian, the tangent vectors $\mathbf{u}, \mathbf{v}$ imply the existence of a *tangent plane* to the surface at a given point, much in the same way the slope at a point implies a straight line in the one-dimensional case. The normal vector $\mathbf{n}$ is always perpendicular to the tangent plane.

If the point (x_0, y_0, z_0) is the base from which the tangent vectors and normal vector are drawn, and (x, y, z) is any other point in space, then the equation of the tangent plane is:

$$\mathbf{n} \cdot \Delta \mathbf{x} = 0 ,$$

where

$$\Delta \mathbf{x} = \langle x, y, z \rangle - \langle x_0, y_0, z_0 \rangle .$$

From what we know about planes, we can also write

$$ax + by + cz + d = 0$$

to represent the tangent plane. To reconcile this with the vector definition, write out the full dot product:

$$n_x (x - x_0) + n_y (y - y_0) + n_z (z - z_0) = 0 ,$$

or

$$n_x x + n_y y + n_z z + d = 0 ,$$

with

$$d = -n_x x_0 - n_y y_0 - n_z z_0 .$$

We can say a bit more about the Cartesian case, as

$$n_x = -f_x$$

$$n_y = -f_y$$

$$n_z = 1$$

would mean

$$-f_x (x - x_0) - f_y (y - y_0) + (z - z_0) = 0 .$$

5 Surface Integral

For a given vector function $\mathbf{F}(\mathbf{r})$, choose any point $\mathbf{r}_0$ in space and place a small patch of area ΔS containing $\mathbf{r}_0$. The small patch is approximately planar, thus we define a unit normal vector $\hat{\mathbf{n}}$ to the plane that, by convention, points away from the surface if closed.

The vector function $\mathbf{F}(\mathbf{r}_0)$ is approximately constant at the surface, and depending on the orientation of $\hat{\mathbf{n}}$, the components of $\mathbf{F}$ may pass into or out of the surface by an amount ϕ given by:

$$\phi_0 = \mathbf{F}(\mathbf{r}_0) \cdot \hat{\mathbf{n}}(\mathbf{r}_0) \Delta S$$

5.1 Notion of Flux

By stitching together many such small patches to form a whole surface S , we generalize the above to:

$$\Phi = \iint_S \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS$$

Note that $\mathbf{F}$ and $\hat{\mathbf{n}}$ are both functions of position, which means both $\mathbf{F}$ and S must be given.

When the integrand is $\mathbf{F} \cdot \mathbf{n}$ or something like it, the result Φ is often called the *flux* of $\mathbf{F}$ through the surface S .

Surface Integral

The quantity $\mathbf{F} \cdot \mathbf{n}$ readily generalizes to an arbitrary function $G(\mathbf{r})$, in where the above becomes the *surface integral*:

$$\Phi = \iint_S G(\mathbf{r}) dS$$

For the special case with $\mathbf{F}(\mathbf{r}) = \hat{\mathbf{n}}$ or $G(\mathbf{r}) = 1$, the result is the surface area of S :

$$A = \iint_S dS$$

5.2 Integration Region

In the setup of the surface integral, the domain of integration S denotes the surface itself. There is a trick, however, to project the surface down to the xy -plane, where the ‘shadow’ of the surface becomes the *integration region*, denoted $\mathcal{R}$. This means we’re integrating over a smaller area, thus we can anticipate a correction term as we recast the integral.

For a given patch dS oriented by normal vector $\hat{\mathbf{n}}$, the xy -projected area is given by dS' such that

$$dS \cos(\theta) = dS' = dx dy,$$

where θ is given by

$$\hat{\mathbf{n}} \cdot \hat{\mathbf{k}} = \cos(\theta).$$

The surface integral can therefore be written

$$\Phi = \iint_{\mathcal{R}} \frac{G(\mathbf{r})}{\hat{\mathbf{n}} \cdot \hat{\mathbf{k}}} dx dy.$$

For surfaces $z = f(x, y)$, the surface integral simplifies again:

$$\Phi = \iint_{\mathcal{R}} G(x, y, z) \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2} dx dy$$

Problem 8

Evaluate the integral of

$$G(\mathbf{r}) = x + z$$

on the surface of the plane

$$x + y + z = 1$$

in the positive octant of Cartesian coordinates (i.e. $x \geq 0$, $y \geq 0$, $z \geq 0$).

Answer:

$$\phi = \sqrt{3} \int_0^1 \int_0^{1-x} (1-y) dy dx = \frac{1}{\sqrt{3}}$$

Problem 9

For the sphere

$$x^2 + y^2 + z^2 = R^2,$$

calculate the surface area with $z > 0$.

Answer:

$$G(x, y, z) = 1$$

$$z = f(x, y) = \sqrt{R^2 - x^2 - y^2}$$

$$\frac{1}{\hat{\mathbf{n}} \cdot \hat{\mathbf{k}}} = \sqrt{1 + \left(\frac{x}{z}\right)^2 + \left(\frac{y}{z}\right)^2} = \frac{R}{z}$$

$$S = R \iint_{x^2+y^2 \leq R^2} \frac{dx dy}{\sqrt{R^2 - x^2 - y^2}}$$

$$S = 2\pi R \int_0^R \frac{r dr}{\sqrt{R^2 - r^2}} = 2\pi R^2$$

Problem 10

Evaluate the integral of

$$G(\mathbf{r}) = z^2$$

on the surface of the sphere

$$x^2 + y^2 + z^2 = R^2$$

with $z > 0$.

Answer:

$$\Phi = R \iint_{x^2+y^2 \leq R^2} \sqrt{R^2 - x^2 - y^2} \, dx \, dy$$

$$\Phi = 2\pi R \int_0^R \sqrt{R^2 - r^2} \, r \, dr = \frac{2}{3}\pi R^4$$

5.3 Closed Integration Surface

It is often required that S be *closed*, which is to say the integration surface encloses a volume. To denote this, the

closed-form of the above can be written:

$$\Phi = \oiint_S \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}} \, dS$$

Equivalently, we speak of the integration volume $\mathcal{V}$ around which S is wrapped. In this situation, S is instead written $\partial\mathcal{V}$. Also, since the notation for multiple integration gets rather clunky, we can use the symbols $\mathcal{V}$, $\partial\mathcal{V}$, etc. to imply the dimension of the integral. Therefore the flux through a closed surface $\partial\mathcal{V}$ enclosing $\mathcal{V}$ is written:

$$\Phi = \oint_{\partial\mathcal{V}} \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}} \, dS$$

6 Fields and Derivatives

6.1 Scalar Field and Vector Field

By now, we know that a scalar is in essence just a number, whereas a vector is a list of numbers. Expanding this notion, we can have a *scalar field*, which is a scalar that exists everywhere in a region, such as the temperature in a room. For this, we might write

$$T = T(\mathbf{r}) ,$$

which takes the position vector $\mathbf{r}$ and reports the magnitude of the temperature T at that location. The tools of calculus require that $T(\mathbf{r})$ be continuous and differentiable - as real temperature would be - in order for useful analysis to be possible.

A *vector field* is a vector that exists everywhere in a given region. Instead of tracking only the magnitude of a given quantity, a vector field also tracks the direction, such as the wind velocity $\mathbf{V}(\mathbf{r})$ in a given room. In Cartesian coordinates, we might write

$$\mathbf{V}(\mathbf{r}) = V_x \hat{\mathbf{x}} + V_y \hat{\mathbf{y}} + V_z \hat{\mathbf{z}} ,$$

where each scalar component V_j is a function of $\mathbf{r}$.

Gradient of a Scalar Field

Given any scalar field $\phi(\mathbf{r})$, where $\mathbf{r}$ is the Cartesian space position vector $\mathbf{r} = \langle x, y, z \rangle$, one can imagine taking partial derivatives with respect to each spatial component and building a vector field from these:

$$\frac{d}{d\mathbf{r}} (\phi(x, y, z)) = \frac{\partial \phi}{\partial x} \hat{\mathbf{x}} + \frac{\partial \phi}{\partial y} \hat{\mathbf{y}} + \frac{\partial \phi}{\partial z} \hat{\mathbf{z}}$$

For shorthand, we can condense all derivative notation via $\partial/\partial_j \rightarrow \partial_j$ to write

$$d_{\mathbf{r}} (\phi(x, y, z)) = \partial_x \phi \hat{\mathbf{x}} + \partial_y \phi \hat{\mathbf{y}} + \partial_z \phi \hat{\mathbf{z}} .$$

Reading right-to-left, we can regard the components of a vector field as the spatial derivative of a scalar field.

Problem 11

In two dimensions, find a vector field that points in the positive radial direction and has unit magnitude.

Answer:

$$\mathbf{V}(x, y) = \frac{x \hat{\mathbf{x}} + y \hat{\mathbf{y}}}{\sqrt{x^2 + y^2}}$$

Problem 12

In three dimensions, find a vector field that points in the positive radial direction and has unit magnitude.

Answer:

$$\mathbf{V}(x, y, z) = \frac{x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}}}{\sqrt{x^2 + y^2 + z^2}}$$

Problem 13

In two dimensions, find a vector field that makes a 45-degree angle with the x -axis and has magnitude $(x + y)^2$.

Answer:

$$\mathbf{V}(x, y) = \frac{\hat{\mathbf{x}} + \hat{\mathbf{y}}}{\sqrt{2}} (x + y)^2$$

Problem 14

In two dimensions, find a vector field that is purely tangential and has magnitude equal to the distance from the origin.

Answer:

$$\mathbf{V}(x, y) = -y \hat{\mathbf{x}} + x \hat{\mathbf{y}}$$

6.2 Field Lines

Given an arbitrary vector field

$$\mathbf{V}(\mathbf{r}) = V_x \hat{\mathbf{x}} + V_y \hat{\mathbf{y}}$$

in two dimensions, one can integrate in the spatial domain to attain information on the position vector(s) $\mathbf{r}$, up to an integration constant, in the form of *field lines*.

From plane curve analysis, we know the ratio V_y/V_x is equivalent to the slope dy/dx , the slope of the field line at a given point:

$$\frac{dy}{dx} = \frac{V_y}{V_x}$$

Then, the equation of the family of field lines takes the form of an indefinite integral:

$$\int V_x dx = \int V_y dy$$

Problem 15

Given

$$\mathbf{V} = y \hat{\mathbf{x}} + x \hat{\mathbf{y}} ,$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = \frac{x}{y} \rightarrow \int y dy = \int x dx$$

$$x^2 - y^2 = C$$

Problem 16

Given

$$\mathbf{V} = \hat{\mathbf{x}} + \hat{\mathbf{y}} ,$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = 1 \rightarrow \int dy = \int dx$$

$$y = x + C$$

Problem 17

Given

$$\mathbf{V} = x \hat{\mathbf{x}} - y \hat{\mathbf{y}},$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = \frac{-y}{x} \rightarrow \int \frac{dy}{y} = - \int \frac{dx}{x}$$

$$\ln(xy) = C' \rightarrow xy = C$$

Problem 18

Given

$$\mathbf{V} = \frac{y \hat{\mathbf{x}} + x \hat{\mathbf{y}}}{\sqrt{x^2 + y^2}},$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = \frac{x/V}{y/V} = \frac{x}{y}$$

$$x^2 - y^2 = C$$

Problem 19

Given

$$\mathbf{V} = y \hat{\mathbf{x}} + xy \hat{\mathbf{y}},$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = x \rightarrow y = \frac{1}{2}x^2 + C$$

Problem 20

Given

$$\mathbf{V} = \hat{\mathbf{x}} + y \hat{\mathbf{y}},$$

find an equation for the field lines.

Answer:

$$\frac{dy}{dx} = y \rightarrow \int \frac{dy}{y} = \int dx$$

$$\ln(y) = x + C \rightarrow y = Ae^x$$

6.3 Del Operator

Differentiating a scalar field to produce a vector field is achieved with the spatial derivative operator

$$d_{\mathbf{r}} = \partial_x \hat{\mathbf{x}} + \partial_y \hat{\mathbf{y}} + \partial_z \hat{\mathbf{z}},$$

which has a special name and a special symbol. Instead of $d_{\mathbf{r}}$, one writes ∇ , known as the *del operator*, also known as the *nabla* operator. Explicitly, one writes:

$$\nabla = \partial_x \hat{\mathbf{x}} + \partial_y \hat{\mathbf{y}} + \partial_z \hat{\mathbf{z}} \quad (19.7)$$

Using the ∇ symbol, the gradient of a scalar field $\phi(\mathbf{r})$ is written

$$\nabla \phi(\mathbf{r}) = \partial_x \phi \hat{\mathbf{x}} + \partial_y \phi \hat{\mathbf{y}} + \partial_z \phi \hat{\mathbf{z}}$$

Recalling the two-dimensional Taylor's theorem given by Equation (19.5), we can immediately generalize to three dimensions while keeping the notation tidy:

$$f(\mathbf{r}) \approx f(\mathbf{r}_0) + \sum_{q=1}^{\infty} \frac{1}{q!} (\Delta \mathbf{r} \cdot \nabla)^q f(\mathbf{r}_0) \quad (19.8)$$

In the above, the scalar operator $\Delta \mathbf{r} \cdot \nabla$ unpacks via

$$\Delta \mathbf{r} \cdot \nabla = \Delta x \partial_x + \Delta y \partial_y + \Delta z \partial_z.$$

There are three ways to apply the del operator, each with its own name. In the following, let $\phi(\mathbf{r})$ be a scalar field and let $\mathbf{V}(\mathbf{r})$ be a vector field:

1. Gradient

$$\nabla \phi(\mathbf{r}) = \partial_x \phi \hat{\mathbf{x}} + \partial_y \phi \hat{\mathbf{y}} + \partial_z \phi \hat{\mathbf{z}}$$

2. Divergence

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \frac{\partial}{\partial x} F_x + \frac{\partial}{\partial y} F_y + \frac{\partial}{\partial z} F_z$$

3. Curl

$$\nabla \times \mathbf{V}(\mathbf{r}) = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \partial_x & \partial_y & \partial_z \\ V_x & V_y & V_z \end{vmatrix}$$

$$\nabla \times \mathbf{V}(\mathbf{r}) = (\partial_y V_z - \partial_z V_y) \hat{\mathbf{x}} + (\partial_z V_x - \partial_x V_z) \hat{\mathbf{y}} + (\partial_x V_y - \partial_y V_x) \hat{\mathbf{z}}$$

As a vector operator, the del operator respects the object onto which it is applied. For instance, you never see $\nabla \cdot \phi$ or $\nabla \mathbf{V}$ in standard vector calculus.

Laplacian Operator

The object $\nabla \cdot \nabla$ resolves to a scalar operator called the *Laplacian*, and is often written without the vector symbol:

$$\nabla \cdot \nabla = \nabla^2 = \partial_{xx} + \partial_{yy} + \partial_{zz}$$

Del Operator Identities

Consider two scalar fields $a(\mathbf{r})$, $b(\mathbf{r})$, along with two vector fields $\mathbf{F}(\mathbf{r})$, $\mathbf{G}(\mathbf{r})$. With these, we may establish a slew of del operator identities, some of which are:

$$\nabla(ab) = a\nabla b + b\nabla a \quad (19.9)$$

$$\begin{aligned} \nabla(\mathbf{F} \cdot \mathbf{G}) &= (\mathbf{G} \cdot \nabla)\mathbf{F} + (\mathbf{F} \cdot \nabla)\mathbf{G} \\ &\quad + \mathbf{F} \times (\nabla \times \mathbf{G}) + \mathbf{G} \times (\nabla \times \mathbf{F}) \end{aligned} \quad (19.10)$$

$$\nabla \cdot (a\mathbf{F}) = a\nabla \cdot \mathbf{F} + \mathbf{F} \cdot \nabla a \quad (19.11)$$

$$\nabla \cdot (\mathbf{F} \times \mathbf{G}) = \mathbf{G} \cdot (\nabla \times \mathbf{F}) - \mathbf{F} \cdot (\nabla \times \mathbf{G}) \quad (19.12)$$

$$\nabla \cdot \nabla \times \mathbf{F} = 0 \quad (19.13)$$

$$\nabla \times (a\mathbf{F}) = a\nabla \times \mathbf{F} + (\nabla a) \times \mathbf{F} \quad (19.14)$$

$$\begin{aligned} \nabla \times (\mathbf{F} \times \mathbf{G}) &= (\mathbf{G} \cdot \nabla)\mathbf{F} - (\mathbf{F} \cdot \nabla)\mathbf{G} \\ &\quad + \mathbf{F}(\nabla \cdot \mathbf{G}) - \mathbf{G}(\nabla \cdot \mathbf{F}) \end{aligned} \quad (19.15)$$

$$\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2 \mathbf{F} \quad (19.16)$$

$$\nabla \times \nabla a = 0 \quad (19.17)$$

Equation (19.9) is straightforward application of the product rule. To see this explicitly, write:

$$\begin{aligned} a\nabla b + b\nabla a &= a\partial_x b \hat{\mathbf{x}} + a\partial_y b \hat{\mathbf{y}} + a\partial_z b \hat{\mathbf{z}} \\ &\quad + b\partial_x a \hat{\mathbf{x}} + b\partial_y a \hat{\mathbf{y}} + b\partial_z a \hat{\mathbf{z}} \\ &= (a\partial_x b + b\partial_x a)\hat{\mathbf{x}} + (a\partial_y b + b\partial_y a)\hat{\mathbf{y}} \\ &\quad + (a\partial_z b + b\partial_z a)\hat{\mathbf{z}} \\ &= \partial_x(ab)\hat{\mathbf{x}} + \partial_y(ab)\hat{\mathbf{y}} + \partial_z(ab)\hat{\mathbf{z}} \\ &= \nabla(ab) \end{aligned}$$

We work the proof right-to-left or left-to-right, depending which is easier.

To go about proving Equation (19.10), there is enough work to do with just the x -component, let alone the y - and z -components. By brute force, the left side expands as:

$$\begin{aligned} \left. \nabla(\mathbf{F} \cdot \mathbf{G}) \right|_x &= \partial_x(\mathbf{F} \cdot \mathbf{G}) \\ &= F_x \partial_x G_x + F_y \partial_x G_y + F_z \partial_x G_z \\ &\quad + G_x \partial_x F_x + G_y \partial_x F_y + G_z \partial_x F_z \end{aligned}$$

Now we have four quantities to expand on the right. Once again, stick to just the x -component so the algebra doesn't explode.

$$\begin{aligned} \left. (\mathbf{G} \cdot \nabla)\mathbf{F} \right|_x &= (\mathbf{G} \cdot \nabla)F_x \\ &= G_x \partial_x F_x + G_y \partial_y F_x + G_z \partial_z F_x \end{aligned}$$

$$\begin{aligned} \left. (\mathbf{F} \cdot \nabla)\mathbf{G} \right|_x &= (\mathbf{F} \cdot \nabla)G_x \\ &= F_x \partial_x G_x + F_y \partial_y G_x + F_z \partial_z G_x \end{aligned}$$

$$\begin{aligned} \left. \mathbf{F} \times (\nabla \times \mathbf{G}) \right|_x &= F_y (\partial_x G_y - \partial_y G_x) \\ &\quad - F_z (\partial_z G_x - \partial_x G_z) \end{aligned}$$

$$\begin{aligned} \left. \mathbf{G} \times (\nabla \times \mathbf{F}) \right|_x &= G_y (\partial_x F_y - \partial_y F_x) \\ &\quad - G_z (\partial_z F_x - \partial_x F_z) \end{aligned}$$

Summing down the total right side, we see all negative terms canceling with matching positive terms. The surviving terms match those on the left of Equation (19.10), and we're done. (You can work out the y - and z -components if you must.)

To prove Equation (19.11), we can work left-to-right easily enough:

$$\begin{aligned} \nabla \cdot (a\mathbf{F}) &= \nabla \cdot (aF_x \hat{\mathbf{x}} + aF_y \hat{\mathbf{y}} + aF_z \hat{\mathbf{z}}) \\ &= \partial_x(aF_x) + \partial_y(aF_y) + \partial_z(aF_z) \\ &= a\nabla \cdot \mathbf{F} + \mathbf{F} \cdot \nabla a \end{aligned}$$

Problem 21

Show that Equation (19.12) is true by any means.

Translated to English, Equation (19.13) tells us *the divergence of the curl of a vector field is zero*. Working with components individually, we have

$$\begin{aligned} \left. \nabla \times \mathbf{F} \right|_x &= \partial_y F_z - \partial_z F_y \\ \left. \nabla \times \mathbf{F} \right|_y &= \partial_z F_x - \partial_x F_z \\ \left. \nabla \times \mathbf{F} \right|_z &= \partial_x F_y - \partial_y F_x \end{aligned}$$

so then:

$$\begin{aligned}\nabla \cdot (\nabla \times \mathbf{F}) \Big|_x &= \partial_x \partial_y F_z - \partial_x \partial_z F_y \\ \nabla \cdot (\nabla \times \mathbf{F}) \Big|_y &= \partial_y \partial_z F_x - \partial_y \partial_x F_z \\ \nabla \cdot (\nabla \times \mathbf{F}) \Big|_z &= \partial_z \partial_x F_y - \partial_z \partial_y F_x\end{aligned}$$

Adding all terms on the right resolves to identically zero and the proof is done.

Problem 22

Show that Equation (19.14) is true by any means.

Problem 23

Show that Equation (19.15) is true by any means.

Problem 24

Show that Equation (19.16) is true by any means.

Problem 25

Show that Equation (19.17) is true by any means.

6.4 Lagrange Multipliers

Suppose we have a two-variable scalar field $f(x, y)$, and there is reason to extremize f subject to a different function $g(x, y) = c$, where c characterizes a level curve on g .

Parallel Gradients

Finding critical points in f subject to g entails noticing that the gradient of each function is the same at a critical point, up to a proportionality constant λ , the Lagrange multiplier:

$$\nabla f(x, y) = \lambda \nabla g(x, y)$$

Said a different way, a constrained optimization problem minimizes a functional F

$$F[x, y, \lambda] = \nabla f(x, y) - \lambda \nabla g(x, y)$$

at critical points (x_0, y_0) .

When there are multiple constraint functions, each introduces a new and different λ . The gradient of the original function and all constraints remain proportional:

$$\nabla f(x, y) = \lambda_1 \nabla g_1(x, y) + \lambda_2 \nabla g_2(x, y) + \dots$$

Example

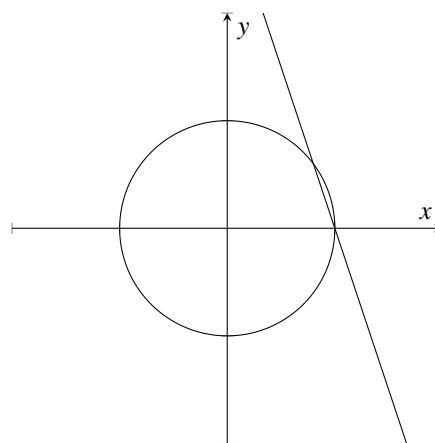
For an example, suppose we have a function

$$f(x, y) = x^2 + y^2,$$

and we are to find any critical points in f subject to the constraining line

$$y + 3x = 3.$$

Sketched below is the constraining function along with a single level curve ($f = \text{constant}$) of $f(x, y)$.



Constructing the functional $F[\lambda]$, we have

$$\begin{aligned}F[\lambda] &= \nabla f(x, y) - \lambda \nabla g(x, y) \\ &= \langle 2x, 2y \rangle - \lambda \langle 3, 1 \rangle.\end{aligned}$$

Setting the left side to zero, we gain two new equations

$$2x_0 = 3\lambda$$

$$2y_0 = \lambda,$$

and a third equation is given by g :

$$y_0 + 3x_0 = 3$$

As a system of three equations and three unknowns, the results for λ, x_0, y_0 must be found simultaneously, resulting in

$$\lambda = 3/5$$

$$x_0 = 9/10$$

$$y_0 = 3/10,$$

and the problem is solved.

Problem 26

For positive x, y, z , find the maximum of

$$f(x, y, z) = xy^2 + yz^2 + 2xyz$$

subject to

$$x + y + z = 5.$$

Answer: $x_0 = 1, y_0 = z_0 = 2, f(1, 2, 2) = 20$

Problem 27

Identify the four critical points of the function $f(x, y) = x^2 + 2y^2$ that coincide with the unit circle $x^2 + y^2 = 1$.

Hint: $\langle 2x, 4y \rangle = \lambda \langle 2x, 2y \rangle$

Problem 28

Find the largest rectangle that fits inside the first quadrant of the ellipse $x^2/a^2 + y^2/b^2 = 1$.

Answer: $x_0 = a/\sqrt{2}, y_0 = b/\sqrt{2}$

Problem 29

Find the largest square that fits inside the first quadrant of the ellipse $x^2/a^2 + y^2/b^2 = 1$.

Hint:

$$\nabla(xy) = \lambda_1 \nabla \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 \right) + \lambda_2 \nabla(x - y)$$

Problem 30

Find the largest rectangular volume that fits inside the positive octant of the ellipsoid $(x/a)^2 + (y/b)^2 + (z/c)^2 = 1$.

Answer: $x_0 = a/\sqrt{3}, y_0 = b/\sqrt{3}, z_0 = c/\sqrt{3}$

Problem 31

A cylinder of radius R and length L is capped on each end by a cone of height H . Maximize the volume for a given surface area.

Hint:

$$\begin{aligned} \nabla \left(\pi R^2 L + \frac{2}{3} \pi R^2 H \right) = \\ \lambda \nabla \left(2\pi RL + 2\pi R \sqrt{R^2 + H^2} \right) \end{aligned}$$

Answer:

$$V = \frac{AR}{3} = A^{3/2} (2\pi)^{-1/2} \frac{5^{-1/4}}{3}$$

Problem 32

Let a and b be positive numbers. Show that not both of the products $a(1-a)$ and $b(1-b)$ can be greater than $1/4$.

Answer: Divide a square of area 1 into four rectangles having areas $A_1 = ab, A_2 = (1-a)(1-b), A_3 = a(1-b), A_4 = b(1-a)$. Using either of $\nabla A_1 = \lambda \nabla A_2$ or $\nabla A_3 = \lambda \nabla A_4$, conclude that $a = b$. Differentiating $f(x) = x(1-x)$, we see $f(x)$ is maximized at $x = 1/2$, therefore each area A_j equals the product $a(1-a)$ and is no greater than $1/4$.

7 Divergence

The divergence

$$\nabla \cdot \mathbf{F}(\mathbf{r}) = \frac{\partial}{\partial x} F_x + \frac{\partial}{\partial y} F_y + \frac{\partial}{\partial z} F_z$$

of a vector field $\mathbf{F}(\mathbf{r})$ resolves to a scalar. To see this intuitively, consider a small box in space having dimensions $\Delta x, \Delta y, \Delta z$ and volume ΔV . If the box centers on a point $\mathbf{r}_0$, then the field $\mathbf{F}(\mathbf{r}_0)$ has a notion of flux through all sides of the box.

Precisely, we consider a surface integral

$$\phi_0 = \lim_{\Delta V \rightarrow 0} \frac{1}{\Delta V} \iint_{\text{box}} \mathbf{F}(\mathbf{r}_0) \cdot \hat{\mathbf{n}}(\mathbf{r}_0) dS,$$

where the unit normal vector $\hat{\mathbf{n}}$ always points away from the surface. In detail, the above breaks into six parts, one for each face of the box. For the two x -faces, we have:

$$\begin{aligned} \frac{1}{\Delta V} \iint_{x^+} \mathbf{F}\left(x_0 + \frac{\Delta x}{2}, y_0, z_0\right) \cdot \hat{\mathbf{x}} dy dz \\ \approx F_x\left(x_0 + \frac{\Delta x}{2}, y_0, z_0\right) \Delta y \Delta z / \Delta V \\ \frac{1}{\Delta V} \iint_{x^-} \mathbf{F}\left(x_0 - \frac{\Delta x}{2}, y_0, z_0\right) \cdot (-\hat{\mathbf{x}}) dy dz \\ \approx -F_x\left(x_0 - \frac{\Delta x}{2}, y_0, z_0\right) \Delta y \Delta z / \Delta V \end{aligned}$$

Observe that $\Delta y \Delta z / \Delta V$ simplifies to $1/\Delta x$. Sum each equation on the right to find, by definition of the partial derivative:

$$\lim_{\Delta V \rightarrow 0} \frac{1}{\Delta V} \iint_{x\text{-faces}} \mathbf{F}(\mathbf{r}_0) \cdot \mathbf{n}(\mathbf{r}_0) dS = \left. \frac{\partial}{\partial x} F_x \right|_{\mathbf{r}_0}$$

Repeating the exercise for the y - and z -faces yields the final result:

$$\phi_0 = \lim_{\Delta V \rightarrow 0} \frac{1}{\Delta V} \iint_{\text{box}} \mathbf{F} \cdot \hat{\mathbf{n}} dS = \nabla \cdot \mathbf{F}(\mathbf{r}_0)$$

Evidently, we find that the flux per volume through the surface of the box equals the divergence inside the box.

Divergence in Cylindrical Coordinates

Problem 33

Find the formula for the divergence of a vector field $\mathbf{F}(\rho, \phi, z)$ in cylindrical coordinates given by:

$$x(\rho, \phi, z) = \rho \cos(\phi)$$

$$y(\rho, \phi, z) = \rho \sin(\phi)$$

$$z(\rho, \phi, z) = z$$

Hint: Let $\Delta V = \rho \Delta \rho \Delta \theta \Delta z$. For the pair of ρ -surfaces, write

$$\begin{aligned} \frac{1}{\Delta V} \iint_{\rho^+} \mathbf{F} \cdot \hat{\boldsymbol{\rho}} \rho d\theta dz \\ \approx F_\rho\left(\rho_0 + \frac{\Delta \rho}{2}, \theta_0, z_0\right) \left(\rho_0 + \frac{\Delta \rho}{2}\right) \frac{\Delta \theta \Delta z}{\Delta V} \\ \frac{1}{\Delta V} \iint_{\rho^-} \mathbf{F} \cdot (-\hat{\boldsymbol{\rho}}) \rho d\theta dz \\ \approx -F_\rho\left(\rho_0 - \frac{\Delta \rho}{2}, \theta_0, z_0\right) \left(\rho_0 - \frac{\Delta \rho}{2}\right) \frac{\Delta \theta \Delta z}{\Delta V} \end{aligned}$$

such that:

$$\lim_{\Delta V \rightarrow 0} \frac{1}{\Delta V} \iint_{\rho\text{-faces}} \mathbf{F} \cdot \hat{\boldsymbol{\rho}} \rho d\theta dz = \left. \frac{1}{\rho_0} \frac{\partial}{\partial \rho} (\rho F_\rho) \right|_{\rho_0}$$

Answer:

$$\nabla \cdot \mathbf{F} = \frac{1}{\rho} \frac{\partial}{\partial \rho} (\rho F_\rho) + \frac{1}{\rho} \frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}$$

Divergence in Spherical Coordinates

Problem 34

Find the formula for the divergence of a vector field $\mathbf{F}(r, \theta, \phi)$ in spherical coordinates given by:

$$x = r \sin(\theta) \cos(\phi)$$

$$y = r \sin(\theta) \sin(\phi)$$

$$z = r \cos(\theta)$$

Answer:

$$\begin{aligned} \nabla \cdot \mathbf{F} = \frac{\partial_r (r^2 F_r)}{r^2} \\ + \frac{\partial_\theta (\sin(\theta) F_\theta)}{r \sin(\theta)} + \frac{\partial_\phi (F_\phi)}{r \sin(\theta)} \end{aligned}$$

7.1 Divergence Theorem

For a vector field $\mathbf{F}(\mathbf{r})$, we showed that the flux through a small closed surface is equal to the divergence of $\mathbf{F}$ inside the contained volume. It's natural, then, to wonder if the same notion extends to non-infinitesimal surfaces and volumes.

To make some progress, consider an arbitrary surface $\mathcal{S}$ enclosing volume $\mathcal{V}$. If the total volume is comprised of N total volume elements, we denote the j th element as ΔV_j , and the total volume $\mathcal{V}$ can be written as the limit of a sum:

$$\mathcal{V} = \lim_{N \rightarrow \infty} \sum_{j=1}^N \Delta V_j = \iiint_{\mathcal{V}} dV$$

The internal slicing of volume elements - such as by cubes or by pyramids - doesn't really matter, as any two adjacent panels have equal and opposite area elements, namely $\hat{\mathbf{n}} dS$ and $-\hat{\mathbf{n}} dS$. It follows that the only relevant flux occurs on the surface $\mathcal{S}$.

Next, recall the equation that applies to a single volume element and assign the index j , replacing index 0. Note that the surface of the j th volume element is written $\mathcal{S}_j$:

$$\lim_{\Delta V_j \rightarrow 0} \frac{1}{\Delta V_j} \iint_{\mathcal{S}_j} \mathbf{F}(\mathbf{r}_j) \cdot \hat{\mathbf{n}}(\mathbf{r}_j) dS_j = \nabla \cdot \mathbf{F}(\mathbf{r}_j)$$

The plan is to multiply each side by the volume element ΔV_j and sum through all elements. Of course, the factor

$1/\Delta V_j$ exactly cancels on the left:

$$\sum_{j=1}^N \iint_{\partial V_j} \mathbf{F}(\mathbf{r}_j) \cdot \hat{\mathbf{n}}(\mathbf{r}_j) dS_j = \sum_{j=1}^N \nabla \cdot \mathbf{F}(\mathbf{r}_j) \Delta V_j$$

In the limit $N \rightarrow \infty$ the left becomes a closed surface integral and the right becomes a volume integral:

$$\oint_{\mathcal{S}} \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \iiint_{\mathcal{V}} \nabla \cdot \mathbf{F}(\mathbf{r}) dV \quad (19.18)$$

Flux and Divergence

Equation (19.18) is called the *divergence theorem*. As suspected, the theorem tells us that *the total flux through a closed surface equals the divergence inside the enclosed volume*. More economically, we write the same theorem as:

$$\oint_{\mathcal{S}} \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \int_{\mathcal{V}} \nabla \cdot \mathbf{F}(\mathbf{r}) d^3x$$

8 Central Forces

There is probably no better way to present vector calculus than in a context relevant to physics. For this we may discuss two regimes - gravitational and electrostatic - both exhibiting the central $1/r^2$ attraction. It will turn out, though, that electrostatics is the richer sibling. First, we may speak of both attractive and repulsive forces based on the signs of interacting charges. Gravity, of course, is a purely attractive force. Moreover, the dynamics of moving charges gives rise to magnetism, for which gravity has no analog.

Electrostatic Force

Consider two point charges q_0, q_1 separated by distance r in free space. The force onto the test charge q_0 due to a given charge q_1 is given by Coulomb's law:

$$\mathbf{F}_{\text{elec}}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{q_0 q_1}{r^2} \hat{\mathbf{r}}, \quad (19.19)$$

where $\hat{\mathbf{r}}$ is the unit vector pointing from q_1 to q_0 .

The permittivity constant ϵ_0 is specifically intended for vacuum conditions, loosely called 'free space'. Within matter, also known as media, the permittivity changes and loses the 0-subscript: $\epsilon_0 \rightarrow \epsilon$

Gravitational Force

Consider two point masses m_0, m_1 separated by distance r in free space. The force onto the test mass m_0 due to a given mass m_1 is given by Newton's law of gravitation:

$$\mathbf{F}_{\text{grav}}(\mathbf{r}) = G \frac{m_0 m_1}{r^2} \hat{\mathbf{r}}, \quad (19.20)$$

where $\hat{\mathbf{r}}$ is the unit vector pointing from m_1 to m_0 .

8.1 Notion of Field

In free space containing two charges q_0, q_1 , we can imagine that q_1 's effect on the space surrounding it would be there whether or not the test charge q_0 is present. Thus, we can divide the force vector by q_0 and forget the test charge altogether, yielding the electric field, denoted $\mathbf{E}(\mathbf{r})$:

$$\frac{\mathbf{F}_{\text{elec}}(\mathbf{r})}{q_0} = \mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{q_1}{r^2} \hat{\mathbf{r}} \quad (19.21)$$

Similarly, we can take Newton's law of gravitation and divide out the test mass m_0 to get the gravitational field, denoted $\mathbf{g}(\mathbf{r})$:

$$\frac{\mathbf{F}_{\text{grav}}(\mathbf{r})}{m_0} = \mathbf{g}(\mathbf{r}) = G \frac{m_1}{r^2} \hat{\mathbf{r}} \quad (19.22)$$

Problem 35

Use Newton's gravitation constant, the mass of the earth, and the radius of the earth to calculate the magnitude of the gravitational field at sea level.

Answer: $g = 9.81 \text{ m/s}^2$

Superposition

An important consequence of the electrostatic force is the principle of vector *superposition*, which applies when a test charge is subject to the electric force from two or more sources. Identical notions apply to the gravitational force. In either regime, the total force is calculated trivially by vector addition:

$$\mathbf{F}_{\text{net}} = \mathbf{F}_1 + \mathbf{F}_2 + \dots$$

Superposition also applies to the electric field and gravitational field. The net field at a point in space is the vector sum of the contributions from each source:

$$\mathbf{E}_{\text{net}} = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \dots$$

$$\mathbf{g}_{\text{net}} = \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{g}_3 + \dots$$

If there are N given charges q_j at positions $\mathbf{r}_j$, we write the force and the field in sigma notation:

$$\mathbf{F}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_{j=1}^N \frac{q_0 q_j}{|\mathbf{r} - \mathbf{r}_j|^2} \hat{\mathbf{r}}_j$$

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_{j=1}^N \frac{q_j}{|\mathbf{r} - \mathbf{r}_j|^2} \hat{\mathbf{r}}_j$$

Note we could rewrite the above in terms of the gravitational force and field, but for brevity we won't always do so. Instead, it's faster to mentally swap charges for masses

$$q_i q_j \leftrightarrow m_i m_j$$

and replace the coefficients:

$$\frac{1}{4\pi\epsilon_0} \leftrightarrow G$$

Note this similarity breaks down eventually, so avoid believing that gravity is mathematically identical to electrostatics.

Problem 36

A charge $+q$ is situated at $\langle 1, 0, 0 \rangle$ and a charge $-q$ is situated at $\langle -1, 0, 0 \rangle$. Find the electric field at any position $\langle 0, y, 0 \rangle$.

Answer:

$$\mathbf{E}(y) = \frac{-\hat{\mathbf{x}}}{2\pi\epsilon_0 (1 + y^2)^{3/2}}$$

8.2 Continuous Matter

As a classical theory, we assume that matter is approximately (or exactly) a smooth differentiable manifold. Thus we may speak of charged matter as carrying a continuous charge distribution depending on position.

A thin wire can be approximated as one-dimensional object with a given charge per unit length, i.e. linear charge density, as $\lambda(\mathbf{r}) = dq/dl$. For a given point $\mathbf{r}$ in space, the electric field due to such a wire is best written:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int_{\text{wire}} \lambda(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} dl'$$

A thin sheet can be approximated as two-dimensional object with a given charge per unit area, i.e. surface charge density, as $\sigma(\mathbf{r}) = dq/dS$. For a given point $\mathbf{r}$ in space, the electric field due to such a surface is:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \iint_{\text{surface}} \sigma(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} dS'$$

A continuous volume of charge has charge density $\rho(\mathbf{r}) = dq/dV$. For a given point $\mathbf{r}$ in space, the electric field due to such a volume is:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \iiint_{\text{solid}} \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3x'$$

Problem 37

An infinite straight wire is oriented along the $\pm x$ axis. Calculate the magnitude of the electric field at any point on the y -axis. Discuss why the electric field does not depend on x .

Answer:

$$E = E_y = \frac{2\lambda}{4\pi\epsilon_0} \int_0^\infty \frac{x dx}{(x^2 + y^2)^{3/2}} = \frac{\lambda}{2\pi\epsilon_0 y}$$

Net Charge

The net charge of an object is readily calculated by integrating the charge density across the extent of the object. In one, two, and three dimensions, we have

$$\begin{aligned} Q_{1D} &= \int_{\text{wire}} \lambda(\mathbf{r}) dl \\ Q_{2D} &= \iint_{\text{surface}} \sigma(\mathbf{r}) dS \\ Q_{3D} &= \iiint_{\text{solid}} \rho(\mathbf{r}) d^3x. \end{aligned}$$

9 Gauss's Law

For a given vector function $\mathbf{F}(\mathbf{r})$, recall that the flux Φ through a surface is given by the surface integral

$$\Phi = \iint_S \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS.$$

9.1 Gaussian Surface

The surface S does not necessarily refer to the surface of a physical object. Rather, the surface on which the vector function is integrated can be arbitrary, which takes the name *Gaussian surface*.

Electric Flux

By slipping the electric field $\mathbf{E}(\mathbf{r})$ into the above surface integral, the left becomes the *electric flux* of $\mathbf{E}$ through the surface S :

$$\Phi_E = \iint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS$$

Gravitational Flux

Similarly, insert the gravitational field $\mathbf{g}(\mathbf{r})$ into the above surface integral to write *gravitational flux* of $\mathbf{g}$ through the surface S :

$$\Phi_G = \iint_S \mathbf{g}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS$$

Divergence in Free Space

Consider a point charge q situated at the origin in otherwise free space. The electric field due to the point charge reads

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{\mathbf{r}}.$$

Having a radial component only, it's natural to regard $\mathbf{E}(\mathbf{r})$ in spherical coordinates, in where the θ - and ϕ - components of the field aren't present.

In the spherical coordinate system, we calculate the divergence of the field everywhere except $\mathbf{r} = 0$:

$$\nabla \cdot \mathbf{E} = \frac{1}{r^2} \partial_r \left(r^2 \left(\frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \right) \right) = 0$$

Interestingly, the divergence of the electric field at all points in space equals zero, except on the point charge itself, in where the derivative operation breaks down.

By the principle of superposition of the electric field, it follows that the field due to a distribution of charges in free space leads to the same result:

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = 0$$

Divergence at Point Charge

To calculate the divergence of the electric field *at* the location of the point charge, we need to tap the notion of electric flux and the divergence theorem.

Consider a spherical Gaussian surface of radius R centered at the origin. The normal vector to the surface is $\hat{\mathbf{r}}$, the same unit vector that steers the electric field at a given location. The electric flux through such a surface is thus

$$\Phi_E = \int_0^{2\pi} \int_0^\pi \left(\frac{1}{4\pi\epsilon_0} \frac{q}{R^2} \right) R^2 \sin(\theta) d\theta d\phi,$$

simplifying to

$$\Phi_E = \iint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \frac{q}{\epsilon_0}.$$

Notice that we integrated over the entire spherical Gaussian surface, so the integral was closed. This compells writing the divergence theorem applied to the electric field of the point charge:

$$\iiint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \frac{q}{\epsilon_0} = \iiint_V \nabla \cdot \mathbf{E}(\mathbf{r}) dV$$

Paying attention to $\nabla \cdot \mathbf{E}(\mathbf{r})$, we see the divergence of the field at the location of the charge must be nonzero. Otherwise we could not have a finite quantity on the left.

Problem 38

Show explicitly that:

$$\int_0^\pi \int_0^{2\pi} \sin(\theta) d\theta d\phi = 4\pi$$

Point Charge Density

To proceed, assume that q , despite being a point charge at the origin, is still the integral of a charge density ρ over the region it occupies:

$$q = \iiint_V \rho(\mathbf{r}) dV$$

Now we introduce the curious form for the charge density

$$\rho(\mathbf{r}) = q\delta^3(\mathbf{r}),$$

where δ^3 is the Dirac delta function in three dimensions. Inside the integrand, the job of $\delta^3(\mathbf{r})$ is such that

$$\begin{aligned}\delta^3(\mathbf{r} = 0) &= 1 \\ \delta^3(\mathbf{r} \neq 0) &= 0.\end{aligned}$$

This also means the integration region can be within an arbitrary volume, as the Dirac delta function flattens everything except $\mathbf{r} = 0$.

With the above, we write

$$\frac{q}{\epsilon_0} = \frac{1}{\epsilon_0} \iiint_V \rho(\mathbf{r}) dV = \iiint_V \nabla \cdot \mathbf{E}(\mathbf{r}) dV,$$

equivalent to

$$0 = \iiint_V \left(\nabla \cdot \mathbf{E}(\mathbf{r}) - \frac{\rho(\mathbf{r})}{\epsilon_0} \right) dV,$$

finally telling us

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\epsilon_0}.$$

9.2 Gauss's Law

The results

$$\begin{aligned}\oiint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS &= \frac{q}{\epsilon_0} \\ \nabla \cdot \mathbf{E}(\mathbf{r}) &= \frac{\rho(\mathbf{r})}{\epsilon_0}\end{aligned}$$

are equivalent statements of Gauss's law for a point charge.

These were derived for a point charge located at the origin, and the extension of the proof for continuous distributions will be done shortly. For now, it's worth pausing to memorize what Gauss's law is telling us:

1. The integral form is a 'big picture' statement that relates the electric flux through a surface to the charge inside the surface.
2. The differential version is a 'small picture' statement about exactly where $\nabla \cdot \mathbf{E}$ is nonzero.

Gauss's Law for Point Mass

Problem 39

Repeat the above derivation with the gravitational force and gravitational field in place of the electrostatic force and field. Let $\rho(\mathbf{r})$ be the density of mass.

Answer:

$$\begin{aligned}\oiint_S \mathbf{g}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS &= -4\pi Gm \\ \nabla \cdot \mathbf{g}(\mathbf{r}) &= -4\pi G\rho(\mathbf{r})\end{aligned}$$

Gauss's Law for Charge Distribution

To prove Gauss's law for a distribution of N total charges, we consider a density function given by

$$\rho(\mathbf{r}) = \sum_{j=1}^N q_j \delta^3(\mathbf{r} - \mathbf{r}_j)$$

which accounts for each charge q_j at position $\mathbf{r}_j$. Then, at a given point $\mathbf{r}$, the electric field is

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \iiint_V \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'_j}{|\mathbf{r} - \mathbf{r}'_j|^3} d^3x'$$

Next, we must calculate $\nabla \cdot \mathbf{E}(\mathbf{r})$. To do so, use the curious identity

$$\nabla \cdot \left(\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right) = 4\pi \delta^3(\mathbf{r} - \mathbf{r}')$$

to write

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \iiint_V \rho(\mathbf{r}') 4\pi \delta^3(\mathbf{r} - \mathbf{r}') d^3x'.$$

The integral on the right resolves to $\rho(\mathbf{r})/\epsilon_0$, or

$$\nabla \cdot \mathbf{E}(\mathbf{r}) = \frac{\rho(\mathbf{r})}{\epsilon_0}.$$

Perhaps not surprisingly, Gauss's law for distributions looks the same as the trivial case, with all complexity contained in $\rho(\mathbf{r})$.

To establish the integral version, integrate both sides of the above with respect to volume:

$$\iiint_V \nabla \cdot \mathbf{E}(\mathbf{r}) d^3x = \iiint_V \frac{\rho(\mathbf{r})}{\epsilon_0} d^3x$$

The volume integral on the left can be replaced with a closed surface integral via the divergence theorem. The right side simply evaluates to the total charge enclosed by the surface divided by ϵ_0 :

$$\oiint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \frac{q_{\text{enc}}}{\epsilon_0}$$

9.3 Exploiting Gauss's Law

Gauss's law in integral form

$$\oiint_S \mathbf{E}(\mathbf{r}) \cdot \hat{\mathbf{n}}(\mathbf{r}) dS = \frac{q_{\text{enc}}}{\epsilon_0}$$

is readily exploited when the electric field $\mathbf{E}(\mathbf{r})$ is not given.

To demonstrate, we can use the above to calculate the electric field $\mathbf{E}(\mathbf{r})$ of a point charge Q at the origin. Assuming only $\mathbf{E} = E(r)\hat{\mathbf{r}}$, center a Gaussian sphere of radius R around the charge. Then Gauss's law simplifies to

$$R^2 \int_0^\pi \int_0^{2\pi} E(r) \sin(\theta) d\theta d\phi = \frac{Q}{\epsilon_0},$$

from which we conclude

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \frac{Q}{R^2} \hat{\mathbf{r}},$$

as expected.

Infinite Line of Charge

Consider an infinite line of charge that carries uniform charge density $\lambda(\mathbf{r})$. Enclose a segment of the line in a Gaussian cylinder of radius r and total length L . The line of charge coincides with the cylinder's axis.

The charge enclosed within the cylinder is simply the density times the length, so we write

$$q_{\text{enc}} = \lambda L.$$

There is zero net electric field through the endcaps of the cylinder due to symmetry arguments. At any point on the wrapping surface, we note the normal vector $\hat{\mathbf{r}}(\mathbf{r})$ is always parallel to $\mathbf{E}(\mathbf{r})$, also by symmetry arguments.

With this, Gauss's law reduces to

$$E \iint_{\text{cyl.}} dS = \frac{\lambda L}{\epsilon_0},$$

where the remaining integral is over the cylinder of the surface, excluding the endcaps. Thus

$$E(2\pi rL) = \frac{\lambda L}{\epsilon_0},$$

and the factor of L cancels. In vector form, we finally write

$$\mathbf{E}(\mathbf{r}) = \frac{\lambda \hat{\mathbf{r}}}{2\pi\epsilon_0 r}$$

for an infinite line of charge.

Infinite Plane of Charge

Consider an infinite xy -plane of charge that carries uniform charge density $\sigma(\mathbf{r})$. Place a Gaussian cylinder of area A on the surface so the end of the cylinder extends height L above the plane. The normal to the plane coincides with the cylinder's axis, namely $\hat{\mathbf{z}}$.

The charge enclosed within the cylinder is the density times the area of one endcap, so we write

$$q_{\text{enc}} = \sigma A.$$

There is zero net electric field through the curved side of the cylinder due to symmetry arguments. At any point on the endcap, we note the normal vector $\hat{\mathbf{z}}(\mathbf{r})$ is always parallel to $\mathbf{E}(\mathbf{r})$, also by symmetry arguments.

With this, Gauss's law reduces to

$$E \iint_{\text{cyl.}} dS = \frac{\sigma A}{\epsilon_0},$$

where the remaining integral is over the endcap of cylinder above the plane. Thus

$$EA = \frac{\sigma A}{\epsilon_0},$$

and the factor of A cancels. In vector form, we finally write

$$\mathbf{E}(\mathbf{r}) = \frac{\sigma \hat{\mathbf{z}}}{\epsilon_0}$$

for an infinite plane of charge.

10 Line Integral

Chapter 20

Complex Analysis

1 Complex Algebra Review

1.1 Complex Number

Let any complex number z be represented as

$$z = x + iy ,$$

where x and y are real numbers, and i is the imaginary unit satisfying

$$i^2 = -1 .$$

The x -component is called the ‘real part’ of z , written

$$x = \operatorname{Re}(z) ,$$

and the y -component is the ‘imaginary part’ of z :

$$y = \operatorname{Im}(z)$$

Complex Conjugate

Any complex number z has a complex conjugate $\bar{z} = z^*$, also a complex number, defined such that

$$\bar{z} = z^* = x - iy .$$

That is, the complex conjugate simply reverses the sign on the imaginary component.

Scalar Multiplication

A complex number z can be scaled by a dimensionless real number λ by multiplying λ into each component of z :

$$\lambda z = \lambda x + i\lambda y$$

1.2 Complex Arithmetic

For two complex numbers z_1, z_2 , the equations of complex arithmetic can be generated from two statements

$$\begin{aligned} z_1 * z_2 &= z_2 * z_1 \\ \overline{z_1 * z_2} &= \overline{z_1} * \overline{z_2} , \end{aligned}$$

where the generalized operator represents either complex addition (+) or complex multiplication ($\cdot$).

From the above, one finds

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2)$$

and

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1) .$$

It’s straightforward to show that complex addition and multiplication follow the standard associative and distributive properties:

$$\begin{aligned} (z_1 z_2) z_3 &= z_1 (z_2 z_3) \\ z_1 (z_2 + z_3) &= (z_1 z_2) + (z_1 z_3) \end{aligned}$$

Complex Magnitude

The complex magnitude

$$|z| = \sqrt{z\bar{z}} = \sqrt{(x + iy)(x - iy)} = \sqrt{x^2 + y^2}$$

is a real number that measures the distance from z to the origin in the complex plane.

Complex Division

For two complex number z_1, z_2 , the ratio is defined as:

$$\frac{z_1}{z_2} = \frac{z_1 \bar{z_2}}{|z_2|^2}$$

One can readily check that this definition readily satisfies $\overline{z_1 * z_2} = \overline{z_1} * \overline{z_2}$, where the generalized arithmetic operator $(*)$ is replaced by the division symbol $(/)$.

1.3 Complex Plane

Polar Representation

Complex numbers lend naturally to polar representation

$$\begin{aligned}x &= r \cos(\phi) \\y &= r \sin(\phi) \\r &= |z| = \sqrt{x^2 + y^2} \\\phi &= \arctan(y/x),\end{aligned}$$

where ϕ is the complex phase of z .

This setup is identical to that of plane polar coordinates with the real numbers along the x -axis and the imaginary numbers along the y -axis. The complex number z can be written

$$z = r (\cos(\phi) + i \sin(\phi)).$$

Rotations

The special complex number z_θ with $|z_\theta| = 1$ and any phase θ is a rotation operator for complex numbers:

$$z_\theta = \cos(\theta) + i \sin(\theta)$$

This is because the product $z_\theta z$ for any $z(r, \phi)$ results in:

$$z_\theta z = r (\cos(\phi + \theta) + i \sin(\phi + \theta))$$

That is multiplying by z_θ has the effect of a change of phase $\phi \rightarrow \phi + \theta$.

1.4 Euler's Formula

Repeated Rotations

Making repeated use of the rotation operator z_θ , suppose we start with a complex number $z(r, \phi)$ and make n identical rotations by the angle θ :

$$z_\theta^n z = r (\cos(\phi + n\theta) + i \sin(\phi + n\theta))$$

Without loss of generality, we can suppose the original complex number z is the real number $z = 1$, meaning $r = 1$ and $\phi = 0$. This yields a new way to write the rotation operator as

$$z_\theta = \left(\cos\left(\frac{\theta}{h}\right) + i \sin\left(\frac{\theta}{h}\right) \right)^h,$$

where $h = 1/n$.

Pressing the limit $h \rightarrow \infty$, the quantity θ/h tends to zero, warranting the small-angle approximation to replace both trigonometry terms:

$$z_\theta = \lim_{h \rightarrow \infty} \left(1 + \frac{i\theta}{h} \right)^h$$

The right side is precisely the definition of Euler's constant e raised to the power $i\theta$. In summary, we have found

$$z_\theta = e^{i\theta} = \cos(\theta) + i \sin(\theta),$$

one of the most useful relationships in mathematics.

Calculus-based Derivation

Begin with the polar representation of a complex number

$$z = r (\cos(\phi) + i \sin(\phi)),$$

and compute the differential dz . From calculus, we know

$$dz = \frac{\partial z}{\partial r} dr + \frac{\partial z}{\partial \phi} d\phi,$$

where the ∂ symbol denotes a partial derivative. Evaluating this and simplifying, find

$$\frac{dz}{z} = \frac{dr}{r} + i d\phi.$$

With all variables separated, integrate the above to find

$$\ln(z) = \ln(r) + i\phi + \mathcal{C},$$

dropping the integration constant. This result is simply the natural log of Euler's formula:

$$z = r e^{i\phi}$$

Multiplication and Division

Euler's formula makes quick work of multiplication and division of complex numbers. For any two complex numbers $z_1(r_1, \phi_1)$, $z_2(r_2, \phi_2)$, we always have

$$\begin{aligned}z_1 z_2 &= r_1 r_2 e^{i(\phi_1 + \phi_2)} \\\frac{z_1}{z_2} &= \frac{r_1}{r_2} e^{i(\phi_1 - \phi_2)}\end{aligned}$$

Complex Logarithm

Consider a general complex number

$$z(r, \phi) = r e^{i\phi},$$

which admits a logarithmic form

$$\ln(z) = \ln(r) + i\phi.$$

Branches

Unlike $z(r, \phi)$, the complex logarithm of z has a complex phase term $i\phi$ that does not ‘reset’ outside the interval $[0 : 2\pi)$. This raises an important subtlety called branches, or branch cuts, which are apparent phase discontinuities in complex functions projected onto the complex plane.

Complex Exponent

For two complex numbers $z(r, \phi)$, $w(\alpha, \beta)$, the exponent calculation z^w proceeds as:

$$\begin{aligned} z^w &= (r e^{i\phi})^{\alpha+i\beta} \\ &= r^\alpha r^{i\beta} e^{i\phi\alpha} e^{-\phi\beta} \\ &= e^{\ln(r)\alpha} e^{\ln(r)i\beta} e^{i\phi\alpha} e^{-\phi\beta} \\ &= e^{\alpha \ln(r) - \beta\phi} e^{i(\beta \ln(r) + \alpha\phi)} \\ &= \exp((\ln(r) + i\phi)(\alpha + i\beta)) \\ &= e^{w \ln(z)} \end{aligned}$$

2 Solving Classic Systems

Complex numbers are a pathway to many abilities some would consider to be unusual.

2.1 Velocity and Acceleration

If a complex number z depends on time via

$$z(t) = r(t) e^{i\theta(t)},$$

we may take derivatives to write equations for ‘velocity’ and ‘acceleration’ in the complex plane. Letting

$$\begin{aligned}\frac{dr}{dt} &= \dot{r} \\ \frac{d\theta}{dt} &= \dot{\theta} = \omega,\end{aligned}$$

one finds:

$$\begin{aligned}\frac{d}{dt}z(t) &= \dot{r}e^{i\theta} + ie^{i\theta}r\omega \\ \frac{d^2}{dt^2}z(t) &= e^{i\theta}(\ddot{r} - r\omega^2) + ie^{i\theta}(2\dot{r}\omega + r\dot{\omega})\end{aligned}$$

The results for $\dot{z}$ and $\ddot{z}$ are each complex numbers, carrying real and imaginary components. By associating

$$\begin{aligned}e^{i\theta} &\rightarrow \hat{r} \\ ie^{i\theta} &\rightarrow \hat{\theta},\end{aligned}$$

we discover a shortcut for the velocity and acceleration vectors in polar coordinates.

2.2 Simple Harmonic Oscillator

Near a stable point, many classical systems are characterized by a position $x(t)$ that obeys

$$\ddot{x} + \omega_0^2 x = 0,$$

where the ‘double-dot’ operator implies two time derivatives, i.e. $\ddot{x} = d^2x/dt^2$, and the angular frequency ω_0 is a real-valued constant. This is the so-called simple harmonic oscillator having known solutions based on trigonometric functions.

Setup

The SHO problem is elegantly solved using complex numbers. Let us pose the same problem for a complex-valued, time-dependent variable $w(t)$:

$$\ddot{w} + \omega_0^2 w = 0$$

As a complex number w , decomposes to

$$w(t) = x(t) + iy(t),$$

or in terms of real and imaginary components

$$\begin{aligned}x(t) &= \text{Re}(w(t)) \\ y(t) &= \text{Im}(w(t)),\end{aligned}$$

specifically

$$\begin{aligned}\ddot{x} + \omega_0^2 x &= 0 \\ \ddot{y} + \omega_0^2 y &= 0.\end{aligned}$$

Solution

Next, we propose solutions to $w(t)$ as

$$w(t) = A e^{\lambda t},$$

where A is the amplitude constant and λ is a frequency constant. Substitution of $w(t)$ into the SHO differential equation quickly reveals

$$\lambda^2 + \omega_0^2 = 0,$$

telling us

$$\lambda = \pm i\omega_0.$$

Since the SHO differential equation is linear, it follows that the general solution is the sum of partial solutions

$$w(t) = A_1 e^{i\omega_0 t} + A_2 e^{-i\omega_0 t}$$

Proceed by writing the complex constants A_1, A_2 in polar form

$$\begin{aligned}A_1 &= a_1 e^{i\phi_1} \\ A_2 &= a_2 e^{i\phi_2},\end{aligned}$$

where a_1, a_2 are real-valued, and ϕ_1, ϕ_2 are phase constants.

Using Euler’s formula to expand the exponential terms, we find

$$\begin{aligned}w(t) &= a_1 \cos(\phi_1 + \omega_0 t) + a_2 \cos(\phi_2 - \omega_0 t) \\ &\quad + ia_1 \sin(\phi_1 + \omega_0 t) + ia_2 \sin(\phi_2 - \omega_0 t),\end{aligned}$$

or, splitting the real from the imaginary parts:

$$\begin{aligned}x(t) &= a_1 \cos(\phi_1 + \omega_0 t) + a_2 \cos(\phi_2 - \omega_0 t) \\ y(t) &= a_1 \sin(\phi_1 + \omega_0 t) + a_2 \sin(\phi_2 - \omega_0 t)\end{aligned}$$

These solutions are identical up to the phase constants ϕ_1, ϕ_2 . For instance, let $\phi_2 \rightarrow \phi_2 + \pi/2$ to transform the sines into cosines. Proceeding with the $x(t)$ -equation, let

$$\begin{aligned} a &= a_1 \cos(\phi_1) + a_2 \cos(\phi_2) \\ b &= -a_1 \sin(\phi_1) + a_2 \sin(\phi_2), \end{aligned}$$

and we get

$$x(t) = a \cos(\omega_0 t) + b \sin(\omega_0 t).$$

Or, to make the solution even tighter, let

$$\begin{aligned} a &= R \cos(\phi_0) \\ b &= R \sin(\phi_0), \end{aligned}$$

and $x(t)$ becomes

$$x(t) = R \cos(\omega_0 t - \phi_0).$$

Note that the number of free constants is down to two, which should be the case for a second-order differential equation.

Damped Harmonic Oscillator

Now we consider the differential equation for the damped harmonic oscillator given by

$$\ddot{x} + b\dot{x} + \omega_0^2 x = 0,$$

where b is the damping coefficient.

Following the same procedure that applies to the SHO case, we replace all $x(t)$ with $w(t)$ and consider complex solutions

$$w(t) = A e^{\lambda t},$$

immediately leading to

$$\lambda^2 + b\lambda + \omega_0^2 = 0.$$

The two solutions for λ come out as

$$\lambda_{\pm} = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - \omega_0^2}.$$

The way b relates to ω_0 dictates the overall character of the solution.

Overdamped Harmonic Oscillator

In the special case $b/2 > \omega_0$, the damping term is strong enough to overwhelm the system's tendency to oscillate, and the solution decays exponentially without oscillating. (All relevant variables in the problem are real-valued.) Up to arbitrary constants determined by initial conditions, the overdamped oscillator obeys

$$x(t) = A_1 e^{\lambda_+ t} + A_2 e^{\lambda_- t}.$$

Underdamped Harmonic Oscillator

If we instead have $b/2 < \omega_0$, the λ -terms become

$$\lambda_{\pm} = -\frac{b}{2} \pm i\sqrt{\omega_0^2 - \frac{b^2}{4}},$$

now including an imaginary component. Utilizing the SHO analysis above, the general solution to this case reads

$$w(t) = e^{-bt/2} \left(a_1 e^{\tilde{\omega}t + \phi_1} + a_2 e^{-\tilde{\omega}t + \phi_2} \right),$$

where

$$\tilde{\omega} = \sqrt{\omega_0^2 - b^2/4}.$$

The damping term causes the amplitude to decay exponentially in time. Note the oscillatory portion of the solution is identical to that of the simple harmonic oscillator, and can be reduced to sine and/or cosine terms with two arbitrary constants. Note that the effective angular frequency $\tilde{\omega}$ depends on the damping term.

Critically-Damped Harmonic Oscillator

The behavior of the damped oscillator depends chiefly on the quantity $\omega_0^2 - b^2/4$, giving rise to either damped or oscillatory motion. The special case $\omega_0^2 = b^2/4$ is the criteria for *critical damping*. Physically, a critically-damped system returns to its equilibrium position in the shortest time.

Returning to the proposed solution $w(t) = A \exp(\lambda t)$ under the condition

$$\omega_0^2 - \frac{b^2}{4} = 0,$$

we see there is only one choice for λ , namely $\lambda = b/2$. As such, this would mean there is only one arbitrary parameter A in the the solution, which is one too few parameters to qualify as a general solution to a second-order differential equation. In other words, something is wrong with our guess for $w(t)$.

Starting the problem over again by introducing the shorthand notation

$$\frac{d}{dt}x(t) = \dot{x} = d_t x(t),$$

the differential equation of the damped oscillator can be written

$$[d_{tt} + b d_t + \omega_0^2] x(t) = 0,$$

where the quantity in square brackets is an object that operates on $x(t)$. Proceed by ‘completing the square’ within the operator to land at

$$\left[d_t + \frac{b}{2}\right]^2 x(t) = \left(\frac{b^2}{4} - \omega_0^2\right) x(t) .$$

For convenience, use the shorthand

$$D = \left[d_t + \frac{b}{2}\right] \\ -\tilde{\omega}^2 = \frac{b^2}{4} - \omega_0^2$$

to write

$$D^2 x(t) = -\tilde{\omega}^2 x(t) .$$

Next, let

$$x(t) = e^{-bt/2} \tilde{x}(t) ,$$

and the above becomes

$$e^{-bt/2} D^2 \tilde{x}(t) + \tilde{x}(t) D^2 e^{-bt/2} = -\tilde{\omega}^2 e^{-bt/2} \tilde{x}(t) .$$

Dealing with the middle term first, notice

$$D^2 e^{-bt/2} = \left[d_{tt} + b d_t + \frac{b^2}{4}\right] e^{-bt/2} \\ = \left(\frac{b^2}{4} - \frac{b^2}{2} + \frac{b^2}{4}\right) e^{-bt/2} = 0 ,$$

and the above simplifies to

$$D^2 \tilde{x}(t) = -\tilde{\omega}^2 \tilde{x}(t) .$$

At this point, we finally apply the case of critical damping, in where $\tilde{\omega} = 0$. This reduces the above to

$$D^2 \tilde{x}(t) = 0 ,$$

having general solution

$$\tilde{x}(t) = a_1 + a_2 t .$$

To see this quickly, you may treat D as a derivative operator and mentally integrate both sides of the equation. If skeptical, formally unpack the operation via $D = d_t + b/2$ and find the same result. Finally, we assemble the solution to the critically-damped oscillator:

$$x(t) = e^{-bt/2} (a_1 + a_2 t)$$

Driven Harmonic Oscillator

It’s also possible to analyze the so-called driven oscillator, generally described by

$$\ddot{x} + b\dot{x} + \omega_0^2 x = f(t) .$$

Due to the linearity in the left hand side, we know already that the solution to the above takes the form

$$x(t) = x_h(t) + x_p(t) ,$$

i.e. the sum of a homogeneous part $x_h(t)$ and a particular part $x_p(t)$. Knowing $x_h(t)$ already to be

$$x_h(t) = R e^{-bt/2} \cos(\omega_0 t - \phi_0) ,$$

the task is reduced to finding a particular solution $x_p(t)$.

In general, finding the particular solution to the above with arbitrary $f(t)$ is as difficult as it sounds, so we make the job easier by considering a sinusoidal driving function

$$f(t) = \gamma \cos(\alpha t) .$$

To solve the problem on hand, it’s convenient to convert all variables to polar form, and then take only the real part of the solution. Proceeding this way, we have, for a complex variable $w(t)$,

$$\ddot{w} + b\dot{w} + \omega_0^2 w = \gamma e^{i\alpha t} .$$

Next, we postulate complex solutions of the form

$$w(t) = A e^{i\alpha t}$$

for some arbitrary constant A , and the above becomes

$$A \left(-\alpha^2 + ib\alpha + \omega_0^2\right) e^{i\alpha t} = \gamma e^{i\alpha t} .$$

Solving for A gives

$$A = \frac{\gamma (\omega_0^2 - \alpha^2) - i\gamma b\alpha}{(\omega_0^2 - \alpha^2)^2 + b^2\alpha^2} ,$$

where if we let

$$u = \omega_0^2 - \alpha^2 \\ v = b\alpha ,$$

a complex number q can be written as

$$q = u - iv = q_0 e^{-i\phi_0} ,$$

with

$$q_0 = \sqrt{(\omega_0^2 - \alpha^2)^2 + b^2\alpha^2} .$$

Rewriting A , we have:

$$A = \frac{\gamma q_0 e^{-i\phi_0}}{q_0^2} = \frac{\gamma e^{-i\phi_0}}{\sqrt{(\omega_0^2 - \alpha^2)^2 + b^2\alpha^2}}$$

Finally, the particular solution to the driven oscillator reads

$$w(t) = \frac{\gamma e^{i(\alpha t - \phi_0)}}{\sqrt{(\omega_0^2 - \alpha^2)^2 + b^2\alpha^2}}$$

$$\phi_0 = \tan^{-1} \left(\frac{b\alpha}{\omega_0^2 - \alpha^2} \right)$$

Note that this particular solution lacks a term like $\exp(bt/2)$, which, much unlike the homogeneous solution, doesn't decay over long times.

The amplitude A depends on how ω_0 relates to α . Calculating

$$\frac{d}{d\alpha} |A(\alpha)| = 0$$

indicates the special α_R for which the amplitude is maximal. Performing this calculation, one finds

$$\alpha_R^2 = \omega_0^2 - \frac{b^2}{2},$$

indicating

$$A_R = \frac{\gamma/b}{\sqrt{\omega_0^2 - b^2/2}}.$$

Note that when $\alpha = \alpha_R$, the system is said to be in *resonance*.

2.3 Particle in Magnetic Field

Consider a particle of mass m and charge q in the presence of a uniform magnetic field $\mathbf{B}$. The force incident on the particle is given by

$$\mathbf{F} = q \mathbf{v}(t) \times \mathbf{B},$$

where $\mathbf{v}(t)$ is the instantaneous velocity. Equations of motion are determined by applying Newton's second law

$$\mathbf{F} = m \frac{d}{dt} \mathbf{v}(t).$$

Being a three-dimensional problem, let us choose to align the magnetic field with the positive z -axis. Then, by eliminating $\mathbf{F}$, we have

$$m \frac{d}{dt} \mathbf{v}(t) = qB \mathbf{v}(t) \times \hat{\mathbf{z}},$$

or

$$\frac{d}{dt} \mathbf{v}(t) = \omega_0 \mathbf{v}(t) \times \hat{\mathbf{z}}$$

$$\omega_0 = \frac{qB}{m}.$$

The unpacks into three equations:

$$\frac{d}{dt} v_x(t) = \omega_0 v_y(t)$$

$$\frac{d}{dt} v_y(t) = -\omega_0 v_x(t)$$

$$\frac{d}{dt} v_z(t) = 0$$

Noting that $v_z(t)$ is constant in time, we immediately know the solution for $z(t)$, namely

$$z(t) = z_0 + v_z t,$$

and we may focus entirely on what occurs in the xy -plane. Defining a complex variable

$$w(t) = v_x(t) + i v_y(t),$$

we capture the information written above using the derivative

$$\frac{d}{dt} w(t) = \omega_0 v_y(t) - i \omega_0 v_x(t),$$

simplifying to

$$i \frac{d}{dt} w(t) = \omega_0 w(t).$$

This first-order differential equation is easily solved by

$$w(t) = A e^{-i\omega_0 t} = |A| e^{i(\phi_0 - \omega_0 t)},$$

where A is an arbitrary complex constant. Taking the real and imaginary parts of the above, the equations for $v_x(t)$, $v_y(t)$ emerge:

$$v_x(t) = \text{Re}(w(t)) = |A| \cos(\omega_0 t - \phi_0)$$

$$v_y(t) = \text{Im}(w(t)) = -|A| \sin(\omega_0 t - \phi_0)$$

While the above can be integrated separately to attain equations of motion $x(t)$, $y(t)$, we can integrate $w(t)$ directly by introducing a variable

$$R = x(t) + i y(t),$$

whose derivative is $w(t)$:

$$\frac{d}{dt}R(t) = w(t)$$

Then, the integral of the above can be written

$$R(t) = R_0 + |A| \int_0^t e^{i(\phi_0 - \omega_0 t')} dt',$$

which can be solved with relative ease:

$$\begin{aligned} R(t) &= R_0 + |A| e^{i\phi_0} \int_0^t e^{-i\omega_0 t'} dt' \\ &= R_0 + |A| e^{i\phi_0} \frac{1}{-i\omega_0} e^{-i\omega_0 t'} \Big|_0^t \\ &= R_0 + |A| e^{i\phi_0} \frac{i}{\omega_0} (e^{-i\omega_0 t} - 1) \\ &= R_0 + |A| e^{i\phi_0 - i\omega_0 t} \frac{i}{\omega_0} - |A| e^{i\phi_0} \frac{i}{\omega_0} \end{aligned}$$

Repacking the integration constants together, we write

$$\tilde{R}_0 = R_0 - |A| e^{i\phi_0} \frac{i}{\omega_0},$$

the solution reads

$$\begin{aligned} R(t) &= \tilde{R}_0 + |A| e^{i(\phi_0 - \omega_0 t)} \frac{i}{\omega_0} \\ &= \tilde{R}_0 + \frac{|A|}{\omega_0} (i \cos(\omega_0 t - \phi_0) + \sin(\omega_0 t - \phi_0)). \end{aligned}$$

Finally, we find

$$\begin{aligned} x(t) &= \text{Re}(R(t)) = \tilde{x}_0 + \frac{|A|}{\omega_0} \sin(\omega_0 t - \phi_0) \\ y(t) &= \text{Im}(R(t)) = \tilde{y}_0 + \frac{|A|}{\omega_0} \cos(\omega_0 t - \phi_0), \end{aligned}$$

where

$$\begin{aligned} \tilde{x}_0 &= x_0 + \frac{|A|}{\omega_0} \sin(\phi_0) \\ \tilde{y}_0 &= y_0 - \frac{|A|}{\omega_0} \cos(\phi_0). \end{aligned}$$

This is the exact circular motion for a charged particle in a uniform magnetic field.

3 Complex Differentiation

A complex function $w(z)$ of a single variable $z = x + iy$ has the structure

$$w(z) = u(x, y) + iv(x, y),$$

where u and v are the respective real and imaginary components of the function.

If the components of w depend on time, we've seen that taking derivatives of $w(z(t))$ have utility in problem solving. However, a treatment of complex derivatives require care in the general case.

3.1 Partial Derivatives

Begin by calculating the differential of a function $w(u, v)$ while substituting dx and dy for their representations in terms of dz and $d\bar{z}$:

$$\begin{aligned} dw(x, y) &= dx \frac{\partial w}{\partial x} + dy \frac{\partial w}{\partial y} \\ &= \frac{1}{2} (dz + d\bar{z}) \frac{\partial w}{\partial x} + \frac{1}{2i} (dz - d\bar{z}) \frac{\partial w}{\partial y} \\ &= \frac{dz}{2} \left(\frac{\partial w}{\partial x} - i \frac{\partial w}{\partial y} \right) + \frac{d\bar{z}}{2} \left(\frac{\partial w}{\partial x} + i \frac{\partial w}{\partial y} \right) \end{aligned}$$

Meanwhile, the same function w can be written $w(z, \bar{z})$, having differential version

$$\delta w(z, \bar{z}) = \delta z \frac{\partial w}{\partial z} + \delta \bar{z} \frac{\partial w}{\partial \bar{z}}.$$

Comparing the two equations provides a definition for $\partial w / \partial z$ and $\partial w / \partial \bar{z}$:

$$\begin{aligned} \frac{\partial w}{\partial z} &= \frac{1}{2} \left(\frac{\partial w}{\partial x} - i \frac{\partial w}{\partial y} \right) \\ \frac{\partial w}{\partial \bar{z}} &= \frac{1}{2} \left(\frac{\partial w}{\partial x} + i \frac{\partial w}{\partial y} \right) \end{aligned}$$

Polar Frame

In place of x and y , we may cast complex functions in terms of r and θ , leading to the following derivative operators:

$$\begin{aligned} r \frac{\partial}{\partial r} &= x \frac{\partial}{\partial x} + y \frac{\partial}{\partial y} = z \frac{\partial}{\partial z} + \bar{z} \frac{\partial}{\partial \bar{z}} \\ \frac{\partial}{\partial \theta} &= x \frac{\partial}{\partial y} - y \frac{\partial}{\partial x} = iz \frac{\partial}{\partial z} - i\bar{z} \frac{\partial}{\partial \bar{z}} \\ z \frac{\partial}{\partial z} &= \frac{1}{2} \left(r \frac{\partial}{\partial r} - i \frac{\partial}{\partial \theta} \right) \\ \bar{z} \frac{\partial}{\partial \bar{z}} &= \frac{1}{2} \left(r \frac{\partial}{\partial r} + i \frac{\partial}{\partial \theta} \right) \end{aligned}$$

To derive these, let $x = r \cos \theta$ and $y = r \sin \theta$, and then use the chain rule on $w(r, \theta)$ and $w(z, \bar{z})$.

3.2 Total Derivative

While partial derivatives of complex functions are straightforward, the total derivative is trouble. Proceeding in a calculus-101 analogy, we write

$$\frac{dw(z, \bar{z})}{dz} = \lim_{\Delta z \rightarrow 0} \frac{\partial w}{\partial z} + \frac{\partial w}{\partial \bar{z}} \frac{\Delta \bar{z}}{\Delta z},$$

where if $\Delta z = |\Delta z| e^{i\theta}$, then the ratio $\Delta \bar{z} / \Delta z$ becomes $e^{-2i\theta}$, which can have *any* phase θ as $\Delta z \rightarrow 0$. The best thing we can do about the total derivative is to restrict w to have no explicit $\bar{z}$ -dependence, eliminating the second term altogether. We therefore take the following two equations as criteria of the total derivative:

$$\begin{aligned} \frac{dw}{dz} &= \frac{\partial w}{\partial z} \\ \frac{\partial w}{\partial \bar{z}} &= 0 \end{aligned}$$

3.3 Analytic Functions

We have seen that a complex function's simultaneous dependence on the complex point z and its complex conjugate $\bar{z}$ can have an ambiguous total derivative. Such functions where z and $\bar{z}$ appear are denoted as $w(z, \bar{z})$, with the letter w reserved.

Many interesting complex functions only depend on z (with $\bar{z}$ absent), denoted $f(z)$. If the derivative df/dz exists then the function is called *analytic*. Points where df/dz does not exist are called *singular*, where isolated singular points are called *poles*. An *entire* function has no singular points in its domain.

To demonstrate, the following three functions are *not* analytic for all z :

$$\begin{aligned} f_1(z) &= x^2 - y^2 \\ f_2(z) &= x^2 + iy^2 \\ f_3(z) &= r^2 (\cos(\theta) + i \sin(\theta)) \end{aligned}$$

Check this by calculating derivatives of each:

$$\begin{aligned} \frac{\partial}{\partial \bar{z}} f_1(z) &= \frac{\partial}{\partial \bar{z}} (z\bar{z}) = z \neq 0 \\ \frac{\partial}{\partial \bar{z}} f_2(z) &= \frac{1}{2} \left(\frac{\partial f_2}{\partial x} + i \frac{\partial f_2}{\partial y} \right) = x - y \neq 0 \\ \frac{\partial}{\partial \bar{z}} f_3(z) &= \frac{1}{2\bar{z}} \left(r \frac{\partial f_3}{\partial r} + i \frac{\partial f_3}{\partial \theta} \right) = \frac{2rz - irz}{2\bar{z}} \neq 0 \end{aligned}$$

On the other hand, the following functions are analytic for all z :

$$\begin{aligned} f_4(z) &= x^2 + 2ixy - y^2 = z^2 \\ f_5(z) &= \ln(r) + i\theta = \ln(z) \\ f_6(z) &= r^\alpha e^{i\alpha\theta} = z^\alpha \end{aligned}$$

You can puzzle these out from algebra alone. If $\bar{z}$ vanishes from the function, the function is likely analytic.

3.4 Cauchy-Riemann Conditions

Complex functions with no explicit $\bar{z}$ -dependence follow an analogy from vector calculus. Start with $\partial w / \partial \bar{z} = 0$, and take the derivative of

$$f(x, y) = u(x, y) + iv(x, y)$$

using

$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2} \left(\frac{\partial w}{\partial x} + i \frac{\partial w}{\partial y} \right)$$

to write the *Cauchy-Riemann conditions*:

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial v}{\partial y} \\ \frac{\partial v}{\partial x} &= -\frac{\partial u}{\partial y} \end{aligned}$$

Level Curves

For a complex function

$$f(x, y) = u(x, y) + iv(x, y)$$

obeying the Cauchy-Riemann conditions, note that for two constants c_1 and c_2 , the level curves

$$\begin{aligned} u(x, y) &= c_1 \\ v(x, y) &= c_2 \end{aligned}$$

are orthogonal, as

$$\begin{aligned} \nabla u \cdot \nabla v &= \left(\frac{\partial u}{\partial x} \hat{x} + \frac{\partial u}{\partial y} \hat{y} \right) \cdot \left(\frac{\partial v}{\partial x} \hat{x} + \frac{\partial v}{\partial y} \hat{y} \right) \\ &= \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \\ &= -\frac{\partial u}{\partial x} \frac{\partial u}{\partial y} + \frac{\partial u}{\partial y} \frac{\partial u}{\partial x} = 0. \end{aligned}$$

Harmonic Functions

For a complex function

$$f(x, y) = u(x, y) + iv(x, y)$$

obeying the Cauchy-Riemann conditions, the Laplacian always vanishes, as

$$\begin{aligned} \nabla^2 f &= \nabla \cdot \nabla f(x, y) \\ &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 v}{\partial y^2} \\ &= \frac{\partial^2 v}{\partial x^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 u}{\partial y^2} = 0, \end{aligned}$$

or more strongly,

$$\begin{aligned} \nabla^2 u(x, y) &= 0 \\ \nabla^2 v(x, y) &= 0. \end{aligned}$$

Conventionally, $u(x, y)$ and $v(x, y)$ that satisfy the above are called *harmonic functions*.

Connection to Electromagnetism

In two dimensions, consider two real vector fields $\mathbf{A}$ and $\mathbf{B}$ defined in terms of harmonic functions $u(x, y)$, $v(x, y)$:

$$\begin{aligned} \mathbf{A} &= u \hat{x} - v \hat{y} \\ \mathbf{B} &= v \hat{x} + u \hat{y} \end{aligned}$$

Imposing the Cauchy-Riemann equations onto $\mathbf{A}$ and $\mathbf{B}$, we see that the divergence and curl of each field resemble Maxwell's equations in charge-free two-dimensional space:

$$\begin{aligned} \nabla \cdot \mathbf{A} &= 0 \\ \nabla \times \mathbf{A} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{B} &= 0 \end{aligned}$$

Approaching this differently, it turns out that two problems from electromagnetism are solved by the complex function:

$$f(z) = \frac{1}{z - z_0}$$

Letting

$$\begin{aligned} x - x_0 &= \rho \cos \theta \\ y - y_0 &= \rho \sin \theta, \end{aligned}$$

$f(z)$ may be written

$$\begin{aligned} f(z) &= \frac{x - x_0}{(x - x_0)^2 + (y - y_0)^2} \\ &\quad - i \frac{y - y_0}{(x - x_0)^2 + (y - y_0)^2} \\ &= \frac{1}{\rho} (\cos \theta - i \sin \theta) \\ &= u + iv, \end{aligned}$$

$$\begin{aligned} u &= \rho^{-1} \cos \theta \\ v &= -\rho^{-1} \sin \theta. \end{aligned}$$

Then, the fields $\mathbf{A} = u \hat{\mathbf{x}} - v \hat{\mathbf{y}}$ and $\mathbf{B} = v \hat{\mathbf{x}} + u \hat{\mathbf{y}}$ respectively tell us

$$\begin{aligned} \mathbf{A} &= \frac{\cos \theta \hat{\mathbf{x}} + \sin \theta \hat{\mathbf{y}}}{\rho} = \frac{\hat{\mathbf{r}}}{\rho} \\ \mathbf{B} &= \frac{-\sin \theta \hat{\mathbf{x}} + \cos \theta \hat{\mathbf{y}}}{\rho} = \frac{\hat{\boldsymbol{\theta}}}{\rho}. \end{aligned}$$

Explicitly, $\mathbf{A}$ is proportional to the electric field vector due to a line of charge, whereas $\mathbf{B}$ is proportional to the magnetic field vector due to a line of current.

4 Contour Integrals

Now we develop the notion of integration in the complex plane. Consider a contour C that begins and ends at the respective points z_a and z_b in the complex plane. The integral of a function $f(z)$ over C can be recast using a real parameter t via the chain rule:

$$\int_C f(z) dz = \int_{t_a}^{t_b} f(z(t)) \frac{dz(t)}{dt} dt$$

Substituting

$$f(z) = u(x, y) + iv(x, y)$$

and using prime notation for derivatives via

$$z' = x'(t) + iy'(t),$$

the contour integral splits into real and imaginary parts:

$$\int_C f(z) dz = \int_C (ux' - vy') dt + i \int_C (vx' + uy') dt$$

Re-using the notation $\mathbf{A} = u\hat{\mathbf{x}} - v\hat{\mathbf{y}}$, $\mathbf{B} = v\hat{\mathbf{x}} + u\hat{\mathbf{y}}$, write the integral in vector notation

$$\int_C f(z) dz = \int_C \mathbf{A} \cdot d\mathbf{l} + i \int_C \mathbf{B} \cdot d\mathbf{l},$$

where $d\mathbf{l} = dx\hat{\mathbf{x}} + dy\hat{\mathbf{y}}$. Evidently, the the integral of a function $f(z)$ in the complex plane decomposes into a pair of integrals involving the fields $\mathbf{A}$, $\mathbf{B}$.

4.1 Cauchy's Integral Theorem

Starting with the result above, we may consider closed contours C and apply Stokes's theorem to transform each line integral into an area integral in the complex plane. Denoting the the off-plane direction $\hat{\mathbf{k}}$, we have

$$\oint_C f(z) dz = \int_{\Omega} \hat{\mathbf{k}} \cdot (\nabla \times \mathbf{A}) dx dy + i \int_{\Omega} \hat{\mathbf{k}} \cdot (\nabla \times \mathbf{B}) dx dy,$$

which, by the rules of vector calculus, resolves to *zero* when region the Ω is completely enclosed by C .

This is the essence (and the proof) of the Cauchy Integral Theorem, formally stating that if a function $f(z)$ is analytic in a simply-connected region R , then the integral along a closed path C in R equals zero.

4.2 Defects

Singular points in the integration region that cause $f(z)$ to become non-analytic must be 'stepped around' to be excluded from the contour C .

Consider the integral

$$I_n^{(0)} = \oint_C \frac{dz}{(z - z_0)^n},$$

where n is an integer and z_0 is a singular point interior to C . Since the integration contour may be arbitrarily shaped, we may choose a unit circular path around the point z_0 , running counter-clockwise by convention, with

$$z(\theta) = z_0 + e^{i\theta}$$

$$z'(\theta) = ie^{i\theta}.$$

Substituting $z(\theta)$ into the above and using the delta function

$$\delta(t) = \frac{1}{2\pi} \int_0^{2\pi} e^{-i\alpha t} d\alpha,$$

the above integral becomes:

$$I_n^{(0)} = 2\pi i \delta(n - 1) = \begin{cases} 2\pi i & n = 1 \\ 0 & n \neq 1 \end{cases}$$

4.3 Cauchy Integral Formula

Generalizing the analysis of defects, we consider the integral

$$I_n = \oint_C \frac{f(z) dz}{(z - z_0)^n}$$

about a circular path of arbitrarily-small radius

$$z(\theta) = \lim_{r \rightarrow 0} z_0 + re^{i\theta}.$$

The act of taking $r \rightarrow 0$ is equivalent to expanding $f(z_0)$ by Taylor series to discard high-order terms, provided that derivatives of $f(z)$ exist. Using the expansion

$$f(z) = \sum_{q=0}^{\infty} \frac{f^{(q)}(z_0)}{q!} (z - z_0)^q,$$

the above simplifies to

$$I_n = \sum_{q=0}^{\infty} \frac{f^{(q)}(z_0)}{q!} I_{n-q}^{(0)}.$$

The term $I_{n-q}^{(0)}$ contains a delta function in the quantity $n - q - 1$, which nukes all terms in the sum except the one satisfying $q = n - 1$, and thus

$$I_n = \frac{2\pi i}{(n-1)!} f^{(n-1)}(z_0).$$

Rewriting the integral gives us the (very important) Cauchy integral formula:

$$\oint_C \frac{f(z) dz}{(z - z_0)^n} = \frac{2\pi i}{(n-1)!} f^{(n-1)}(z_0)$$

Speaking to the pedantic reader, it's possible to prove that the existence of all derivatives $f^{(n)}(z)$ is guaranteed by the Cauchy integral formula.

4.4 Analytic Continuation

Taylor series is convergent until the contour C touches a singular point z_0 , where the *radius of convergence* corresponds to the largest contour C_0 . In a process called *analytic continuation*, we may choose a point z_1 inside C_0 where Taylor expansion is valid, implying a new contour C_1 centered on z_1 with its own radius of convergence, in which another Taylor expansion for $f(z)$ applies. The non-overlapping part of C_1 is new 'territory' that the z_0 -centered approximation doesn't cover. Iterating this process, we may cover the whole complex plane, as long as singular points (and regions) are stepped around.

It readily follows that a closed contour integral that encloses singularities is equal to the sum of elementary integrals around the singularities. If the region of analyticity is a multiply-connected surface due to singularities, $f(z)$ may be multi-valued.

4.5 Laurent Series

A generalization of the Taylor series that includes both positive and negative exponents is the *Laurent series*:

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n$$

Bringing the Cauchy integral formula

$$I_n = \oint_C \frac{f(z) dz}{(z - z_0)^n}$$

into the mix, we find, by direct substitution:

$$\begin{aligned} I_n &= \oint_C \sum_{m=-\infty}^{\infty} \frac{a_m (z - z_0)^m dz}{(z - z_0)^n} \\ &= \sum_{m=-\infty}^{\infty} a_m \oint_C \frac{dz}{(z - z_0)^{n-m}} \end{aligned}$$

The remaining integral is simply $I_{n-m}^{(0)}$, so we have

$$\begin{aligned} I_n &= \sum_{m=-\infty}^{\infty} a_m I_{n-m}^{(0)} \\ &= \sum_{m=-\infty}^{\infty} a_m 2\pi i \delta(n - m - 1), \end{aligned}$$

so the surviving term satisfies $m = n - 1$. Simplifying quickly gives $I_n = a_{n-1} 2\pi i$, or $I_{n+1} = a_n 2\pi i$.

Solving for a_n , we find a formula for the Laurent series coefficients:

$$\begin{aligned} a_n &= \frac{1}{2\pi i} \oint_{\Gamma} \frac{f(z) dz}{(z - z_0)^{n+1}} \\ n &= 0, \pm 1, \pm 2, \dots \end{aligned}$$

Note that Γ is a contour topologically equivalent to C_0 .

Example: Annulus

Consider a function $f(z)$ that is analytic in the annulus

$$R_1 \leq |z - z_0| \leq R_0$$

centered on z_0 . Begin by writing the $n = 0$ case of the Cauchy integral formula

$$2\pi i f(z) = \oint_C \frac{f(\tilde{z}) d\tilde{z}}{\tilde{z} - z},$$

where the function $f(z)$ is approximated by Laurent series, and contour C encloses z .

Next, stretch C so as to wrap the inside of the annulus, with a tight 'bridge' of canceling paths that connect the enclosing radii. The resulting contours are C_1 and C_0 with opposing directions of integration. The contour integral along C_0 corresponding to R_0 was solved previously and generates the $n \geq 0$ terms:

$$a_{n \geq 0} = \frac{1}{2\pi i} \oint_{C_0} \frac{f(z) dz}{(z - z_0)^{n+1}}$$

Along contour C_1 , the fraction $1/(\tilde{z} - z)$ may be expanded via geometric series

$$\begin{aligned} \frac{1}{\tilde{z} - z} &= \frac{1}{\tilde{z} - z_0 + z_0 - z} \\ &= \frac{1}{z_0 - z} \frac{1}{1 + \frac{\tilde{z} - z_0}{z_0 - z}} = \sum_{m=0}^{\infty} \frac{(z_0 - \tilde{z})^m}{(z_0 - z)^{m+1}}, \end{aligned}$$

which guarantees convergence as

$$|z - z_0| > |\tilde{z} - z_0| = R_1.$$

To discover the restriction on a_n along C_1 , replace $\tilde{z} - z_0$ and further $f(\tilde{z})$ in the integral formula as follows:

$$\begin{aligned} 2\pi i f(z) &= \oint_{C_1} \frac{f(\tilde{z}) d\tilde{z}}{\tilde{z} - z_0} \\ &= \oint_{C_1} \sum_{n=-\infty}^{\infty} a_n (\tilde{z} - z_0)^n \\ &\quad \times \sum_{m=0}^{\infty} \frac{(z_0 - \tilde{z})^m}{(z_0 - z)^{m+1}} d\tilde{z} \end{aligned}$$

Keep condensing terms to write

$$\begin{aligned} 2\pi i f(z) &= \sum_{n=-\infty}^{\infty} a_n \sum_{m=0}^{\infty} (z - z_0)^{-(m+1)} \\ &\quad \times \oint_{C_1} (\tilde{z} - z_0)^{m+n} d\tilde{z}, \end{aligned}$$

$$\begin{aligned} 2\pi i f(z) &= 2\pi i \sum_{n=-\infty}^{\infty} a_n \\ &\quad \times \sum_{m=0}^{\infty} (z - z_0)^{-(m+1)} \delta(-(m+n)-1). \end{aligned}$$

The delta functions tells us $m + n = -1$, and we find

$$f(z) = \sum_{n=-1}^{-\infty} a_n (z - z_0)^n.$$

Evidently, only the negative n -terms have survived on contour C_1 . We conclude that

$$a_{n<0} = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z) dz}{(z - z_0)^{n+1}}.$$

5 Residue Calculus

Our study of contour integrals in the complex plane has yielded several useful results. First, the Cauchy integral theorem tells us that an analytic function $f(z)$ integrated along a closed contour C free of singularities always resolves to zero.

Isolated singularities (poles) $z_0^{(p)}$ are handled by expanding $f(z)$ as a Laurent series

$$f(z) = \sum_{n=-\infty}^{\infty} a_n^{(p)} (z - z_0^{(p)})^n,$$

where the coefficients a_n were found to be

$$a_n^{(p)} = \frac{1}{2\pi i} \oint_{C_0^{(p)}} \frac{f(z) dz}{(z - z_0^{(p)})^{n+1}}$$

$$n = 0, \pm 1, \pm 2, \dots$$

Take the special case $n = -1$ to write

$$2\pi i a_{-1}^{(p)} = \oint_{C_0^{(p)}} f(z) dz,$$

telling us that the integral of $f(z)$ around a pole is equal to the constant $2\pi i a_{-1}^{(p)}$.

This process may repeat for each pole inside the contour C , resulting in the sum

$$2\pi i \sum_p a_{-1}^{(p)} = \oint_C f(z) dz.$$

This amazing result says that solving contour integrals is reduced to finding the Laurent coefficients $a_{-1}^{(p)}$ at each pole $z_0^{(p)}$. The coefficient $a_{-1}^{(p)}$ is called the *residue* at $z_0^{(p)}$:

$$2\pi i \sum_p \text{Res} \left[f \left(z_0^{(p)} \right) \right] = \oint_C f(z) dz$$

Calculating Residue(s)

The *order* of a pole $z_0^{(p)}$ is the lowest (most negative) index of the Laurent series expansion of $f(z)$ around the pole. A *simple* pole has a lowest index of -1 . In general, a pole z_0 of order m has corresponding Laurent series

$$f(z) = \sum_{n=-m}^{\infty} a_n (z - z_0)^n.$$

Now we introduce the function $g(z)$ such that

$$g(z) = (z - z_0)^m f(z) = \sum_{q=0}^{\infty} a_{q-m} (z - z_0)^q,$$

which bumps z_0 to the numerator. Since the sum index begins at zero, $g(z)$ is simply a Taylor series, meaning

$$a_{q-m} = \frac{g^{(q)}(z_0)}{(q)!},$$

where the case $q - m = -1$ gives the residue of $f(z)$ at z_0 :

$$\text{Res} [f(z_0)] = \frac{g^{(m-1)}(z_0)}{(m-1)!}$$

For functions containing only simple poles, the above reduces to

$$\text{Res} [f(z_0)] = g(z_0).$$

What we see is an integral on the left, and a simple function evaluation on the right. In practice, this means that whole families of integrals can be cheated by choosing an integration contour that makes the residue easy to calculate.

5.1 Ratios

For functions of the form

$$f(z) = \frac{p(z)}{q(z)},$$

containing a simple pole in the denominator, i.e. $q(z_0) = 0$, it's simple to show that the residue calculation always resolves to

$$\text{Res} [f(z_0)] = \frac{p(z_0)}{q'(z_0)}.$$

To prove this, write the definition of $q'(z)$ and simplify to produce

$$\lim_{z \rightarrow z_0} \frac{z - z_0}{q(z)} = \frac{1}{q'(z)},$$

and then eliminate $z - z_0$ using

$$g(z) = (z - z_0) f(z) = (z - z_0) \frac{p(z)}{q(z)}$$

to get

$$\lim_{z \rightarrow z_0} g(z) = \frac{p(z)}{q'(z)} = \text{Res} [f(z_0)].$$

5.2 Polynomial Functions

Consider the integral

$$I = \int_{-\infty}^{\infty} \frac{dx}{1+x^2}$$

whose domain is the real number line. If we connect $x = \infty$ to $x = -\infty$ with a (counterclockwise) semicircular arc, the resulting contour encloses the upper half of the complex plane.

Factoring the denominator to clearly see the singular points, the integral becomes

$$I = \oint_C \frac{dz}{(z-i)(z+i)},$$

which encloses one simple pole $z_0 = i$. The pole at $z = -i$ is outside the integration contour and is ignored.

Proceed by writing

$$g(z) = (z-i)f(z)$$

and quickly find

$$\text{Res}[f(i)] = \frac{1}{2i},$$

and the integral resolves to

$$I = 2\pi i \text{Res}[f(i)] = \pi$$

Problem 1

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{dx}{(x^2+1)(x^2+4)}$$

Answer: Contour contains two poles.

$$I = \oint_C \frac{dz}{(z+i)(z-i)(z+2i)(z-2i)}$$

$$g_1(z) = \frac{(z-i)}{(z-i)(z+i)(z^2+4)}$$

$$g_2(z) = \frac{(z-2i)}{(z^2+1)(z-2i)(z+2i)}$$

$$I = 2\pi i (g_1(i) + g_2(2i)) = \frac{\pi}{6}$$

(Or use partial fractions.)

Problem 2

Evaluate:

$$I = \int_0^{\infty} \frac{x^2 dx}{(x^2+4)(x^2+9)}$$

Answer: Contour contains two poles.

$$2I = \oint_C \frac{z^2 dz}{(z+2i)(z-2i)(z+3i)(z-3i)}$$

$$g_1(z) = \frac{z^2(z-2i)}{(z+2i)(z-2i)(z^2+9)}$$

$$g_2(z) = \frac{z^2(z-3i)}{(z^2+4)(z+3i)(z-3i)}$$

$$I = \frac{2\pi i}{2} (g_1(2i) + g_2(3i)) = \frac{\pi}{10}$$

Problem 3

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{dx}{(x^2+1)^2}$$

Answer: Contour contains one order-two pole.

$$I = \oint_C \frac{dz}{((z+i)(z-i))^2}$$

$$g(z) = \frac{(z-i)^2}{((z+i)(z-i))^2}$$

$$I = 2\pi i \left(\frac{d}{dz} g(z) \Big|_{z=i} \right) = \frac{\pi}{2}$$

Problem 4

Evaluate:

$$I = \int_0^{\infty} \frac{dx}{(4x^2+1)^3}$$

Answer: Contour contains one order-3 pole.

$$2I = \oint_C \frac{dz}{4^3 (z + \frac{i}{2})^3 (z - \frac{i}{2})^3}$$

$$g(z) = \frac{(z - \frac{i}{2})^3}{4^3 (z + \frac{i}{2})^3 (z - \frac{i}{2})^3}$$

$$I = \frac{2\pi i}{2} \left(\frac{1}{2} \frac{d^2}{dz^2} g(z) \Big|_{z=i/2} \right) = \frac{3\pi}{32}$$

5.3 Jordan's Lemma

Starting with the Fourier transform integral

$$I = \int_{-\infty}^{\infty} f(x) e^{ikx} dx,$$

we carry the problem to the complex plane under three assumptions: (i) $k > 0$ and is a real number, (ii) $f(z)$ is analytic in the upper-half plane with the exception of simple poles, (iii) $\lim_{|z| \rightarrow \infty} f(z) = 0$.

Jordan's lemma for Fourier transform states that the integration path can be closed by an infinite semicircle in the upper-half plane. For the $k < 0$ case, the path would enclose the lower half-plane.

5.4 Sine and Cosine in Polynomial

For a real variable $a > 0$, the integral

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \frac{\cos kx}{x^2 + a^2} dx \\ &= \operatorname{Re} \int_{-\infty}^{\infty} \frac{e^{ikx}}{x^2 + a^2} dx = \oint_C \frac{e^{ikz}}{z^2 + a^2} dz \end{aligned}$$

is easily recast as a contour integral using Jordan's lemma.

Using $g(z) = e^{ikz}/(z + ia)$, we quickly find $I = (\pi/a) e^{-ka}$. Generalizing this, one finds:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) \cos(kx) dx &= \operatorname{Re} \int_{-\infty}^{\infty} f(x) e^{ikx} dx \\ \int_{-\infty}^{\infty} f(x) \sin(kx) dx &= \operatorname{Im} \int_{-\infty}^{\infty} f(x) e^{ikx} dx \end{aligned}$$

Problem 5

Evaluate:

$$I = \int_0^{\infty} \frac{\cos 2x}{9x^2 + 4} dx$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= \frac{1}{18} \operatorname{Re} \oint_C \frac{e^{2iz}}{\left(z + \frac{2i}{3}\right) \left(z - \frac{2i}{3}\right)} dz \\ g(z) &= \frac{e^{2iz}}{z + \frac{2i}{3}} \\ g(2i/3) &= \frac{3}{4i} e^{-4/3} \\ I &= \frac{2\pi i}{18} \frac{3}{4i} e^{-4/3} = \frac{\pi}{12} e^{-4/3} \end{aligned}$$

Problem 6

Evaluate:

$$I = \int_0^{\infty} \frac{\cos 2x}{(9x^2 + 4)^2} dx$$

Answer: Contour contains one order-two pole.

$$\begin{aligned} I &= \frac{1}{2 \cdot 9^2} \operatorname{Re} \oint_C \frac{e^{2iz}}{\left(z + \frac{2i}{3}\right)^2 \left(z - \frac{2i}{3}\right)^2} dz \\ g(z) &= \frac{e^{2iz} \left(z - \frac{2i}{3}\right)^2}{\left(z + \frac{2i}{3}\right)^2 \left(z - \frac{2i}{3}\right)^2} \\ g^{(1)}(2i/3) &= -i \left(\frac{14}{3} \cdot \frac{3^3}{4^3} \right) e^{-4/3} \\ I &= 2\pi i \left(g^{(1)}(2i/3) \right) = \frac{7\pi}{288} e^{-4/3} \end{aligned}$$

Problem 7

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{\cos kx}{(x^2 + a^2)^2} dx$$

Answer: Contour contains one order-two pole.

$$\begin{aligned} I &= \operatorname{Re} \oint_C \frac{e^{ikz}}{(z^2 + a^2)^2} dz \\ g(z) &= \frac{e^{ikz}}{(z + ia)^2} \\ g^{(1)}(ia) &= \frac{i}{4} \frac{ka + 1}{e^{ka} a^3} \\ I &= \frac{\pi}{2} \frac{ka + 1}{e^{ka} a^3} \end{aligned}$$

Problem 8

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{x \sin kx}{x^2 + a^2} dx$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= \operatorname{Im} \oint_C \frac{z e^{ikz}}{z^2 + a^2} dz \\ g(z) &= \frac{z e^{ikz}}{z + ia} \\ g(ia) &= \frac{1}{2} e^{-ka} \\ I &= \operatorname{Im} \frac{2\pi i}{2 e^{ka}} = \frac{\pi}{e^{ka}} \end{aligned}$$

Problem 9

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{x \sin kx}{(x^2 + a^2)^2} dx$$

Answer: Exploit a previous example to ease calculations.

$$\begin{aligned} I &= -\frac{\partial}{\partial k} \int_{-\infty}^{\infty} \frac{\cos kx}{(x^2 + a^2)^2} dx \\ &= -\frac{\partial}{\partial k} \left(\frac{\pi}{2} \frac{e^{ka}}{e^{ka}} \right) = \frac{\pi k}{2a e^{ka}} \end{aligned}$$

or, by standard means:

$$I = \operatorname{Im} \oint_C \frac{z e^{ikz} dz}{(z^2 + a^2)^2} = \operatorname{Im} \frac{2\pi i k}{4a e^{ka}} = \frac{\pi k}{2a e^{ka}}$$

Problem 10

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{x \sin x}{x^2 + 4x + 5} dx$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= \operatorname{Im} \oint_C \frac{z e^{ikz} dz}{(z + 2 + i)(z + 2 - i)} \\ g(z) &= \frac{z e^{ikz} (z + 2 - i)}{(z + 2 + i)(z + 2 - i)} \\ g(-2 + i) &= \frac{(-2 + i) e^{i(-2+i)}}{2i} \\ I &= \frac{\pi}{e} (2 \sin 2 + \cos 2) \end{aligned}$$

5.5 Trigonometric Functions

Integrals of the form

$$I = \int_0^{2\pi} f(\cos(\theta), \sin(\theta)) d\theta$$

can be recast by expressing all θ -terms in terms of z and choosing an integration contour C_0 as the unit circle centered on $z = 0$. Using the equations of complex trigonometry, and also noting that $dz = ir e^{i\theta} d\theta$, the above becomes

$$I = -i \oint_{C_0} f\left(\frac{z + z^{-1}}{2}, \frac{z - z^{-1}}{2i}\right) \frac{dz}{z}.$$

Problem 11

Evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{13 + 5 \sin \theta}$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= -i \oint_{C_0} \frac{dz}{z} \frac{1}{13 + (5/2i)(z - \bar{z})} \\ &= \oint_{C_0} \frac{dz}{\frac{5}{2}(z + \frac{i}{5})(z + 5i)} \\ g(z) &= \frac{(z + \frac{i}{5})}{\frac{5}{2}(z + \frac{i}{5})(z + 5i)} \\ g\left(\frac{-i}{5}\right) &= \frac{-i}{12} \\ I &= 2\pi i \left(\frac{-i}{12}\right) = \frac{\pi}{6} \end{aligned}$$

Problem 12

Evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{5 - 4 \sin \theta}$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= -i \oint_{C_0} \frac{dz}{z} \frac{1}{5 + 2i(z - \bar{z})} \\ &= -\oint_{C_0} \frac{dz}{2(z - \frac{i}{2})(z - 2i)} \\ g(z) &= \frac{-1 \cdot (z - \frac{i}{2})}{(z - \frac{i}{2})2(z - 2i)} \\ g\left(\frac{i}{2}\right) &= \frac{-i}{3} \\ I &= 2\pi i \left(\frac{-i}{3}\right) = \frac{2\pi}{3} \end{aligned}$$

Problem 13

For $a > |b|$, evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{a + b \cos \theta}$$

Answer: Contour contains one pole.

$$\begin{aligned} I &= -i \oint_{C_0} \frac{2 dz}{2az + b(1 + z^2)} \\ &= -i \oint_{C_0} \frac{2 dz}{b\left(z + \frac{a}{b} - \frac{\sqrt{a^2 - b^2}}{b}\right)\left(z + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}\right)} \end{aligned}$$

$$z_0 = -\frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}$$

$$g(z) = \frac{2(z - z_0)}{b\left(z + \frac{a}{b} - \frac{\sqrt{a^2 - b^2}}{b}\right)\left(z + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}\right)}$$

$$g(z_0) = \frac{1}{\sqrt{a^2 - b^2}}$$

$$I = 2\pi i \left(\frac{-i}{\sqrt{a^2 - b^2}} \right) = \frac{2\pi}{\sqrt{a^2 - b^2}}$$

Problem 14

For $a > 1$, evaluate:

$$I = \int_0^\pi \frac{d\theta}{(a + \cos \theta)^2}$$

Answer: Contour contains one order-two pole.

$$I = -i \oint_{C_0} \frac{2z dz}{(2az + 1 + z^2)^2}$$

$$= -i \oint_{C_0} \frac{2z dz}{\left(z + a - \sqrt{a^2 - 1}\right)^2 \left(z + a + \sqrt{a^2 - 1}\right)^2}$$

$$z_0^+ = -a + \sqrt{a^2 - 1}$$

$$z_0^- = -a - \sqrt{a^2 - 1}$$

$$g(z) = \frac{2z(z - z_0^+)^2}{(z - z_0^+)^2(z - z_0^-)^2}$$

$$g^{(1)}(z) = \frac{-2(z + z_0^-)}{(z - z_0^-)^3}$$

$$g^{(1)}(z_0^+) = \frac{a}{2(a^2 - 1)^{3/2}}$$

$$I = 2\pi i \left(\frac{-ia}{2(a^2 - 1)^{3/2}} \right) = \frac{\pi a}{(a^2 - 1)^{3/2}}$$

5.6 Two-Contour Trick

The infinite complex plane (or a fraction of it) need not be enclosed by a semicircular contour. Rectangles are just as valid, which are an ideal application of the *two-contour trick*. This entails noticing when the integral of $f(z)$ on two enclosing contours C_1 and C_2 is the same up to a complex factor.

To illustrate, the integral

$$I = \int_{-\infty}^{\infty} \frac{e^{ax} dx}{1 + e^x}$$

with $0 > a > 1$ may be rewritten

$$I = \int_{C_1} \frac{e^{az} dz}{1 + e^z},$$

where contour C_1 is the real number line. Next, introduce a second contour C_2 that is shifted upward into the imaginary numbers but still parallel to the real line such that

$$z = x + 2\pi i.$$

Integrating ‘backwards’ along C_2 , we have

$$-I e^{2i\pi a} = \int_{C_2} \frac{e^{az} dz}{1 + e^z}.$$

Of course, any contributions to the integral at $x = \pm\infty$ are zero, so we combine C_1 and C_2 to close the integration contour:

$$I(1 - e^{2i\pi a}) = \oint_C \frac{e^{az} dz}{1 + e^z}$$

To finish the calculation above, we note the pole at $z_0 = i\pi$, and then

$$g(z) = \frac{(z - i\pi) e^{az}}{1 + e^z}.$$

However, $g(z_0)$ leads to $0/0$, thus L’hopital’s rule is needed, leading to

$$g(z_0) = \frac{e^{i\pi a}}{e^{i\pi}}.$$

Finally,

$$I(1 - e^{2i\pi a}) = 2\pi i \frac{e^{i\pi a}}{e^{i\pi}},$$

and

$$I = \frac{\pi}{\sin(\pi a)}.$$

Problem 15

Use a pizza slice contour bounded by the positive real line and $z = r e^{2\pi i/n}$ (with vanishing crust at infinity) to evaluate

$$I = \int_0^\infty \frac{dx}{1 + x^n}.$$

Answer: Contour contains one pole.

$$I(1 - e^{2\pi i/n}) = \oint_C \frac{dz}{1 + z^n}$$

$$z_0^n = -1 \rightarrow z_0 = (-1)^{-n} = e^{i\pi/n}$$

$$g(z) = \frac{(z - e^{i\pi/n})}{1 + z^n}$$

$$g(z_0) \propto \frac{0}{0} \rightarrow \text{Need L'hôpital.}$$

$$g(z_0) = -\frac{e^{i\pi/n}}{n}$$

$$I(1 - e^{2\pi i/n}) = -2\pi i \frac{e^{i\pi/n}}{n}$$

$$I = \frac{\pi/n}{\sin(\pi/n)}$$

5.7 Regularization

Principal Value

Consider the *principal value* integral

$$I = \text{P} \int_{-\infty}^{\infty} \frac{f(x) dx}{x - x_0},$$

where by shifting $x \rightarrow z$, we assume that $f(z)$ is analytic except for a finite number of poles, and that $|f| \rightarrow 0$ on the upper (or lower) infinite semicircle in the complex plane.

Since the pole x_0 lies on the real axis, the integration contour cuts directly through x_0 . This is handled by *regularization* of the denominator, which entails introducing a small factor $\delta > 0$ as

$$\begin{aligned} I &= \text{P} \int_{-\infty}^{\infty} \frac{f(x) dx}{x - x_0} \\ &= \lim_{\delta \rightarrow 0} \int_{-\infty}^{\infty} \frac{(x - x_0) f(x) dx}{(x - x_0)^2 + \delta^2} \\ &= \lim_{\delta \rightarrow 0} \oint_C \frac{(z - x_0) f(z) dz}{(z - x_0)^2 + \delta^2}. \end{aligned}$$

After a little complex algebra, find

$$\begin{aligned} I &= \lim_{\delta \rightarrow 0} \oint_C \frac{f(z) dz}{z - x_0 + i\delta} \\ &\quad + \lim_{\delta \rightarrow 0} \oint_C i\delta \frac{f(z) dz}{(z - x_0 - i\delta)(z - x_0 + i\delta)}, \end{aligned}$$

which indicates one simple pole $z_0 = x_0 + i\delta$ inside the upper-half plane. The first integral in fact *excludes* the pole, so x_0 is skipped in subsequent residue calculations. (Use the δ -term as a reminder to skip x_0 .) The second integral is solved by standard residue calculus, i.e., let $g(z) = f(z)/(z - x_0 + i\delta)$, resulting in $\pi i f(x_0)$.

Pulling the results together, we write

$$I^+ = \pi i f(x_0) + \lim_{\delta \rightarrow 0} \oint_C \frac{f(z) dz}{z - x_0 + i\delta},$$

where if we started with $\delta < 0$ instead, the integration contour would flip to the lower-half plane, resulting in

$$I^- = -\pi i f(x_0) + \lim_{\delta \rightarrow 0} \oint_C \frac{f(z) dz}{z - x_0 - i\delta}.$$

In tighter notation (regardless of path or the sign of δ), one may write

$$I = \text{P} \int_{-\infty}^{\infty} \frac{f(x) dx}{x - x_0} = \text{P} \oint_C \frac{f(z) dz}{z - x_0},$$

reminding us to *include* x_0 inside integration contour, but take the residue with a factor of $1/2$.

Problem 16

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{\sin x}{x} dx$$

Answer: This is a straightforward principal value integral:

$$I = \text{Im} \left(\pi i e^{i \cdot 0} + \lim_{\delta \rightarrow 0} \oint_C \frac{e^{iz} dz}{z + i\delta} \right) = \pi$$

Problem 17

Evaluate each of:

$$I = \text{P} \int_{-\infty}^{\infty} \frac{\cos kx}{(x - x_0)(x^2 + 2)} dx$$

$$J = \text{P} \int_{-\infty}^{\infty} \frac{\sin kx}{(x - x_0)(x^2 + 2)} dx$$

Answer:

$$\begin{aligned} K &= \text{P} \oint_C \frac{e^{ikz}}{(z - x_0)(z^2 + 2)} dz \\ &= i\pi \frac{e^{ikx_0}}{x_0^2 + 2} + 2\pi i \left(\frac{e^{-k\sqrt{2}}}{(\sqrt{2}i - x_0)(2\sqrt{2}i)} \right) \\ &= \left(\frac{-\pi}{x_0^2 + 2} \left(\sin kx_0 + \frac{x_0 e^{-\sqrt{2}k}}{\sqrt{2}} \right) \right) \\ &\quad + i \left(\frac{\pi}{x_0^2 + 2} \left(\cos kx_0 - e^{-\sqrt{2}k} \right) \right) \\ &= I + iJ \end{aligned}$$

Dispersion Relations

One special case for $f(z)$ occurs when the upper-half plane contains no singularities, making the contour integral the I^+ -equation resolve to zero. By decomposing f into real and imaginary components u and v , respectively, we derive the *Kramers-Kroing* relations:

$$u(x_0) = \frac{1}{\pi} \text{P} \int_{-\infty}^{\infty} \frac{v(x)}{x - x_0} dx$$

$$v(x_0) = \frac{-1}{\pi} \text{P} \int_{-\infty}^{\infty} \frac{u(x)}{x - x_0} dx$$

5.8 Branch Cuts

Non-Integer Powers and Logarithms

Complex numbers involving exponents and logarithms follow plainly from Euler's formula:

$$z^a = r^a e^{ai\theta}$$

$$\ln z = \ln r + i\theta$$

Of course, the periodicity of θ leads to certain functions behaving non-smoothly as the line defined by $\theta = 0$ is crossed. For instance, the value of the logarithm

$$\ln z(r, 0) = \ln r$$

$$\ln z(r, 2\pi) = \ln r + 2\pi i$$

at two equal points in the complex plane can disagree with itself by (at least) $2\pi i$.

For another example, the square root $z^{1/2} = r^{1/2} e^{i\theta/2}$ is also multi-valued, as

$$z^{1/2}(r, 0) = r^{1/2}$$

$$z^{1/2}(r, 2\pi) = -r^{1/2}.$$

All of this is troublesome for contour integrals, so the line on which $f(z)$ is ill-behaved, called a *branch cut*, must be stepped around. To proceed generally we denote an initial phase θ_0 that defines a branch cut $z = r e^{i\theta_0}$, and then define

$$\theta_0 + 2\pi N \leq \theta < \theta_0 + 2\pi(N + 1)$$

for an integer N , called the *branch*, that indexes multiples of 2π .

Problem 18

Use

$$\bar{z} \frac{\partial}{\partial \bar{z}} = \frac{1}{2} \left(r \frac{\partial}{\partial r} + i \frac{\partial}{\partial \theta} \right)$$

to show that the complex power and logarithm functions are analytic everywhere except for the branch θ_0 .

$$\bar{z} \frac{\partial}{\partial \bar{z}} (z^a) = \frac{a}{2} e^{ia\theta} (r^a - r^a) = 0$$

$$\bar{z} \frac{\partial}{\partial \bar{z}} (\ln z) = \left(\frac{r}{r} + i^2 \right) = 0$$

Products with Powers

We next consider the integral

$$I = \int_0^\infty f(x) x^a dx,$$

where $f(x)$ is well-behaved and non-singular on the real line, and the presence of x^a demands a branch but on $\theta_0 = 0$.

To proceed we move the integral to the complex plane and go along three contours: (i) C_+ , corresponding to $z = x + i\delta$ just above the real line, (ii) C_R , a nearly-full trip around the complex plane with $R \rightarrow \infty$, and (iii) C_- , coming from infinity back zero just below the real line with $z = x - i\delta$.

Only the $C_\pm$ contours contribute to the integral, so in the limit $\delta \rightarrow 0$ we have:

$$I = \int_{C_+} f(z) z^a dz$$

$$-e^{2i\pi a} I = \int_{C_-} f(z) z^a dz$$

Solving for I , the final answer pops out:

$$\int_0^\infty f(x) x^a dx = \frac{1}{1 - e^{2i\pi a}} \oint_C f(z) z^a dz$$

Note that the integration contour surrounds the whole complex plane minus the branch cut. Don't forget to include all poles in residue calculations.

Problem 19

Evaluate:

$$I = \int_0^\infty \frac{x^a dx}{(1+x)^2}$$

Answer: Contour contains one order-two pole.

$$\oint_C \frac{z^a dz}{(1+z)^2} = 2\pi i \sum_p \text{Res} \left[f(z_0^{(p)}) \right]$$

$$g(z) = \frac{z^a (1+z)^{-2}}{(1+z)^2}$$

$$\begin{aligned}
 g^{(1)}(z) &= a z^{a-1} \\
 g^{(-1)}(z) &= a 1^{a-1} e^{i\pi a} e^{-i\pi} = -a e^{i\pi a} \\
 I &= \frac{-2\pi i a e^{i\pi a}}{1 - e^{2i\pi a}} = \frac{\pi a}{\sin(\pi a)}
 \end{aligned}$$

Positive Real Domain

The general problem

$$I = \int_0^\infty f(x) dx$$

exploits a branch cut spectacularly. We begin by considering a different integral

$$\tilde{I} = \int_0^\infty f(x) \ln(x) dx$$

on the same contours C_+ and C_- used above, on the branch $0 \leq \theta < 2\pi$. This gives us

$$\begin{aligned}
 \int_{C_+} f(z) \ln(z) dz &= \tilde{I} \\
 \int_{C_-} f(z) z^a dz &= -\tilde{I} - 2\pi i \int_0^\infty f(x) dx,
 \end{aligned}$$

which sum together to perfectly cancel $\tilde{I}$, provided the usual assumptions that allow the integral with $R \rightarrow \infty$ to vanish. Solving for the original I , we find:

$$\int_0^\infty f(x) dx = \frac{i}{2\pi} \int_C f(z) \ln(z) dz$$

Problem 20

Evaluate:

$$I = \int_0^\infty \frac{dx}{(x+2)(x+1)^2}$$

Answer: Contour contains two poles.

$$\begin{aligned}
 I &= \frac{i}{2\pi} \cdot 2\pi i \left(\left. \frac{d}{dz} \frac{\ln z}{z+2} \right|_{z=-1} + \left. \frac{\ln z}{(z+1)^2} \right|_{z=-2} \right) \\
 &= 1 - \ln 2
 \end{aligned}$$

Chapter 21

Variational Calculus

1 Introduction

Any well-rounded student of natural philosophy or STEM field has at least one thorough encounter with the tenets of classical mechanics. A relatable picture of reality is constructed from Newton's laws of motion, along with conservation laws handling energy, momentum, and angular momentum. Toss in a few revelations from electromagnetism and thermodynamics and pre-1905 physics is complete, right?

Not quite. What if you learned that there is another tenant in the building living among the classical laws? That there is another principle of mechanics at play that not only contains, but runs a bit deeper than Newton's seventeenth-century picture? This is in fact the case, and it is called the *principle of least action*.

The tool kit used to grapple with 'least action' is called the *calculus of variations*. Variational calculus, as it's also called, is all about solving for *critical curves* rather than critical points as done in ordinary calculus.

As it pertains to classical mechanics, the principle of least action tells something very profound: *The true path on which a body moves is the one that minimizes the difference between kinetic and potential energy along that path.* Using symbols T for kinetic energy and U for potential, the above means

$$S = \int (T(v(t)) - U(x(t))) dt$$

is always minimized if $x(t)$ and $v(t)$ represent the correct path of position and velocity. The quantity S is called the *action*. The action evaluates to a larger number if the wrong $x(t)$ or $v(t)$ are fed into the integral.

This is undoubtedly a strange way to think about mechanics, namely because there is no need for forces, momentum, or any vectors at all. As a matter of strictness, the idea of 'least' action is sometimes a misnomer, as sometimes the case of 'most' action is more applicable. In either case, we are safe saying 'principle of stationary action'.

2 Euler-Lagrange Equation

To begin we will work strictly on the xy plane without mentioning physics until the first solid result is gained. Consider two fixed points in the Cartesian plane

$$(x_0, y_0), (x_1, y_1)$$

that represent the initial and final position of any well-behaved curve $y(x)$. The curve $y(x)$ shall be considered the 'true' path that connects the initial and final fixed points:

$$y(x) = \text{true path from } x_0 \text{ to } x_1$$

Varied Path

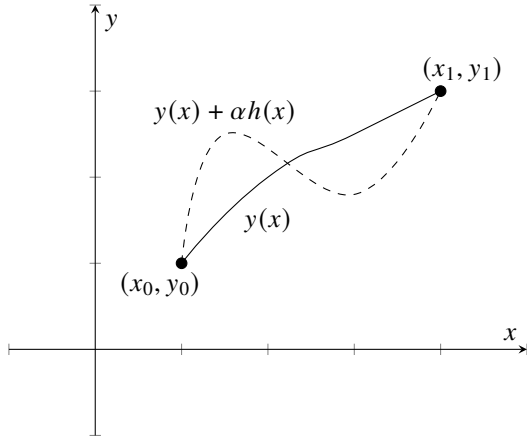
To accompany $y(x)$ in the plane, introduce a second curve that starts and finishes at the same endpoints, but is allowed to meander in the plane due to an additional term $\alpha h(x)$ such that

$$f(x) = y(x) + \alpha h(x) .$$

The variable α is a dimensionless *variational parameter*. The curve $h(x)$ is an 'incorrectness' that gives structure to the varied path, and this is decidedly zero at each endpoint:

$$h(x_0) = h(x_1) = 0$$

Summarizing this construction, we sketch the following figure in the Cartesian plane:



Writing the Action

Now, conceive of a new function Λ that depends on x , $f(x)$, and also the derivative $f'(x) = df(x)/dx$. Such a function $\Lambda(x, f(x), f'(x))$ can be arbitrary in most respects, whether it represent a physical quantity or a conceptual one is a mere formality as of now.

To construct an action from this, integrate the function Λ along the path from x_0 to x_1 :

$$S = \int_{x_0}^{x_1} \Lambda(x, f(x), f'(x)) dx$$

Keeping the notation from getting out of control, we hide the explicit mention of x from the f -related terms:

$$S = \int_{x_0}^{x_1} \Lambda(x, f, f') dx$$

Derivative of the Action

Buried in f and f' is the dependence on the variational parameter α , where we explicitly have

$$\begin{aligned} f &= y + \alpha h \\ f' &= y_x + \alpha h' , \end{aligned}$$

which means the action S is also a function of the parameter α . In traditional calculus-101 fashion, we then ask what happens by taking derivative of S with respect to α ? Setting this up, the problem means to calculate

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} \frac{d}{d\alpha} (\Lambda(x, f, f')) dx .$$

The parameter α does not depend on x itself (that's what h is for), which is why the α -derivative penetrates the integral without fuss.

To grapple with what $dA/d\alpha$ means, reach for partial derivatives to write

$$\frac{dA}{d\alpha} = \frac{\partial \Lambda}{\partial x} \frac{dx}{d\alpha} + \frac{\partial \Lambda}{\partial f} \frac{df}{d\alpha} + \frac{\partial \Lambda}{\partial f'} \frac{df'}{d\alpha} ,$$

where the first term is identically zero, and the derivative terms $df/d\alpha$ and $df'/d\alpha$ are nothing but h and h' , respectively. So far then, we have:

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} \left(\frac{\partial \Lambda}{\partial f} h + \frac{\partial \Lambda}{\partial f'} h' \right) dx .$$

Integrating by Parts

To proceed, we'll focus on the primed term in the above, and choose the following substitutions for integration by parts

$$\begin{aligned} u &= \partial \Lambda / \partial f' \\ dv &= h' dx , \end{aligned}$$

with

$$\begin{aligned} du &= \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f'} \right) dx \\ v &= h . \end{aligned}$$

Then, by the standard form

$$\int u \, dv = uv \Big| - \int v \, du ,$$

the derivative of the action looks like

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} \left(\frac{\partial \Lambda}{\partial f} h - h \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f'} \right) \right) dx + \frac{\partial \Lambda}{\partial f'} h \Big|_{x_0}^{x_1} ,$$

where the boundary term wholly vanishes because $h(x)$ is zero at each endpoint.

Minimizing the Action

Summarizing our progress, the derivative of the action has taken the form

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} h(x) \left(\frac{\partial \Lambda}{\partial f} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f'} \right) \right) dx .$$

Now comes the crucial observation regarding the variational parameter α . The derivative $dS/d\alpha$ goes to zero as α itself goes to zero:

$$\begin{aligned} dS/d\alpha &\rightarrow 0 \\ \alpha &\rightarrow 0 \end{aligned}$$

By making such a change, the varied path flattens down to the true path, which means $f(x)$ flattens down to $y(x)$:

$$f(x) \rightarrow y(x)$$

Euler-Lagrange Equation

In the zero-variation limit, the above becomes

$$0 = \int_{x_0}^{x_1} h(x) \underbrace{\left(\frac{\partial \Lambda}{\partial y} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial y_x} \right) \right)}_{=0} dx ,$$

and the integral must clearly evaluate to zero. Since $h(x)$ is generally nonzero except for the endpoints, the parenthesized quantity must therefore be zero along the whole path. Plucking out this item from the integral, we arrive at the famed *Euler-Lagrange* equation:

$$0 = \frac{\partial \Lambda}{\partial y} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial y_x} \right) \quad (21.1)$$

The Euler-Lagrange equation doesn't look like much at fist, perhaps a nice accident of the chain rule. It is interesting to fathom, though, that the above holds for whatever arbitrary function $\Lambda(x, y(x), y_x(x))$ is chosen.

Change of Domain

The whole derivation of Equation (21.1) can be repeated in the time domain, in which case t takes the place of x . Following this through, let us substitute

$$\begin{aligned} x &\rightarrow t \\ y &\rightarrow x(t) \\ y_x &\rightarrow v(t) , \end{aligned}$$

where $v(t)$ is the velocity $x'(t)$. While we're at it, relabel the arbitrary function Λ with the letter L :

$$\Lambda \rightarrow L(t, x(t), v(t))$$

With all this, the Euler-Lagrange equation takes a form in the time domain:

$$0 = \frac{\partial L}{\partial x} - \frac{d}{dt} \left(\frac{\partial L}{\partial v} \right) \quad (21.2)$$

2.1 The Lagrangian

By choosing the proper function for L , something curious happens with the Euler-Lagrange equation (21.2). After some fiddling, one readily stumbles on the combination

$$L = T(v) - U(x) , \quad (21.3)$$

called the *Lagrangian*. The functions $T(v)$, $U(x)$ are the respective kinetic and potential energies of the body being considered.

By inserting Equation (21.3) into Equation (21.2), we write

$$0 = \frac{\partial (T - U)}{\partial x} - \frac{d}{dt} \left(\frac{\partial (T - U)}{\partial v} \right) ,$$

simplifying to

$$0 = -\frac{\partial U}{\partial x} - \frac{d}{dt} \left(\frac{\partial T}{\partial v} \right) .$$

From here, a deeply rich formulation of classical physics called *Lagrangian mechanics* can be developed. This topic won't be formally indulged here, but its foundations are explored here nonetheless.

Recovering Newton's Law

For the Newtonian regime, we have that the kinetic energy $T(v)$ is given by

$$T(v) = \frac{1}{2}mv^2,$$

where m is the mass of the body or particle. Then, the Euler-Lagrange equation with the choice $L = T - U$ becomes

$$0 = -\frac{\partial U}{\partial x} - \frac{d}{dt} \left(\frac{\partial}{\partial v} \frac{1}{2}mv^2 \right),$$

readily simplifying to Newton's second law:

$$m \frac{d}{dt} v(t) = -\frac{\partial}{\partial x} U(x)$$

Astonishingly, the principle of least action seemingly *does* contain the classic laws of motion, and the true path followed by an object really is the one that minimizes the integral of $T - U$ along the path.

3 Formalism

While the above derivation of the Euler-Lagrange equation took place in two dimensions, the same reasoning applies when there are more variables at play. To this end, it helps to have an efficient symbol for certain derivative terms. If we have a function f , and we need the partial derivative with respect to x , we should be able to write

$$\begin{aligned} f_x &= \partial f / \partial x \\ f_{xx} &= \partial^2 f / \partial x^2 \end{aligned}$$

without ambiguity. To have the operator by itself with no mention of which function it's acting upon, we use the notation

$$\partial_x = \frac{\partial}{\partial x} = \partial / \partial x .$$

Of course, the same notation is useful for full derivatives, for instance

$$d_t f = \frac{d}{dt} f .$$

With such shortcuts, the Euler-Lagrange equation can be written in the most minimal way:

$$0 = L_x - d_t L_v$$

Formal Derivation

With the main ideas established, let us re-state the action calculation

$$S = \int_{x_0}^{x_1} \Lambda(x, f, f') dx$$

in slightly different terms. The function $f(x)$ is still the varied path, except variations are represented by

$$\begin{aligned} f &\rightarrow f + \Delta f \\ f_x &\rightarrow f_x + (\Delta f)' , \end{aligned}$$

where Δf replaces the $\alpha h(x)$ construction, and Δf is zero at each endpoint.

The action S is no ordinary function, but is formally called a 'functional' depending on $f(x)$, denoted $S[f]$. So far then, we write

$$S[f] = \int_{x_0}^{x_1} \Lambda(x, f, f_x) dx .$$

Then a variation in S is written

$$\Delta S = \int_{x_0}^{x_1} \Lambda(x, f + \Delta f, f_x + (\Delta f)') dx - S[f] .$$

By Taylor-expanding the inner quantity to first order, we further have

$$\begin{aligned} \Lambda(x, f + \Delta f, f_x + (\Delta f)') &= \\ \Lambda(x, f, f_x) &+ \frac{\partial \Lambda}{\partial f} \Delta f + \frac{\partial \Lambda}{\partial f_x} (\Delta f)' , \end{aligned}$$

and ΔS simplifies to

$$\Delta S = \int_{x_0}^{x_1} \left(\frac{\partial \Lambda}{\partial f} \Delta f + \frac{\partial \Lambda}{\partial f_x} (\Delta f)' \right) dx .$$

From here, the derivation looks much like the 'informal' one we started with, and the steps to finish are the same. Letting ΔS and Δf go to zero simultaneously after integrating by parts, one finds the now-familiar Euler-Lagrange equation

$$0 = \frac{\partial \Lambda}{\partial f} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f_x} \right) .$$

More Dimensions

The formal derivation of the Euler-Lagrange equation that generalizes to N simultaneous functions

$$f \rightarrow \{f^{(1)}, f^{(2)}, \dots, f^{(N)}\} ,$$

each with a different variation

$$f^{(j)} \rightarrow f^{(j)} + \Delta f^{(j)} ,$$

with vanishing variation at the endpoints

$$\Delta f^{(j)}(x_0) = \Delta f^{(j)}(x_1) = 0$$

is straightforwardly written.

The function Λ becomes

$$\Lambda(x, f^{(1)}, f^{(2)}, \dots, f^{(N)}, f_x^{(1)}, f_x^{(2)}, \dots, f_x^{(N)}) ,$$

and gives rise to m Euler-Lagrange equations that all look the same up to index number:

$$0 = \frac{\partial \Lambda}{\partial f^{(j)}} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f_x^{(j)}} \right) \quad (21.4)$$

3.1 Special Form

Starting with $\Lambda(x, f, f')$, let us calculate the total derivative in x , which is

$$\frac{d\Lambda}{dx} = \frac{\partial\Lambda}{\partial x} + \frac{\partial\Lambda}{\partial f} \frac{df}{dx} + \frac{\partial\Lambda}{\partial f_x} \frac{df_x}{dx},$$

and then replace $\partial\Lambda/\partial f$ using the Euler-Lagrange equation

$$\frac{d\Lambda}{dx} = \frac{\partial\Lambda}{\partial x} + \frac{d}{dx} \left(\frac{\partial\Lambda}{\partial f_x} \right) f_x + \frac{\partial\Lambda}{\partial f_x} f_{xx}.$$

Note that there is an equivalence between

$$\frac{df}{dx} \leftrightarrow \frac{\partial f}{\partial x}$$

because f is a function of only x . (This is certainly not true for Λ or any multivariate function.)

Off to the side, calculate the total derivative of $f_x \partial\Lambda/\partial f_x$, which looks like

$$\frac{d}{dx} \left(f_x \frac{\partial\Lambda}{\partial f_x} \right) = f_{xx} \frac{\partial\Lambda}{\partial f_x} + f_x \frac{d}{dx} \left(\frac{\partial\Lambda}{\partial f_x} \right),$$

containing the same f_{xx} -term as the previous result. Eliminating this between the two, we have, after simplifying:

$$\frac{d\Lambda}{dx} = \frac{\partial\Lambda}{\partial x} + \frac{d}{dx} \left(f_x \frac{\partial\Lambda}{\partial f_x} \right)$$

Putting everything on one side gives special form of the Euler-Lagrange equation:

$$0 = \frac{\partial\Lambda}{\partial x} - \frac{d}{dx} \left(\Lambda - f_x \frac{\partial\Lambda}{\partial f_x} \right) \quad (21.5)$$

The above is especially informative if the function Λ has no explicit dependence on x , for if this is the case then we can only have

$$\Lambda - f_x \frac{\partial\Lambda}{\partial f_x} = \text{constant},$$

where f_x is not zero. This result is sometimes called the Beltrami identity.

3.2 Constant of Motion

It's impossible to resist jumping back into time domain and make another connection to Newtonian mechanics. Momentarily make the same swap of variables

$$\begin{aligned} x &\rightarrow t \\ f &\rightarrow x(t) \\ L &\rightarrow T(v) - U(x), \end{aligned}$$

and so on.

Since the Lagrangian has no explicit time dependence, we immediately use the Beltrami identity to write

$$L - v \frac{\partial L}{\partial v} = \text{constant}.$$

Knowing that the kinetic energy $T(v)$ takes the form $mv^2/2$, we find (as before) that

$$\frac{\partial L}{\partial v} = \frac{\partial T}{\partial v} = \frac{\partial}{\partial v} \left(\frac{1}{2} mv^2 \right) = mv,$$

meaning

$$v \frac{\partial L}{\partial v} = mv^2 = 2T(v).$$

Putting it all together, we find

$$L - 2T = T - U - 2T = \text{constant},$$

or in other words,

$$T + U = -\text{constant},$$

therefore conservation of energy also emerges from the principle of least action.

3.3 Conservation of Energy

When it comes to having more than one dimension, one wonders what the generalization of Equation (21.5) may be, and whether there is one new constant per added dimension. This is in fact *not* the case, but the real answer is more beautiful anyway. Anticipating the outcome of this analysis, consider a function E (for 'energy') that has all the same dependencies as Λ :

$$E = E(x, f^{(1)}, f^{(2)}, \dots, f^{(N)}, f_x^{(1)}, f_x^{(2)}, \dots, f_x^{(N)})$$

Then, taking inspiration from the constant of motion found in the one-dimensional case, consider the relationship

$$E = \sum_{j=1}^N f_x^{(j)} \frac{\partial\Lambda}{\partial f_x^{(j)}} - \Lambda.$$

If E is to be constant, then the total derivative of E had better resolve to zero. Pursuing this, we write:

$$\frac{dE}{dx} = \sum_{j=1}^N f_{xx}^{(j)} \frac{\partial\Lambda}{\partial f_x^{(j)}} + \sum_{j=1}^N f_x^{(j)} \frac{d}{dx} \left(\frac{\partial\Lambda}{\partial f_x^{(j)}} \right) - \frac{d\Lambda}{dx}$$

From the chain rule we also have

$$\frac{d\Lambda}{dx} = \frac{\partial\Lambda}{\partial x} + \sum_{j=1}^N \frac{\partial\Lambda}{\partial f^{(j)}} f_x^{(j)} + \sum_{j=1}^N \frac{\partial\Lambda}{\partial f_x^{(j)}} f_{xx}^{(j)},$$

and note that the f_{xx} -sums occur in both equations, and so does $d\Lambda/dx$. Eliminating the common terms in each, we get the result

$$\frac{dE}{dx} = -\frac{\partial\Lambda}{\partial x} + \sum_{j=1}^N f_x^{(j)} \left(\frac{d}{dx} \left(\frac{\partial\Lambda}{\partial f_x^{(j)}} \right) - \frac{\partial\Lambda}{\partial f^{(j)}} \right).$$

In the above, the parenthesized portion is none other than the Euler-Lagrange equation, and is identically zero. Evidently, the whole concern of energy conservation reduces to the statement

$$\frac{dE}{dx} = -\frac{\partial\Lambda}{\partial x}.$$

Note that we've worked with the unspecified function Λ depending fundamentally on x . To turn the above into a statement about physics, make the following replacement:

$$\begin{aligned} x &\rightarrow t \\ \Lambda &\rightarrow L = T(v) - U(x) \end{aligned}$$

3.4 Generalized Coordinates

As it pertains to physical systems, The Euler-Lagrange equation (21.4) occurs in the form

$$0 = \frac{\partial L}{\partial q^{(j)}} - \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}^{(j)}} \right), \quad (21.6)$$

where the terms $q^{(j)}$ are *generalized coordinates*. The physical units of a given $q^{(j)}$ need not be spatial. The readiest example of this would be polar coordinates, in where the components of q are represented by

$$\mathbf{q} = \langle r, \theta \rangle.$$

Proof

To establish this firmly, let us pull no punches and do a proper proof. Supposing a system evolving by variable t depends on coordinates $x^{(j)}(t)$ and their derivatives $v^{(j)}(t)$, the Euler-Lagrange equation takes the form

$$0 = \frac{\partial\Lambda}{\partial x^{(j)}} - \frac{d}{dt} \left(\frac{\partial\Lambda}{\partial v^{(j)}} \right).$$

Next, suppose the list of coordinates $\{x(t)\}$ can be defined in terms of another list $\{q(t)\}$ such that

$$x^{(j)} = x^{(j)}(t, q^{(1)}, q^{(2)}, \dots, q^{(N)})$$

if there are N members in $\{q(t)\}$. To first order, the two sets of coordinates further relate by

$$v^{(j)} = \frac{\partial x^{(j)}}{\partial t} + \sum_{k=1}^N \frac{\partial x^{(j)}}{\partial q^{(k)}} \dot{q}^{(k)}.$$

By the chain rule, we further find

$$\frac{\partial x^{(j)}}{\partial q^{(k)}} = \frac{\partial x^{(j)}}{\partial q^{(k)}} \frac{dt}{dt} = \frac{\partial v^{(j)}}{\partial \dot{q}^{(k)}}.$$

With the second set of coordinates, the proposed Euler-Lagrange equation reads

$$0 \stackrel{?}{=} \frac{\partial\Lambda}{\partial q^{(j)}} - \frac{d}{dt} \left(\frac{\partial\Lambda}{\partial \dot{q}^{(j)}} \right),$$

which motivates picking on the inner term:

$$\frac{\partial\Lambda}{\partial \dot{q}^{(j)}} = \sum_{k=1}^N \frac{\partial\Lambda}{\partial v^{(k)}} \frac{\partial v^{(k)}}{\partial \dot{q}^{(j)}} = \sum_{k=1}^N \frac{\partial\Lambda}{\partial v^{(k)}} \frac{\partial x^{(k)}}{\partial q^{(j)}}$$

Take the time derivative of both sides to get

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial\Lambda}{\partial \dot{q}^{(j)}} \right) &= \sum_{k=1}^N \frac{d}{dt} \left(\frac{\partial\Lambda}{\partial v^{(k)}} \right) \frac{\partial x^{(k)}}{\partial q^{(j)}} \\ &\quad + \sum_{k=1}^N \frac{\partial\Lambda}{\partial v^{(k)}} \frac{d}{dt} \left(\frac{\partial x^{(k)}}{\partial q^{(j)}} \right), \end{aligned}$$

and simplify carefully to finish the proof:

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial\Lambda}{\partial \dot{q}^{(j)}} \right) &= \\ &\sum_{k=1}^N \left(\frac{\partial\Lambda}{\partial x^{(k)}} \frac{\partial x^{(k)}}{\partial q^{(j)}} + \frac{\partial\Lambda}{\partial v^{(k)}} \frac{\partial v^{(k)}}{\partial q^{(j)}} \right) \\ &= \frac{\partial\Lambda}{\partial q^{(j)}} \end{aligned}$$

4 Unconstrained Systems

4.1 Motion on a Curve

In uniform gravity, consider a frictionless particle of mass m that sits on the curve

$$y(x) = \frac{k}{\alpha} (x - a)^\alpha$$

without departure. If the velocity of the particle is $\mathbf{v} = \langle \dot{x}, \dot{y} \rangle$, find the equations of motion of this system.

With the information provided, there is enough to write the kinetic and potential of this system all in terms of x -variables

$$T = \frac{1}{2} m \dot{x}^2 \left(1 + k^2 (x - a)^{2\alpha-2} \right)$$

$$U = \frac{mgk}{\alpha} (x - a)^\alpha,$$

and the Lagrangian, of course, is $L = T - U$.

Applying the Euler-Lagrange equation, we must pursue

$$\frac{\partial L}{\partial x} - \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) = 0.$$

Doing so and simplifying, arrive at a juicy differential equation:

$$0 = \ddot{x} \left(1 + k^2 (x - a)^{2\alpha-3} \right) + \dot{x} \frac{k^2}{2} (2\alpha - 2) (x - a)^{2\alpha-3} + gk (x - a)^{\alpha-1}$$

Particle in a Well

For the cases with $k > 0$ and $\alpha > 1$ is even, the particle is stuck in a ‘well’ centered at $x = a$. The simplest of these has $\alpha = 2$, where the above reduces to the differential equation

$$0 = \ddot{w} \left(1 + w^2 \right) + \dot{w}^2 w + gkw,$$

where $w = kz$ and $z = x - a$.

Despite the $\alpha = 2$ simplification, the above is still difficult to treat in the general case. We can do a quick reality check for small motions, which has $1 \gg w^2$ and $gk \gg \dot{w}^2$. For this we recover the setup for the simple harmonic oscillator, as expected:

$$\ddot{w} = -gkw$$

Particle on an Incline

An interesting modification to this setup has $\alpha = 1$, with $y(x)$ representing a straight line. Restarting the analysis from here and applying the Euler-Lagrange equation leads to

$$\ddot{x} = \frac{-gk}{1 + k^2},$$

and similarly,

$$\ddot{y} = k\ddot{x}.$$

The total acceleration is the sum of square of each:

$$|a| = \sqrt{\ddot{x}^2 + \ddot{y}^2} = \frac{gk}{\sqrt{1 + k^2}}$$

Letting θ denote the angle of incline, we further have

$$k = \tan(\theta)$$

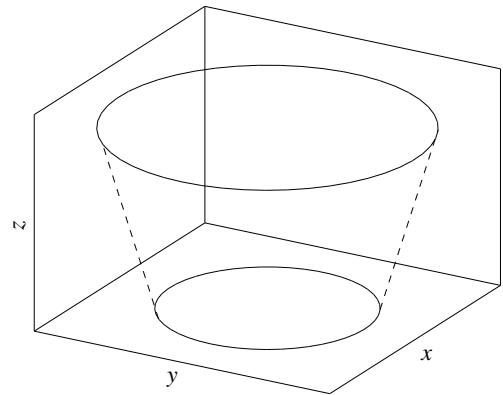
and

$$|a| = g \sin(\theta).$$

4.2 Minimal Surface

Soap Film Problem

A stretchable membrane, such as a thin soap film, is trapped two hoops to form an axially-symmetric surface. One hoop is a circle laying on the xy -plane at $z = z_0$. the other hoop is parallel to the first and suspended at $z = z_1$. Ignoring gravity, determine the shape of the membrane that minimizes surface area. In other words, determine the proper shape of the dotted line profile indicated in the figure below.



The minimal surface problem is characterized by the surface area (conveniently named S) of the whole film. Using

elementary methods, or by exploiting the axial symmetry of the system, the surface area reads

$$S = \int_{z_0}^{z_1} 2\pi r(z) \sqrt{1 + r_z^2} dz,$$

where r_z is the derivative of r with respect to z .

Without needing to finish the integral, we pick out the working quantity to be

$$\Lambda = r \sqrt{1 + r_z^2},$$

which is a function of r and r_z , but not z itself. This warrants use of the special form of the Euler-Lagrange equation (21.5), which means, for this problem,

$$\Lambda - r_z \frac{\partial \Lambda}{\partial r_z} = C_0,$$

where C_0 is a constant.

Plugging Λ into the above and turning the crank leads to a separable differential equation

$$dz = \frac{dr}{\sqrt{r^2/C_0^2 - 1}}.$$

By the substitution

$$r = C_0 \cosh(\beta)$$

with

$$dr = C_0 \sinh(\beta) d\beta,$$

the differential equation simplifies to

$$dz = C_0 d\beta,$$

revealing the simple relationship

$$z = C_0 \beta + C_1,$$

where C_1 is an integration constant. After eliminating β , suddenly we have an equation for $r(z)$:

$$r(z) = C_0 \cosh\left(\frac{z - C_1}{C_0}\right)$$

The constants C_0, C_1 are specified by the hoop radii, namely $r(z_0) = R_0$ and $r(z_1) = R_1$.

Fleshing out an easy example, the symmetric case with $R_1 = R_2$ and

$$-z_0 = z_1 = L$$

leads to

$$C_1 = \frac{z_0 + z_1}{2} = \frac{-L + L}{2} = 0.$$

The $r(z)$ equation then says

$$R = C_0 \cosh\left(\frac{L}{C_0}\right),$$

containing one unknown.

Straight Line

A much easier minimal surface problem is to prove that the shortest path connecting two points is a straight line. Setting up the classic ‘arc length’ integral looks like

$$S = \int_{x_0}^{x_1} \sqrt{1 + y_x^2} dx,$$

from which we pick out

$$\Lambda = \sqrt{1 + y_x^2}.$$

Unsurprisingly there is no explicit x -dependence in Λ , warranting the identity

$$\Lambda - y_x \frac{\partial \Lambda}{\partial y_x} = C,$$

with C constant. Simplifying, we find

$$\frac{1}{\sqrt{1 + y_x^2}} = C,$$

which can only mean y_x is a constant, or $y(x)$ is a straight line.

4.3 Motion on a Cycloid

Brachistochrone

The Ancient Greeks were interested in a curious problem that has widespread practical application. In uniform gravity, suppose a body at an initial height needs to slide down some kind of plank, ramp, or other curve so as to reach a lower height in the *shortest* time possible, ignoring friction. The curve that solves this problem is called the *brachistochrone*.

Starting from first principles, the time T taken to slide down such a curve is given by

$$T = \int_{y(t_0)}^{y(t_1)} dt,$$

where T is the quantity to minimize. The differential arc length ds relates to dt by

$$ds = v(t) dt,$$

where $v(t)$ is the velocity of the body at time t .

From geometry we have that

$$ds^2 = dx^2 + dy^2 ,$$

and meanwhile from energy conservation we know

$$v(x) = \sqrt{2g(y_0 - y(x))} .$$

Updating the T -integral with this information gives

$$T = \int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \frac{dx}{\sqrt{2g(y_0 - y(x))}} .$$

Now, we're trying to minimize T as a way of solving for y , which seems completely backwards until utilizing the calculus of variations. The above can be regarded as a functional

$$T = \int_{y_0}^{y_1} \Lambda(x, y, y_x) dx ,$$

with no explicit x -dependence. Reasoning from the special form of the Euler-Lagrange equation (21.5), we have

$$\Lambda - y_x \frac{\partial \Lambda}{\partial y_x} = C ,$$

where C is constant. Calculating this out, we find

$$\frac{1}{\sqrt{2g(y_0 - y(x))}} \frac{1}{\sqrt{1 + (dy/dx)^2}} = C ,$$

which can be turned into an integral for $x(y)$:

$$\int dx = \int \sqrt{\frac{2gC^2(y_0 - y)}{1 - 2gC^2(y_0 - y)}} dy$$

Now introduce a peculiar trigonometric substitution in the variable θ such that

$$2gC^2(y_0 - y) = \frac{1 - \cos(\theta)}{2}$$

and

$$dy = -\frac{\sin(\theta)}{4gC^2} d\theta .$$

Running this substitution through the above integral, we have

$$\int dx = \frac{-1}{4gC^2} \int (1 - \cos(\theta)) d\theta ,$$

and evidently the combination $1/4gC^2$ has units of space, and this will be renamed to R , as in 'radius'.

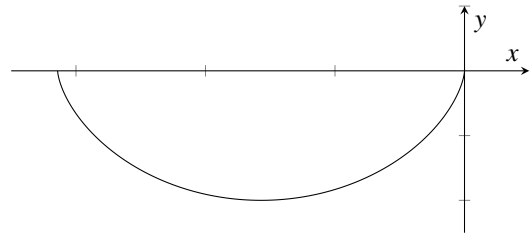
Finishing up the calculation for x , we finally have

$$x = x_0 - R(\theta - \sin(\theta)) , \quad (21.7)$$

and solving similarly for y ,

$$y = y_0 - R(1 - \cos(\theta)) . \quad (21.8)$$

This pair of parametric equations is a surprising result, namely because a unified equation $y(x)$ is not straightforwardly attained. The shape described is known as a *cycloid*.



In the domain $0 \leq \theta \leq 2\pi$, we see the cycloid generates right-to-left. To rectify this, one can change $\theta \rightarrow -\theta$ to reverse the evolution of the x -equation while leaving the y -equation unchanged. To put the entire picture above the x -axis, one may flip the sign on R as if reversing the sign on g .

Tautochrone

A question similar to the brachistochrone is to seek a curve called the *tautochrone*, the shape with the property that a body can start sliding from rest anywhere on the curve and reach the lowest point in a fixed time.

It may be no surprise that the cycloid also solves this problem, so we start with this assumption and check that the claim is satisfied. To this end, let us take as a starting point

$$T = \int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \frac{dx}{\sqrt{2g(y_0 - y(x))}} ,$$

where x_1 is located at the bottom of the cycloid (upturned as sketched above).

Knowing the solution to $y(\theta)$, we can write

$$y_0 - y = R(\cos(\theta_0) - \cos(\theta)) ,$$

where θ_0 characterizes the initial position of a body sliding from rest. Then, the equation for velocity as a function of θ reads

$$v(\theta) = \sqrt{2gR(\cos(\theta_0) - \cos(\theta))} .$$

The remaining quantities in the integral must also be expressed in terms of θ . Doing so carefully, the T -integral becomes

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sqrt{1 - \cos(\theta)} d\theta}{\sqrt{\cos(\theta_0) - \cos(\theta)}}.$$

Note that the change of variables $\theta \rightarrow -\theta$ has been invoked, simply reversing the sign on the integral to make sure the result is positive.

The task is to solve the integral and ultimately show that θ_0 does not affect the result. For this we use a pair of trigonometric identities

$$\begin{aligned} \sin\left(\frac{\theta}{2}\right) &= \sqrt{\frac{1 - \cos(\theta)}{2}} \\ \cos(\theta) &= 2 \cos^2\left(\frac{\theta}{2}\right) - 1, \end{aligned}$$

so now

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sin(\theta/2) d\theta}{\sqrt{\cos^2(\theta_0/2) - \cos^2(\theta/2)}}.$$

Proceed with the u -substitution

$$\begin{aligned} u &= \cos(\theta/2) / \cos(\theta_0/2) \\ du &= -d\theta \sin(\theta/2) / 2 \cos(\theta_0/2), \end{aligned}$$

and all θ_0 -dependence vanishes from the integral, leaving

$$T = \sqrt{\frac{R}{g}} \int_1^0 \frac{-2 du}{\sqrt{1 - u^2}}.$$

To finish the integral one could proceed by elementary methods, however notice in our final u -substitution that the initial condition θ_0 has been divided out, and thus never mattered. Choosing a θ_0 that trivializes the integral, namely $\theta_0 = 0$, makes the hard part vanish. One way or the other, the answer boils down to

$$T = \pi \sqrt{\frac{R}{g}},$$

affirming the cycloid as the solution to the tautochrone problem.

5 Constrained Systems

5.1 Constrained Lagrangian

Let us return to Equation (21.4) representing a system of several variables, namely:

$$\frac{\partial \Lambda}{\partial f^{(j)}} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f_x^{(j)}} \right) = 0$$

With zero on the right side of the equation, one speaks of this as an ‘unconstrained’ case.

Lagrange multipliers are an elegant means for enforcing constraints of motion by modifying the right side:

$$\frac{\partial \Lambda}{\partial f^{(j)}} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f_x^{(j)}} \right) = \sum_{k=1}^n \lambda_k \frac{\partial g_k}{\partial f^{(j)}} \quad (21.9)$$

Replacing zero is any number n total constraint terms. Each involves a Lagrange multiplier λ and also a function g to specify the constraint itself. In particular, g_k must evaluate to zero when the constraint is satisfied.

Generalized Force

From a mechanical point of view, the right side of Equation (21.9) is called the *generalized force*. This is justified by taking a modified Lagrangian

$$L = T(v) - U(x) - \lambda g(x) ,$$

where $g(x)$ is a potential energy term. Then, the gradient factor

$$Q = -\lambda \frac{\partial}{\partial x} g(x)$$

when $g(x) = 0$ corresponds to the body obeying the constraint.

In terms of generalized coordinates, the mechanical analog to Equation (21.9) reads

$$\frac{\partial L}{\partial q^{(j)}} - \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}^{(j)}} \right) = Q_{(j)} = \sum_{k=1}^n \lambda_k \frac{\partial g_k}{\partial f^{(j)}} .$$

5.2 Sagging Cable

In gravity, a homogeneous cable of fixed length suspended between two points will sag downward to minimize the gravitational potential energy throughout the cable. The exact shape of the cable is straightforwardly attained using the calculus of variations with a Lagrange multiplier.

Setup

If the linear mass density of the cable is a constant ρ , then a functional representing the gravitational potential energy of the cable is written

$$U[y] = \rho g \int_{x_0}^{x_1} y(x) \sqrt{1 + \left(\frac{dy}{dx} \right)^2} dx .$$

To instill the notion that the length of the cable is constant, a second functional is constructed:

$$V[y] = \int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx} \right)^2} dx$$

The ‘grand’ functional with which we must work is one that relates U and V by a Lagrange multiplier such that

$$F[y] = U[y] - \lambda V[y] ,$$

and the working quantity becomes

$$F[y] = \int_{x_0}^{x_1} (\rho g y(x) - \lambda) \sqrt{1 + \left(\frac{dy}{dx} \right)^2} dx .$$

Shape of the Cable

Introduce the substitution

$$u(x) = y(x) - \frac{\lambda}{\rho g}$$

such that the functional changes to

$$F[u] = \rho g \int_{x_0}^{x_1} u(x) \sqrt{1 + \left(\frac{du}{dx} \right)^2} dx ,$$

or in more symbolic notation,

$$F[u] = \int_{x_0}^{x_1} \Lambda(x, u, u') dx .$$

There is no explicit x -dependence in Λ , so the identity

$$\Lambda - u_x \frac{\partial \Lambda}{\partial u_x} = C_0$$

must hold, where C_0 is a constant. Substituting B into the above and simplifying results in a differential equation

$$\left(\frac{du}{dx} \right)^2 = \left(\frac{u}{C_0/\rho g} \right)^2 - 1 ,$$

which can be separated into x - and u -integrals:

$$\int dx = \int \frac{du}{\sqrt{u^2/(C_0/\rho g)^2 - 1}}$$

These are straightforwardly solved as

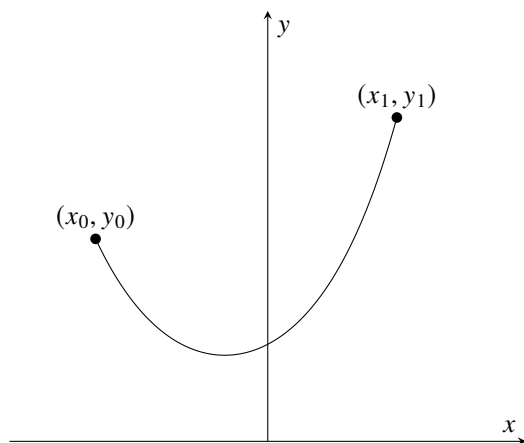
$$x = \frac{C_0}{\rho g} \cosh^{-1} \left(\frac{u}{C_0/\rho g} \right) + C_1 ,$$

where C_1 is an arbitrary constant.

Restoring the original y variable through the layers of substitutions, we end up with, for the final shape of the cable:

$$y(x) = \frac{\lambda}{\rho g} + \frac{C_0}{\rho g} \cosh \left(\frac{x - C_1}{C_0/\rho g} \right)$$

The answer you can walk away with is, ‘the sagging cable makes the shape of a hyperbolic cosine’. There are three constants in the solution that grant all the flexibility for adjusting endpoints and cable length, but the shape is always governed by cosh as depicted:



Length of the Cable

None of λ , C_0 , or C_1 alone specify the length of the cable. To determine this, we write

$$L = \int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx} \right)^2} dx$$

and use the dy/dx derived from the shape $y(x)$. Carrying this out results in

$$L = \frac{C_0}{\rho g} \sinh \left(\frac{x - C_1}{C_0/\rho g} \right) \Big|_{x_0}^{x_1} .$$

Lowest Point

The location x^* at which the cable sags lowest is the point satisfying $dy/dx = 0$, a criteria easily written:

$$0 = \sinh \left(\frac{x^* - C_1}{C_0/\rho g} \right)$$

This is satisfied by $x^* = C_1$, thus C_1 is equal to the x with the lowest y . The lowest point reached by the cable is

$$y^* = \frac{\lambda + C_0}{\rho g} .$$

Having a complete description of the sagging cable doesn't always mean that solving problems is an easy chore. The constant C_0 , for instance, is brutally tangled into the y -equation, and also makes an appearance in the formula for L . Once the problem is set up, the number crunching is best left to a computer.

Worked Example

To have an example, consider a cable hung between two endpoints of equal height and equal distance x_0 to the origin. If the cable length obeys $L = 4x_0$, determine the height difference H between the lowest point and the highest point.

The maximum height difference $H = y(\pm x_0) - y^*$ is straightforwardly written

$$H = \frac{C_0}{\rho g} \left(\cosh \left(\frac{x_0}{C_0/\rho g} \right) - 1 \right) ,$$

and the length factors in via

$$L = 2 \frac{C_0}{\rho g} \sinh \left(\frac{x_0}{C_0/\rho g} \right) ,$$

which is given as $4x_0$. The above is characterized by

$$2q = \sinh(q) ,$$

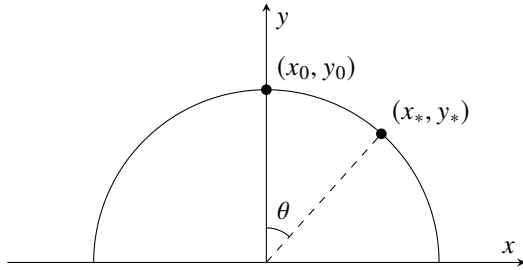
approximately solved by

$$\frac{x_0}{C_0/\rho g} = q \approx 2.177 .$$

With this, H comes out to

$$H \approx x_0 \left(\frac{\cosh(2.177) - 1}{2.177} \right) \approx 1.59 x_0 .$$

5.3 Sliding Down a Sphere



A particle of mass m sits at rest on a spherical surface of radius R . Receiving a very small nudge to the right, the particle slides down the sphere due to gravity. Measuring the particle's evolution using angle θ , determine the critical angle θ_* that corresponds to the particle leaving the sphere's surface and entering free-fall.

Newtonian Analysis

By standard Newtonian analysis, it's straightforward to write an equation in polar coordinates for the normal force N that keeps the particle on the surface

$$N = -\frac{mv^2}{R} + mg \cos(\theta) ,$$

where $-mv^2/R$ is the angular acceleration. The moment when the sliding particle enters free-fall, characterized by θ_* , v_* , occurs exactly when the normal force reaches zero:

$$0 = -\frac{mv_*^2}{R} + mg \cos(\theta_*)$$

Meanwhile, energy conservation gives us

$$E = \frac{1}{2}mv^2 + mgR \cos(\theta) ,$$

where E is constant and consists of a kinetic term and a potential term. This also holds until the condition θ_* , v_* is met:

$$mgR = \frac{1}{2}mv_*^2 + mgR \cos(\theta_*)$$

Eliminating v_* between the two equations and simplifying, we conclude easily that

$$\cos(\theta_*) = \frac{2}{3} ,$$

which means the particle leaves the sphere's surface at angle:

$$\theta_* = \arccos\left(\frac{2}{3}\right) \approx 0.841 \text{ rad} \approx 48.2^\circ$$

Constrained Motion Analysis

It is illustrative to solve the problem using constraints. To this end we will write the Lagrangian of the system with the assumption that r is allowed to vary in time:

$$L = \frac{1}{2}m \left((r_t)^2 + (r\theta_t)^2 \right) - mgr \cos(\theta)$$

Then, the so-called constraint simply makes sure that r is a constant:

$$g(r) = r - R$$

With two variables in play, namely r and θ , the constrained Euler-Lagrange equation (21.9) yields two items:

$$\begin{aligned} \frac{\partial L}{\partial r} - \frac{d}{dt} \left(\frac{\partial L}{\partial r_t} \right) &= \lambda \frac{\partial}{\partial r} (r - R) \\ \frac{\partial L}{\partial \theta} - \frac{d}{dt} \left(\frac{\partial L}{\partial \theta_t} \right) &= \lambda \frac{\partial}{\partial \theta} (r - R) \end{aligned}$$

Carrying these calculations through gives a pair of equations

$$\begin{aligned} mr^2 \omega_t &= -2mrr_t \omega + mgr \sin(\theta) \\ mr_{tt} &= mr\omega^2 - mg \cos(\theta) - \lambda , \end{aligned}$$

where λ is the normal force that keeps the particle outside the sphere, and the angular speed θ_t is selectively replaced by another Greek letter ω :

$$\frac{\partial \theta}{\partial t} = \theta_t = \omega$$

The moment that λ goes to zero is the moment the particle leaves contact with the sphere. Motion is characterized by $r_{tt} = r_t = 0$ because $r = R$. At the critical point, the angular speed θ_t takes on a special value ω_* . Updating the above and simplifying, we have

$$\begin{aligned} \omega_t &= \frac{g}{R} \sin(\theta) \\ \omega_*^2 &= \frac{g}{R} \cos(\theta_*) . \end{aligned}$$

The top equation can be manipulated to write

$$R \int \omega d\omega = \int g \sin(\theta) d\theta ,$$

which is indeed a statement of energy conservation:

$$E = mgR = \frac{1}{2}mR^2 \omega^2 + mgR \cos(\theta)$$

The critical case $\theta = \theta_*$, $\omega = \omega_*$, produces the same answer as above.

5.4 Sliding Down a Curve

A particle of mass m begins at position (x_0, y_0) on the curve

$$y = -kx^3$$

where $x_0 < 0$. The particle slides without friction due to gravity down the curve. Let us determine the condition for which the particle leaves the curve, if it does so at all.

Setting up the Lagrangian with a constraint, we write

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(y + kx^3).$$

Apply the Euler-Lagrange equation (21.9) for each variable x, y to find

$$\begin{aligned} m\ddot{x} &= 3k\lambda x^2 \\ mg + m\ddot{y} &= \lambda. \end{aligned}$$

The normal force N on the particle has components

$$\begin{aligned} N_x &= 3k\lambda x^2 \\ N_y &= \lambda, \end{aligned}$$

or

$$N = \lambda\sqrt{1 + 9k^2x^4}.$$

The last job is to eliminate λ . Using

$$\dot{y} = -3kx^2\dot{x},$$

take a time derivative to find

$$\ddot{y} = -6kx\dot{x}^2 - 3kx^2\ddot{x},$$

so

$$\lambda = mg - 6mkx\dot{x}^2 - 3mkx^2\ddot{x}$$

We learn more about $\dot{x}$ using energy conservation to write

$$mgy_0 = \frac{1}{2}m\dot{x}^2(1 + 9k^2x^4) + mgy,$$

meaning

$$\dot{x}^2 = -2g\left(\frac{y - y_0}{1 + 9k^2x^4}\right).$$

Replacing $\ddot{x}$ with $3\lambda x^2$, find

$$\lambda = \frac{mg}{1 + 9k^2x^4} \left(1 + 12k\frac{x(y - y_0)}{1 + 9k^2x^4}\right),$$

so the normal force on the particle is given by

$$\frac{N}{mg} = \frac{1 + 9k^2x^4 + 12k^2x(-x^3 + x_0^3)}{(1 + 9k^2x^4)^{3/2}}.$$

The criteria for the particle leaving the curve is $N = 0$.

5.5 Maximal Area

About Enough Length

Two points fixed on the x -axis are separated by Δx . The points are connected by a length L of string that is longer than Δx but shorter than $\pi\Delta x/2$:

$$\Delta x < L < \frac{\pi}{2}\Delta x$$

Maximize the area contained above the x -axis and under the string.

Supposing the path of the string is $y(x)$, the system is described by an area integral and an arc length integral:

$$\begin{aligned} A[y] &= \int_{x_0}^{x_1} y \, dx \\ L[y] &= \int_{x_0}^{x_1} \sqrt{1 + y_x^2} \, dx \end{aligned}$$

Introduce a Lagrange multiplier λ to combine each quantity:

$$F[y] = \int_{x_0}^{x_1} \left(y - \lambda\sqrt{1 + y_x^2}\right) dx$$

The working quantity

$$\Lambda = y - \lambda\sqrt{1 + y_x^2}$$

has no explicit x -dependence, thus we reason from Equation (21.5) that

$$\Lambda - y_x \frac{\partial \Lambda}{\partial y_x} = C_1,$$

where C_1 is a constant. Running our function Λ through this results in

$$y - \frac{\lambda}{\sqrt{1 + y_x^2}} = C_1,$$

which can be written as a separable differential equation:

$$\frac{dy}{dx} = \sqrt{\left(\frac{\lambda}{y - C_1}\right)^2 - 1}$$

Letting $w = y - C_1$ we quickly transform the above into an integral

$$\int dx = \int \frac{w \, dw}{\sqrt{\lambda^2 - w^2}},$$

motivating another substitution

$$\begin{aligned} q &= \lambda^2 - w^2 \\ dq &= -2w dw . \end{aligned}$$

From here the integration results in

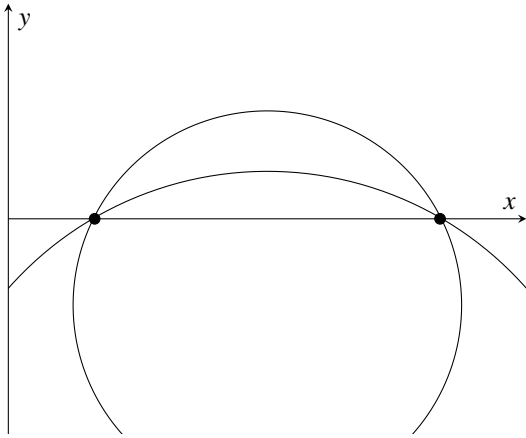
$$x = -\sqrt{q} + C_0 ,$$

where C_0 is a constant.

Restoring all of the substitutions back to y , we get a result that is undoubtedly a circle:

$$(x - C_0)^2 + (y - C_1)^2 = \lambda^2$$

Evidently, the shape that solves the problem is a semicircular arc with radius λ centered at (C_0, C_1) as shown:



Too Much Length

To solve the maximal area problem above, each integral is framed in terms of y and its derivative y_x , which is to assume that the curve never ‘doubles back’ in itself, i.e. $y(x)$ is a function. Clearly this won’t work in the case where there is too much string, i.e.

$$L > \frac{\pi}{2} \Delta x ,$$

which is the case we ponder now.

By intuitive arguments, we could speculate (but not rely on) that the solution for the modified case is also semicircular. If so, it seems that the figure above already depicts the answer, as the quadrant under the x -axis happens to display circles that satisfy the criteria.

As it turns out, it’s a bit easier to proceed by studying a closed loop which has no endpoints. To rationalize this we may assume the special case

$$\Delta x \ll L ,$$

so that the flat part of the resulting curve will be negligible. In effect, we’re finding the otherwise-unconstrained finite closed loop that encloses the most area.

Cartesian Analysis

One way to frame the problem is to write integrals for A and L that do not depend on y being a function. To this end, we borrow from vector calculus and Green’s theorem to write the following area formula for a closed curve:

$$\begin{aligned} A[x, y] &= \oint \frac{1}{2} (x dy - y dx) \\ L[x, y] &= \oint \sqrt{dx^2 + dy^2} \end{aligned}$$

To proceed, frame each of x and y as parametric equations in the variable t . Then, using the chain rule, the above become

$$\begin{aligned} A[x, y] &= \oint \frac{1}{2} (x y_t - y x_t) dt \\ L[x, y] &= \oint \sqrt{x_t^2 + y_t^2} dt , \end{aligned}$$

and from here we combine these using a Lagrange multiplier:

$$F[x, y] = \oint \left(\frac{1}{2} (x y_t - y x_t) - \lambda \sqrt{x_t^2 + y_t^2} \right) dt$$

Picking out the working quantity

$$\Lambda = \left(\frac{1}{2} (x y_t - y x_t) - \lambda \sqrt{x_t^2 + y_t^2} \right)$$

and counting the variables, we can apply two instances of the Euler-Lagrange equation:

$$\begin{aligned} \frac{d}{dt} \left(\frac{\partial \Lambda}{\partial x_t} \right) - \frac{\partial \Lambda}{\partial x} &= 0 \\ \frac{d}{dt} \left(\frac{\partial \Lambda}{\partial y_t} \right) - \frac{\partial \Lambda}{\partial y} &= 0 \end{aligned}$$

Substituting Λ into each gives

$$\begin{aligned} \frac{d}{dt} \left(\frac{-1}{2} y - \frac{\lambda x_t}{\sqrt{x_t^2 + y_t^2}} \right) &= \frac{1}{2} y_t \\ \frac{d}{dt} \left(\frac{1}{2} x - \frac{\lambda y_t}{\sqrt{x_t^2 + y_t^2}} \right) &= \frac{-1}{2} x_t , \end{aligned}$$

both of which are easily integrated. Introducing respective integration constants C_0, C_1 , we find, after simplifying:

$$\begin{aligned} x - C_0 &= \frac{\lambda y_t}{\sqrt{x_t^2 + y_t^2}} \\ y - C_1 &= \frac{-\lambda x_t}{\sqrt{x_t^2 + y_t^2}} \end{aligned}$$

Square both sides and add to recover the formula for a circle:

$$(x - C_0)^2 + (y - C_1)^2 = \lambda^2$$

Polar Analysis

The same result can be framed more ‘naturally’ in polar coordinates, however the calculation that follows isn’t any easier than the Cartesian analysis. The area and length functionals respectively take the form

$$\begin{aligned} A[r, \theta] &= \oint \frac{1}{2} r^2 d\theta \\ L[r, \theta] &= \oint r \sqrt{1 + \left(\frac{1}{r} r_\theta\right)^2} d\theta, \end{aligned}$$

so the total functional using a Lagrange multiplier reads

$$F[r, \theta] = \oint \left(\frac{1}{2} r^2 - \lambda r \sqrt{1 + \left(\frac{1}{r} r_\theta\right)^2} \right) d\theta.$$

The working quantity

$$\Lambda = \frac{1}{2} r^2 - \lambda \sqrt{r^2 + r_\theta^2}$$

is a function of r and r_θ , but there is no explicit θ -dependence. By Equation (21.5), we additionally have

$$\Lambda - r_\theta \frac{\partial \Lambda}{\partial r_\theta} = C,$$

where C is constant. Substituting Λ into the above, after simplifying, gives

$$\frac{1}{2} r^2 - C = \frac{\lambda r^2}{\sqrt{r^2 + r_\theta^2}}.$$

Polar Frame

In a polar coordinate system, the differential arc length is given by

$$ds = \langle dr, r d\theta \rangle,$$

having two components that form the sides of a right triangle with hypotenuse

$$ds = \sqrt{dr^2 + r^2 d\theta^2}.$$

The angle formed between the hypotenuse and $r d\theta$ shall be denoted ϕ and is given by:

$$\sin(\phi) = \frac{r d\theta}{\sqrt{dr^2 + r^2 d\theta^2}} = \frac{r}{\sqrt{r^2 + r_\theta^2}}$$

In terms of ϕ , the differential equation for this system now reads

$$\frac{1}{2} r^2 - C = \lambda r \sin(\phi).$$

Now, if the path of integration is to make a closed loop somewhere in the plane, then the angle ϕ will (at least) hit all of its values in the domain $[0, 2\pi]$. From this we can write a pair of relations

$$\begin{aligned} \frac{1}{2} r_-^2 - C &= \lambda r_- \sin\left(\frac{-\pi}{2}\right) \\ \frac{1}{2} r_+^2 - C &= \lambda r_+ \sin\left(\frac{\pi}{2}\right), \end{aligned}$$

readily implying

$$\lambda = \frac{1}{2} (r_+ + r_-).$$

The distances r_+, r_- are interpreted as the respective furthest and nearest distances from the origin to the extremes of the path.

The case $r^2 = 2C$ corresponds to two points along the path where the $d\theta$ -component of the arc length is zero, i.e. the displacement is momentarily parallel to the line made by r .

Cartesian Frame

It’s possible to begin with the polar analysis of the problem, namely

$$\frac{1}{2} r^2 - C = \frac{\lambda r^2}{\sqrt{r^2 + r_\theta^2}},$$

and end up with a Cartesian result.

Looking at the case $r^2 = 2C$, this corresponds to the scenario where the position vector is tangent to the curve. By specially tuning $C = 0$, we move the origin to somewhere on the curve itself, with the polar axis being along

the tangent to the curve at that point. With this choice, we then have

$$r^2 = \frac{2\lambda r^2}{\sqrt{r^2 + r_\theta^2}},$$

readily simplifying to a separable differential equation

$$\int d\theta = \int \frac{dr}{\sqrt{4\lambda^2 - r^2}},$$

having solution

$$\theta = \theta_0 + \arcsin\left(\frac{r}{2\lambda}\right),$$

where the integration constant $\theta_0 = 0$ by construction.

Rearranging the above, we ultimately find

$$r = 2\lambda \sin(\theta).$$

Making use of the identities

$$x = r \cos(\theta)$$

$$y = r \sin(\theta),$$

we swiftly make out the fingerprint of a circle:

$$r^2 = 2\lambda r \sin(\theta)$$

$$x^2 + y^2 = 2\lambda y$$

$$x^2 + (y - \lambda)^2 = \lambda^2$$

Chapter 22

Differential Equations

1 Separable Equation

A differential equation of the form

$$n(y) \frac{dy}{dt} + m(t) = 0$$

can be solved analytically by means of *separation*.

To proceed, suppose $n(y)$ is the derivative of some unknown function $N(y)$, and also suppose a similar thing for $m(t)$:

$$n(y) = \frac{d}{dy} N(y)$$

$$m(t) = \frac{d}{dt} M(t)$$

Then, by the chain rule, the differential equation is written

$$\frac{dN}{dt} + \frac{dM}{dt} = 0,$$

easily solved by

$$N(t) + M(t) = C,$$

where C is an integration constant.

Problem 1

Solve:

$$x \, dy - y \, dx = 2x^2 y^2 \, dy$$

Hint: Let $f = y/x$ so that

$$df = \frac{dy}{x} - \frac{y \, dx}{x^2} = \frac{x \, dy - y \, dx}{x^2},$$

and then

$$df = 2y^2 \, dy.$$

Answer:

$$x = \frac{y}{2y^3/3 + C}$$

1.1 Freefall Kinematics

Freefall with Linear Air Damping

A simple model for the velocity $v(t)$ of a freefalling object with linear air damping is given by

$$\frac{dv}{dt} = -g - bv,$$

where g is the local gravity constant and b is the linear damping constant.

Separating variables, one writes

$$\frac{dv}{kv/g + 1} = -g \, dt.$$

Let $q = bv/g$ such that $dq = b \, dv/g$, and the above becomes

$$\frac{dq}{1 + q} = -b \, dt.$$

Integrate both sides to find, up to an integration constant C :

$$\ln(1 + q) = -bt + C.$$

Isolate q and restore the v -variable to find

$$v = \tilde{C} e^{-bt} - \frac{g}{b},$$

where $\tilde{C}$ is arbitrary. To impose an initial condition on $v(t)$, let v_0 correspond to $v(t=0)$, giving

$$v_0 = \tilde{C} - \frac{g}{b}.$$

Eliminate $\tilde{C}$ to write the final solution for the velocity:

$$v(t) = \left(v_0 + \frac{g}{b} \right) e^{-bt} - \frac{g}{b}$$

Problem 2

Integrate once more with respect to time and make corresponding arguments about integration constants and the initial position x_0 to get the position $x(t)$:

$$x(t) = x_0 - \frac{gt}{b} + \frac{1}{b} \left(v_0 + \frac{g}{b} \right) (1 - e^{-bt})$$

Freefall with Nonlinear Air Damping

The velocity $v(t)$ of an object in freefall with air resistance can be also modeled by

$$\frac{dv}{dt} = -g - kv^2,$$

where g is the local gravity constant and k is the nonlinear damping constant.

Separating variables, one writes

$$\frac{dv}{kv^2/g + 1} = -g dt.$$

Let $b = k/g$ and let $q = \sqrt{b}v$ to get

$$\frac{dq}{q^2 + 1} = -\sqrt{b}g dt.$$

Proceed by letting

$$q = \tan(\theta)$$

such that

$$dq = \sec^2(\theta) d\theta.$$

The integrand cancels, and the above simplifies to

$$\theta = -\sqrt{kgt} + \phi,$$

where ϕ is an integration constant. Restore the velocity variable to finally arrive at:

$$v(t) = \sqrt{\frac{g}{k}} \tan\left(\phi - \sqrt{kgt}\right)$$

Modified Air Drag Kinematics

The velocity $v(t)$ of an object in gravity that is pushed upward with a force proportional to v^2 is modeled by

$$\frac{dv}{dt} = -g + kv^2,$$

where g is the local gravity constant and k is the propulsion coefficient.

Separating variables, one writes

$$\frac{dv}{kv^2/g - 1} = g dt.$$

Let $b = k/g$ to write the above as

$$\frac{dv}{(1 - \sqrt{b}v)(1 + \sqrt{b}v)} = -g dt,$$

and split the left using partial fractions:

$$\frac{dv}{1 - \sqrt{b}v} + \frac{dv}{1 + \sqrt{b}v} = -2g dt$$

Let $q = \sqrt{b}v$ to simplify again

$$\frac{dq}{1 - q} + \frac{dq}{1 + q} = -2\sqrt{b}g dt,$$

and integrate across the whole equation to find, up to an arbitrary constant C :

$$\ln\left(\frac{1+q}{1-q}\right) = -2\sqrt{kgt} + C$$

Let $\tau = \sqrt{kgt}$ so the above becomes

$$\frac{1+q}{1-q} = e^{-2\tau+C}.$$

Since C is arbitrary, it hurts nothing to recast C by letting $A = e^C$, making A the new constant:

$$\frac{1+q}{1-q} = Ae^{-2\tau}$$

The last hard job is to solve for q . Working the above, we first land at

$$q = \frac{Ae^{-2\tau} - 1}{Ae^{-2\tau} + 1},$$

which is workable, but we can do better. Recast the constant A in terms of yet another arbitrary constant ϕ such that $A = e^{-2\phi}$. Then, multiply the whole equation through by $e^{\phi+\tau}$. This should get you

$$q = \frac{e^{-(\phi+\tau)} - e^{\phi+\tau}}{e^{-(\phi+\tau)} + e^{\phi+\tau}},$$

which simplifies beautifully:

$$q = \tanh(\phi + \tau)$$

Restoring all original variables, the above reads

$$v(t) = \sqrt{\frac{g}{k}} \tanh\left(\phi + \sqrt{kgt}\right)$$

Problem 3

In the above derivation, let $k \rightarrow -k$ to recover the solution for freefall with nonlinear air damping.

Answer: All steps are essentially the same until choosing τ , which instead becomes

$$\tau = i\sqrt{kg}t = i\tau_0,$$

thus

$$q = \frac{Ae^{-2i\tau_0} - 1}{Ae^{-2i\tau_0} + 1},$$

Recast A such that $A = e^{2i\phi}$ and multiply the whole equation through by $e^{i(\phi-\tau_0)}$ to get

$$q = \frac{e^{-i(\phi-\tau_0)} - e^{i(\phi-\tau_0)}}{e^{-i(\phi-\tau_0)} + e^{i(\phi-\tau_0)}},$$

simplify to

$$q = -i \tan(\phi - \tau_0).$$

Restore the v -variable and simplify to confirm:

$$v(t) = -\sqrt{\frac{g}{k}} \tan(\phi - \sqrt{kg}t)$$

Note the overall minus sign leading this equation; it arises naturally from the time reversal implicit in the substitution $k \rightarrow -k$.

2 Integrating Factor

Looking at the differential equation

$$\frac{dy}{dt} + n(t) y(t) = m(t) ,$$

it would be convenient to turn the left side into a total derivative. To this end, multiply through by a so-far unknown function $u(t)$ so write

$$\frac{dy}{dt} u(t) + n(t) u(t) y(t) = m(t) u(t) .$$

For the left side to be the total derivative of the quantity

$y(t) u(t)$, it must be true that

$$\frac{du}{dt} = n(t) u(t) ,$$

which is itself a separable differential equation. Rearranging, we write

$$\frac{du}{u} = n(t) ,$$

solved by

$$u(t) = e^{\int n(t) dt} .$$

The term $u(t)$ is called an *integrating factor*.

Chapter 23

Fourier Analysis

Part IV

Advanced Mathematics

Chapter 24

Special Functions

1 Gamma Function

1.1 Applications

Stirling's Approximation

For very large numbers, a crude approximation for the factorial of such a number is given by a weak version of Stirling's approximation, namely

$$n! \approx \left(\frac{n}{e}\right)^n,$$

which can be derived from elementary calculus.

While sufficient for many cases, the weak Stirling's approximation is missing an essential term. The full Stirling's approximation reads

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}, \quad (24.1)$$

which has an extra $\sqrt{2\pi n}$ tacked onto the right. Innocent as it appears, accounting for this term is not trivial.

To begin, write the definition of the gamma function

$$\Gamma(z) = (z-1)! = \int_0^\infty t^{z-1} e^{-t} dt,$$

from which we write

$$n! = \int_0^\infty e^{-x} x^n dx.$$

Looking at the $e^{-x}x^n$ -term, notice that this is very sharply-peaked at $x = n$, thus we approximate $x \approx n$ inside the integral. In particular, we write

$$x = n + \epsilon,$$

where $\epsilon \ll n$. With this, we further write

$$\begin{aligned} \ln(e^{-x}x^n) &= \ln(e^{-(n+\epsilon)}(n+\epsilon)^n) \\ &= -n - \epsilon + n \ln(n + \epsilon) \\ &= -n - \epsilon + n \left(\ln(n) + \ln\left(1 + \frac{\epsilon}{n}\right) \right). \end{aligned}$$

Use Taylor expansion to simplify the last term via

$$n \ln\left(1 + \frac{\epsilon}{n}\right) = \epsilon - \frac{\epsilon^2}{2n} + \frac{\epsilon^3}{3n^2} - \dots,$$

and write, after simplifying:

$$e^{-x}x^n \approx \exp\left(-n + \ln(n^n) - \frac{\epsilon^2}{2n} + \frac{\epsilon^3}{3n^2} - \dots\right)$$

Looking again at the integral for $n!$, and since the off-peak values are diminished, it makes no harm to extend the lower integration limit to $-\infty$. With this, we have

$$n! \approx \left(\frac{n}{e}\right)^n \int_{-\infty}^\infty \exp\left(-\frac{\epsilon^2}{2n} + \frac{\epsilon^3}{3n^2} - \dots\right) d\epsilon.$$

Keeping only the order-two term in ϵ , the remaining Gaussian integral resolves via

$$\int_{-\infty}^\infty e^{-\epsilon^2/2n} d\epsilon = \sqrt{2\pi n},$$

and Equation (24.1) emerges.

Hypersphere

In n dimensions, a hypersphere of 'radius' R is given by

$$\sum_{j=1}^n x_j^2 < R^2.$$

In Cartesian coordinates, the volume of the hypersphere is given by

$$V_n = \int_{\sum_{j=1}^n x_j^2 < R^2} dx_1 dx_2 \dots dx_n .$$

In a spherical coordinate system, the integral separates into radial and angular components. The radial component is easy to solve, but the angular part is not:

$$V_n = \int_0^R r^{n-1} dr \int d\Omega_{n-1} = \frac{R^n}{n} \int d\Omega_{n-1}$$

To proceed, recall the identity

$$\sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} dx ,$$

and multiply n copies together to write

$$\begin{aligned} (\sqrt{\pi})^n &= \left(\int_{-\infty}^{\infty} e^{-x_1^2} dx_1 \right) \dots \left(\int_{-\infty}^{\infty} e^{-x_n^2} dx_n \right) \\ &= \int_{-\infty}^{\infty} e^{-r^2} r^{n-1} dr \int d\Omega_{n-1} . \end{aligned}$$

The r -integral resembles a gamma function identity

$$\int_0^{\infty} e^{-r^2} r^{n-1} dr = \frac{\Gamma(n/2)}{2} ,$$

implying

$$\pi^{n/2} = \frac{\Gamma(n/2)}{2} \int d\Omega_{n-1} .$$

With two ways to express the angular component of the volume integral, eliminate it entirely to find

$$V_n = \frac{\pi^{n/2} R^n}{n} \frac{2}{\Gamma(n/2)} .$$

Use

$$\frac{\Gamma(n/2)}{2} = \frac{\Gamma(n/2 + 1)}{n}$$

$$\Gamma(n) = (n-1)!$$

to get the final answer for the volume of a hypersphere:

$$V_n = \frac{\pi^{n/2} R^n}{(n/2)!} \quad (24.2)$$

Testing for $n = 1, \dots, 4$, we have:

$$\begin{aligned} V_1 &= \frac{\sqrt{\pi} R}{(1/2)!} = 2R \\ V_2 &= \pi R^2 \\ V_3 &= \frac{\pi^{3/2} R^3}{(3/2)!} = \frac{4\pi R^3}{3} \\ V_4 &= \frac{\pi^2 R^4}{2} \end{aligned}$$

As for the surface area of the hypersphere, we can rely on the fact that works clearly in lower dimensions. In particular we write

$$A_{n-1} = \frac{d}{dR} V_n ,$$

where A_n is the area of n -dimensional hypersphere. Substituting Equation (24.2), the above can be written

$$A_n = \frac{n+1}{R} V_{n+1} . \quad (24.3)$$

Testing for $n = 1, \dots, 4$, we have:

$$\begin{aligned} A_1 &= \frac{2}{R} V_2 = 2\pi R \\ A_2 &= \frac{3}{R} V_3 = 4\pi R^2 \\ A_3 &= \frac{4}{R} V_4 = 2\pi^2 R^3 \\ A_4 &= \frac{5}{R} V_5 \end{aligned}$$

Problem 1

Calculate the surface area A_4 in terms of R .

Chapter 25

Systems and Distributions

1 Large-N Systems

Continuous Systems

For a stochastic process that produces events A_k in a continuous range instead of a discrete set, the normalization condition

$$\sum_{k=1}^n P(A_k) = 1$$

becomes an infinite sum. When confronted with this, the sum becomes an integral according to

$$\sum_{k=1}^n P(A_k) \rightarrow \int dP(A_k) = 1.$$

Probability Distribution Function

At this point, we abbreviate $A_k \rightarrow k$, and then use the chain rule to write

$$1 = \int_n \frac{dP(k)}{dk} dk = \int_n w(k) dk.$$

The continuous function $w(k)$ is called the *probability density*, or *probability distribution function* (although ‘w’ stands for *weight*). Specifically, $w(k)$ is the probability of an event occurring within a window $[k, k + dk]$.

In the continuous limit, the equations for the statistical average, general expectation value, and standard deviation

generalize to

$$\langle k \rangle = \int_n k \cdot w(k) dk \quad (25.1)$$

$$\langle f \rangle = \int_n f(k) w(k) dk \quad (25.2)$$

$$\sigma_f = \sqrt{\int_n (f(k) - \langle f \rangle)^2 w(k) dk} \quad (25.3)$$

in accordance with the definitions:

$$\sigma_f = \sqrt{\langle f^2 \rangle - 2 \langle f \rangle \langle f \rangle + \langle f \rangle^2}$$
$$\sigma_f = \sqrt{\langle f^2 \rangle - \langle f \rangle^2}$$

Problem 1

What is the expected area of a right triangle with a hypotenuse of k whose non-right angles are uniformly distributed over the interval $(0, \pi/2)$? Answer:

$$\begin{aligned} \langle A \rangle &= \frac{k^2/2}{\pi/2} \int_0^{\pi/2} \cos(\theta) \sin(\theta) d\theta \\ &= \frac{k^2/2}{\pi/2} \int_0^1 x dx = \frac{k^2}{2\pi} \end{aligned}$$

Problem 2

Divide a given line segment into two other line segments. Then, cut each of these new line segments into two more line segments. What is the probability that the resulting four line segments are the sides of a quadrilateral?

Answer: Let the total length be L , and require that no one side be longer than $L/2$. After the initial cut, let the longer segment have length x , and the shorter segment

$L - x$. Diving the longer segment at point z (from the start of x), it is required that $z < L/2$ and simultaneously $x - z < L/2$. Therefore, the window of allowed z has width $L/2 - (x - L/2) = L - x$. The normalized probability of an allowed z along x is:

$$P = N \int_{L/2}^L \frac{L-x}{x} dx = \frac{(L \ln x - x) \Big|_{L/2}^L}{L/2} = 2 \ln 2 - 1 \approx 38.6\%$$

1.1 Random Variables

Consider a set $\{A_k\}$ of random (not necessarily independent) variables.

Sum of Random Variables

Suppose that the sum of random variables comes to A :

$$A = \sum_{k=1}^n A_k$$

In the continuous large- n limit, the average value of A can be written as an n -dimensional integral

$$\langle A \rangle = \int A \cdot w(A_1, \dots, A_n) dA_1 \dots dA_n.$$

Replace A in the above with its sum representation:

$$\langle A \rangle = \sum_{k=1}^n \int A_k \cdot w(A_1, \dots, A_n) dA_1 \dots dA_n,$$

where the ‘sum’ symbol has been harmlessly pulled outside all n of the integrals.

Simplifying the above is a straightforward exercise, with the majority of integrals satisfying the normalization condition and resolving to one. After the dust settles, one finds

$$\langle A \rangle = \langle A_1 \rangle + \langle A_2 \rangle + \dots + \langle A_n \rangle,$$

which, strictly translated, means *the average of the sum is the sum of the averages*:

$$\langle A \rangle = \sum_{k=1}^n \langle A_k \rangle \quad (25.4)$$

Problem 3

Prove explicitly the $n = 2$ case of the above by letting $A = x + y$ to show that $\langle A \rangle = \langle x \rangle + \langle y \rangle$. Answer:

$$\begin{aligned} \langle A \rangle &= \int \int (x + y) w(x, y) dx dy \\ &= \int x \left(\int w(x, y) dy \right) dx \\ &\quad + \int y \left(\int w(x, y) dx \right) dy \\ &= \int x \cdot w(x) dx + \int y \cdot w(y) dy \\ &= \langle x \rangle + \langle y \rangle \end{aligned}$$

Independent Random Variables

More can be said about the weight function $w(k)$ in the regime of independent random variables. In the same sense that $P(A \cap B) = P(A)P(B)$ applies to independent events, we write

$$w(A_1, A_n, \dots, A_n) = w(A_1)w(A_2) \dots w(A_n)$$

when all probability distribution values $w(A_k)$ are independent.

Product of Independent Random Variables

Suppose that the product of random variables $\{B_k\}$ of n comes to

$$B = \prod_{k=1}^n B_k = B_1 \cdot B_2 \dots B_n,$$

and let us calculate the average value $\langle B \rangle$. Going by definition, this amounts to

$$\langle B \rangle = \prod_{k=1}^n \int B_k \cdot w(B_1) \dots w(B_n) dB_1 \dots dB_n,$$

the ‘product’ symbol has been pulled outside all n of the integrals, and the probability distribution is factored to accommodate independent B_k .

From this, we see the right side is the product of n independent integrals, and conclude

$$\langle B \rangle = \langle B_1 \rangle \cdot \langle B_2 \rangle \dots \langle B_n \rangle,$$

which, strictly translated, means *the average of the product is the product of the averages*:

$$\langle B \rangle = \prod_{k=1}^n \langle B_k \rangle \quad (25.5)$$

1.2 Variance

Starting from the sum

$$A = \sum_{k=1}^n A_k ,$$

square both sides and convince yourself that

$$A^2 = \left(\sum_{i=1}^n A_i \right) \left(\sum_{j=1}^n A_j \right) = \sum_{k=1}^n A_k^2 + \sum_{i \neq j} c_{ij} A_i A_j ,$$

where c_{ij} are the binomial coefficients to represent all cross terms.

Meanwhile, the square of the average $\langle A \rangle$ comes out to

$$\langle A \rangle^2 = \sum_k \langle A_k \rangle^2 + \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle .$$

We can also calculate $\langle A^2 \rangle$ by exploiting the independence among A_k , resulting in

$$\langle A^2 \rangle = \sum_{k=1}^n \langle A_k^2 \rangle + \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle .$$

Taking the difference $\langle A^2 \rangle - \langle A \rangle^2$, the cross terms cancel and we arrive at a simple relation connecting A to its members:

$$\begin{aligned} \langle A^2 \rangle - \langle A \rangle^2 &= \sum_{k=1}^n \langle A_k^2 \rangle - \langle A_k \rangle^2 \\ &+ \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle - \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle \end{aligned}$$

The quantity $\langle A^2 \rangle - \langle A \rangle^2$ is defined as the *variance* in A :

$$\text{Var} (A) = \langle A^2 \rangle - \langle A \rangle^2 \quad (25.6)$$

As we've built it, the variance has some more handy expressions:

$$\text{Var} (A) = \sum_{k=1}^n \langle A_k^2 \rangle - \langle A_k \rangle^2 = \sum_{k=1}^n \text{Var} (A_k)$$

1.3 Dispersion

A variation in the sum A of independent variables, denoted ΔA , is also known as *dispersion*, defined as:

$$\Delta A = A - \langle A \rangle = \sum_{k=1}^n (A_k - \langle A_k \rangle) \quad (25.7)$$

From this, it's easy to show that the average dispersion is zero:

$$\langle \Delta A \rangle = \langle A \rangle - \langle A \rangle = 0$$

The expectation value $\langle \Delta A^2 \rangle$, however, is more telling. By brute force, first write

$$\begin{aligned} \Delta A^2 &= ((A_1 - \langle A_1 \rangle) + (A_2 - \langle A_2 \rangle) + \dots)^2 \\ &= \sum_{k=1}^n (A_k - \langle A_k \rangle)^2 + \sum_{i \neq j} c_{ij} \Delta A_i \Delta A_j , \end{aligned}$$

so then:

$$\langle \Delta A^2 \rangle = \sum_{k=1}^n \langle (A_k - \langle A_k \rangle)^2 \rangle + \sum_{i \neq j} c_{ij} \langle \Delta A_i \rangle \langle \Delta A_j \rangle$$

This result is perhaps not surprising, telling us the total ΔA^2 is the sum of its constituents:

$$\langle \Delta A^2 \rangle = \sum_{k=1}^n \langle \Delta A_k^2 \rangle \quad (25.8)$$

In the large- n limit, the average $\langle A \rangle$ scales with n , and meanwhile we see $\langle \Delta A^2 \rangle$ also scales with n . The ratio of the RMS dispersion over the average thus tends to zero, as

$$\frac{\sqrt{\langle \Delta A^2 \rangle}}{\langle A \rangle} \approx \frac{1}{\sqrt{n}} \rightarrow 0 ,$$

telling us that fluctuations in A become negligibly small.

2 Discrete Distributions

2.1 Two-State System

Consider a balanced coin that is tossed to generate n random events resulting in either H (eads) or T (ails). If we are interested in the portion m 'heads' events that occur without the order of events being important, the combination number

$$C_n^m = \frac{n!}{m! (n-m)!}$$

summarizes the system. Said another way, the multiplicity of the system Ω is 'n choose m':

$$C_n^m = \binom{n}{m}$$

The sum of all C_n^m across the whole range of m , namely from 0 to n , must resolve to the total multiplicity of events, namely 2^n for a coin tossing game:

$$2^n = \sum_{m=0}^n \frac{n!}{m! (n-m)!}$$

Multiplicity

A synonym for C_n^m is the *multiplicity*, denoted Ω (Greek *Omega*), which for the two-state system means

$$\Omega(m, n) = C_n^m.$$

The total multiplicity Ω_T is the sum of all $\Omega(m, n)$ for the available states m . For the two-state system with n trials, we simply have

$$\Omega_T = 2^n.$$

One can readily imagine that the notion of multiplicity generalizes to more than the two-state system. In a general case, the probability of observing the system in state m is the ratio of the corresponding Ω to Ω_T :

$$\Pr(m, n) = \frac{1}{\Omega_T} \Omega(m, n)$$

Normalized Probability Distribution

Knowing the total states available to the two-state system, we can write the probability of attaining m events among n trials in any two-state system as:

$$P(m, n) = \frac{1}{2^n} \frac{n!}{m! (n-m)!} \quad (25.9)$$

In the above definition, we divide by the factor 2^n so that the sum of all probabilities - accounting for all outcomes - sums to one.

The combination number C_n^m can be interpreted nicely by spotting the pattern that emerges in trivial cases. A single toss can result in T or H , which we denote

$$\omega_1 = (T, H).$$

Denoting m as the number of H -events, we write

$$C_1^0 = 1 \quad C_1^1 = 1$$

For a game of $n = 2$ tosses, the list of possible events is

$$\omega_2 = (TT, TH, HT, HH).$$

Again denoting m as the number of H -events, we write

$$C_2^0 = 1 \quad C_2^1 = 2 \quad C_2^2 = 1$$

Similarly, a game of three tosses has

$$\omega_3 = (TTT, TTH, THT, THH, HTT, HTH, HHH)$$

with combinations

$$C_3^0 = 1 \quad C_3^1 = 3 \quad C_3^2 = 3 \quad C_3^3 = 1.$$

The pattern in C_n^m (stand back and look at the page) matches the rows of Pascal's triangle.

Heuristic Derivation

There is a (perhaps) intuitive way to derive C_n^m . Take n coins and lay them all down showing T , represented by $C_n^0 = 1$. Turn any one of the coins to H and find $C_n^1 = n$. Turn any two of the coins to H and find

$$C_n^2 = n \frac{(n-1)}{2},$$

and for three,

$$C_n^3 = n \frac{(n-1)}{2} \frac{(n-2)}{3},$$

and so on. Building this up for m total H -faces, we find

$$C_n^m = \frac{n!}{m! (n-m)!},$$

the familiar combination number.

2.2 Multi-State System

A generalization of the two-state system is the *multi-state* system. Going for a modest example, consider a three-sided coin with faces A, B, C . Flipping such a coin to generate n total events, let:

- $n_A \rightarrow$ Number of outcomes A
- $n_B \rightarrow$ Number of outcomes B
- $n_C \rightarrow$ Number of outcomes C
- $p_A \rightarrow$ Probability of outcome A

- $p_B \rightarrow$ Probability of outcome B
- $p_C \rightarrow$ Probability of outcome C

With this, we can write the multiplicity of the system exhibiting the state (n_A, n_B, n_C, n) :

$$\Omega(n_A, n_B, n_C, n) = \frac{n!}{n_A! n_B! n_C!} p_A^{n_A} p_B^{n_B} p_C^{n_C}$$

In the special case $C = 0$, the above reduces to the non-normalized two-state distribution.

3 Binomial Distribution

Derivation

Consider an *unbalanced* coin having inherent probability p to land on H (eads), and correspondingly $1 - p$ to land on T (ails). As a generalized two-state system, a game of n tosses generates the same potential outcomes:

$$\Omega_1 = T, H$$

$$\Omega_2 = TT, TH, HT, HH$$

$$\Omega_3 = TTT, TTH, THT, THH, HTT, HTH, HHH$$

Of course, the probability P of generating m Heads-events requires an extra argument to account for the imbalance p . Denoting the modified combination symbol $P_n^m(p)$, the two-state analysis generalizes by:

$$P_1^0(p) = 1 - p$$

$$P_1^1(p) = p$$

$$P_2^0(p) = (1 - p)^2$$

$$P_2^1(p) = 2 \cdot p (1 - p)$$

$$P_2^2(p) = p^2$$

$$P_3^0(p) = (1 - p)^3$$

$$P_3^1(p) = 3 \cdot p (1 - p)^2$$

$$P_3^2(p) = 3 \cdot p^2 (1 - p)$$

$$P_3^3(p) = p^3$$

Evidently, the factors of p and $1 - p$ compound into the terms p^m and $(1 - p)^{n-m}$, but otherwise this analysis traces that of the two-state system exactly. Scanning for a pattern in the above, we evidently have

$$P_n^m(p) = \binom{n}{m} (1 - p)^{n-m} p^m.$$

This result is known as the *binomial distribution*, and gives the probability of attaining, in general, m events of weight p among n trials:

$$P(m, n, p) = \frac{n!}{m! (n - m)!} (1 - p)^{n-m} p^m \quad (25.10)$$

Note there is no need to divide by 2^n . The binomial distribution as written is unit-normalized already.

Problem 4

Evaluate $P(m, n, 1/2)$ to reproduce Equation (25.9).

Average Values

Define a random variable z_k that is equal to one if the event H with weight p occurs in the k -th trial, and is equal to zero otherwise. The average value of z_k is then

$$\begin{aligned} \langle z_k \rangle &= P(H) z(H) + P(T) z(T) \\ &= p \cdot 1 + (1 - p) \cdot 0 = p, \end{aligned}$$

and, simply enough, the average of z_k^2 reads

$$\langle z_k^2 \rangle = p \cdot 1^2 + (1 - p) \cdot 0^2 = p.$$

Standard Deviation

The standard deviation in z , denoted σ_z , is evidently

$$\sigma_z = \sqrt{\langle z_k^2 \rangle - \langle z_k \rangle^2} = \sqrt{p - p^2} = \sqrt{p(1 - p)}.$$

Next, note that the number m of H -events among the n independent trials is the sum

$$m = \sum_{k=1}^n z_k,$$

implying

$$\langle \Delta m^2 \rangle = \sum_{k=1}^n \langle \Delta z_k^2 \rangle = n \langle \Delta z^2 \rangle,$$

or, in tighter notation for large- n systems,

$$\sigma_m = \sqrt{n \sigma_z^2} = \sqrt{np(1 - p)}.$$

Mode

The mode of the binomial distribution corresponds to the value m that maximizes $P(m, n, p)$. To gain on this question, consider the ratio

$$R = \frac{P(m + 1, n, p)}{P(m, n, p)},$$

readily simplifying to

$$R = \left(\frac{n - m}{m + 1} \right) \frac{p}{1 - p}.$$

The special value $R = 1$ indicates where the derivative of $P(m, n, p)$ is zero, i.e. the probability transitions from increasing with m to decreasing with m . Setting $R = 1$ at the critical m^* , we further find

$$m^* = np - (1 - p).$$

In the large- n limit, the mode m^* approximates to:

$$m^* \approx np$$

Problem 5

Haldor the Viking has slain sixteen ooze creatures in the swamp. After a thorough forensic analysis, Haldor finds a single gold cup among the corpses. He remembers from swamp lore that a slain ooze has a $1/3$ chance to drop a gold cup. What are the chances he found just one cup after slaying sixteen oozees? Repeat the calculation for finding two cups, three cups, etc., up to sixteen cups. Also account for zero cups.

Answer: Model a slain ooze as a weighted coin with a Heads probability of $1/3$, and a Tails probability of $2/3$, which calls for a straightforward application of the binomial distribution. For finding one gold cup, we have

$$\begin{aligned} P(16, 1, 1/3) &= \frac{16!}{1!(16-1)!} (2/3)^{16-1} (1/3)^1 \\ &= \frac{16}{3} \left(\frac{2}{3}\right)^{15} \approx 0.01218, \end{aligned}$$

and then for other numbers of gold cups:

$$\begin{aligned} P(16, 2, 1/3) &\approx 0.04567 \\ P(16, 3, 1/3) &\approx 0.1066 \\ P(16, 4, 1/3) &\approx 0.1732 \\ P(16, 5, 1/3) &\approx 0.2078 \end{aligned}$$

$$\begin{aligned} P(16, 6, 1/3) &\approx 0.1905 \\ P(16, 7, 1/3) &\approx 0.1361 \\ P(16, 8, 1/3) &\approx 0.07654 \\ P(16, 9, 1/3) &\approx 0.03402 \end{aligned}$$

$$\begin{aligned} P(16, 10, 1/3) &\approx 0.01191 \\ P(16, 11, 1/3) &\approx 0.003247 \\ P(16, 12, 1/3) &\approx 0.0006765 \\ P(16, 13, 1/3) &\approx 0.0001041 \end{aligned}$$

$$\begin{aligned} P(16, 14, 1/3) &\approx 0.00001115 \\ P(16, 15, 1/3) &\approx 0.0000007434 \\ P(16, 16, 1/3) &\approx 0.00000002323 \\ P(16, 0, 1/3) &\approx 0.001522 \end{aligned}$$

Problem 6

Monique is practicing netball. She knows from experience that the probability of her making any one shot is 70%. Her coach has asked her to keep practicing until she scores 50 goals. How many shots would she need to attempt to ensure that the probability of making at least 50 shots is more than 99%?

Answer: This problem is analogous to flipping a weighted coin with bias p . The probability of scoring k shots in N tosses is

$$P(k, N, p) = \frac{N!}{k!(N-k)!} (1-p)^{N-k} p^k,$$

where summing over k gives the cumulative distribution:

$$99\% = \sum_{k=50}^N \frac{N!}{k!(N-k)!} (0.3)^{N-k} (0.7)^k$$

This is best solved by a computer, where one should find $N = 86$.

4 Continuous Distributions

4.1 Gaussian Distribution

Recall that the probability of generating k results among n total trials in a two-state system is given by

$$P(k, n) = \frac{1}{2^n} \frac{n!}{k! (n-k)!}.$$

Introduce the shift

$$k \rightarrow k + \frac{n}{2},$$

which modifies the above:

$$P(k, n) = \frac{1}{2^n} \frac{n!}{\left(\frac{n}{2} + k\right)! \left(\frac{n}{2} - k\right)!}$$

In the large- k limit, making k a continuous variable, it makes sense to describe the system solely in terms of expectation values and their deviations, a notion formally called the *central limiting theorem*. Here we develop this idea on a two-state system to derive a central equation in probability theory called the *Gaussian distribution*.

To proceed in the large n -limit, we deploy Stirling's approximation for large numbers

$$\begin{aligned} \ln(n!) &\approx n \ln(n) - n + \ln(\sqrt{2\pi n}) \\ n! &\approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}, \end{aligned}$$

and the probability density reduces to

$$w(k) = e^{-2k^2/n} \sqrt{\frac{2}{\pi n}}. \quad (25.11)$$

The result $w(k)$ is the famed normalized Gaussian distribution centered at $k = 0$. Introducing a nonzero shift of base-point value a , the generalized equation is

$$w(k) = e^{-2(k-a)^2/n} \sqrt{\frac{2}{\pi n}}.$$

Using Gaussian integrals, the average values and standard deviation are readily calculated:

$$\begin{aligned} \langle k \rangle &= \int_n k \cdot w(k) dk = a \\ \langle k^2 \rangle &= \int_n k^2 \cdot w(k) dk = \frac{n}{4} + a^2 \\ \sigma_k &= \sqrt{\langle k^2 \rangle - \langle k \rangle^2} = \sqrt{\frac{n}{4}} \end{aligned}$$

4.2 Poisson Distribution

Imagine trying to count the number of water molecules that pass a point in a river flowing at average speed v . Over time interval t , the average molecule count is directly proportional to vt . To reduce notation clutter, let us ignore the proportionality constant and take vt as a dimensionless quantity. Due to local random fluctuations in the river, an actual measurement would never precisely land on vt , but instead on an interval surrounding vt . Naturally we wonder, what is the time-varying probability $P_k(t)$ that k molecules are measured over the interval t ?

To begin, partition the elapsed time t into n identical bins of width Δt such that $\Delta t \rightarrow 0$, and observe that each $P_k(\Delta t)$ relates to its $k-1$ and $k+1$ neighbors as:

$$\lim_{\Delta t \rightarrow 0} P_0(\Delta t) \gg P_1(\Delta t) \gg P_2(\Delta t) \gg P_3(\Delta t) \gg \dots$$

This means it's more likely to measure few molecules in a small Δt -interval as opposed to many. We may proceed using weighted two-state analysis, wherein a Δt -interval may either be unfilled with zero molecules, or filled with one or more molecules. Borrowing the apparatus developed previously, we write

$$P(k, n, v\Delta t) = \frac{n!}{k! (n-k)!} (1 - v\Delta t)^{n-k} (v\Delta t)^k,$$

where n and k are integers. Substituting $t = n\Delta t$, we have

$$P(k, n, vt) = \frac{(vt)^k}{k!} \left(\frac{n!}{(n-k)!} \frac{1}{n^k} \right) \left(1 - \frac{vt}{n} \right)^{n-k}.$$

In the large- n limit, the approximations

$$\begin{aligned} \frac{n!}{(n-k)!} &\approx n^k \\ \left(1 - \frac{vt}{n} \right)^{n-k} &\approx e^{-vt} \end{aligned}$$

are valid, and re-casting vt as a dimensionless variable q lands us at the anticipated *Poisson distribution*:

$$P_k(q) = \frac{q^k}{k!} e^{-q} \quad (25.12)$$

Summing over the variable k tells us $P_k(t)$ is already normalized:

$$\sum_{k=0}^{\infty} \frac{q^k}{k!} e^{-q} = e^{-q} \left(\sum_{k=0}^{\infty} \frac{q^k}{k!} \right) = e^{-q} e^q = 1$$

With $P_k(t)$ on hand, we may calculate $\langle k \rangle$, $\langle k^2 \rangle$, and the standard deviation:

$$\begin{aligned}\langle k \rangle &= \sum_{k=0}^{\infty} k \frac{q^k}{k!} e^{-q} = e^{-q} \sum_{k=1}^{\infty} \frac{q^k}{(k-1)!} \\ &= e^{-q} \sum_{p=0}^{\infty} \frac{q^{(p+1)}}{p!} = e^{-q} q e^q = q \\ \langle k^2 \rangle &= \sum_{k=0}^{\infty} k^2 \frac{q^k}{k!} e^{-q} = e^{-q} q \sum_{p=0}^{\infty} (p+1) \frac{q^p}{p!} \\ &= q + q^2 \\ \sigma_k &= \sqrt{q^2 + q - q^2} = \sqrt{q}\end{aligned}$$

4.3 Random Product Problem

Consider the real numbers in the interval $(0 : 2)$. Let $\tilde{x}_1$ be a random number chosen from this interval, let $\tilde{x}_2$ be a second random number, and so on up to $\tilde{x}_n$. (Repeats are allowed but unlikely.)

Expectation

With this setup, suppose we are interested in the product of numbers in the list:

$$X_n = \prod_{j=1}^n \tilde{x}_j = \tilde{x}_1 \cdot \tilde{x}_2 \cdot \tilde{x}_3 \cdots \tilde{x}_n$$

Sampling from $(0 : 2)$, it is true that the average random value is one:

$$\langle \tilde{x}_j \rangle = 1.$$

This should mean right away that the average product is also one:

$$\langle X_n \rangle = 1 \cdot 1 \cdot 1 \cdots = 1$$

Disaster

All seems well until we try to verify $\langle X_n \rangle = 1$ on a calculator. To illustrate, take the contrived list with five members

$$\{\tilde{x}_j\} = \{0.8, .9, 1.0, 1.1, 1.2\},$$

so the product is

$$X_5 = (0.8)(0.9)(1.0)(1.1)(1.2) = 0.9504,$$

which is less than one.

The effect gets worse for increasing n , for if we continue the pattern so the list includes 0.7, 1.3, the product is

$$X_7 \approx 0.8648.$$

The members $\tilde{x}_j < 1$ weigh down the product X_n more than members $\tilde{x}_j > 1$ weigh the product up. After many trials, the net result is $X_n \rightarrow 0$, in contradiction to $\langle X_n \rangle = 1$.

You're encouraged to verify this on a computer with a variety of $\tilde{x}_j$ and a variety of n -values to see there is clearly something wrong with the way X_n is expected to behave. It seems that X_n reliably *decreases* for increasing n , so we inevitably conclude $X_n \rightarrow 0$.

Modified Interval

Going back to the beginning, adjust the interval to $(0 : 2.5)$ so that

$$\langle \tilde{x}_j \rangle = 1.25,$$

and run similar experiments. Now we're multiplying a list of numbers whose average is clearly greater than one. However, much like the previous setup, the product X_n still goes to zero.

Adjust the interval once more to $(0 : 3)$ and start over. This time, we have

$$\langle \tilde{x}_j \rangle = 1.5,$$

and pattern finally breaks. One can check that product X_n tends to grow for increasing n , and for large n , the trend $X_n \rightarrow \infty$ occurs.

Tuning the Interval

Given the evidence on hand, there should be some interval $(0 : p)$, where p is some number between 2.5 and 3 such that X_n does not tend to zero and does not tend to infinity:

$$X_n \propto \langle X_n \rangle$$

To estimate p , one may write a simple trial-and-error program that allows p to vary:

1. Choose an initial value for p .
2. Choose a sufficiently large sample of n values from the interval $(0 : p)$ and calculate the corresponding X_n .
3. If X_n goes to zero, increase p .
4. If X_n goes to infinity, decrease p .
5. Goto step 2.

Doing this, one finds, after many trials:

$$p \approx 2.718 \dots$$

This answer is tantalizingly close to Euler's constant. Who saw that coming?

Proper Analysis

To reconcile the random product problem, begin with the natural logarithm of the product X_n :

$$\ln(X_n) = \ln(\tilde{x}_1) + \ln(\tilde{x}_2) + \ln(\tilde{x}_3) + \dots$$

In the limit $n \rightarrow \infty$, it stands to reason that *every* real number in the interval $(0 : p)$ is represented by some $\tilde{x}_j$ or another. Rearranging the sum to write these in order, we have

$$\ln(X) = \lim_{n \rightarrow \infty} \sum_{j=1}^n \ln\left(p \frac{j}{n}\right).$$

The total interval $(0 : p)$ can be made from n copies of a small interval Δx , which means $p/n = \Delta x$. Also substituting $x = j/n$, the above becomes

$$\ln(X) = \lim_{\Delta x \rightarrow 0} \frac{1}{p} \sum_{x>0}^1 \ln(px) \Delta x.$$

The sum becomes an integral in the continuous limit

$$\ln(X) = \frac{1}{p} \int_0^1 (\ln(p) + \ln(x)) dx,$$

and the solution is straightforward:

$$\begin{aligned} p \ln(X) &= (x \ln(p) + x \ln(x) - x) \Big|_0^1 \\ p \ln(X) &= \ln(p) - 1 \end{aligned}$$

Now comes the final argument. By avoiding $\ln(0) \rightarrow -\infty$ and also $\ln(\infty) \rightarrow \infty$, we're asking for X to be a finite number. In the infinite limit, it can only be that $X \rightarrow 1$:

$$p \ln(1) = 0 = \ln(p) - 1$$

The only solution to $\ln(p) = 1$ is $p = e$ and we're done.

Problem 7

Consider the real numbers in the interval $(0 : 1)$, and let $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$, etc. represent random samples from this interval. How many times n must a random $\tilde{x}_j$ be multiplied into a very large number $A \gg 1$ until the product is approximately one? In other words, solve for n in the following:

$$1 \approx A \cdot \tilde{x}_1 \cdot \tilde{x}_2 \cdots \tilde{x}_n$$

Hint:

$$0 \approx \ln(A) + \sum_{j=1}^n (\ln(x) + 1) - \sum_{j=1}^n (1)$$

The answer is $n \approx \ln(A)$.

4.4 Random Sums Problem

Accumulating random values $0 < r_k < 1$ in a sum, how many iterations $\langle n \rangle$ until the total is greater than one, on average?

Geometric Analysis

Begin by interpreting each interval $0 \leq r_k \leq 1$ as an independent 'number line' for each of the n variables needed. For $n = 2$, r_1, r_2 lie on orthogonal axes of a two-dimensional plane. For $n = 3$, r_1, r_2, r_3 lie on orthogonal axes of a three-dimensional volume, and so on.

Geometrically, the criteria

$$\sum_{j=1}^n r_j > 1$$

thus defines a triangular area in two dimensions, a pyramid-like volume in three dimensions, a hyper-volume in four-dimensions, and so on. The space enclosed by each 'volume' is defined by

$$\sum_{j=1}^n r_j \leq 1.$$

For convenience, let us label $r_1 \rightarrow z$, $r_2 \rightarrow y$, $r_3 \rightarrow x$, $r_4 \rightarrow t$, $r_5 \rightarrow u$.

Examining $n = 2$, the line $z + y = 1$ encloses half of the unit square, formally shown via

$$\begin{aligned} V_2 &= \int_0^1 \int_0^{1-z} dy dz \\ &= \int_0^1 (1-z) dz = \left(z - \frac{z^2}{2} \right) \Big|_0^1 = \frac{1}{2}. \end{aligned}$$

For $n = 3$, the plane $z + y + x = 1$ encloses one sixth of the unit cube:

$$V_3 = \int_0^1 \int_0^{1-z} \int_0^{1-z-y} dx dy dz = \frac{1}{6}$$

Jumping to $n = 4$ is impossible to visualize, however the required integral is easy enough to write and solve:

$$V_4 = \int_0^1 \int_0^{1-z} \int_0^{1-z-y} \int_0^{1-z-y-x} dt dx dy dz = \frac{1}{24}$$

Evidently, the enclosed volume is always the inverse of the factorial of the number of dimensions,

$$V_n = \frac{1}{n!}.$$

Probabilistic Calculation

We ultimately seek the expectation value $\langle n \rangle$, given by

$$\langle n \rangle = \sum_{n=2}^{\infty} n \cdot P(n) ,$$

where $P(n)$ is the probability of satisfying

$$\sum_{j=1}^n r_j < 1 .$$

By the geometric analysis, observe that $P(n)$ corresponds to the ‘window’ of volume bounded between V_n and V_{n-1} :

$$P(n) = \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}$$

With this, we can calculate the expectation value

$$\langle n \rangle = \sum_{n=2}^{\infty} n \cdot \frac{n-1}{n!} = \sum_{n=2}^{\infty} \frac{1}{(n-2)!} = \sum_0^{\infty} \frac{1}{n!} ,$$

which indeed converges to Euler’s constant:

$$e = \sum_0^{\infty} \frac{1}{n!}$$

Amazingly, we conclude:

$$\langle n \rangle = e$$

Chapter 26

Vector Spaces

1 Foundations

1.1 Ket Notation

After a tour through calculus, vectors, and their holy marriage in vector calculus, one is well aware that a vector $\mathbf{a}$ is defined as a list of numbers or variables. We begin by replacing the ‘arrow’ notation with *ket* notation popularized by Paul Dirac:

$$\mathbf{a} = |a\rangle$$

To write out the vector explicitly, list the components inside the $|\rangle$ symbol:

$$\mathbf{a} = \langle a_1, a_2, a_3 \rangle = |a_1, a_2, a_3\rangle$$

Of course, the number of components need not be three, and the coordinate system implied need not be Cartesian.

1.2 Complex Components

All vector components are assumed to be complex numbers unless restricted by circumstance. A complex number z has two components, real and imaginary, such that

$$z = \alpha + i\beta,$$

where α, β are real numbers, and i is the imaginary unit:

$$i = \sqrt{-1}$$

The same complex number z can be expressed in polar form

$$z = r e^{i\phi},$$

where

$$r = |z| = \sqrt{\alpha^2 + \beta^2}$$

is the magnitude, and

$$\phi = \arctan(\beta/\alpha)$$

is the complex phase of z . Every complex number z a complex conjugate $\bar{z} = z^*$ that inverts the imaginary component:

$$\bar{z} = z^* = \alpha - i\beta = r e^{-i\phi}$$

Complex numbers obey special operations for addition, multiplication, and division. For two complex numbers z_j with $j = 1, 2$, we have

$$z_1 \pm z_2 = (\alpha_1 + \alpha_2) \pm i(\beta_1 + \beta_2)$$

$$z_1 \cdot z_2 = (\alpha_1\alpha_2 - \beta_1\beta_2) + i(\alpha_1\beta_2 + \alpha_2\beta_1)$$

$$z_1/z_2 = z_1 \cdot z_2^*/|z_2|^2$$

1.3 Sets

A *set* is generally defined as a collection of distinct, well-defined objects. Perhaps the most common set is the real numbers, denoted $\mathbb{R}$. Distinguishing the set of integers $\mathbb{Z}$ from the irrational numbers $\mathbb{Q}'$, we can relate each set using the *union* operator:

$$\mathbb{R} = \mathbb{Z} \cup \mathbb{Q}'$$

An individual member of a set is called an *element*. For example, the set $\mathbb{C}$ of complex numbers is comprised of all elements z . This is formally denoted using the *in* symbol $\in$ as

$$z \in \mathbb{C}.$$

Mathematical statements can be shortened further by introducing the *for all* symbol $\forall$, along with the *there exists* symbol $\exists$. For instance, the idea that ‘for all complex numbers z there exists a complex conjugate z^* ’ can be written as:

$$\forall z \in \mathbb{C}, \exists z^* \in \mathbb{C}$$

1.4 Spaces

A *space* is a set with some kind of ordered structure. For instance, the space of all ordered pairs of real numbers, i.e., all two-dimensional vectors with real components, is denoted $\mathbb{R}_2$.

For a less trivial example, we may define a space $\mathbb{L}_2[a, b]$ of all functions $\{f(x)\}$ obeying

$$\int_a^b |f(x)|^2 dx < \infty.$$

2 Vector Space

A *vector space*, for a given vector $|A\rangle$, contains the set of all allowed vectors that $|A\rangle$ could have been. Formally, we say that a vector space $\mathcal{V}$ is comprised of complex elements $\{|a\rangle\}$ that obeys the *vector space axioms*.

2.1 Vector Space Axioms

In the axioms that follow, consider any three vectors $|a\rangle$, $|b\rangle$, $|c\rangle$ in the vector space $\mathcal{V}$. Let α, β be two nonzero complex scalars.

Addition

Vectors still obey the familiar rules for addition. Embedded in the definition are the notions of commutativity and associativity:

$$\begin{aligned} |a\rangle + |b\rangle &= |b\rangle + |a\rangle \\ |a\rangle + (|b\rangle + |c\rangle) &= (|a\rangle + |b\rangle) + |c\rangle \end{aligned}$$

Scalar Multiplication

Multiplying a vector by a complex number results in a new vector that is parallel to the original. In particular, this means:

$$\forall |a\rangle \in \mathcal{V}, \forall \alpha \in \mathbb{C} : \exists \alpha |a\rangle \in \mathcal{V}$$

As expected, scalar multiplication follows the rules of commutativity and associativity:

$$\begin{aligned} \alpha (\beta |a\rangle) &= (\alpha\beta) |a\rangle \\ \alpha (|a\rangle + |b\rangle) &= \alpha |a\rangle + \alpha |b\rangle \\ (\alpha + \beta) |a\rangle &= \alpha |a\rangle + \beta |a\rangle \end{aligned}$$

Zero Vector

There exists a *zero vector* in the vector space that does not contribute to any sum. In the language of symbols, this precisely means

$$\exists |0\rangle \in \mathcal{V} : \forall |a\rangle \in \mathcal{V},$$

or in practice, for addition:

$$|a\rangle + |0\rangle = |a\rangle$$

The zero vector plays an expected role in scalar multiplication:

$$0 |a\rangle = |0\rangle$$

Additive Inverse

Every vector has a ‘negative’ version of itself called the *additive inverse*. That is:

$$\begin{aligned} \forall |a\rangle \in \mathcal{V} : \exists |-a\rangle \in \mathcal{V}, \\ |a\rangle + |-a\rangle = |0\rangle \end{aligned}$$

2.2 Uniqueness

Uniqueness of Zero Vector

The first non-axiomatic issue to address is whether there exist multiple zero vectors in a given vector space. To capture this concern, take two vectors $|a\rangle$, $|b\rangle$ and add a unique zero vector to each:

$$\begin{aligned} |a\rangle + |0\rangle &= |a\rangle \\ |b\rangle + |0\rangle' &= |b\rangle \end{aligned}$$

Using the shorthand $|a\rangle + |b\rangle = |c\rangle$, add the two equations to get

$$|c\rangle + |0\rangle + |0\rangle' = |c\rangle.$$

As it appears, the combination $|0\rangle + |0\rangle'$ can only be the zero vector itself:

$$|0\rangle + |0\rangle' = |0\rangle$$

Evidently, $|0\rangle'$ plays an indistinguishable role from $|0\rangle$. In conclusion, we find there exists exactly one (abbreviated $\exists!$) zero vector per vector space:

$$\exists! |0\rangle \in \mathcal{V}$$

Uniqueness of Additive Inverse

In a similar spirit, we can show that the additive inverse of a vector is unique. Take two copies of a vector $|a\rangle$ and add a unique additive inverse vector to each:

$$\begin{aligned} |a\rangle + |-a\rangle &= |0\rangle \\ |a\rangle + |-a\rangle' &= |0\rangle \end{aligned}$$

Adding the two equations, we have

$$2|a\rangle + |-a\rangle + |-a\rangle' = |0\rangle,$$

which is only true if

$$\begin{aligned} |-a\rangle + |-a\rangle' &= 2|-a\rangle \\ |-a\rangle' &= |-a\rangle, \end{aligned}$$

telling us the additive inverse is unique. Our declaration of the additive inverse becomes more specific:

$$\forall |a\rangle \in \mathcal{V} : \exists! |-a\rangle \in \mathcal{V}$$

Subtraction

The notion of subtraction can be formally introduced after establishing uniqueness of the additive inverse:

$$|a\rangle - |b\rangle = |a\rangle + |-b\rangle$$

2.3 Identities**Multiplication by One**

From the above axioms, it's easy to show that the only scalar that multiplies into a vector to return the same vector is one:

$$1 |a\rangle = |a\rangle$$

Multiplication by Zero

Similarly we can ask which scalar multiplies into a vector to return the zero vector:

$$\alpha |a\rangle = |0\rangle$$

The answer is obviously zero, but can we prove it? To do so, take the above statement as a proposition, and then let $\alpha \rightarrow -\alpha$:

$$-\alpha |a\rangle = |0\rangle$$

If both statements are to be true, then it can only be that $\alpha = -\alpha$, which is only satisfied by $\alpha = 0$.

Condensed Notation

For a vector $|a\rangle$ and a scalar α , scalar multiplication is often represented as:

$$\alpha |a\rangle = |\alpha a\rangle$$

Likewise, the addition of two vectors $|a\rangle, |b\rangle$ is written:

$$|a\rangle + |b\rangle = |a + b\rangle$$

2.4 Applications**Cartesian Plane**

The set of vectors based at the origin in the Cartesian plane qualifies as a vector space, provided that the 'usual' rules (from trigonometry) of vector addition and scalar multiplication are allowed.

Real n-tuples

Another vector space is the set of real-valued n -tuples

$$a = (a_1, a_2, \dots, a_n)$$

obeying the addition rule

$$a + b = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n)$$

and the multiplication rule

$$\lambda a = (\lambda a_1, \lambda a_2, \dots, \lambda a_n) .$$

3 Inner Product

3.1 Bra Notation

There is another way to write vectors as they pertain to vector spaces using the so-called *bra notation*:

$$\langle a|$$

The so-called bra-vector is related to $\mathbf{a} = |a\rangle$, but is not identical.

Particularly, the bra-vector is called the *dual vector* of $|a\rangle$, also called a *linear functional*. The vector space occupied by $\langle a|$ is called the *dual space* to that occupied by $|a\rangle$. On their own, bra-vectors obey the same axioms as ket-vectors.

3.2 Inner Product

As an operator, a bra-vector $\langle b|$ can ‘act on’ a ket-vector $|a\rangle$ to produce a scalar:

$$\langle b|a\rangle = \alpha$$

Assuming $|a\rangle, |b\rangle$ are of the same vector space, the quantity $\langle b|a\rangle$ is called the *inner product* of vectors $|b\rangle$ and $|a\rangle$.

Conjugate

As an axiom, let us use up some available freedom to require that swapping $a \leftrightarrow b$ results in the complex conjugate of the original product:

$$\langle b|a\rangle = (\langle a|b\rangle)^* = \overline{\langle a|b\rangle}$$

Negative Product

It’s possible for the inner product to yield a negative result. For two nonzero vectors $|a\rangle, |b\rangle$ satisfying

$$\langle b|a\rangle = \alpha,$$

calculate the inner product $\langle b| - a\rangle$ to find:

$$\begin{aligned}\langle b| - a\rangle &= \langle b|0 - a\rangle \\ &= \cancel{\langle b|0\rangle} - \langle b|a\rangle \\ &= -\alpha\end{aligned}$$

3.3 Linearity

The inner product obeys linearity rules much as an ordinary operator would. For the vectors $|a\rangle, |u\rangle, |v\rangle$, along with scalars α, β , we first have:

$$\langle a|\alpha u + \beta v\rangle = \alpha \langle a|u\rangle + \beta \langle a|v\rangle$$

Furthermore:

$$\begin{aligned}\langle \alpha u + \beta v|a\rangle &= \langle \alpha u|a\rangle + \langle \beta v|a\rangle \\ &= \overline{\langle a|\alpha u\rangle} + \overline{\langle a|\beta v\rangle} \\ &= \alpha^* \overline{\langle a|u\rangle} + \beta^* \overline{\langle a|v\rangle}\end{aligned}$$

Comparing each result, we have:

$$\overline{\langle a|\alpha u + \beta v\rangle} = \langle \alpha u + \beta v|a\rangle$$

3.4 Norm

By calculating $\langle a|a\rangle = \overline{\langle a|a\rangle}$, we readily find $\alpha = \alpha^* = \bar{\alpha}$, telling us the self-inner product always yields a real number:

$$\langle a|a\rangle \in \mathcal{R}$$

The square root the self-inner product is called the *norm* of the vector, axiomatically a positive number:

$$\|a\| = \sqrt{\langle a|a\rangle} > 0$$

Zero Norm

It immediately follows that the norm of any non-zero vector cannot itself be zero, with the only exception being the zero vector:

$$\langle a|a\rangle = 0 \iff |a\rangle = |0\rangle$$

3.5 Applications

Complex Vector Space

Consider the vector space C_n whose elements are vectors containing n individual complex numbers, i.e.

$$\exists |x\rangle \in C_n : |x\rangle = |x_1, x_2, \dots, x_n\rangle$$

For two vectors $|a\rangle$ and $|b\rangle$ in C_n , the inner product can be defined as

$$\langle a|b\rangle = \sum_{j=1}^n a_j^* b_j,$$

or more generally, the definition can include weighting coefficients

$$\langle a|b\rangle = \sum_{j=1}^n a_j^* b_j w_j$$

for $w_j > 0 \in \mathcal{R}$.

Complex Function Space

By analogy to the inner product for vectors, a similar equation can be written for two complex functions $f(z)$, $g(z)$ defined in the interval $z \in [a, b]$ as

$$\langle f|g \rangle = \int_a^b f^*(z) g(z) dz .$$

Of course, the above can be generalized with a weighting function $w(z) > 0 \in \mathcal{R}$ such that

$$\langle f|g \rangle = \int_a^b f^*(z) g(z) w(z) dz .$$

4 Linear Combinations

Consider a vector space $\mathcal{V}$ admitting a set of n vectors $\{|\phi_j\rangle\}$ with $j = 1, 2, \dots, n$. Introducing a set of n complex coefficients $\{c_j\}$, we construct a *linear combination*:

$$|a\rangle = \sum_{j=1}^n c_j |\phi_j\rangle$$

4.1 Span

The linear combination vector $|a\rangle$, along with all other linear combinations of $\{|\phi_j\rangle\}$, occupy a sub-space $\mathcal{V}' \in \mathcal{V}$. In tighter terms, we say the vectors $\{|\phi_j\rangle\}$ *span* the vector space $\mathcal{V}'$.

4.2 Basis

If it turns out that $\mathcal{V}' = \mathcal{V}$, any vector allowed in $\mathcal{V}$ can be expressed as some linear combination of its elements. In this case, vectors $\{|\phi_j\rangle\}$ are called a *basis*, and the number n is a positive non-infinite integer called the *dimension* of the space.

4.3 Linear Independence

While the notion of ‘span’ makes sure there are ‘not too few’ basis vectors, we introduce *linear independence* to assure there aren’t too many. That is, any basis vector $|\phi_k\rangle$ that can be expressed as a linear combination is *not* really a basis vector, and the dimension of the space may shrink by one.

Equivalently, we may argue that a set of linearly independent basis vectors only satisfies

$$\sum_{j=1}^n c_j |\phi_j\rangle = |0\rangle$$

when all coefficients $c_j = 0$. To show this we choose any two nonzero c_k and $c_{k'}$ (with the rest zero), reducing the above to

$$c_k |\phi_k\rangle = -c_{k'} |\phi_{k'}\rangle .$$

Clearly, the vector $|\phi_{k'}\rangle$ is not independent from $|\phi_k\rangle$ and either can be excluded from the basis.

4.4 Uniqueness of Coefficients

We can show that the coefficients c_j are unique for a given linear combination. Supposing we have a resultant vector $|a\rangle$ that that is ‘arrived at’ by two different sets of coefficients

$$|a\rangle = \sum_{j=1}^n c_j |\phi_j\rangle$$

$$|a\rangle = \sum_{j=1}^n c'_j |\phi_j\rangle .$$

Adding each equation and dividing by 2, we quickly find

$$|a\rangle = \sum_{j=1}^n \left(\frac{c_j + c'_j}{2} \right) |\phi_j\rangle ,$$

which only holds if every c_j is equal to c'_j .

5 Orthonormal Basis

5.1 Orthogonal Vectors

Two vectors $|\phi_j\rangle, |\phi_k\rangle$ are *orthogonal vectors* if their inner product is zero:

$$\langle\phi_j|\phi_k\rangle = 0$$

If *all* basis vectors $\{|\phi_j\rangle\}$ are mutually orthogonal, they constitute an *orthogonal basis*.

5.2 Normalized Basis

A basis vector is *normalized* if its self-inner product resolves to one:

$$\langle\phi_j|\phi_j\rangle = 1 ,$$

in which case the change of notation

$$|\phi_j\rangle \rightarrow |e_j\rangle$$

is made. Of course, one can always normalize each vector in an orthogonal basis by dividing out the norm:

$$|e_j\rangle = \frac{1}{\sqrt{\langle\phi_j|\phi_j\rangle}} |\phi_j\rangle$$

Orthonormal Basis

If all basis vectors are mutually orthogonal and have a norm of one, the set $\{|e_j\rangle\}$ is called an *orthonormal basis*. We summarize this by writing

$$\langle e_j|e_k\rangle = \delta_{jk} ,$$

where δ_{jk} is the Kronecker delta symbol:

$$\delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$$

5.3 Vector Components

Equipped with the notion of the orthonormal basis, let us reconsider the linear combination

$$|a\rangle = \sum_{j=1}^n a_j |e_j\rangle ,$$

and solve for the coefficients a_j .

Using what's sometimes called Fourier's trick, notice that projecting any bra-vector $\langle e_k|$ will trigger one inner

product on the left, and n inner products on the right. However $n - 1$ of these will be *zero*, and this plucks out the k th coefficient from the sum:

$$\langle e_k|a\rangle = \sum_{j=1}^n a_j \langle e_k|e_j\rangle = a_j \delta_{kj} = a_k$$

Evidently, any coefficient c_j can be reverse-engineered from a linear combination by the relation

$$a_j = \langle e_j|a\rangle .$$

The coefficients a_j are synonymous with the *components* of a vector. In pure bra-ket notation, a linear combination reads

$$|a\rangle = \sum_{j=1}^n \langle e_j|a\rangle |e_j\rangle ,$$

reminding us that the components of a vector are strictly related to the choice of basis.

5.4 Isomorphism

Consider a vector space $\mathcal{V}$ admitting a basis $\{|e_j\rangle\}$. A linear combination vector $|a\rangle$, in component form, can be written

$$|a\rangle = |a_1, a_2, \dots, a_n\rangle .$$

On the right side, we see that the (complex) components form an n -dimensional space of their own, namely C_n .

To capture the 'one-to-oneness' between the original vector space and that occupied by its components, we say the n -dimensional inner product space is *isomorphic* with C_n , or

$$\mathcal{V}_{(n)} \cong C_n .$$

5.5 Gram-Schmidt Procedure

An arbitrary basis $\{|\phi_j\rangle\}$ can always be transformed into an orthonormal basis by the *Gram-Schmidt procedure*.

Denote $\{|e_j\rangle\}$ as the desired set of unit-normalized orthogonal vectors, and $\{|e'_j\rangle\}$ as a non-normalized version (a notational convenience). Starting with the $j = 1$ vector, we write the easy result

$$|e'_1\rangle = |\phi_1\rangle$$

$$|e_1\rangle = |e'_1\rangle / \sqrt{\langle e'_1|e'_1\rangle} .$$

Next, we need a new vector $|e'_2\rangle$ that involves $|\phi_2\rangle$ and is orthogonal to $|e_1\rangle$. This is achieved by writing

$$\begin{aligned} |e'_2\rangle &= |\phi_2\rangle - \langle e_1|\phi_2\rangle |e_1\rangle \\ |e_2\rangle &= |e'_2\rangle / \sqrt{\langle e'_2|e'_2\rangle} . \end{aligned}$$

Continuing for $j = 3$, we need a vector $|e'_3\rangle$ that involves $|\phi_3\rangle$ and is orthogonal to $|e_1\rangle, |e_2\rangle$, satisfied by

$$|e'_3\rangle = |\phi_3\rangle - \langle e_1|\phi_3\rangle |e_1\rangle - \langle e_2|\phi_3\rangle |e_2\rangle ,$$

subject to the same normalization rule.

In the general $j = n$ case, this pattern extends to

$$\begin{aligned} |e'_n\rangle &= |\phi_n\rangle - \langle e_1|\phi_n\rangle |e_1\rangle \\ &\quad - \langle e_2|\phi_n\rangle |e_2\rangle - \cdots - \langle e_{n-1}|\phi_n\rangle |e_{n-1}\rangle , \end{aligned}$$

normalized by

$$|e_n\rangle = \frac{1}{\sqrt{\langle e'_n|e'_n\rangle}} |e'_n\rangle .$$

Arbitrary Basis

Let us use the Gram-Schmidt procedure to produce an orthonormal basis from:

$$|\phi_1\rangle = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, |\phi_2\rangle = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, |\phi_3\rangle = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Starting with $|e_1\rangle$, we have

$$|e_1\rangle = \frac{1}{\sqrt{\langle \phi_1|\phi_1\rangle}} |\phi_1\rangle = \frac{1}{\sqrt{2}} |\phi_1\rangle = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix} .$$

Next, $|e'_2\rangle$ is given by:

$$\begin{aligned} |e'_2\rangle &= |\phi_2\rangle - \langle e_1|\phi_2\rangle |e_1\rangle = \begin{bmatrix} -1/2 \\ 1/2 \\ 1 \end{bmatrix} \\ |e_2\rangle &= \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \end{aligned}$$

Finally, for $|e'_3\rangle$, we have

$$|e'_3\rangle = |\phi_3\rangle - \langle e_1|\phi_3\rangle |e_1\rangle - \langle e_2|\phi_3\rangle |e_2\rangle = \begin{bmatrix} 2/3 \\ -2/3 \\ 2/3 \end{bmatrix} ,$$

normalizing to

$$|e_3\rangle = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} .$$

Legendre Polynomials

Consider the set of real-valued polynomial functions of order no greater than four

$$P_4(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + c_4x^4 ,$$

known as *Legendre polynomials*. Confining the x -domain to the window $-1 \leq x \leq 1$, we may introduce the inner product of two such polynomials $f(x), g(x)$ as

$$\langle f|g\rangle = \int_{-1}^1 f(x) g(x) dx .$$

Given $P_4(x)$, there are five vectors $|\phi_j\rangle = x^j$ with $j = 0, 1, 2, 3, 4$ form the basis of a five-dimensional vector space.

By the Gram-Schmidt procedure, we can normalize the basis $\{|\phi_j\rangle\}$ starting with

$$|e_0\rangle = \frac{1}{\sqrt{\int_{-1}^1 dx}} |\phi_0\rangle = \frac{1}{\sqrt{2}} |\phi_0\rangle = \frac{1}{\sqrt{2}} .$$

Proceeding for $|e_1\rangle$, we have

$$\begin{aligned} |e'_1\rangle &= |\phi_1\rangle - \langle e_0|\phi_1\rangle |e_0\rangle \\ |e_1\rangle &= \frac{1}{\sqrt{\langle e'_1|e'_1\rangle}} |e'_1\rangle , \end{aligned}$$

reducing to

$$\begin{aligned} |e'_1\rangle &= |\phi_1\rangle - \langle e_0|\phi_1\rangle |e_0\rangle \\ |e_1\rangle &= \sqrt{\frac{3}{2}} |\phi_1\rangle = \sqrt{\frac{3}{2}} x . \end{aligned}$$

Continuing for $|e_2\rangle$, begin with

$$\begin{aligned} |e'_2\rangle &= |\phi_2\rangle - \langle e_0|\phi_2\rangle |e_0\rangle \\ &\quad - \langle e_1|\phi_2\rangle |e_1\rangle = |\phi_2\rangle - \frac{\sqrt{2}}{3} |e_0\rangle , \end{aligned}$$

where normalization requires calculating

$$\langle e'_2|e'_2\rangle = \int_{-1}^1 \left(x^2 - \frac{1}{3}\right)^2 dx = \frac{8}{45} ,$$

landing us at

$$|e_2\rangle = \sqrt{\frac{5}{8}} (3x^2 - 1) .$$

Turning the same crank, it's straightforwardly shown that the remaining normalized basis vectors resolve to three calculations to perform:

$$\begin{aligned} |e_3\rangle &= \sqrt{\frac{7}{8}} (5x^3 - 3x) \\ |e_4\rangle &= \frac{3}{8\sqrt{2}} (35x^4 - 30x^2 + 3) . \end{aligned}$$

With an orthonormal basis on hand, we can expand an arbitrary function, such as $h(x) = x^4$, in terms of the basis as a linear combination:

$$|h\rangle = \sum_{j=0}^4 h_j |e_j\rangle ,$$

where the components h_j are given by

$$h_j = \langle e_j | h \rangle .$$

By symmetry of $h(x) = x^4$, all odd h_j are zero, leaving

$$h_0 = \frac{1}{\sqrt{2}} \int_{-1}^1 x^4 dx = \frac{\sqrt{2}}{5}$$

$$h_2 = \sqrt{\frac{5}{8}} \int_{-1}^1 (3x^6 - x^4) dx = \sqrt{\frac{8}{5}} \frac{2}{7}$$

$$h_4 = \frac{3}{8\sqrt{2}} \int_{-1}^1 (35x^8 - 30x^6 + 3x^4) dx = \frac{1}{35} \frac{8\sqrt{2}}{3}$$

As a reality check, we can readily verify that $|h\rangle$ still corresponds to x^4 , as all other x^n terms cancel out:

$$\begin{aligned} |h\rangle &= \sum_{j=0}^4 h_j |e_j\rangle \\ &= h_0 |e_0\rangle + h_2 |e_2\rangle + h_4 |e_4\rangle = |\phi_4\rangle \end{aligned}$$

6 Normed Vector Space

6.1 Normed Vector Space

The self-inner product of a vector $|a\rangle$, namely

$$\|a\| = \sqrt{\langle a|a\rangle} \in \mathcal{R} \geq 0$$

with

$$\|a\| = 0 \iff |a\rangle = 0$$

is the norm of the vector $|a\rangle$. It turns out that the notion of ‘norm’ extends to vector spaces.

A vector space $\mathcal{V}$ is said to be *normed* if two of its elements $|a\rangle, |b\rangle$, obey the triangle inequality:

$$\|a + b\| \leq \|a\| + \|b\|$$

A normed vector space must also contain the linearity relation

$$\|\alpha a\| = \sqrt{\langle \alpha a|\alpha a\rangle} = |\alpha| \|a\|$$

for a complex scalar α .

6.2 Two-Dimensional Systems

Maximum as Norm

Consider the vector space $\mathcal{R}_2$, i.e. pairs of real numbers (x, y) . Let us show that the ‘maximum’ function

$$\|(x, y)\|_m = \max\{|x|, |y|\}$$

is a norm on $\mathcal{R}_2$.

Taking two vectors

$$|a\rangle = (x_1, y_1)$$

$$|b\rangle = (x_2, y_2) ,$$

the ‘max’ function tells us

$$\|a + b\|_m = \max(|x_1 + x_2|, |y_1 + y_2|) .$$

Note from the triangle inequality that the arguments sent to max $\{\}$ function obey

$$|x_1 + x_2| \leq |x_1| + |x_2|$$

$$|y_1 + y_2| \leq |y_1| + |y_2| .$$

Also observe that $|a\rangle, |b\rangle$ are subject to

$$|x_1| \leq \|a\|_m$$

$$|y_1| \leq \|a\|_m$$

$$|x_2| \leq \|b\|_m$$

$$|y_2| \leq \|b\|_m .$$

Summing the x -equations and the y -equations, we find

$$|x_1| + |x_2| \leq \|a\|_m + \|b\|_m$$

$$|y_1| + |y_2| \leq \|a\|_m + \|b\|_m .$$

Tracing back the inequality symbols, we may finally write

$$\begin{aligned} \|a + b\|_m &= \max(|x_1 + x_2|, |y_1 + y_2|) \\ &\leq \max(|x_1| + |x_2|, |y_1| + |y_2|) \\ &\leq \max(\|a\|_m + \|b\|_m, \|a\|_m + \|b\|_m) \\ &\leq \|a\|_m + \|b\|_m , \end{aligned}$$

satisfying a requirement of a norm.

To complete the job we also establish a linearity relation:

$$\begin{aligned} \|\alpha a\|_m &= \max(|\alpha x_1|, |\alpha y_1|) \\ &= |\alpha| \max(|x_1|, |y_1|) \\ &= |\alpha| \|a\|_m \end{aligned}$$

Sum as Norm

Consider the (same) vector space $\mathcal{R}_2$, i.e. pairs of real numbers (x, y) . Let us show that the ‘sum’ function

$$\|(x, y)\|_s = |x| + |y|$$

is a norm on $\mathcal{R}_2$.

Taking the two vectors

$$|a\rangle = (x_1, y_1)$$

$$|b\rangle = (x_2, y_2) ,$$

the ‘sum’ function gives, using the same identities as above,

$$\begin{aligned} \|a + b\|_s &= |a| + |b| \\ &= |x_1 + y_1| + |x_2 + y_2| \\ &\leq |x_1| + |y_1| + |x_2| + |y_2| \\ &\leq \|a\|_m + \|b\|_m . \end{aligned}$$

To check for linearity, we write

$$\begin{aligned} \|\alpha a\|_s &= |\alpha x_1| + |\alpha y_1| \\ &= |\alpha| (|x_1| + |y_1|) \\ &= |\alpha| \|a\|_s . \end{aligned}$$

Unit Ball

In any vector space $\mathcal{V}$, the unit ‘ball’ $\mathcal{B}_1$ is defined as

$$\mathcal{B} = \{|a\rangle \in \mathcal{V} : \|a\| \leq 1\}.$$

Plotting the ‘max’ function in the xy -plane, the unit ball resolves to a square frame of side 1, as $\max(a) = 1$ in the unit ball. In terms of the ‘sum’ function, the ball resolves to a filled diamond with points at $(0, \pm 1)$ and $(\pm 1, 0)$, generated by $|x| + |y| \leq 1$.

6.3 Identities

Cauchy-Bunyakovsky-Schwarz Inequality

There is an important fact called the *Cauchy-Bunyakovsky-Schwarz Inequality* that must be established.

For two vectors $|a\rangle, |b\rangle$, consider the nonzero sum $|a\rangle - \lambda |b\rangle$, or for short, $|a - \lambda b\rangle \neq |0\rangle$. Now expand the norm of this vector:

$$\begin{aligned} \|a - \lambda b\|^2 &= \langle a - \lambda b | a - \lambda b \rangle \\ &= \|a\|^2 - \lambda \langle a | b \rangle - \lambda^* \langle b | a \rangle + \lambda^2 \|b\|^2 \end{aligned}$$

Next choose

$$\lambda = \frac{\langle b | a \rangle}{\|b\|^2},$$

and the above becomes

$$\begin{aligned} \|a - \lambda b\|^2 &= \|a\|^2 - \frac{\langle b | a \rangle}{\|b\|^2} \langle a | b \rangle \\ &\quad - \frac{\langle a | b \rangle}{\|b\|^2} \langle b | a \rangle + \frac{\langle a | b \rangle \langle b | a \rangle}{\|b\|^2 \|b\|^2} \|b\|^2. \end{aligned}$$

The last two terms cancel exactly, and the norm on the left is defined to be positive and real. What’s left is

$$\|a\|^2 - \frac{\langle a | b \rangle \langle b | a \rangle}{\|b\|^2} > 0,$$

which simplifies to a very powerful result:

$$|\langle a | b \rangle| \leq \|a\| \|b\|$$

Triangle Inequality

For two vectors $|a\rangle, |b\rangle$, consider the nonzero sum $|a + \lambda b\rangle \neq |0\rangle$. Expand the norm of this vector,

$$\begin{aligned} \|a + \lambda b\|^2 &= \langle a + \lambda b | a + \lambda b \rangle \\ &= \|a\|^2 + \lambda \langle a | b \rangle + \lambda^* \langle b | a \rangle + \lambda^2 \|b\|^2 \\ &= \|a\|^2 + 2\text{Re}(\lambda \langle a | b \rangle) + \lambda^2 \|b\|^2, \end{aligned}$$

and let $\lambda = 1$:

$$\|a + b\|^2 = \|a\|^2 + 2\text{Re}(\langle a | b \rangle) + \|b\|^2$$

By the Cauchy-Bunyakovsky-Schwarz inequality, the middle term is less than $\|a\| \|b\|$:

$$\begin{aligned} \|a + b\|^2 &\leq \|a\|^2 + 2\|a\| \|b\| + \|b\|^2 \\ &\leq (\|a\| + \|b\|)^2 \end{aligned}$$

Take a final square root of both sides and the proof is done.

Pythagorean Theorem

In the special case $\langle a | b \rangle = 0$, the triangle inequality reduces to the Pythagorean theorem:

$$\|a + b\|^2 = \|a\|^2 + \|b\|^2$$

Parallelogram Law

For two vectors $|a\rangle, |b\rangle$ and the linear combinations $|a + b\rangle, |a - b\rangle$, there exists a *parallelogram law* concerning vector addition:

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

To prove this, write out the norm of $|a + b\rangle$ and $|a - b\rangle$

$$\begin{aligned} \|a + b\|^2 &= \|a\|^2 + \|b\|^2 + \langle a | b \rangle + \langle b | a \rangle \\ \|a - b\|^2 &= \|a\|^2 + \|b\|^2 - \langle a | b \rangle - \langle b | a \rangle, \end{aligned}$$

and add the resulting equations to finish the proof.

A complementary result comes from subtracting the equations:

$$\|a + b\|^2 - \|a - b\|^2 = 4 \text{Re}(\langle a | b \rangle)$$

One may imagine what combination on the left yields $\text{Im}(\langle a | b \rangle)$ on the right. It’s straightforward to show that the job is done by:

$$i\|a + ib\|^2 - i\|a - ib\|^2 = 4 \text{Im}(\langle a | b \rangle)$$

Polar Identity

The sum of $4 \operatorname{Re} (\langle a|b \rangle)$ and $4 \operatorname{Im} (\langle a|b \rangle)$, each given above, leads to the *polar identity*. First note that

$$4 \operatorname{Re} (\langle a|b \rangle) + 4 \operatorname{Im} (\langle a|b \rangle) = 4 (\langle a|b \rangle) ,$$

which yields the identity

$$\begin{aligned} 4 (\langle a|b \rangle) &= \|a + b\|^2 - \|a - b\|^2 \\ &\quad + i\|a + ib\|^2 - i\|a - ib\|^2 \end{aligned}$$

Problem 1

Use the Cauchy-Bunyakovsky-Schwarz inequality to show that

$$1 < \int_0^{\pi/2} \sqrt{\sin(x)} \, dx < \sqrt{\frac{\pi}{2}} .$$

Hint: For the left, establish

$$\sin(x) \leq \sqrt{\sin(x)} \leq \sqrt{x}$$

and integrate, remembering

$$\int_0^{\pi/2} \sin(x) \, dx = 1 .$$

For the right, associate $a = \sqrt{\sin(x)}$ and $b = 1$ so that $\langle a|a \rangle = 1$ and $\langle b|b \rangle = \pi/2$.

7 Countably Finite System

7.1 Convergent Sequence

Suppose $\mathcal{V}$ is a normed vector space. The sequence of vectors

$$\{|a_j\rangle \in \mathcal{V} : j = (1, 2, 3, \dots)\}$$

is said to be *convergent* to the vector $|a\rangle$ if

$$\forall k_\epsilon : \exists \epsilon > 0$$

such that if $k > k_\epsilon$, then

$$\|a - a_k\| < \epsilon.$$

This statement inspires a definition of a convergent vector:

$$|a\rangle = \lim_{k \rightarrow \infty} |a_k\rangle$$

7.2 Cauchy Sequence

The (same) sequence of vectors $\{|a_j\rangle \in \mathcal{V}\}$ qualifies as a *Cauchy sequence* if

$$\forall k_\epsilon : \exists \epsilon > 0,$$

then

$$\|a_m - a_n\| < \epsilon$$

provided that $m, n > k_\epsilon$.

It's easy to show that any convergent sequence qualifies as a Cauchy sequence. For two vectors $|a_m\rangle, |a_n\rangle$, we know

$$\|a - a_m\| < \alpha\epsilon$$

$$\|a - a_n\| < \beta\epsilon$$

for two parameters $\alpha, \beta > 0 \in \mathcal{R}$. Adding each, we have

$$\|a - a_m\| + \|a_n - a\| < (\alpha + \beta)\epsilon,$$

which becomes, by the triangle inequality,

$$\|a - a_m + a_n - a\| < \epsilon,$$

reducing to the fingerprint of a Cauchy sequence:

$$\|a_m - a_n\| < \epsilon$$

7.3 Complete Space

A normed vector spaces in which all Cauchy sequences converge is called a *complete space*, also known as a Banach space. In terms of an orthonormal basis, an arbitrary vector is given by

$$|a\rangle = \sum_{j=1}^n a_j |e_j\rangle,$$

where clearly

$$\|a\| = \sqrt{|a|_1^2 + |a|_2^2 + \dots + |a|_n^2}.$$

We may further consider a vector $|b\rangle$ that is itself a Cauchy sequence of vectors $|a^{(k \leq m)}\rangle$ such that

$$|b\rangle = \sum_{k=1}^m |a^{(k)}\rangle = \sum_{j=1}^n \sum_{k=1}^m a_j^{(k)} |e_j\rangle = \sum_{j=1}^n b_j |e_j\rangle,$$

where each coefficient b_j is itself a Cauchy sequence of the complex coefficients a_j :

$$b_j = \sum_{k=1}^m a_j^{(k)}$$

7.4 Supremum

Consider the complete infinite-dimensional space C_{ab} of complex-valued functions $f(x) : x \in [a, b]$. Here we define the 'supremum' function

$$\|f\|_{sup} = \max \{|f(x)| : x \in [a, b]\}.$$

The 'sup' norm guarantees homogeneous convergence of a Cauchy sequence of functions $f^{(k)}(x)$ to a single function $f(x)$.

Unlike other norms we've encountered, the $\|f\|_{sup}$ does not bear a notion of inner product. Choosing a trivial example $f(x) = \cos x$, $g(x) = x$ in the interval $x \in [a, b]$, we have

$$\|f\|_{sup} = 1$$

$$\|g\|_{sup} = \pi$$

$$\|f + g\|_{sup} = 1 + \pi$$

$$\|f - g\|_{sup} = -1 - \pi,$$

which violates the parallelogram law:

$$\begin{aligned} \|f + g\|_{sup}^2 + \|f - g\|_{sup}^2 &\neq 2 \left(\|f\|_{sup}^2 + \|g\|_{sup}^2 \right) \\ (1 + \pi)^2 + (-1 - \pi)^2 &\neq 2 \left(1 + \pi^2 \right) \\ 4\pi &\neq 0 \end{aligned}$$

7.5 Hilbert Space

A complete inner product space is called a *Hilbert space*, and we have shown that all finite-dimensional vector spaces are Hilbert spaces. The 'supremum' norm is a unique example of a complete space that is *not* a Hilbert space.

8 Countably Infinite System

The results of the previous section readily generalize to handle a *countably infinite* basis.

8.1 Fourier Series

Consider an orthonormal basis $\{|e_j\rangle\}$ containing an *infinite* number of basis vectors. An infinite linear combination

$$|x\rangle = \sum_{j=1}^{\infty} \langle e_j|x\rangle |e_j\rangle$$

is the *Fourier series* of the vector $|x\rangle$ in the basis $\{|e_j\rangle\}$, where $\langle x|e_j\rangle$ are *Fourier coefficients*.

8.2 Bessel Inequality

Let us show that any partial sum of a Fourier series is a Cauchy sequence.

Truncating the series at the n th term gives

$$|x^{(n)}\rangle = \sum_{j=1}^n \langle e_j|x\rangle |e_j\rangle$$

as a partial sum. Another truncation of the series with $m > n$ can be written

$$|x^{(m)}\rangle = |x^{(n)}\rangle + \sum_{j=n+1}^m \langle e_j|x\rangle |e_j\rangle ,$$

where the norm of the difference of the two vectors reads

$$\|x^{(m)} - x^{(n)}\|^2 = \sum_{j=n+1}^m |\langle e_j|x\rangle|^2 .$$

The above series is assured to be a positive real number, reducing the problem to showing that

$$\sum_{j=1}^{\infty} |\langle e_j|x\rangle|^2$$

does not diverge.

Proceed by writing out the $m \rightarrow \infty$ case, giving

$$\langle x - x^{(n)} | x - x^{(n)} \rangle = \|x\|^2 - \sum_{j=1}^n |\langle e_j|x\rangle|^2 \geq 0 .$$

Reading the equation from the right, we arrive at the *Bessel inequality*

$$\sum_{j=1}^n |\langle e_j|x\rangle|^2 \leq \|x\|^2 ,$$

proving that a partial sum of a Fourier series is a Cauchy sequence. Reading the above from left to right, we also have

$$\|x - x^{(n)}\| = \sqrt{\|x\|^2 - \sum_{j=1}^n |\langle e_j|x\rangle|^2} ,$$

which in the case of convergence ($n \rightarrow \infty$), we get the *Parseval relation*:

$$\|x\|^2 = \sum_{j=1}^{\infty} |\langle e_j|x\rangle|^2$$

8.3 Inner Product Space of Functions

The space $C_2[a, b]$ of the inner product of two complex functions was written as

$$\langle f|g\rangle = \int_a^b f^*(z) g(z) dz .$$

For certain cases of the function $f(z)$, for instance a discontinuous function, the space $C_2[a, b]$ is easily shown to not be complete with respect to the usual notion of norm,

$$\|f\| = \sqrt{\int_a^b |f|^2 dz} ,$$

implying that a Hilbert space of functions must be carefully discerned.

According to the Riesz-Fisher theorem, we may define a Hilbert of functions with a countably infinite basis, denoted $\mathcal{L}_2[a, b]$. Furthermore, the Stone-Weierstrass theorem states that the set of polynomials $\{|x_j\rangle\}$ with $j = 1, 2, \dots$ forms a basis in $\mathcal{L}_2[a, b]$. Of course, we found such vectors to be non-orthogonal, corrected by the Gram-Schmidt process to produce the Legendre polynomials.

9 Operators

9.1 Definition

An *operator* L is a function that ‘acts on’ a vector $|x\rangle \in \mathcal{V}$ to create a new vector

$$L|x\rangle = |Lx\rangle = |y\rangle$$

that may or may not live in $\mathcal{V}$.

9.2 Linear Operator

An operator that maps a vector to its own vector space $\mathcal{V}$ is said to be *linear* if the relation

$$L|\alpha u + \beta v\rangle = \alpha L|u\rangle + \beta L|v\rangle$$

is satisfied, where α, β are complex scalars. Needless to mention, scalar multiplication is a special case of a linear operator.

Linearity Check

Interpreting vectors as functions, we can check whether certain operations for a function $f(x)$ qualify as linear operators. For example, the transformation

$$L(f(x)) = \sin(f(x))$$

fails when tested for linearity:

$$\begin{aligned} L(\alpha f(x)) &= \sin(\alpha f(x)) = \sin \alpha \cos(f(x)) \\ &\quad + \cos \alpha \sin(f(x)) \\ &\neq \alpha L(f(x)) \end{aligned}$$

The less trivial example

$$L(f(x)) = \int_0^1 \sin(xy) f(y) dy$$

does qualify as a linear operator, as the function f enters the integral linearly. Explicitly, we have

$$\begin{aligned} L(\alpha f(x) + \beta g(x)) &= \int_0^1 \sin(xy) (\alpha f(y) + \beta g(y)) dy \\ &= \alpha \int_0^1 \sin(xy) f(y) dy + \beta \int_0^1 \sin(xy) g(y) dy \\ &= \alpha L(f(x)) + \beta L(g(x)) . \end{aligned}$$

9.3 Adjoint Operator

Given an operator L , the *adjoint* operator $L^\dagger$, is defined such that

$$\langle b|L|a\rangle = \overline{\langle a|L^\dagger|b\rangle}$$

readily implying:

$$\begin{aligned} \langle b|L|a\rangle &= \langle L^\dagger b|a\rangle \\ (L^\dagger)^\dagger &= L \end{aligned}$$

Two linear operators A and B always obey the relation

$$(AB)^\dagger = B^\dagger A^\dagger ,$$

proven by writing $\langle u|AB|v\rangle$ two different ways and comparing each right-hand result:

$$\begin{aligned} \langle u|AB|v\rangle &= \langle (AB)^\dagger u|v\rangle \\ \langle u|AB|v\rangle &= \langle A^\dagger u|Bv\rangle = \langle B^\dagger A^\dagger u|v\rangle \end{aligned}$$

9.4 Hermitian Operator

An operator that is its own adjoint operator is called *self-adjoint*, also known as *Hermitian*:

$$L = L^\dagger \rightarrow \langle b|La\rangle = \langle Lb|a\rangle$$

If L is Hermitian, the operator $L^\dagger$ is called the *Hermitian conjugate* to L .

For self-adjoint operators, the quantity

$$\lambda = \langle b|L|a\rangle$$

is always real-valued.

We may also inquire whether the product AB of two Hermitian operators is itself Hermitian. Starting with $(AB)^\dagger = B^\dagger A^\dagger$, let $A = A^\dagger$ and $B = B^\dagger$ to find

$$AB = (B^\dagger A^\dagger)^\dagger = (BA)^\dagger ,$$

telling us AB is Hermitian only if $AB = BA$.

9.5 Anti-Hermitian Operator

An *Anti-Hermitian* operator is one that obeys

$$L = -L^\dagger .$$

For two Hermitian operators A, B , it turns out that $AB - BA$ is anti-Hermitian:

$$\begin{aligned} (AB - BA)^\dagger &= (AB)^\dagger - (BA)^\dagger \\ &= B^\dagger A^\dagger - A^\dagger B^\dagger = BA - AB \end{aligned}$$

Partial Derivative Operator

The partial derivative operator

$$A = \frac{\partial}{\partial x}$$

is an anti-Hermitian operator for certain boundary conditions. Consider two arbitrary function $f(x)$, $g(x)$ in the domain Ω where each function is zero on the boundary $\partial\Omega$.

Then, writing out $\langle f | A | g \rangle$ two different ways gives

$$\begin{aligned} \langle f | A | g \rangle &= \langle A^\dagger f | g \rangle \\ \int_{\Omega} f^* \partial_x g \, dx &= \int_{\Omega} (\partial_x)^\dagger f^* g \, dx . \end{aligned}$$

Integrating the left side by parts, we write

$$f^* g \Big|_{\partial\Omega} - \int_{\Omega} \partial_x f^* g \, dx = \int_{\Omega} (\partial_x)^\dagger f^* g \, dx ,$$

where the boundary term equals zero by construction, indicating A to be anti-Hermitian.

9.6 Projector

For any fixed vector $|a\rangle$, the combination

$$P_a = |a\rangle \langle a|$$

is called the *projector* of $|a\rangle$. The projector does nothing on its own, but waits for a bra- or ket-vector to interact with the left or the right side, respectively. Acting on a vector $|x\rangle$, we have

$$P_a |x\rangle = |a\rangle \langle a|x\rangle = \langle a|x\rangle |a\rangle ,$$

which is $|a\rangle$ multiplied by a scalar.

Properties

The projector is a linear operator, easily verified by

$$\begin{aligned} P_a |\alpha u + \beta v\rangle &= |a\rangle \langle a| (|\alpha u\rangle + |\beta v\rangle) \\ &= |a\rangle (\alpha \langle a|u\rangle + \beta \langle a|v\rangle) \\ &= \alpha \langle a|u\rangle |a\rangle + \beta \langle a|v\rangle |a\rangle \\ &= \alpha P_a |u\rangle + \beta P_a |v\rangle , \end{aligned}$$

and is also Hermitian:

$$\begin{aligned} \langle u | P_a | v \rangle &= \langle u | (\langle a | v \rangle | a \rangle) \\ &= \langle a | v \rangle \langle u | a \rangle \\ &= (\langle u | a \rangle \langle a |) | v \rangle \\ &= \langle P_a u | v \rangle \end{aligned}$$

9.7 Identity Operator

Consider any vector $|x\rangle$ in the vector space $\mathcal{V}$ spanned by the basis vectors $\{|e_k\rangle\}$. Choosing any basis vector $|e_j\rangle$, apply a projector

$$P_{e_j} = |e_j\rangle \langle e_j|$$

to $|x\rangle$ to get

$$P_{e_j} |x\rangle = \langle e_j | x \rangle |e_j\rangle = x_j |e_j\rangle .$$

Summing over the index j , we find

$$\left(\sum_j P_{e_j} \right) |x\rangle = \sum_j x_j |e_j\rangle = |x\rangle .$$

The parenthesized quantity that leaves the vector unchanged is called the *identity operator* I . That is,

$$I |x\rangle = |x\rangle ,$$

where

$$I = \sum_j |e_j\rangle \langle e_j|$$

is also called the *completeness relation* for the basis. (It is possible for a basis to be *incomplete*, in which case the above sum is not equivalent to an identity operator.)

9.8 Commutator

For any two operators A and B , the difference

$$[AB] = AB - BA$$

defines their *commutator*, also known as the *commutation relation*. The result of $[AB]$ tells us which ‘extra terms’ emerge when swapping two operators. When the commutator evaluates to zero, the operators are said to *commute*.

For example, consider two operators

$$\begin{aligned} A &= \frac{\partial}{\partial x} \\ B &= x \end{aligned}$$

that can act on a function $f(x)$. Allowing the commutator to act on $f(x)$, we write

$$\begin{aligned} (AB - BA) |f\rangle &= \partial_x (x |f\rangle) - x (\partial_x |f\rangle) \\ &= |f\rangle + x \partial_x |f\rangle - x \partial_x |f\rangle \\ &= |f\rangle , \end{aligned}$$

telling us that that operators on hand do not commute, but instead obey

$$AB - BA = I .$$

10 Eigen-Calculations

If an operator L applied to a vector $|x\rangle$ results in a parallel vector $\lambda |x\rangle$, then $|x\rangle$ is called an *eigenvector* of L , and λ is the corresponding *eigenvalue*:

$$L |x\rangle = \lambda |x\rangle$$

10.1 Real Eigenvalues

It's straightforward to show that eigenvalues are always real if L is self-adjoint (Hermitian):

$$\langle x | L | x \rangle = \lambda \langle x | x \rangle \rightarrow \lambda = \frac{\langle x | L | x \rangle}{\langle x | x \rangle}$$

To establish this, write the eigenvalue problem

$$L |x\rangle = \lambda |x\rangle ,$$

and project an arbitrary vector $\langle y |$ to write

$$\langle y | L | x \rangle = \lambda \langle y | x \rangle ,$$

equivalent to

$$\langle L^\dagger y | x \rangle = \lambda \langle y | x \rangle .$$

Take the complex conjugate of each side to get

$$\overline{\langle L^\dagger y | x \rangle} = \lambda^* \overline{\langle y | x \rangle} ,$$

or

$$\langle x | L^\dagger y \rangle = \langle x | L^\dagger | y \rangle = \lambda^* \langle x | y \rangle$$

Finally, let us set $|y\rangle = |x\rangle$, and without loss of generality assume $|x\rangle$ is normalized, so we may take $\langle y | x \rangle = \langle x | y \rangle = 1$. After simplifying, we are left with

$$\begin{aligned} \langle x | L | x \rangle &= \lambda \\ \langle x | L^\dagger | x \rangle &= \lambda^* \end{aligned}$$

If L is self-adjoint, we automatically have $L = L^\dagger$. This can only mean $\lambda = \lambda^*$, thus all λ are real.

10.2 Orthogonal Eigenvectors

For a linear self-adjoint operator L , we can show that two distinct eigenvalues λ_1, λ_2 correspond to two orthogonal eigenvectors.

Start with

$$\begin{aligned} \langle x_2 | L | x_1 \rangle &= \lambda_1 \langle x_2 | x_1 \rangle \\ \langle x_1 | L | x_2 \rangle &= \lambda_2 \langle x_1 | x_2 \rangle , \end{aligned}$$

and complex-conjugate the second equation to eliminate the $\langle x_2 | x_1 \rangle$ -term:

$$\langle x_2 | L | x_1 \rangle = \left(\frac{\lambda_1}{\lambda_2} \right) \overline{\langle x_1 | L | x_2 \rangle} = \left(\frac{\lambda_1}{\lambda_2} \right) \langle x_2 | L | x_1 \rangle$$

For $\lambda_1 \neq \lambda_2$, the only reasonable conclusion is

$$\langle x_2 | L | x_1 \rangle = 0 \rightarrow \langle x_1 | x_2 \rangle = 0 ,$$

meaning $|x_1\rangle, |x_2\rangle$ must be orthogonal.

10.3 Equal Eigenvalues

If n eigenvalues are equal, one speaks of *n-fold degeneracy*. In this case, the corresponding eigenvectors are not necessarily orthogonal, in which case the vectors form a vector subspace of the original vector space that might admit its own orthonormal basis.

10.4 Calculating Eigenvectors

When the form of L is given, it's usually possible to solve for all eigenvalues λ_j . With n of these established, the next move is to calculate the eigenvectors directly using

$$\begin{aligned} L |x^{(j)}\rangle &= \lambda_j |x^{(j)}\rangle \\ j &= 1, 2, \dots, n \end{aligned}$$

Whether or not the eigenvectors form an orthonormal basis, we may express an arbitrary vector $|u\rangle$ as a linear combination:

$$|u\rangle = \sum_j C_j |x^{(j)}\rangle$$

Recall that the eigenvectors $|x^{(j)}\rangle$ are only orthogonal if L is self-adjoint, i.e. Hermitian.

10.5 Time Derivative Operators

Consider a vector space $\mathcal{V}$ of dimension n admitting a fixed orthonormal basis $\{|e_j\rangle\}$ where $j = 1, \dots, n$. A time-varying vector $|u(t)\rangle$ is a linear combination of time-varying coefficients such that

$$|u(t)\rangle = \sum_j u_j(t) |e_j\rangle .$$

Single Time Derivative Operator

Now we introduce the single time-derivative operator

$$L = \frac{\partial}{\partial t} = \partial_t .$$

If we let L act on an eigenvector $|x^{(j)}(t)\rangle$, the result is equal to $|x(t)\rangle$ multiplied by its corresponding eigenvalue λ . That is,

$$L |x(t)\rangle = \lambda |x(t)\rangle ,$$

or

$$\sum_j \partial_t x_j(t) |e_j\rangle = \sum_j \lambda x_j(t) |e_j\rangle ,$$

implying n copies of the same separable differential equation

$$\partial_t x_j(t) = \lambda x_j(t)$$

for each index j .

Elementary methods give the solution for each equation

$$x_j(t) = x_j(t=0) e^{\lambda t} = x_{0j} e^{\lambda t} ,$$

telling us

$$\begin{aligned} |x(t)\rangle &= \sum_j x_{0j} e^{\lambda t} |e_j\rangle \\ &= e^{\lambda t} \sum_j x_{0j} |e_j\rangle = e^{\lambda t} |x_0\rangle . \end{aligned}$$

Perhaps not surprisingly, the time dependence evolves exponentially in time.

As a matter of technicality, a vector $|x(t)\rangle$ is best described as an *eigenfunction*, as the operator $L = \partial_t$ is trivial for time-independent vectors.

Double Time Derivative Operator

We also consider the double time-derivative operator $L = \partial_{tt}$. Using the same setup, it follows that each x_j is governed by the differential equation

$$\partial_{tt} x_j(t) = \lambda x_j(t) ,$$

whose solution is governed by λ .

For $\lambda = 0$, the coefficients evolve linearly in time:

$$\lambda = 0$$

$$x_j(t) = x_{0j} + x_{1j} t$$

For $\lambda \neq 0$, we have a linear combination of exponential terms:

$$\lambda \neq 0$$

$$x_j(t) = x_{0j} e^{\sqrt{\lambda} t} + x_{1j} e^{-\sqrt{\lambda} t}$$

11 Operator as Matrix

Consider a vector $|x\rangle$ living in vector space $\mathcal{V}$ that admits an orthonormal basis $\{|e_j\rangle\}$. As a linear combination of coefficients $\{x_j\}$, such a vector is written

$$|x\rangle = \sum_j x_j |e_j\rangle .$$

Suppose another vector $|y\rangle$, which is itself a linear combination of coefficients $\{y_j\}$ in the same basis, arises by applying a linear operator A onto $|x\rangle$:

$$|y\rangle = A |x\rangle = \sum_j y_j |e_j\rangle$$

The question now is, what can we discern about the operator A ?

11.1 Matrix Elements

We proceed by ‘solving for’ any component y_j , which entails taking the inner product with a basis vector $\langle e_{k \neq j}|$ to get

$$\langle e_k | A | x \rangle = \sum_j \langle e_k | y_j | e_j \rangle = y_j \delta_{jk} = y_k ,$$

implying

$$\begin{aligned} y_j &= \langle e_j | A | x \rangle = \langle e_j | A \sum_k x_k | e_k \rangle \\ &= \sum_k \langle e_j | A | e_k \rangle x_k . \end{aligned}$$

That is, any y_j depends on each member of $\{x_j\}$ multiplied by a number

$$A_{jk} = \langle e_j | A | e_k \rangle$$

called a *matrix element*.

The set of matrix elements $\{A_{jk}\}$ is the *matrix* represented by the operator A . To restore the operator A in terms of its elements, begin with the identity $A = IAI$ and write out each identity operator explicitly to get

$$\begin{aligned} A &= IAI = \sum_j \sum_k |e_j\rangle \langle e_j| A | e_k\rangle \langle e_k| \\ &= \sum_j \sum_k |e_j\rangle A_{jk} \langle e_k| . \end{aligned}$$

Here we emphasize that the choice of basis vectors has direct impact on the components A_{jk} .

11.2 Matrix Algebra

Matrix Addition

Two operators A and B readily add to form a new operator C such that

$$\begin{aligned} A |x\rangle + B |x\rangle &= C |x\rangle \\ A_{jk} + B_{jk} &= C_{jk} , \end{aligned}$$

which of course requires A, B to be of equal dimension.

Scalar Multiplication

A scalar λ can be ‘multiplied into’ an operator A by scaling each component to create another operator B :

$$\begin{aligned} B &= \lambda A \\ B_{jk} &= \lambda A_{jk} \end{aligned}$$

Matrix Multiplication

Two operators A and B can ‘multiply’ to form a new operator C such that

$$\begin{aligned} A (B |x\rangle) &= C |x\rangle \\ AB &= C , \end{aligned}$$

which is generally an associative operation, but not commutative:

$$\begin{aligned} (AB) C &= A (BC) \\ AB &\neq BA \end{aligned}$$

The formula for matrix multiplication is calculated by brute force. $C = AB$ expands to

$$\begin{aligned} C &= \sum_j \sum_k \sum_{j'} \sum_m |e_j\rangle A_{jk} \langle e_k | e_{j'}\rangle B_{j'm} \langle e_m| \\ C &= \sum_j \sum_m |e_j\rangle \left(\sum_k A_{jk} B_{km} \right) \langle e_m| , \end{aligned}$$

telling us

$$C_{jm} = \sum_k A_{jk} B_{km} .$$

11.3 Matrix Transpose

For a given operator A with components A_{jk} , the *transpose* of A , denoted A^T , has components A_{kj} . That is, the transpose swaps rows $\leftrightarrow$ columns. Using the matrix multiplication rule, it's straightforward to show that the transpose

of the product of two matrices equals the product of the individually transposed matrices in reversed order:

$$(AB)^T = B^T A^T$$

It's also straightforward to show the following determinant identity:

$$\det(A^\dagger) = \det((A^*)^T) = (\det(A))^*$$

11.4 Matrix Trace

For operators A represented by a square ($n \times n$) matrix, a special quantity exists called the *trace* of the matrix. The trace is defined as the sum of the components along the diagonal:

$$\text{tr} A = A_{11} + A_{22} + \cdots + A_{nn} = \sum_{k=1}^n A_{kk}$$

12 Hermitian Matrix

Recall that an operator A that is its own adjoint operator, meaning $A = A^\dagger$ as appearing in the definition

$$\langle y | A | x \rangle = \overline{\langle x | A^\dagger | y \rangle},$$

where

$$|x\rangle, |y\rangle \in \mathcal{V},$$

is a Hermitian operator, synonymous with *Hermitian matrix*, where $A^\dagger$ is the *Hermitian conjugate*.

In component form, we note that

$$\langle y | Ax \rangle = \sum_{ij} A_{ij} y_i^* x_j,$$

where meanwhile, an equivalent statement is

$$\overline{\langle x | A^\dagger | y \rangle} = \sum_{ij} (A_{ji}^\dagger)^* x_j y_i^*,$$

telling us that the components of a Hermitian operator obey

$$A_{ij}^* = A_{ji}^\dagger.$$

Anti-Hermitian Matrix

The properties of anti-Hermitian operators also extend to matrices. An anti-Hermitian matrix satisfies

$$\begin{aligned} A^\dagger &= -A \\ A_{ij}^* &= -A_{ji}. \end{aligned}$$

As a corollary, we note that if a matrix A is Hermitian, then iA is anti-Hermitian, and vice-versa.

Symmetric Matrix

When the components of a Hermitian matrix are all real-valued, the matrix is symmetric. Work this from the identity

$$A_{ij}^* = A_{ji}^\dagger,$$

and notice that the complex conjugate A^* is just A again:

$$A_{ij} = A_{ji}^\dagger$$

Since $A = A^\dagger$ by definition, we further have

$$A_{ij} = A_{ji},$$

telling us A is symmetric.

As a special case of the Hermitian operator, we automatically have that the eigenvectors corresponding to non-equal eigenvalues of a symmetric matrix are orthogonal.

12.1 Commuting Operators

Now we derive the important fact that commuting Hermitian operators share an orthonormal basis. To begin, consider an operator A admitting an eigenvector $|\psi_n\rangle$ with corresponding eigenvalue a_n , and also a second operator B admitting an eigenvector $|\phi_n\rangle$ with corresponding eigenvalue b_n :

$$A |\psi_n\rangle = a_n |\psi_n\rangle$$

$$B |\phi_n\rangle = b_n |\phi_n\rangle$$

Our key assumption is that any $|\psi_n\rangle$ can be written as a linear combination of $\{|\phi_n\rangle\}$, and vice-versa, which assumes each vector is a member of the same basis:

$$|\psi_n\rangle = \sum_m \gamma_{mn} |\phi_m\rangle$$

$$|\phi_n\rangle = \sum_m \tilde{\gamma}_{mn} |\psi_m\rangle.$$

In the above, $\gamma, \tilde{\gamma}$ are matrix coefficients. We gain a restriction on $\gamma, \tilde{\gamma}$ by the substitution

$$\begin{aligned} |\psi_n\rangle &= \sum_m \gamma_{mn} \sum_{m'} \tilde{\gamma}_{m'm} |\phi_{m'}\rangle \\ &= \sum_{m'} \left(\sum_m \gamma_{mn} \tilde{\gamma}_{m'm} \right) |\phi_{m'}\rangle, \end{aligned}$$

implying the delta relation

$$\delta_{m'n} = \sum_m \gamma_{mn} \tilde{\gamma}_{m'm}.$$

To gain two more delta relations, compute $A |\psi_n\rangle$ two different ways to write

$$\begin{aligned} A |\psi_n\rangle &= A \sum_m \gamma_{mn} |\phi_m\rangle \\ &= A \sum_m \sum_{m'} \gamma_{mn} \tilde{\gamma}_{m'm} |\psi_{m'}\rangle \end{aligned}$$

and

$$a_n |\psi_n\rangle = \sum_{m'} \sum_m \gamma_{mn} \tilde{\gamma}_{m'm} a_{m'} |\psi_{m'}\rangle.$$

Comparing each right side we find

$$\frac{1}{a_n} \sum_m \gamma_{mn} \tilde{\gamma}_{m'm} a_{m'} = \delta_{m'n},$$

and a similar exercise for $B|\phi_n\rangle$ yields

$$\frac{1}{b_n} \sum_m \gamma_{mn} \tilde{\gamma}_{m'n} b_{m'} = \delta_{m'n} .$$

Anticipating a commutation calculation, let the operator B act on $A|\psi_n\rangle$

$$\begin{aligned} BA|\psi_n\rangle &= B \sum_{m'} \sum_m \gamma_{mn} \tilde{\gamma}_{m'm} a_{m'} |\psi_{m'}\rangle \\ &= \sum_{m'} \sum_{\alpha} \sum_m \gamma_{mn} \tilde{\gamma}_{m'm} a_{m'} \gamma_{\alpha m'} b_{\alpha} |\phi_{\alpha}\rangle \\ &= \sum_{m'} \sum_{\rho} \sum_{\alpha} \tilde{\gamma}_{\rho\alpha} \gamma_{\alpha m'} b_{\alpha} \delta_{nm'} a_n |\psi_{\rho}\rangle \\ &= \sum_{m'} \sum_{\rho} \delta_{\rho m'} \delta_{nm'} a_n b_{\rho} |\psi_{\rho}\rangle \\ &= a_n b_n |\psi_n\rangle , \end{aligned}$$

which simplifies nicely.

Similarly, we must let A act on $B|\psi_n\rangle$. Begin by calculating $B|\psi_n\rangle$ to get

$$\begin{aligned} B|\psi_n\rangle &= B \sum_m \gamma_{mn} |\phi_m\rangle \\ &= \sum_m \gamma_{mn} b_m |\phi_m\rangle \\ &= \sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} |\psi_{m'}\rangle . \end{aligned}$$

The rest goes as

$$\begin{aligned} AB|\psi_n\rangle &= A \sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} |\psi_{m'}\rangle \\ &= \sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} a_{m'} |\psi_{m'}\rangle \\ &= \sum_{m'} \delta_{nm'} b_n a_{m'} |\psi_{m'}\rangle \\ &= a_n b_n |\psi_n\rangle . \end{aligned}$$

To finish, let us write the commutator of A and B to conclude

$$\begin{aligned} [AB]|\psi_n\rangle &= (AB - BA)|\psi_n\rangle \\ &= (a_n b_n - a_n b_n)|\psi_n\rangle = 0 , \end{aligned}$$

which evaluates to *zero*. That is, we get a zero commutator of two operators whose eigenvectors share an orthonormal basis.

13 Matrix in Hilbert Subspace

13.1 Laplacian Operator

In the Hilbert space of functions $\mathcal{L}_2[-1, 1]$, one can determine the components of the Laplacian operator

$$B = \partial_{xx}$$

of a subspace spanned by an orthonormal basis.

For instance, taking

$$\begin{aligned} |e_1\rangle &= \frac{1}{\sqrt{2}} \\ |e_2\rangle &= \frac{1}{\sqrt{2}} \left(\sin\left(\frac{\pi}{2}x\right) + \cos(\pi x) \right) \\ |e_3\rangle &= \frac{1}{\sqrt{2}} \left(\sin\left(\frac{\pi}{2}x\right) - \cos(\pi x) \right), \end{aligned}$$

the corresponding matrix B is calculated from

$$B = \begin{bmatrix} \langle e_1|B|e_1\rangle & \langle e_1|B|e_2\rangle & \langle e_1|B|e_3\rangle \\ \langle e_2|B|e_1\rangle & \langle e_2|B|e_2\rangle & \langle e_2|B|e_3\rangle \\ \langle e_3|B|e_1\rangle & \langle e_3|B|e_2\rangle & \langle e_3|B|e_3\rangle \end{bmatrix},$$

where

$$\langle e_j|B|e_k\rangle = \int_{-1}^1 e_j^*(x) \partial_{xx} e_k(x) dx.$$

Carrying out each integral, find

$$B = \frac{-\pi^2}{8} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 5 & -3 \\ 0 & -3 & 5 \end{bmatrix},$$

which is a Hermitian matrix by inspection.

13.2 Two Operators

In the (same) Hilbert space of functions $\mathcal{L}_2[-1, 1]$, one can determine the components of a derivative operator $A = \partial_x$ and a Laplacian operator $B = \partial_{xx}$ of a subspace spanned by an orthonormal basis.

Using an example orthonormal basis

$$\begin{aligned} |e_1\rangle &= \frac{1}{\sqrt{2}} \\ |e_2\rangle &= \frac{1}{\sqrt{2}} \sin(\pi x) \\ |e_3\rangle &= \frac{1}{\sqrt{2}} \cos(\pi x), \end{aligned}$$

we first check that the subset is closed under operations A :

$$\begin{aligned} A|e_1\rangle &= \partial_x \left(\frac{1}{\sqrt{2}} \right) = 0 \\ A|e_2\rangle &= \partial_x \left(\frac{1}{\sqrt{2}} \sin(\pi x) \right) \\ &= \frac{\pi}{\sqrt{2}} \cos(\pi x) = \pi |e_3\rangle \\ A|e_3\rangle &= \partial_x \left(\frac{1}{\sqrt{2}} \cos(\pi x) \right) \\ &= -\frac{\pi}{\sqrt{2}} \sin(\pi x) = -\pi |e_2\rangle \end{aligned}$$

Each nontrivial result can be written in terms of the original basis vectors. Therefore $A|x\rangle$, where $|x\rangle$ is an arbitrary linear combination of the basis vectors, will result in a vector in the same subspace. Expressing A as a matrix requires calculating $A_{jk} = \langle e_j|A|e_k\rangle$, resulting in

$$A = \pi \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix},$$

which is a *not* a Hermitian matrix.

Repeating the same exercise using $B = \partial_{xx}$ results (of course) in a Hermitian matrix

$$B = -\pi^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Interestingly, since the operation ∂_{xx} is two instances of ∂_x , it should follow that $AA = B$, easily checked by matrix multiplication:

$$\pi^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} = -\pi^2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

13.3 Pauli Matrices

Consider the set of three 2×2 Hermitian matrices

$$\begin{aligned} \sigma_1 &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ \sigma_2 &= \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \\ \sigma_3 &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \end{aligned}$$

called the *Pauli matrices*, where $i = \sqrt{-1}$. Interestingly, each matrix ($k = 1, 2, 3$) has the property

$$\sigma_k \sigma_k = I,$$

where I is the two-dimensional identity matrix. As a consequence, it follows that

$$\begin{aligned}\sigma_k^{2m} &= I \\ \sigma_k^{2m+1} &= \sigma_k\end{aligned}$$

for integer m . Furthermore, we have

$$\begin{aligned}\sigma_2\sigma_3 &= i\sigma_1 \\ \sigma_3\sigma_1 &= i\sigma_2 \\ \sigma_1\sigma_2 &= i\sigma_3.\end{aligned}$$

Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues of the Pauli matrices are determined by

$$\sigma_k |x\rangle = \lambda_k |x\rangle.$$

Working with σ_1 as an example, we write

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

giving two equations

$$\begin{aligned}x_2 &= \lambda_1 x_1 \\ x_1 &= \lambda_1 x_2,\end{aligned}$$

having two nontrivial branches $\lambda_1 = 1$ and $\lambda_1 = -1$, implying either $x_1 = x_2$, or respectively, $x_1 = -x_2$.

By standard means, find one normalized eigenvector per eigenvalue:

$$\begin{aligned}\lambda_1 = +1 &\rightarrow |x^{(+)}\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ \lambda_1 = -1 &\rightarrow |x^{(-)}\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}\end{aligned}$$

Repeating the exercise on $\sigma_2 |y\rangle = \lambda_2 |y\rangle$, and a third time on $\sigma_3 |z\rangle = \lambda_3 |z\rangle$, we find the eigenvalues are always $\lambda_k = \pm 1$. The corresponding normalized eigenvectors turn out to be

$$\begin{aligned}|y^{(+)}\rangle &= \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ 1 \end{bmatrix} \\ |y^{(-)}\rangle &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix} \\ |z^{(+)}\rangle &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ |z^{(-)}\rangle &= \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \end{bmatrix}.\end{aligned}$$

14 Functions of Operators

14.1 Motivation

Recall that the single time derivative operator $L = \partial_t$ as it appears in the problem $L|x\rangle = \lambda|x\rangle$ has exponentially-evolving eigenvectors $|x(t)\rangle = e^{\lambda t}|x_0\rangle$, where $|x_0\rangle$ is the initial state in a fixed orthonormal basis. The more general statement

$$|x(t)\rangle = e^{Lt}|x_0\rangle$$

leads to the same solution, easily shown by taking a time derivative:

$$\partial_t |x(t)\rangle = L e^{Lt} |x_0\rangle = L |x(t)\rangle$$

14.2 Time Evolution Operator

The combination e^{Lt} is generally known as a *time evolution* operator. The exponential function, despite having an operator in its argument, readily expands as

$$\begin{aligned} e^{Lt} &= \sum_{k=0}^{\infty} \frac{1}{k!} (Lt)^k \\ &= I + (Lt) + \frac{1}{2!} (Lt)^2 + \frac{1}{3!} (Lt)^3 + \cdots \end{aligned}$$

Grouping even terms and odd terms separately, the above reads

$$\begin{aligned} e^{Lt} &= \left(I + \frac{1}{2!} (Lt)^2 + \frac{1}{4!} (Lt)^4 + \cdots \right) \\ &\quad + \left(Lt + \frac{1}{3!} (Lt)^3 + \frac{1}{5!} (Lt)^5 + \cdots \right) \end{aligned}$$

Making the substitution $L = -i\tilde{H}$, we further have

$$\begin{aligned} e^{-i\tilde{H}t} &= \left(I - \frac{1}{2!} (\tilde{H}t)^2 + \frac{1}{4!} (\tilde{H}t)^4 - \cdots \right) \\ &\quad - i \left(\tilde{H}t - \frac{1}{3!} (\tilde{H}t)^3 + \frac{1}{5!} (\tilde{H}t)^5 - \cdots \right), \end{aligned}$$

simplifying nicely to

$$e^{-i\tilde{H}t} = \cos(\tilde{H}t) - i \sin(\tilde{H}t),$$

giving away two more functions where an operator may naturally embed.

14.3 Pauli Matrix Operators

In the special case that $\tilde{H}$ is equal to any of the Pauli matrices $\{\sigma_k\}$ with $k = 1, 2, 3$ up to a proportionality constant μ such that $\tilde{H} = \mu\sigma_k$, the above reduces to

$$e^{-i\sigma_k\mu t} = I \cos(\mu t) - i\sigma_k \sin(\mu t),$$

which removes the operator from any infinite series. Explicitly:

$$\begin{aligned} e^{-i\sigma_1\mu t} &= \begin{bmatrix} \cos(\mu t) & -i \sin(\mu t) \\ -i \sin(\mu t) & \cos(\mu t) \end{bmatrix} \\ e^{-i\sigma_2\mu t} &= \begin{bmatrix} \cos(\mu t) & -\sin(\mu t) \\ \sin(\mu t) & \cos(\mu t) \end{bmatrix} \\ e^{-i\sigma_3\mu t} &= \begin{bmatrix} e^{-i\mu t} & 0 \\ 0 & e^{i\mu t} \end{bmatrix} \end{aligned}$$

15 Unitary Operators

Consider a vector space $\mathcal{V}$ admitting two different sets of basis vectors $\{|e_j\rangle\}$ and $|\tilde{e}_j\rangle$. In terms of coefficients x_j and $\tilde{x}_j$, a given vector $|x\rangle$ is a linear combination in each basis:

$$|x\rangle = \sum_j x_j |e_j\rangle = \sum_j \tilde{x}_j |\tilde{e}_j\rangle$$

Any coefficient(s) x_j , which exist in the vector space $\mathcal{V}$, can be isolated by exploiting the orthogonality between each $|e_j\rangle$ such that

$$x_j = \langle e_j | x \rangle = \sum_k \langle e_j | x_k | \tilde{e}_k \rangle = \sum_k \tilde{x}_k \langle e_j | \tilde{e}_k \rangle ,$$

which applies similarly to $\tilde{x}_j$:

$$\tilde{x}_j = \langle \tilde{e}_j | x \rangle = \sum_k \langle \tilde{e}_j | x_k | e_k \rangle = \sum_k x_k \langle \tilde{e}_j | e_k \rangle$$

15.1 Unitary Matrix

In component notation, the above has implicated two matrices

$$\begin{aligned} U_{jk} &= \langle \tilde{e}_j | e_k \rangle \\ \tilde{U}_{jk} &= \langle e_j | \tilde{e}_k \rangle , \end{aligned}$$

which can be interpreted as a set of coordinate transformation matrices that carry $\{x_j\} \in \mathcal{V} \rightarrow \{\tilde{x}_j\} \in \mathcal{V}$, and vice versa.

These matrices are intricately related, which we first notice by the observation

$$\tilde{U}_{jk} = \langle e_j | \tilde{e}_k \rangle = \overline{\langle \tilde{e}_k | e_j \rangle} = U_{kj}^* = U_{jk}^\dagger ,$$

telling us each matrix is the other's Hermitian conjugate (dropping component notation):

$$\begin{aligned} \tilde{U} &= U^\dagger \\ U &= \tilde{U}^\dagger \end{aligned}$$

Now, the double transformation $\{x_j\} \rightarrow \{\tilde{x}_j\} \rightarrow \{x_j\}$ (and vice versa) tells us that each combination $\tilde{U}U$ and $U\tilde{U}$ is an identity matrix:

$$\begin{aligned} \tilde{U}U &= I \\ U\tilde{U} &= I \end{aligned}$$

Combining the two previous results yields four identities:

$$\begin{aligned} U^\dagger U &= I \\ UU^\dagger &= I \\ \tilde{U}\tilde{U}^\dagger &= I \\ \tilde{U}^\dagger \tilde{U} &= I \end{aligned}$$

Any operator satisfying the above equations is called *unitary*.

15.2 Effect on Inner Product

An important consequence of unitary operators arises when dealing with the inner product of two vectors $|x\rangle, |y\rangle \in \mathcal{V}$. Calculating the inner product of $U|x\rangle$ and $U|y\rangle$, we find

$$\langle Ux | Uy \rangle = \langle U^\dagger Ux | y \rangle = \langle x | y \rangle = \langle x | y \rangle ,$$

indicating that the inner product is unaffected by the operations.

In most applications (namely in two- and three-dimensional space), unitary operations correspond to rotations and reflections of the coordinates $\{x_j\} \in \mathcal{V}$.

15.3 Effect on Basis Vectors

Next, we examine how operator U acts directly on the basis vectors $\{|e_j\rangle\} \in \mathcal{V}$. Supposing we set $|x\rangle = |e_j\rangle$ and $|y\rangle = |e_k\rangle$, we find

$$\langle Ue_j | Ue_k \rangle = \langle e_j | e_k \rangle = \delta_{jk} ,$$

telling us that the combinations $|Ue_j\rangle$ and $|Ue_k\rangle$ are members of a second orthogonal basis $\{|g_j\rangle\}$ that can be generated from the original $\{|e_j\rangle\}$ via

$$\begin{aligned} |g_j\rangle &= U |e_j\rangle \\ j &= 1, 2, 3, \dots, N . \end{aligned}$$

As a matrix, recall that the operator U can be written as

$$U = \sum_{jk} |e_j\rangle U_{jk} \langle e_k| = \sum_{jk} |e_j\rangle \langle \tilde{e}_j | e_k \rangle \langle e_k| ,$$

where the completeness relation

$$I = \sum_k |e_k\rangle \langle e_k|$$

reduces the above to

$$U = \sum_j |e_j\rangle \langle \tilde{e}_j| .$$

Right away, we find

$$U |\tilde{e}_j\rangle = \sum_j |e_j\rangle \langle \tilde{e}_j | \tilde{e}_j \rangle = |e_j\rangle .$$

Similarly, we deduce

$$\tilde{U} = \sum_j |\tilde{e}_j\rangle \langle e_j| ,$$

which leads to

$$\tilde{U} |e_j\rangle = |\tilde{e}_j\rangle .$$

In other words, the matrix U takes the j th vector from the primed basis and churns out the j th vector from the unprimed basis. Or, the $\tilde{U}$ matrix takes the j th vector from the unprimed basis and computes the corresponding primed vector.

In component form, these results read:

$$|e_k\rangle = U |\tilde{e}_k\rangle = \sum_{ij} |\tilde{e}_i\rangle U_{ij} \langle \tilde{e}_j | \tilde{e}_k \rangle = \sum_j U_{jk} |\tilde{e}_j\rangle$$

$$|\tilde{e}_k\rangle = \tilde{U} |e_k\rangle = \sum_{ij} |e_i\rangle \tilde{U}_{ij} \langle e_j | e_k \rangle = \sum_j \tilde{U}_{jk} |e_j\rangle$$

Project $\langle \tilde{e}_j |$ into the first equation and $\langle e_j |$ into the second to recover the component form of each matrix:

$$U_{jk} = \langle \tilde{e}_j | e_k \rangle = \langle \tilde{e}_j | U | \tilde{e}_k \rangle$$

$$\tilde{U}_{jk} = \langle e_j | \tilde{e}_k \rangle = \langle e_j | \tilde{U} | e_k \rangle$$

Sanity Check

For a sanity check, let us apply U to a vector $|x\rangle \in \mathcal{V}$. Calculating this, we have

$$\begin{aligned} U |x\rangle &= \sum_{jk} x_k |e_j\rangle \langle \tilde{e}_j | e_k \rangle \\ &= \sum_j \left(\sum_k U_{jk} x_k \right) |e_j\rangle = |x'\rangle , \end{aligned}$$

where

$$x'_j = \sum_k U_{jk} x_k$$

is the rotated vector component in the same basis $\{|e_j\rangle\}$.

15.4 Rotations

Two Dimensions

The simplest nontrivial case involving unitary operators addresses rotations in the two-dimensional plane. Consider a Cartesian space spanned by the orthonormal basis $|e_1\rangle = \hat{x}$, $|e_2\rangle = \hat{y}$. A second orthonormal basis $\{|\tilde{e}_j\rangle\}$ is oriented

at angle ϕ with respect with respect to the original. In particular:

$$|\tilde{e}_1\rangle = \cos(\phi) |e_1\rangle + \sin(\phi) |e_2\rangle$$

$$|\tilde{e}_2\rangle = -\sin(\phi) |e_1\rangle + \cos(\phi) |e_2\rangle$$

Using the above formula for U_{jk} applied to standard two-dimensional geometry, we quickly find the components of U to be

$$U_{xx} = \langle \tilde{e}_1 | e_1 \rangle = \cos \phi$$

$$U_{xy} = \langle \tilde{e}_1 | e_2 \rangle = -\sin \phi$$

$$U_{yx} = \langle \tilde{e}_2 | e_1 \rangle = \sin \phi$$

$$U_{yy} = \langle \tilde{e}_2 | e_2 \rangle = \cos \phi ,$$

and similarly for $U^\dagger = \tilde{U}$:

$$\tilde{U}_{xx} = \langle e_1 | \tilde{e}_1 \rangle = \cos \phi$$

$$\tilde{U}_{xy} = \langle e_1 | \tilde{e}_2 \rangle = \sin \phi$$

$$\tilde{U}_{yx} = \langle e_2 | \tilde{e}_1 \rangle = -\sin \phi$$

$$\tilde{U}_{yy} = \langle e_2 | \tilde{e}_2 \rangle = \cos \phi$$

There is an important different between the two operators. Matrix U calculates the components of a rotated vector in a fixed basis. Matrix $U^\dagger = \tilde{U}$ transforms the coordinates of a fixed vector when the basis is rotated.

Three Dimensions

In three dimensions, we extend the two-dimensional case to write three matrices (presumably named after aviation terms)

$$R_z(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix} = \text{'yaw'}$$

$$R_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) \\ 0 & 1 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) \end{bmatrix} = \text{'pitch'}$$

$$R_x(\gamma) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\gamma) & -\sin(\gamma) \\ 0 & \sin(\gamma) & \cos(\gamma) \end{bmatrix} = \text{'roll'} ,$$

where a general rotation of a vector in three dimensions is the product

$$R = R_z(\alpha) R_y(\beta) R_x(\gamma) .$$

Due to commutivity rules, the order in which the matrices are applied *does* affect the result. Formally, the

above matrices correspond to an *intrinsic* rotation, having *Tait–Bryan* angles α, β, γ . The domain restriction on each angle is as follows:

$$\begin{aligned} 0 &\leq \alpha < 2\pi \\ 0 &\leq \beta < \pi \\ 0 &\leq \gamma < 2\pi \end{aligned}$$

15.5 Effect on Operator

Now we examine what happens to the components of an operator A when undergoing a change of basis vectors summarized by

$$|x\rangle = \sum_j x_j |e_j\rangle = \sum_j \tilde{x}_j |\tilde{e}_j\rangle .$$

For some vector $|x\rangle$ in the vector space $\mathcal{V}$, the operation $A|x\rangle$ yields a vector $|y\rangle$, also in $\mathcal{V}$. Expressing this calculation in two different orthonormal bases, we have

$$\begin{aligned} \sum_j A_{ij} x_j &= y_i \\ \sum_j \tilde{A}_{ij} \tilde{x}_j &= \tilde{y}_i . \end{aligned}$$

Substituting

$$\begin{aligned} x_j &= \sum_k \tilde{x}_k \langle e_j | \tilde{e}_k \rangle \\ y_i &= \sum_k \tilde{y}_k \langle e_i | \tilde{e}_k \rangle \end{aligned}$$

into the first equation, we end up with

$$A\tilde{U} |\tilde{x}\rangle = \tilde{U} |\tilde{y}\rangle ,$$

where multiplying both sides by U gives

$$UA\tilde{U} |\tilde{x}\rangle = U\tilde{U} |\tilde{y}\rangle = I |\tilde{y}\rangle = |\tilde{y}\rangle .$$

Meanwhile, we already know $|\tilde{y}\rangle = \tilde{A} |\tilde{x}\rangle$ by construction, and we conclude

$$\tilde{A} = UA\tilde{U} .$$

Of course, this result can be attained more directly by substituting

$$A = \sum_{i'} \sum_{j'} |e_{i'}\rangle A_{i'j'} \langle e_{j'}|$$

into $\tilde{A}_{ij} = \langle \tilde{e}_i | A | \tilde{e}_j \rangle$.

16 Differential Equations

16.1 Schrodinger Equation

The chief equation of quantum mechanics is conveniently framed as an eigenvalue problem. As such, the famous Schrodinger equation reads

$$i\hbar \frac{d}{dt} |\psi(t)\rangle = H |\psi(t)\rangle ,$$

where $i = \sqrt{-1}$, $\hbar$ is Planck's constant, and H is a Hermitian operator called the *Hamiltonian* of size $n \times n$. The symbol $|\psi(t)\rangle$ is the *quantum state vector*, which like any other vector, resolves to components provided an orthonormal basis exists:

$$|\psi(t)\rangle = (\psi_1(t), \psi_2(t), \psi_3(t), \dots)$$

Naming the supporting orthonormal basis $|e_j\rangle$, we explicitly have

$$|\psi(t)\rangle = \sum_j \psi_j(t) |e_j\rangle$$

and

$$H = \sum_j \sum_k |e_j\rangle H_{jk} \langle e_k| ,$$

letting us state the problem in component form:

$$\begin{aligned} i\hbar \sum_j \frac{d}{dt} \psi_j(t) |e_j\rangle &= \sum_j \sum_k \sum_{j'} |e_j\rangle H_{jk} \psi_{j'}(t) \langle e_k | e_{j'} \rangle \\ &= \sum_j \sum_k H_{jk} \psi_k(t) |e_j\rangle \end{aligned}$$

With the j -sum present on each side, the above reduces to:

$$i\hbar \frac{d}{dt} \psi_j(t) = \sum_k H_{jk} \psi_k(t)$$

TISE

Assuming that H admits n eigenvalues E_j and n eigenvectors represented by $|\phi^{(j)}\rangle$, we may also write the *time-independent Schrodinger equation*, or *TISE*:

$$H |\phi^{(j)}\rangle = E_j |\phi^{(j)}\rangle$$

Next, since H is a Hermitian operator, its eigenvectors $|\phi^{(j)}\rangle$ form an orthonormal basis for which the quantum state vector can be expressed as a linear combination:

$$|\psi(t)\rangle = \sum_j C_j(t) |\phi^{(j)}\rangle$$

Applying the H -operator to both sides of the above, thereby writing the Schrodinger equation, gives for any given index j ,

$$i\hbar \frac{d}{dt} C_j(t) = E_j C_j(t) ,$$

solved by

$$C_j(t) = C_j(t=0) e^{-iE_j t/\hbar} .$$

Note that the initial value of each $C_n(0)$ is calculated from the initial condition $|\psi(0)\rangle$ according to

$$\langle \phi^{(k)} | \psi(0) \rangle = \sum_j C_j(0) \langle \phi^{(k)} | \phi^{(j)} \rangle = C_k ,$$

where

$$C_k = \langle \phi^{(k)} | \psi(0) \rangle = \sum_j \left(\phi_j^{(k)} \right)^* \psi_j(0) .$$

The evolution of the quantum state vector can thus be written

$$|\psi(t)\rangle = \sum_j C_j(0) e^{-iE_j t/\hbar} |\phi^{(j)}\rangle .$$

16.2 Hamiltonian Matrix

Consider a two-component vector that presumes the existence of an orthonormal basis

$$|u(t)\rangle = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$$

that relates to the time derivative operator by

$$i\partial_t |u(t)\rangle = \hat{H} |u(t)\rangle ,$$

where $\hat{H}$ is the dimensionalized *Hamiltonian matrix*, having form

$$\hat{H} = \begin{bmatrix} 0 & \delta \\ \delta & 0 \end{bmatrix} ,$$

with δ being constant.

Solving the eigenvalue problem

$$\hat{H} |x^{(j)}\rangle = \lambda_j |x^{(j)}\rangle$$

for this case, we quickly find two eigenvalues

$$\begin{aligned} \lambda_+ &= \delta \\ \lambda_- &= -\delta , \end{aligned}$$

and two corresponding eigenvectors

$$|x^{(+)}\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$|x^{(-)}\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

With the eigenvectors for the operator on hand, the vector $|u(t)\rangle$ can be expressed as a linear combination

$$|u(t)\rangle = C_+(t) |x^{(+)}\rangle + C_-(t) |x^{(-)}\rangle$$

$$= \sum_j C_j(t) |x^{(j)}\rangle.$$

Apply $\hat{H}$ to both sides to distill a differential equation

$$i \frac{\partial}{\partial t} C_j(t) = C_j(t) \lambda_j,$$

solved by

$$C_j(t) = C_j(t=0) e^{-i\lambda_j t}.$$

The updated general solution now reads

$$|u(t)\rangle = C_+(0) e^{-i\delta t} |x^{(+)}\rangle + C_-(0) e^{i\delta t} |x^{(-)}\rangle,$$

where the coefficients $C_{\pm}(0)$ are determined by the initial conditions of the system. Since the eigenvectors $|x^{(\pm)}\rangle$ form an orthonormal basis, the coefficients are easily isolated from $|u(t=0)\rangle$:

$$C_j(0) = \langle x^{(j)} | u(0) \rangle$$

Note finally that the exponential terms can be traded for trigonometric terms by Euler's formula to give

$$|u(t)\rangle = \frac{1}{\sqrt{2}} \cos(\delta t) \begin{bmatrix} C_+ + C_- \\ C_+ - C_- \end{bmatrix}$$

$$+ \frac{i}{\sqrt{2}} \sin(\delta t) \begin{bmatrix} -C_+ + C_- \\ -C_+ - C_- \end{bmatrix}.$$

16.3 Damped Harmonic Oscillator

The differential equation that governs the damped harmonic oscillator reads

$$\frac{d^2}{dt^2} x(t) - b \frac{d}{dt} x(t) + \omega_0^2 x(t) = 0,$$

which is a second-order equation. However, this problem can be turned into a system of first-order equations by defining a vector

$$|u(t)\rangle = \begin{bmatrix} x(t) \\ dx(t)/dt \end{bmatrix}$$

along with a matrix

$$A = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -b \end{bmatrix},$$

so the problem may be rewritten

$$\frac{d}{dt} |u(t)\rangle = A |u(t)\rangle.$$

Solving the eigenvalue problem

$$A |q^{(j)}\rangle = \lambda_j |q^{(j)}\rangle,$$

we quickly find

$$\lambda = -\frac{b}{2} \pm \sqrt{\frac{b^2}{4} - \omega_0^2},$$

with corresponding eigenvectors

$$|q^{(+)}\rangle = \frac{1}{\sqrt{\lambda_+^2 + 1}} \begin{bmatrix} 1 \\ \lambda_+ \end{bmatrix}$$

$$|q^{(-)}\rangle = \frac{1}{\sqrt{\lambda_-^2 + 1}} \begin{bmatrix} 1 \\ \lambda_- \end{bmatrix}.$$

Note that the eigenvectors $|q^{(\pm)}\rangle$ are linearly independent but not orthogonal. Regardless of this, the general solution can be written as a linear combination

$$|u(t)\rangle = C_+(t) |q^{(+)}\rangle + C_-(t) |q^{(-)}\rangle$$

$$= \sum_j C_j(t) |q^{(j)}\rangle.$$

Applying the A -operator to both sides of the above produces

$$\sum_j \frac{d}{dt} (C_j(t)) |q^{(j)}\rangle = A \sum_j C_j(t) |q^{(j)}\rangle$$

$$= \sum_j C_j(t) \lambda_j |q^{(j)}\rangle,$$

immediately implying

$$\frac{d}{dt} C_j(t) = C_j(t) \lambda_j,$$

solved by

$$C_j(t) = C_j(t=0) e^{\lambda_j t}.$$

The general solution now reads

$$|u(t)\rangle = \sum_j C_j(0) e^{\lambda_j t} |q^{(j)}\rangle.$$

Setting $t = 0$ in the above gives

$$|u(0)\rangle = \sum_j C_j(0) |q^{(j)}\rangle .$$

Note however that the basis vectors $|q^{(k)}\rangle$ are not orthogonal, thus the coefficients $C_j(0)$ cannot be isolated by taking the inner product with $\langle q^{(k)}|$. To proceed, write the above in component form to get

$$u_k(0) = \sum_j C_j(0) q_k^{(j)} .$$

This is a linear system of the form $A|x\rangle = |b\rangle$ with $|u(0)\rangle$ playing the role of $|b\rangle$, the vector components $q_k^{(j)}$ serving as matrix components, and the components $C_j(0)$ corresponding to $|x\rangle$.

17 Normal Operators

Chapter 27

Differential Geometry

1 Tensors and Index Notation

Flat Space

Ordinary calculus typically deals in Cartesian coordinates with one, two, or three dimensions, where it is understood that the x -, y -, and z -axes always remain mutually perpendicular without ever bending. In such *flat space*, the rules of geometry (and thereby trigonometry, vectors, etc.) are taken everywhere for granted.

Curved Space

It is useful, however, to modify the rules of calculus and geometry so as to embed on a *curved surface* or *curved space*, which we loosely define as any situation where the rules of Euclid won't work. This is a geometric disaster, let alone what happens to trigonometry, vectors, the laws of physics, etc.

The framework for adapting to curved space comes from differential geometry, and in particular, Riemannian geometry.

1.1 Index Notation

A vector A needs one index to register its components, and usually no worry is assigned to how the index appears. As a list, A is equivalent to

$$A = \langle A_x, A_y, A_z \rangle = \langle A^x, A^y, A^z \rangle$$

with the distinction between A_x , A^x being a stylistic choice. The first thing we need to do is dispense with such sloppiness.

Type

To advance on a cleaner index notation, let us adopt the burden that vertical placement of the index *does* matter to a vector, and to quantify this we use the word *type*.

The type of any object can be denoted (m, n) for whole numbers $m > 0$, $n > 0$. When reported as a single number, the term 'type' represents the sum $m + n$. Any type-0 tensor is simply a scalar.

Contravariant and Covariant

Using the type convention, we define a type $(1, 0)$ vector with N components as

$$x^\mu = (x^1, x^2, x^3, \dots, x^N) .$$

Somewhat like a column vector, x^μ is called *contravariant*.

For a vector of type $(0, 1)$ with N components, we write

$$x_\mu = (x_1, x_2, x_3, \dots, x_N) ,$$

which is somewhat like a row vector, and is called *covariant*, also known as a *dual vector*.

1.2 Tensors

Going beyond vectors, it's conceivable to work with mathematical objects that have any number of indices, called *tensors*, of which vectors and matrices are special cases. There are only two type-1 tensors, which are simply the vectors x^μ , x_μ written above.

For a type-2 tensor A , there are three flavors:

$$\begin{aligned}\text{type } (2, 0) &= A^{\mu\nu} \\ \text{type } (1, 1) &= A^\mu_\nu \\ \text{type } (0, 2) &= \Lambda_{\mu\nu}\end{aligned}$$

Each tensor $A^{\mu\nu}$, Λ^μ_ν , $\Lambda_{\mu\nu}$ represents a two-dimensional ‘block’ of terms.

1.3 Linear System

From linear algebra, recall that at linear system $A\mathbf{x} = \mathbf{y}$ may be understood as a matrix A acting on a column vector $\mathbf{x}$ resulting in a new column vector $\mathbf{y}$. In detail, one may encounter

$$\sum_{k=1}^N A_{jk} x_k = y_j .$$

The aim is to update the above to tensor language. Taking $\mathbf{x}$, $\mathbf{y}$ as column vectors, each is replaced by x^k , y^j , respectively. This leaves the puzzle of what to do with the indices on A , as each of the following may look appealing:

$$\begin{aligned}\sum_{k=1}^N A^{jk} x^k &\stackrel{?}{=} y^j \\ \sum_{k=1}^N A_k^j x^k &\stackrel{?}{=} y^j \\ \sum_{k=1}^N A_{jk} x^k &\stackrel{?}{=} y^j\end{aligned}$$

Type Continuity

In the above, the correct option is singled out by type arguments. On the right side of each, the final type of y^j is $(1, 0)$. Only the middle case satisfies this on the left, thus we write the linear system as

$$\sum_{k=1}^N A_k^j x^k = y^j .$$

Einstein Summation Convention

By writing the linear system $A\mathbf{x} = \mathbf{y}$ in index notation, observe on the left that the k -index appears twice, once in the up-position and once in the down-position. It just happens that k is the variable we’re summing over and the summation symbol can be omitted according to the *Einstein summation convention*:

$$A_k^j x^k = y^j$$

The linear system can be re-analyzed so the resulting vector y_j is type $(0, 1)$. The beauty of index notation is we can write down the answer immediately:

$$A_j^k x_k = y_j$$

In this case, the repeated index k vanishes from the left via the Einstein sum convention.

1.4 Tensor Product

Any pair of tensors A , B can be ‘multiplied’ in several ways. To keep things entirely general, we will assume each tensor of arbitrary type, i.e.

$$\begin{aligned}A &= A^{\alpha_1 \alpha_2 \dots \alpha_p}_{\beta_1 \beta_2 \dots \beta_q} \\ B &= B^{\mu_1 \mu_2 \dots \mu_r}_{\nu_1 \nu_2 \dots \nu_s}\end{aligned}$$

That is, tensor A is of type (p, q) , and tensor B is of type (r, s) .

Outer product

The most explosive tensor product is the *outer product*, which results in a tensor of rank $(p + r, q + s)$:

$$A \otimes B = A^{\alpha_1 \alpha_2 \dots \alpha_p}_{\beta_1 \beta_2 \dots \beta_q} B^{\mu_1 \mu_2 \dots \mu_r}_{\nu_1 \nu_2 \dots \nu_s}$$

For a special case, the product of a $(p, 0)$ tensor and a $(0, s)$ tensor constitute a tensor of type (p, s) :

$$A \otimes B = A^{\alpha_1 \alpha_2 \dots \alpha_p} B_{\nu_1 \nu_2 \dots \nu_s}$$

Supposing we also have $p = s = 1$, the above reduces to the product of a column vector and a row vector:

$$A \otimes B = A^\alpha B_\nu$$

If A , B have n components each, the above is equivalent to:

$$\begin{aligned}\begin{bmatrix} A_1 \\ A_2 \\ \dots \\ A_n \end{bmatrix} \begin{bmatrix} B_1 & B_2 & \dots & B_n \end{bmatrix} \\ = \begin{bmatrix} A_1 B_1 & A_1 B_2 & \dots & A_1 B_n \\ A_2 B_1 & A_2 B_2 & \dots & A_2 B_n \\ \dots & \dots & \dots & \dots \\ A_n B_1 & A_n B_2 & \dots & A_n B_n \end{bmatrix}\end{aligned}$$

Contraction

Start with the structure of the outer product and let any one of the contravariant indices in A equal any one of the covariant indices in B , or vice-versa. The resulting object is called a *contraction* between A and B . Supposing the shared index is k , one could have:

$$C = A^{\alpha_1 \alpha_2 \dots k \dots \alpha_p} B^{\mu_1 \mu_2 \dots \mu_r}_{\nu_1 \nu_2 \dots k \dots \nu_s}$$

Using the Einstein sum convention, the contraction C is a tensor of rank $p + q + r + s - 2$. That is, the total contravariant rank is reduced by one, and the total covariant rank is reduced by one.

Problem 1

For a type-2 tensor A , which of Λ_b^a , $\Lambda^{\nu\nu}$, $\Lambda^{\mu\mu}$, Λ_b^b , Λ_ν^μ , Λ^{12} is most like the matrix trace of A ? Answer: Λ_b^b

Inner Product

If contraction is performed recursively on all indices, the rank of the resulting quantity eventually reaches zero, leaving just a scalar $w = A \cdot B$. This is called the *inner product* of the two tensors:

$$w = A \cdot B = A^{\alpha_1 \alpha_2 \dots \alpha_p}_{\mu_1 \mu_2 \dots \mu_r} B^{\mu_1 \mu_2 \dots \mu_r}_{\alpha_1 \alpha_2 \dots \alpha_p}$$

Norm

The *norm* of a type-1 tensor is defined as the self inner product:

$$\|A\|^2 = A^\mu A_\mu$$

The same notion is true for tensors of any type. The type-2 case involves the double sum:

$$\|B\|^2 = B^{\mu\nu} B_{\mu\nu}$$

1.5 Formal Tensor Definition

To formally qualify as a tensor, the components of any indexed object A must scale linearly with respect to a generalized coordinate transformation. For this, we let $q^{\mu'}$ represent generalized coordinates in a transformed system. An object of type (m, n) is a tensor only if:

$$A^{\mu'_1 \dots \mu'_n}_{\nu'_1 \dots \nu'_m} = \frac{\partial q^{\mu'_1}}{\partial q^{\mu_1}} \dots \frac{\partial q^{\mu'_n}}{\partial q^{\mu_n}} \frac{\partial q^{\nu_1}}{\partial q^{\nu'_1}} \dots \frac{\partial q^{\nu_m}}{\partial q^{\nu'_m}} A^{\mu_1 \dots \mu_n}_{\nu_1 \dots \nu_m}$$

For a special case, the type-1 vectors V^μ and V_μ transform as:

$$V^{\mu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} V^\mu$$

$$V_{\mu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} V_\mu$$

The above manifests three ways for type-2 tensors

$$A^{\mu'\nu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} \frac{\partial q^{\nu'}}{\partial q^\nu} A^{\mu\nu}$$

$$A^{\mu'}_{\nu'} = \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} A^\mu_\nu$$

$$A_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} A_{\mu\nu},$$

and so on for those of higher type. For every ‘up’ index on the left, tack on a term like $\partial q^{\mu'}/\partial q^\mu$ to the right. Similarly for every ‘down’ index on the left, tack on a term like $\partial q^\mu/\partial q^{\mu'}$ to the right.

Cartesian and Polar Coordinates

Recall that the Cartesian plane is mapped by ordered pairs (x, y) . Also, a polar coordinate system (r, θ) covers the same plane via

$$x = r \cos(\theta)$$

$$y = r \sin(\theta)$$

A type $(0, 1)$ tensor V with Cartesian components V_x , V_y can be expressed in polar coordinates by associating:

$$q^\mu = x$$

$$q^\nu = y$$

$$q^{\mu'} = r$$

$$q^{\nu'} = \theta$$

Then, using the definition of a tensor we have:

$$V_r = \frac{\partial q^\mu}{\partial q^r} V_\mu = \frac{\partial x}{\partial r} V_x + \frac{\partial y}{\partial r} V_y$$

$$V_\theta = \frac{\partial q^\mu}{\partial q^\theta} V_\mu = \frac{\partial x}{\partial \theta} V_x + \frac{\partial y}{\partial \theta} V_y$$

Or, one can solve for the Cartesian components via

$$V_x = \frac{\partial q^\mu}{\partial x} V_\mu = \frac{\partial r}{\partial x} V_r + \frac{\partial \theta}{\partial x} V_\theta$$

$$V_y = \frac{\partial q^\mu}{\partial y} V_\mu = \frac{\partial r}{\partial y} V_r + \frac{\partial \theta}{\partial y} V_\theta,$$

which requires inverting the formulas for r, θ :

$$r = \sqrt{x^2 + y^2}$$

$$\theta = \arctan\left(\frac{y}{x}\right)$$

From these, we gain:

$$\partial r / \partial x = x/r$$

$$\partial r / \partial y = y/r$$

$$\partial \theta / \partial x = -y/r^2$$

$$\partial \theta / \partial y = x/r^2$$

The story is similar when V is a type $(1, 0)$ tensor. For this, the tensor transformation law gives

$$V^x = \frac{\partial q^x}{\partial q^\mu} V^\mu = \frac{\partial x}{\partial r} V^r + \frac{\partial x}{\partial \theta} V^\theta$$

$$V^y = \frac{\partial q^y}{\partial q^\mu} V^\mu = \frac{\partial y}{\partial r} V^r + \frac{\partial y}{\partial \theta} V^\theta,$$

and correspondingly

$$V^r = \frac{\partial q^r}{\partial q^\mu} V^\mu = \frac{\partial r}{\partial x} V^x + \frac{\partial r}{\partial y} V^y$$

$$V^\theta = \frac{\partial q^\theta}{\partial q^\mu} V^\mu = \frac{\partial \theta}{\partial x} V^x + \frac{\partial \theta}{\partial y} V^y.$$

Problem 2

Show that:

$$V^r = \cos(\theta) V^x + \sin(\theta) V^y$$

$$V^\theta = -\frac{\sin(\theta)}{r} V^x + \frac{\cos(\theta)}{r} V^y.$$

1.6 Scalar Product

Using the definitions above, let us calculate the inner product of two vectors

$$w = A_\mu B^\mu,$$

which is similar to the special case of the outer product above up to a swap of type.

Next, write the transformed tensors

$$B^{\mu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} B^\mu$$

$$A_{\mu'} = \frac{\partial q^\nu}{\partial q^{\mu'}} A_\nu,$$

and multiply the two equations:

$$A_{\mu'} B^{\mu'} = \left(\frac{\partial q^\nu}{\partial q^{\mu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \right) A_\nu B^\mu$$

Kronecker Delta

The parenthesized quantity is only nonzero if $\beta = \mu$, represented using the Kronecker delta relation (not a tensor):

$$\delta_\mu^\nu = \frac{\partial q^\nu}{\partial q^{\mu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} = \begin{cases} 1 & \text{if } \nu = \mu \\ 0 & \text{if } \nu \neq \mu \end{cases}$$

Invariance

In light of the Kronecker delta, the scalar product calculation simplifies to

$$w = A_{\mu'} B^{\mu'} = A_\mu B^\mu.$$

The fact that the scalar product $w = A_\mu B^\mu$ is the same in any coordinate system means w is an *invariant* quantity.

In the special case $A = B$, the scalar w is the norm of the vector:

$$\|A\|^2 = A_\mu A^\mu$$

2 Manifold and the Metric

Let S denote the position vector on a manifold parameterized by generalized coordinates $q^1, q^2, q^3, \dots, q^N$, where N is the number of dimensions on the manifold:

$$S = (q^1, q^2, \dots, q^N)$$

Note that the physical units of any given q^μ are not limited to length.

2.1 Affine Parameter

Each coordinate q^μ can occur as a function of a parameter λ , i.e. $q^\mu = q^\mu(\lambda)$. It turns out, for our studies, that some choices of λ are more ‘natural’ than others. Reaching for the most natural parameter, it makes sense to associate small changes $d\lambda$ with the change in arc length dS on the manifold itself. For this reason we decide

$$d\lambda \propto dS,$$

where λ is called an *affine parameter*.

A convenient affine parameter invokes the notion of constant speed c on the manifold, along with a time parameter $d\tau$ during which motion takes place. From these we make the association

$$c \, d\tau \propto dS.$$

2.2 Differential Line Element

Denoting the differential line element vector as dS , the chain rule dictates

$$dS = \frac{\partial S}{\partial q^1} dq^1 + \frac{\partial S}{\partial q^2} dq^2 + \dots + \frac{\partial S}{\partial q^N} dq^N.$$

Generalized Basis Vectors

The differential line element is abbreviated using the Einstein sum convention

$$dS = a_{(\mu)} dq^\mu$$

provided we define the *generalized basis vectors* $a_{(\mu)}$ such that:

$$a_{(\mu)} = \frac{\partial S}{\partial q^\mu} = \partial_\mu S$$

In component form, the above contains the information

$$a_{(\mu)} = \left(\frac{\partial q^1}{\partial q^\mu}, \frac{\partial q^2}{\partial q^\mu}, \dots, \frac{\partial q^N}{\partial q^\mu} \right).$$

To pick out individual components in the basis vector, write:

$$(a_{(\mu)})^\nu = \frac{\partial q^\nu}{\partial q^\mu} = \delta_\nu^\mu$$

As a type (0, 1) tensor, the basis vector $a_{(\mu)}$ adheres to the transformation law:

$$a_{(\mu')} = \frac{\partial q^\mu}{\partial q^{\mu'}} a_{(\mu)}$$

Dual Line Element

For a second approach to the problem one can imagine down-shifting the index on dq^μ , thereby up-shifting the index on $a_{(\mu)}$. With the calculation structurally the same, store the result in a vector dS^T , which is the dual vector to dS , via

$$dS^T = \frac{\partial S}{\partial q^1} dq_1 + \frac{\partial S}{\partial q^2} dq_2 + \dots + \frac{\partial S}{\partial q^N} dq_N,$$

or

$$dS^T = a^{(\mu)} dq_\mu,$$

suggesting:

$$a^{(\mu)} = \frac{\partial S}{\partial q_\mu} = \partial^\mu S$$

2.3 Metric Tensor

In terms of generalized basis vectors, we can write a generalized product

$$a_{(\mu)} \cdot a^{(\nu)} = \frac{\partial S}{\partial q^\mu} \cdot \frac{\partial S}{\partial q_\nu},$$

which uses the notation for the dot product for illustration. Correspondingly, we have:

$$\begin{aligned} a_{(\mu)} \cdot a_{(\nu)} &= \frac{\partial S}{\partial q^\mu} \cdot \frac{\partial S}{\partial q^\nu} \\ a^{(\mu)} \cdot a^{(\nu)} &= \frac{\partial S}{\partial q_\mu} \cdot \frac{\partial S}{\partial q_\nu} \end{aligned}$$

Under the coordinate transformation $q^\mu \rightarrow q^{\mu'}$, the above is written

$$a_{(\mu')} \cdot a^{(\nu')} = \frac{\partial S}{\partial q^{\mu'}} \cdot \frac{\partial S}{\partial q_{\nu'}},$$

and similar applies for the other two. Treating each term on the right as a type-1 tensor, we further have:

$$\begin{aligned} a_{(\mu')} \cdot a^{(\nu')} &= \left(\frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q_\nu}{\partial q_{\nu'}} \right) \frac{\partial S}{\partial q^\mu} \cdot \frac{\partial S}{\partial q_\nu} \\ a_{(\mu')} \cdot a^{(\nu')} &= \left(\frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q_\nu}{\partial q_{\nu'}} \right) a_{(\mu)} \cdot a^{(\nu)} \end{aligned}$$

Evidently, the quantity $\mathbf{a}_{(\mu)} \cdot \mathbf{a}^{(\nu)}$ is a bona fide type $(1, 1)$ tensor as a matter of definition.

Similar exercises are straightforward for the other two tensor products. In summary, we conclude:

$$\text{type } (2, 0) = \mathbf{a}^{(\mu)} \cdot \mathbf{a}^{(\nu)}$$

$$\text{type } (1, 1) = \mathbf{a}_{(\mu)} \cdot \mathbf{a}^{(\nu)}$$

$$\text{type } (0, 2) = \mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}$$

The Metric

Any of the type-2 tensors above can be called the *metric tensor*, or ‘the metric’ for shorthand, designated g , for the manifold S . Explicitly, each form is given by:

$$g^{\mu\nu} = \mathbf{a}^{(\mu)} \cdot \mathbf{a}^{(\nu)}$$

$$g_{\mu}^{\nu} = \mathbf{a}_{(\mu)} \cdot \mathbf{a}^{(\nu)}$$

$$g_{\mu\nu} = \mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}$$

The ‘mixed’ type $(1, 1)$ metric tensor g_{μ}^{ν} is almost always synonymous with the Kronecker delta δ_{μ}^{ν} . With this distinction in mind, we’ll use the form g_{μ}^{ν} in calculations for extra clarity.

Using the g notation, the transformation law for the metric is:

$$g_{\mu'\nu'} = \frac{\partial q^{\mu}}{\partial q^{\mu'}} \frac{\partial q^{\nu}}{\partial q^{\nu'}} g_{\mu\nu}$$

Metric Symmetry

By construction, namely commutation of the dot product, the metric tensor $g^{\mu\nu}$ is symmetric, which means any pair of indices can be swapped and the components remain the same:

$$g^{\mu\nu} = g^{\nu\mu}$$

$$g_{\mu\nu} = g_{\nu\mu}$$

Jacobian Matrix

The derivative $\partial q^{\mu} / \partial q^{\mu'}$ is the $\mu\mu'$ th component of the standard Jacobian matrix. Switching momentarily to matrix notation, we can write

$$\mathbf{a}_{(\mu')} = J_{\mu'}^{\mu} \mathbf{a}_{(\mu)},$$

and correspondingly,

$$\mathbf{a}_{(\nu')} = J_{\nu'}^{\nu} \mathbf{a}_{(\nu)}.$$

Combining the two, we have

$$\mathbf{a}_{(\mu')} \cdot \mathbf{a}_{(\nu')} = \left(J_{\mu'}^{\mu} J_{\nu'}^{\nu} \right) \mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)},$$

or

$$g_{\mu'\nu'} = \left(J_{\mu'}^{\mu} J_{\nu'}^{\nu} \right) g_{\mu\nu},$$

equivalent to the index-free notation

$$g' = J^T g J.$$

Next we need the determinant of both sides using the general identity

$$\det(ABA^T) = (\det A)^2 (\det B),$$

which for us means

$$\det(g') = (\det J)^2 (\det g),$$

or

$$|\det J| = \frac{\sqrt{|g'|}}{\sqrt{|g|}}.$$

If the original metric is Cartesian, then $g_{\mu\nu} = \delta_{\mu\nu}$ and $g = 1$.

2.4 Gradient

Consider a scalar field $f = f(q^{\mu})$ that depends on the generalized coordinates q^{μ} that cover the manifold S . If $f(q^{\mu})$ is differentiable, we define the *gradient* of $f(q^{\mu})$ as:

$$\nabla f = \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^{\mu}}$$

The gradient operator ∇ only makes sense in terms of the basis vector $\mathbf{a}^{(\mu)}$. One can attain a vector-free version of the above by distributing $\mathbf{a}_{(\alpha)}$ into each side:

$$\mathbf{a}_{(\alpha)} \cdot \nabla f = \mathbf{a}_{(\alpha)} \cdot \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^{\mu}}$$

Recognizing g_{α}^{μ} on the right, we further have

$$\mathbf{a}_{(\mu)} \cdot \nabla f = \frac{\partial f}{\partial q^{\mu}},$$

and the left side is given the shorthand $\partial_{\mu} f$. This suggests a clean vector-free notation

$$\partial_{\mu} f = \frac{\partial f}{\partial q^{\mu}},$$

also known as the gradient.

The gradient $\partial_\mu f$ is a type $(0, 1)$ tensor, i.e. a covariant vector, which in this case is synonymous with a vector field. To check this, we should have by definition that

$$\partial_{\mu'} f = \frac{\partial q^\mu}{\partial q^{\mu'}} \partial_\mu f ,$$

which in traditional notation unpacks to

$$\frac{\partial f}{\partial q^{\mu'}} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial f}{\partial q^\mu} .$$

This is simply a statement of the chain rule and is automatically true, therefore $\partial_\mu f$ is indeed a tensor.

Covariant Gradient

The gradient $\partial_\mu f = \partial f / \partial q^\mu$ may be denoted V_μ such that

$$V_\mu = \partial_\mu f .$$

As a covariant $(0, 1)$ tensor, V_μ satisfies the definition of a tensor when imposing the change of coordinates $q^\mu \rightarrow q^{\mu'}$:

$$V_{\mu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} V_\mu$$

Standard vector notation can be restored by tacking on the basis vectors $\mathbf{a}^{(\mu)}$. Note that the covariant form V_μ is a dual vector, thus we assign appropriate notation:

$$\mathbf{V}^T = \mathbf{a}^{(\mu)} V_\mu$$

The low index on V_μ reminds that the gradient is a dual vector.

Problem 3

Let V_x, V_y be the Cartesian components of the gradient of the function

$$f(x, y) = \ln(\sqrt{x^2 + y^2}) .$$

Calculate the gradient in Cartesian coordinates and convert the result to polar coordinates to write V_r, V_θ . Answer:

$$V_x = \frac{x}{x^2 + y^2}$$

$$V_y = \frac{y}{x^2 + y^2}$$

$$V_r = \frac{\partial x}{\partial r} V_x + \frac{\partial y}{\partial r} V_y = \frac{x \cos(\theta) + y \sin(\theta)}{x^2 + y^2} = \frac{1}{r}$$

$$V_\theta = \frac{\partial x}{\partial \theta} V_x + \frac{\partial y}{\partial \theta} V_y = \frac{-yx + xy}{x^2 + y^2} = 0$$

Contravariant Gradient

A variation on the standard gradient raises the index on the derivative subscript, to give a type $(1, 0)$ contravariant tensor

$$\partial^\mu f = \frac{\partial f}{\partial q_\mu} .$$

As a type $(1, 0)$ tensor, the gradient is represented as the vector $\mathbf{V}$ such that

$$\mathbf{V} = \mathbf{a}_{(\mu)} V^\mu .$$

Reconciliation between $\mathbf{V}^T$ and $\mathbf{V}$ is done using tensor products:

$$\begin{aligned} \mathbf{V}^T \cdot \mathbf{V} &= (\mathbf{a}^{(\mu)} \cdot \mathbf{a}_{(\nu)}) V_\mu V^\nu \\ &= (g_\nu^\mu V_\mu) V^\nu \\ &= V_\nu V^\nu \\ &= \|\mathbf{V}\|^2 \end{aligned}$$

2.5 Differential Interval

In terms of

$$\begin{aligned} dS &= \mathbf{a}_{(\mu)} dq^\mu \\ dS^T &= \mathbf{a}^{(\mu)} dq_\mu , \end{aligned}$$

the norm of the differential line element yields the *differential interval* on the manifold:

$$dS^2 = \|dS\|^2 = (\mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}) dq^\mu dq_\nu$$

As a scalar product, the differential interval is invariant with respect to a coordinate transformation, for instance:

$$dS^2 = dq^\mu dq_\mu = dq^{\mu'} dq_{\mu'}$$

Utilizing (1,1) Metric

In terms of the type $(1, 1)$ metric tensor, the differential interval is

$$dS^2 = (g_\mu^\nu dq^\mu) dq_\nu .$$

The parenthesized term can *only* be equal to dq^ν , thus we see g_μ^ν behaving as a Kronecker delta:

$$g_\mu^\nu dq^\mu = dq^\nu$$

Conversely, we could have written

$$dS^2 = dq^\mu (g_\mu^\nu dq_\nu) ,$$

where the parenthesized term can only be dq_μ , thus:

$$g_\mu^\nu dq_\nu = dq_\mu$$

For a consistency check, use the invariance of dS^2 to write

$$\begin{aligned} dS^2 &= g_\mu^\nu dq^\mu dq_\nu \\ &= g_\mu^\nu (g_\alpha^\mu dq^\alpha) (g_\beta^\nu dq_\beta) \\ &= (g_\alpha^\mu g_\mu^\nu g_\beta^\nu) dq^\alpha dq_\beta. \end{aligned}$$

The quantity dS^2 remains invariant if

$$g_\alpha^\mu g_\mu^\nu g_\beta^\nu = g_\alpha^\beta,$$

availing another Kronecker delta-like identity:

$$g_\alpha^\mu g_\nu^\alpha = g_\nu^\mu$$

Lowering an Index

Another formula for the differential interval can be made from only contravariant lengths:

$$dS^2 = (\mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}) dq^\mu dq^\nu,$$

implying

$$dS^2 = g_{\mu\nu} dq^\mu dq^\nu.$$

From the above, we reason that

$$\begin{aligned} g_{\mu\nu} dq^\mu &= dq_\nu \\ g_{\mu\nu} dq^\nu &= dq_\mu \end{aligned}$$

are mutually true. Evidently, the metric tensor $g_{\mu\nu}$ applied to a contravariant form returns a covariant form. This maneuver is called ‘lowering an index’.

Raising an Index

Yet another formula for the differential interval leads to

$$dS^2 = g^{\mu\nu} dq_\mu dq_\nu.$$

From this, we decide

$$\begin{aligned} g^{\mu\nu} dq_\mu &= dq^\nu \\ g^{\mu\nu} dq_\nu &= dq^\mu \end{aligned}$$

must also hold. That is, the metric tensor $g^{\mu\nu}$ applied to a covariant form returns a contravariant form. This is ‘raising an index’.

Condensation

Another identity involving the metric starts with $dS^2 = dq^\alpha dq_\alpha$ and shifting the vertical position of each index on the dq -terms. For this we have

$$dS^2 = (g^{\mu\alpha} g_{\alpha\nu}) dq_\mu dq^\nu,$$

which brings out another identity synonymous to the Kronecker delta:

$$g^{\mu\alpha} g_{\alpha\nu} = g_\nu^\mu$$

2.6 Metric-Tensor Interactions

Contractions with Vectors

Recall that for two vectors A, B , the scalar product w is defined as

$$w = A_\mu B^\mu.$$

Something we should check is whether the metric tensor successfully changes the index position on the vectors A, B in addition to displacements dq . For this, we postulate:

$$\begin{aligned} A_\mu &= g_{\mu\nu} A^\nu \\ B^\mu &= g^{\mu\nu} B_\nu \end{aligned}$$

Substituting each into the formula for w , one finds

$$w = (g_{\mu\alpha} g^{\mu\nu}) A^\alpha B_\nu = g_\alpha^\nu A^\alpha B_\nu = A^\nu B_\nu.$$

The original w is restored provided that

$$A_\mu B^\mu = A^\mu B_\mu,$$

along with:

$$\begin{aligned} A^\nu &= g_\alpha^\nu A^\alpha \\ B_\alpha &= g_\alpha^\nu B_\nu \end{aligned}$$

This is important news, as we see that the notion of shifting, raising, and lowering indices applies to any type-1 tensor.

To wring out another pair of identities, write for any vector A :

$$\begin{aligned} A_\mu &= g_{\mu\alpha} A^\alpha = (g_{\mu\alpha} g_\nu^\alpha) A^\nu \\ A^\mu &= g^{\mu\alpha} A_\alpha = (g^{\mu\alpha} g_\alpha^\nu) A_\nu \end{aligned}$$

The parenthesized quantities can only be $g_{\mu\nu}, g^{\mu\nu}$, respectively. Evidently then:

$$\begin{aligned} g_{\mu\nu} &= g_{\mu\alpha} g_\nu^\alpha \\ g^{\mu\nu} &= g^{\mu\alpha} g_\alpha^\nu \end{aligned}$$

Contractions with Tensors

The metric can be used to calculate the contravariant gradient of a scalar field $f(q^\mu)$ from the standard (covariant) gradient:

$$\partial^\mu f = g^{\mu\nu} \partial_\nu f$$

The same is true in reverse:

$$\partial_\mu f = g_{\mu\nu} \partial^\nu f$$

When a tensor undergoes contraction, only the common index plays into the calculation and the other indices come for the ride. For instance, if A is a type-3 tensor, we can have any of

$$\begin{aligned} A_k^{\mu j} &= g^{\mu\nu} A_{\nu k}^j \\ A_k^{\mu j} &= g_\nu^\mu A_k^{\nu j} \\ A_{\mu k}^j &= g_{\mu\nu} A_k^{\nu j}, \end{aligned}$$

and so on.

Cartesian and Polar Coordinates

In plane polar coordinates, it's straightforward to show that the nonzero components of the metric are

$$\begin{aligned} g_{rr} &= 1 \\ g_{\theta\theta} &= r^2, \end{aligned}$$

and correspondingly

$$\begin{aligned} g^{rr} &= 1 \\ g^{\theta\theta} &= 1/r^2. \end{aligned}$$

The metric is especially useful for transforming a vector V between coordinate systems. Previously, we found

$$\begin{aligned} V_x &= \frac{\partial r}{\partial x} V_r + \frac{\partial \theta}{\partial x} V_\theta \\ V_y &= \frac{\partial r}{\partial y} V_r + \frac{\partial \theta}{\partial y} V_\theta, \end{aligned}$$

along with

$$\begin{aligned} V^r &= \frac{\partial r}{\partial x} V^x + \frac{\partial r}{\partial y} V^y \\ V^\theta &= \frac{\partial \theta}{\partial x} V^x + \frac{\partial \theta}{\partial y} V^y, \end{aligned}$$

both requiring the inversion of the standard parameterization

$$\begin{aligned} x &= r \cos(\theta) \\ y &= r \sin(\theta). \end{aligned}$$

The r, θ inversion can be avoided by first calculating the 'easy' transformations

$$\begin{aligned} V_r &= \frac{\partial x}{\partial r} V_x + \frac{\partial y}{\partial r} V_y \\ &= \cos(\theta) V_x + \sin(\theta) V_y \\ V_\theta &= \frac{\partial x}{\partial \theta} V_x + \frac{\partial y}{\partial \theta} V_y \\ &= -r \sin(\theta) V_x + r \cos(\theta) V_y. \end{aligned}$$

To proceed, we note $V^x = V_x$, $V^y = V_y$, and then V^r , V^θ are calculated using the metric:

$$\begin{aligned} V^r &= g^{rr} V_r \\ V^\theta &= g^{\theta\theta} V_\theta \end{aligned}$$

Thus we find

$$\begin{aligned} V^r &= V_r = \cos(\theta) V^x + \sin(\theta) V^y \\ V^\theta &= \frac{1}{r^2} V_\theta = -\frac{\sin(\theta)}{r} V^x + \frac{\cos(\theta)}{r} V^y. \end{aligned}$$

2.7 Tensor Symmetry

A tensor is *symmetric* in a pair of indices if it obeys

$$A^{\mu\nu} = A^{\nu\mu},$$

and is *antisymmetric* if it obeys

$$A^{\mu\nu} = -A^{\nu\mu}.$$

Note the above comments apply to tensors of any type (p, q) . Symmetry or antisymmetry need only exist in a set of indices, not in the entire tensor.

A general tensor may be expressed in terms of symmetric and antisymmetric parts via

$$A^{\mu\nu} = \frac{1}{2} (A^{\mu\nu} + A^{\nu\mu}) + \frac{1}{2} (A^{\mu\nu} - A^{\nu\mu}).$$

Introducing the condensed notation

$$\begin{aligned} A^{(\mu\nu)} &= \frac{1}{2} (A^{\mu\nu} + A^{\nu\mu}) \\ A^{[\mu\nu]} &= \frac{1}{2} (A^{\mu\nu} - A^{\nu\mu}), \end{aligned}$$

the above can be written

$$A^{\mu\nu} = A^{(\mu\nu)} + A^{[\mu\nu]}.$$

2.8 Minkowski Metric

A special case of the metric tensor g is the four-dimensional *Minkowski metric*, written η , defined by

$$\begin{aligned}\eta_{00} &= -1 \\ \eta_{11} &= \eta_{22} = \eta_{33} = 1 \\ \eta_{\mu\nu} &= 0 \text{ if } \mu \neq \nu\end{aligned}$$

Akin to a diagonal matrix, off-diagonal components of the flat metric $\eta_{\mu\mu}$ are zero. The 00th component is -1 by convention. All other components on the diagonal are 1.

The same description for η works for the $(0, 2)$ version $\eta^{\mu\nu}$. Having the same utility as g to raise and lower indices, write

$$\eta^{\mu\nu} = (\eta^{\mu\alpha} \eta_{\beta\nu}) \eta_{\alpha\beta}.$$

Since the off-diagonal components of η are zero, it quickly follows that $\mu = \alpha$ and $\beta = \nu$. The parenthesized term can only be $+1$, and we conclude $\eta^{\mu\nu}$ and $\eta_{\alpha\beta}$ are numerically the same.

Worked Examples

Consider two tensors $X^{\mu\nu}$ and V^μ given by

$$\begin{aligned}X^{\mu\nu} &= \begin{bmatrix} X^{00} & X^{01} & X^{02} & X^{03} \\ X^{10} & X^{11} & X^{12} & X^{13} \\ X^{20} & X^{21} & X^{22} & X^{23} \\ X^{30} & X^{31} & X^{32} & X^{33} \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 1 & -1 \\ -1 & 0 & 3 & 2 \\ -1 & 1 & 0 & 0 \\ -2 & 1 & 1 & -2 \end{bmatrix} \\ V^\mu &= (-1, 2, 0, -2),\end{aligned}$$

respectively. Find the quantities (i) X^μ_ν , (ii) X^ν_μ , (iii) $X^{(\mu\nu)}$, (iv) $X_{[\mu\nu]}$, (v) X^λ_λ , (vi) $V^\mu V_\mu$, (vii) $V_\mu X^{\mu\nu}$.

Solution (i): Begin with

$$\eta_{\alpha\nu} X^{\mu\alpha} = X^\mu_\nu$$

and let $\nu = \mu$. The above becomes $\eta_{\mu\mu} X^{\mu\mu} = X^\mu_\mu$, thus the 00-component on the diagonal flips sign (namely from -2 to 2), and the components along the rest of the diagonal stay the same.

Let j, k be nonzero indices and replace $\mu = j, \nu = k$ to find $X^{jk} = X^j_k$, which says the lower 3×3 body of the result X^μ_ν is the same as that of the original $X^{\mu\nu}$.

For the left-most column with $\nu = 0$, the above reduces to $\eta_{00} X^{\mu 0} = X^\mu_0$. This reaffirms that the 00-component switches sign, but also tells us the terms X^μ_0 are the negations of $X^{\mu 0}$.

For the top-most row with $\mu = 0, \nu = k$ we get $X^{0k} = X^0_k$. This means all terms in the top row except the 00-component retain sign. All together, we conclude:

$$X^\mu_\nu = \begin{bmatrix} -2 & 0 & 1 & -1 \\ 1 & 0 & 3 & 2 \\ 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & -2 \end{bmatrix}$$

Solution (ii): Begin with

$$\eta_{\mu\alpha} X^{\alpha\nu} = X^\nu_\mu$$

and let $\nu = \mu$. The above becomes $\eta_{\mu\mu} X^{\mu\mu} = X^\mu_\mu$, thus the 00-component on the diagonal flips sign, and the components along the rest of the diagonal stay the same.

Let j, k be nonzero indices and replace $\mu = j, \nu = k$ to find $X^{jk} = X^k_j$, which says the lower 3×3 body of the result X^ν_μ is the same as that of the original $X^{\mu\nu}$.

For the left column with $\nu = 0, \mu = j$ we get $X^{j0} = X^0_j$. This means all terms in the left column don't change sign.

For the top row with $\mu = 0$, the above reduces to $\eta_{00} X^{0\nu} = X^\nu_0$. This reaffirms that the 00-component switches sign, but also tells us the terms X^ν_0 are the negations of $X^{0\nu}$. All together, we conclude:

$$X^\nu_\mu = \begin{bmatrix} -2 & 0 & -1 & 1 \\ -1 & 0 & 3 & 2 \\ -1 & 1 & 0 & 0 \\ -2 & 1 & 1 & -2 \end{bmatrix}$$

Solution (iii): The transpose of $X^{\mu\nu}$, particularly $X^{\nu\mu}$, is written:

$$X^{\nu\mu} = \begin{bmatrix} 2 & -1 & -1 & -2 \\ 0 & 0 & 1 & 1 \\ 1 & 3 & 0 & 1 \\ -1 & 2 & 0 & -2 \end{bmatrix}$$

By the formula

$$A^{(\mu\nu)} = \frac{1}{2} (A^{\mu\nu} + A^{\nu\mu}),$$

we calculate:

$$X^{(\mu\nu)} = \begin{bmatrix} 2 & -1/2 & 0 & -3/2 \\ -1/2 & 0 & 2 & 3/2 \\ 0 & 2 & 0 & 1/2 \\ -3/2 & 3/2 & 1/2 & -2 \end{bmatrix}$$

Solution (iv): Lower the indices on $X^{\mu\nu}$ via

$$X_{\mu\nu} = \eta_{\mu\alpha}\eta_{\beta\nu}X^{\alpha\beta}.$$

To avoid a brute-force calculation let j, k be nonzero indices and replace $\mu = j, \nu = k$ to find $X_{jk} = X^{jk}$, which says the lower 3×3 body of the result $X_{\mu\nu}$ is the same as that of the original $X^{\mu\nu}$.

For the left column with $\nu = 0$, the above tells us $X_{\mu 0} = -\eta_{\mu\alpha}X^{\alpha 0}$. For $\mu = 0$ we quickly find $X_{00} = X^{00}$, so the 00-component doesn't change. Instead taking $\mu = j > 0$, we have $X_{j0} = -X^{j0}$, meaning the terms in the left column (besides 00) change sign.

For the top row with $\mu = 0$, the above tells us $X_{0\nu} = -\eta_{b\nu}X^{0b}$. Letting $\nu = k > 0$, we have $X_{0k} = -X^{0k}$, meaning the terms in the top row (besides 00) change sign.

Putting it all together, we find

$$X_{\mu\nu} = \begin{bmatrix} 2 & 0 & -1 & 1 \\ 1 & 0 & 3 & 2 \\ 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & -2 \end{bmatrix},$$

along with:

$$X_{\nu\mu} = \begin{bmatrix} 2 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ -1 & 3 & 0 & 1 \\ 1 & 2 & 0 & -2 \end{bmatrix}$$

By the formula

$$A_{[\mu\nu]} = \frac{1}{2} (A_{\mu\nu} - A_{\nu\mu}),$$

we calculate:

$$X_{[\nu\mu]} = \begin{bmatrix} 0 & -1/2 & -1 & -1/2 \\ 1/2 & 0 & 1 & 1/2 \\ 1 & -1 & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 0 \end{bmatrix}$$

Solution (v): Begin with

$$\eta_{\mu\alpha}X^{\alpha\nu} = X_{\mu}^{\nu}$$

and let $\lambda = \nu = \mu$ to write $\eta_{\lambda\alpha}X^{\alpha\lambda} = X_{\lambda}^{\lambda}$, which unpacks as:

$$\begin{aligned} X_{\lambda}^{\lambda} &= \eta_{00}X^{00} + \eta_{11}X^{11} + \eta_{22}X^{22} + \eta_{33}X^{33} \\ &= -2 + 0 + 0 - 2 = -4 \end{aligned}$$

Solution (vi): First calculate the components of V_{μ} using $V_{\mu} = g_{\mu\alpha}V^{\alpha}$ to find

$$V_{\mu} = (1, 2, 0, -2).$$

Then by the Einstein sum convention, we find

$$V^{\mu}V_{\mu} = -1 + 4 + 0 + 4 = 7.$$

Solution (vii):

$$\begin{aligned} V_{\mu}X^{\mu\nu} &= \begin{cases} V_0X^{00} + V_1X^{10} + V_2X^{20} + V_3X^{30} = 4 \\ V_0X^{01} + V_1X^{11} + V_2X^{21} + V_3X^{31} = -2 \\ V_0X^{02} + V_1X^{12} + V_2X^{22} + V_3X^{32} = 5 \\ V_0X^{03} + V_1X^{13} + V_2X^{23} + V_3X^{33} = 7, \end{cases} \end{aligned}$$

or:

$$V_{\mu}X^{\mu\nu} = (4, -2, 5, 7)$$

3 Covariant Derivative

Let $\mathbf{S}$ denote the position vector on an N -dimensional manifold parameterized by generalized coordinates $q^1, q^2, q^3, \dots, q^N$.

3.1 Two Gradients Problem

Given how easily the gradient operator turns a scalar field into a type-1 tensor, it's natural to try the gradient twice to see if a type (0, 2) tensor comes out.

Starting with one gradient

$$\partial_{\nu'} f = \frac{\partial q^\nu}{\partial q^{\nu'}} \partial_\nu f,$$

apply another gradient $\partial_{\mu'}$ to both sides:

$$\partial_{\mu'} \partial_{\nu'} f = \frac{\partial q^\nu}{\partial q^{\mu'}} \partial_{\mu'} \partial_\nu f + \partial_{\mu'} \left(\frac{\partial q^\nu}{\partial q^{\nu'}} \right) \partial_\nu f$$

Simplifying with the chain rule, the above reads

$$\partial_{\mu'} \partial_{\nu'} f = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \partial_\mu \partial_\nu f + \partial_{\mu'} \left(\frac{\partial q^\nu}{\partial q^{\nu'}} \right) \partial_\nu f,$$

which contains two terms.

Alternatively, the gradient $\partial_\nu f$ can be written V_ν and the above becomes:

$$\partial_{\mu'} V_{\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \partial_\mu V_\nu + \partial_{\mu'} \left(\frac{\partial q^\nu}{\partial q^{\nu'}} \right) V_\nu$$

The same pattern arises from the contravariant version of V . Starting with the derivative $\partial_{\nu'} V^{\mu'}$, use the definition of a tensor on V to write

$$\partial_{\nu'} V^{\mu'} = \partial_{\nu'} \left(\frac{\partial q^{\mu'}}{\partial q^\mu} V^\mu \right),$$

expanding to, using the chain rule:

$$\partial_{\nu'} V^{\mu'} = \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \partial_\nu V^\mu + \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial^2 q^{\mu'}}{\partial q^\nu \partial q^\mu} V^\mu$$

From these calculations we see that two successive gradients do *not* result in a tensor, otherwise the second term would not be present on the right. The non-tensorial terms contain second derivatives of the position q^μ , analogous to an acceleration-like term that points to the center of curvature, which may be outside the tangent space defined by the basis vectors.

Antisymmetrized Gradient

In light of the two-gradient calculation not resulting in a tensor, one could wonder if some configuration of operators gets a the job done, where the non-tensorial term can be somehow subtracted off. Pursuing this, consider a quantity

$$B_{\mu\nu} = \partial [\mu V_\nu] = \frac{1}{2} (\partial_\mu V_\nu - \partial_\nu V_\mu),$$

and apply the change of coordinates $q^\mu \rightarrow q^{\mu'}$:

$$B_{\mu'\nu'} = \partial [\mu' V_{\nu'}] = \frac{1}{2} (\partial_{\mu'} V_{\nu'} - \partial_{\nu'} V_{\mu'})$$

We already know how the first term turns out. As for the second term, swap $\mu' \leftrightarrow \nu'$ and apply a negative sign:

$$-\partial_{\nu'} V_{\mu'} = -\frac{\partial q^\mu}{\partial q^{\nu'}} \frac{\partial q^\nu}{\partial q^{\mu'}} \partial_\mu V_\nu - \partial_{\nu'} \left(\frac{\partial q^\nu}{\partial q^{\mu'}} \right) V_\nu$$

Note that the second term in the above is the exact negative of the troublesome term in $\partial_{\mu'} V_{\nu'}$. Therefore the result $B_{\mu'\nu'}$ reads

$$B_{\mu'\nu'} = \frac{1}{2} \left(\frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \partial_\mu V_\nu - \frac{\partial q^\mu}{\partial q^{\nu'}} \frac{\partial q^\nu}{\partial q^{\mu'}} \partial_\mu V_\nu \right).$$

In the second term, the indices μ, ν can be swapped without changing $B_{\mu'\nu'}$, so

$$\begin{aligned} B_{\mu'\nu'} &= \frac{1}{2} \left(\frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \partial_\mu V_\nu - \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^\mu}{\partial q^{\mu'}} \partial_\nu V_\mu \right) \\ &= \left(\frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \right) \frac{1}{2} (\partial_\mu V_\nu - \partial_\nu V_\mu) \\ &= \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} B_{\mu\nu}, \end{aligned}$$

therefore $B_{\mu\nu} = \partial [\mu V_\nu]$ is a type (0, 2) tensor.

3.2 Connection Coefficients

For a given vector $\mathbf{V} = V^\mu \mathbf{a}_{(\mu)}$, the 'two gradients' problem means that the standard derivative of $\mathbf{V}$ is not a tensor. Setting up a similar calculation, write the differential of $\mathbf{V}$ as

$$d\mathbf{V} = d(V^\mu \mathbf{a}_{(\mu)}) = dV^\mu \mathbf{a}_{(\mu)} + V^\mu d\mathbf{a}_{(\mu)},$$

and apply the chain rule to get

$$d\mathbf{V} = \frac{\partial V^\mu}{\partial q^\nu} dq^\nu \mathbf{a}_{(\mu)} + V^\mu \frac{\partial \mathbf{a}_{(\mu)}}{\partial q^\nu} dq^\nu.$$

The problematic second term shall be coerced to remain in the $\mathbf{a}_{(\mu)}$ basis while remaining proportional to V^μ . For this sleight of hand we need new proportionality constants called *connection coefficients*, denoted Γ , so we take:

$$\frac{\partial \mathbf{a}_{(\mu)}}{\partial q^\nu} = \Gamma_{\nu\mu}^\gamma \mathbf{a}_{(\gamma)}$$

Transformation Properties

Imposing a change of coordinates, the above is written

$$\frac{\partial \mathbf{a}_{(\mu')}}{\partial q^{\nu'}} = \Gamma_{\nu'\mu'}^{\gamma'} \mathbf{a}_{(\gamma')}.$$

Substituting what we know about the basis vectors in transformed coordinates, the above is

$$\frac{\partial}{\partial q^{\nu'}} \left(\frac{\partial q^\alpha}{\partial q^{\mu'}} \mathbf{a}_{(\alpha)} \right) = \Gamma_{\nu'\mu'}^{\gamma'} \frac{\partial q^\alpha}{\partial q^{\gamma'}} \mathbf{a}_{(\alpha)}.$$

Simplifying carefully, one finds the transformation law for connection coefficients:

$$\Gamma_{\nu'\mu'}^{\gamma'} = \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial q^\beta}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \Gamma_{\nu\beta}^\alpha + \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial^2 q^\alpha}{\partial q^{\nu'} \partial q^{\mu'}}$$

An alternative to the above can also be derived. Begin with a statement of the chain rule

$$\frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial q^\alpha}{\partial q^{\mu'}} = \frac{\partial q^{\gamma'}}{\partial q^{\mu'}} = \delta_{\mu'}^{\gamma'},$$

and differentiate with respect to $q^{\nu'}$, giving

$$\frac{\partial}{\partial q^{\nu'}} \left(\frac{\partial q^{\gamma'}}{\partial q^\alpha} \right) \frac{\partial q^\alpha}{\partial q^{\mu'}} + \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial^2 q^\alpha}{\partial q^{\nu'} \partial q^{\mu'}} = 0,$$

and use the chain rule to find

$$\frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial^2 q^\alpha}{\partial q^{\nu'} \partial q^{\mu'}} = - \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial^2 q^{\gamma'}}{\partial q^\nu \partial q^\alpha} \frac{\partial q^\alpha}{\partial q^{\mu'}}.$$

Notice the left side of this is present in the transformation law for connection coefficients. Eliminating the common term gives an alternative:

$$\Gamma_{\nu'\mu'}^{\gamma'} = \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial q^\beta}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \Gamma_{\nu\beta}^\alpha - \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial^2 q^{\gamma'}}{\partial q^\nu \partial q^\alpha} \frac{\partial q^\alpha}{\partial q^{\mu'}}$$

There is a tendency for indices to become a mess during derivations. Cleaning up the above using

$$\begin{aligned} \gamma' &\rightarrow \alpha' \\ \nu' &\rightarrow \beta' \\ \mu' &\rightarrow \gamma', \end{aligned}$$

while relabeling dummy indices, the same transformation law reads:

$$\Gamma_{\beta'\gamma'}^{\alpha'} = \frac{\partial q^{\alpha'}}{\partial q^\mu} \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \Gamma_{\nu\gamma}^\mu - \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma}$$

To summarize, we see that Γ is not a tensor, else the second term wouldn't be present. Nonetheless, we now see what needs to happen with Γ with a coordinate transformation.

Torsion

Define the *torsion* T as:

$$T_{\mu\nu}^\alpha = \Gamma_{\mu\nu}^\alpha - \Gamma_{\nu\mu}^\alpha$$

Despite Γ not being a tensor, it turns out the torsion satisfies the definition of an asymmetric type (1, 2) tensor. Using the transformation law, we write:

$$\begin{aligned} T_{\beta'\gamma'}^{\alpha'} &= \Gamma_{\beta'\gamma'}^{\alpha'} - \Gamma_{\gamma'\beta'}^{\alpha'} \\ &= \frac{\partial q^{\alpha'}}{\partial q^\mu} \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \Gamma_{\nu\gamma}^\mu - \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma} \\ &\quad - \frac{\partial q^{\alpha'}}{\partial q^\mu} \frac{\partial q^\nu}{\partial q^{\gamma'}} \frac{\partial q^\gamma}{\partial q^{\beta'}} \Gamma_{\nu\gamma}^\mu + \frac{\partial q^\nu}{\partial q^{\gamma'}} \frac{\partial q^\gamma}{\partial q^{\beta'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma} \end{aligned}$$

The second-order terms cancel out from the above, leaving

$$T_{\beta'\gamma'}^{\alpha'} = \frac{\partial q^{\alpha'}}{\partial q^\mu} \left(\frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \Gamma_{\nu\gamma}^\mu - \frac{\partial q^\nu}{\partial q^{\gamma'}} \frac{\partial q^\gamma}{\partial q^{\beta'}} \Gamma_{\nu\gamma}^\mu \right),$$

where swapping the indices ν, γ in the last term finishes the hard work

$$T_{\beta'\gamma'}^{\alpha'} = \frac{\partial q^{\alpha'}}{\partial q^\mu} \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} (\Gamma_{\nu\gamma}^\mu - \Gamma_{\gamma\nu}^\mu),$$

as the parenthesized pair is just $T_{\nu\gamma}^\mu$ again.

3.3 Covariant Derivative

In terms of connection coefficients, the differential

$$d\mathbf{V} = \frac{\partial V^\mu}{\partial q^\nu} dq^\nu \mathbf{a}_{(\mu)} + V^\mu \frac{\partial \mathbf{a}_{(\mu)}}{\partial q^\nu} dq^\nu$$

of the vector $\mathbf{V} = V^\mu \mathbf{a}_{(\mu)}$ becomes

$$d\mathbf{V} = \frac{\partial V^\mu}{\partial q^\nu} dq^\nu \mathbf{a}_{(\mu)} + \Gamma_{\nu\mu}^\gamma V^\mu dq^\nu \mathbf{a}_{(\gamma)}.$$

Swap the dummy indices μ, γ on the right and simplify to find

$$dV = (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) dq^\nu a_{(\mu)}.$$

We've uncovered a new formula for 'derivative' by replacing the non-tensorial term with the one involving the connection coefficients. For this we use the term *covariant derivative*, denoted D_ν , as

$$D_\nu V = (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) a_{(\mu)}.$$

The parenthesized quantity expresses the covariant derivative acting on the contravariant vector V^μ such that

$$D_\nu V^\mu = \partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma,$$

or in tighter notation

$$D_\nu V = (D_\nu V^\mu) a_{(\mu)},$$

and furthermore:

$$D_\nu V^\mu = a^{(\mu)} \cdot D_\nu V$$

Linearity

By construction, the covariant derivative is a linear operator, obeying

$$D_\alpha (A^\mu + B^\nu) = D_\alpha A^\mu + D_\alpha B^\nu,$$

and also the product rule:

$$D_\alpha (A^\mu B^\nu) = (D_\alpha A^\mu) B^\nu + A^\nu (D_\alpha B^\mu)$$

Alternate Motivation

The same operator can be motivated starting with the dual vector $V^T = V_\mu a^{(\mu)}$, with subsequent steps essentially the same. As a precaution we'll use a different symbol $\tilde{\Gamma}$ for the connection coefficients:

$$D_\nu V_\mu = \partial_\nu V_\mu + \tilde{\Gamma}_{\nu\mu}^\gamma V_\gamma$$

Scalar Field Derivative

A scalar field is equivalent to a type (0, 0) tensor $f = A^\mu B_\mu$. Calculating the covariant derivative of f , we find:

$$\begin{aligned} D_\nu f &= D_\nu (A^\mu B_\mu) \\ &= (D_\nu A^\mu) B_\mu + A_\mu (D_\nu B^\mu) \\ &= (\partial_\nu A^\mu + \Gamma_{\nu\gamma}^\mu A^\gamma) B_\mu + A^\mu (\partial_\nu B_\mu + \tilde{\Gamma}_{\nu\mu}^\gamma B_\gamma) \\ &= \partial_\nu f + \Gamma_{\nu\gamma}^\mu A^\gamma B_\mu + \tilde{\Gamma}_{\nu\mu}^\gamma A^\mu B_\gamma \\ &= \partial_\nu f + A^\gamma B_\mu (\Gamma_{\nu\gamma}^\mu + \tilde{\Gamma}_{\nu\gamma}^\mu) \end{aligned}$$

As a matter of technicality, the derivative of a scalar field does not involve connection coefficients and only $\partial_\mu f$ survives. With this we conclude that $\tilde{\Gamma}$ is the negative of the original Γ as

$$\Gamma_{\nu\gamma}^\mu = -\tilde{\Gamma}_{\nu\gamma}^\mu,$$

letting us rewrite:

$$D_\nu V_\mu = \partial_\nu V_\mu - \Gamma_{\nu\mu}^\gamma V_\gamma$$

Tensor Field Derivative

Consider a type (2, 0) tensor $C^{\mu\nu}$ that is the tensor product $C^{\mu\nu} = A^\mu B^\nu$. Calculating the covariant derivative of C , we find

$$\begin{aligned} D_\alpha (C^{\mu\nu}) &= D_\alpha (A^\mu B^\nu) \\ &= (D_\alpha A^\mu) B^\nu + A^\mu (D_\alpha B^\nu) \\ &= \partial_\alpha (A^\mu B^\nu) + \Gamma_{\alpha\gamma}^\mu A^\gamma B^\nu + \Gamma_{\alpha\gamma}^\nu A^\mu B^\gamma, \end{aligned}$$

or in summary:

$$D_\alpha C^{\mu\nu} = \partial_\alpha C^{\mu\nu} + \Gamma_{\alpha\gamma}^\mu C^{\gamma\nu} + \Gamma_{\alpha\gamma}^\nu C^{\mu\gamma}$$

By the same token, one can calculate the covariant derivative of the type (0, 2) tensor $C_{\mu\nu} = A_\mu B_\nu$ by a similar process, giving:

$$D_\alpha C_{\mu\nu} = \partial_\alpha C_{\mu\nu} - \Gamma_{\alpha\mu}^\gamma C_{\gamma\nu} - \Gamma_{\alpha\nu}^\gamma C_{\mu\gamma}$$

The same pattern continues for more contravariant or covariant indices. For each contravariant index, include a term proportional to $+\Gamma$. For each covariant index, include a term proportional to $-\Gamma$. Mixed cases apply both rules, for instance:

$$D_\alpha C_\nu^\mu = \partial_\alpha C_\nu^\mu + \Gamma_{\alpha\gamma}^\mu C_\nu^\gamma - \Gamma_{\alpha\nu}^\gamma C_\gamma^\mu$$

Tensor Qualification

For $D_\nu V^\mu$ to qualify as a tensor, we require

$$D_{\nu'} V^{\mu'} = \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} D_\nu V^\mu,$$

which is calculated by brute force:

$$\begin{aligned} D_{\nu'} V^{\mu'} &= \partial_{\nu'} V^{\mu'} + \Gamma_{\nu'\gamma'}^{\mu'} V^{\gamma'} \\ &= \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \partial_\nu V^\mu \\ &\quad + \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial^2 q^{\mu'}}{\partial q^\nu \partial q^\gamma} V^\gamma + \Gamma_{\nu'\gamma'}^{\mu'} \frac{\partial q^{\gamma'}}{\partial q^\gamma} V^\gamma \end{aligned}$$

Proceed by factoring $\partial q^\nu / \partial q^{\nu'}$ and $\partial q^{\mu'} / \partial q^\mu$ from each term using the chain rule. Working this out, we find

$$\begin{aligned} D_{\nu'} V^{\mu'} &= \partial_{\nu'} V^{\mu'} + \Gamma_{\nu' \gamma'}^{\mu'} V^{\gamma'} \\ &= \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \partial_\nu V^\mu \\ &\quad + \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \left(\frac{\partial q^\mu}{\partial q^{\alpha'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma} \right) V^\gamma \\ &\quad + \frac{\partial q^\nu}{\partial q^{\nu'}} \frac{\partial q^{\mu'}}{\partial q^\mu} \left(\Gamma_{\beta' \gamma'}^{\alpha'} \frac{\partial q^{\beta'}}{\partial q^\nu} \frac{\partial q^\mu}{\partial q^{\alpha'}} \frac{\partial q^{\gamma'}}{\partial q^\gamma} \right) V^\gamma. \end{aligned}$$

In order to satisfy the definition of a tensor, the two parenthesized terms must sum to $\Gamma_{\nu\gamma}^\mu$, or:

$$\Gamma_{\nu\gamma}^\mu = \frac{\partial q^\mu}{\partial q^{\alpha'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma} + \Gamma_{\beta' \gamma'}^{\alpha'} \frac{\partial q^{\beta'}}{\partial q^\nu} \frac{\partial q^\mu}{\partial q^{\alpha'}} \frac{\partial q^{\gamma'}}{\partial q^\gamma}$$

Rearrange to isolate the Γ with the primed indices:

$$\Gamma_{\beta' \gamma'}^{\alpha'} = \frac{\partial q^{\alpha'}}{\partial q^\mu} \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \Gamma_{\nu\gamma}^\mu - \frac{\partial q^\nu}{\partial q^{\beta'}} \frac{\partial q^\gamma}{\partial q^{\gamma'}} \frac{\partial^2 q^{\alpha'}}{\partial q^\nu \partial q^\gamma}$$

This result exactly matches the transformation law for the connection coefficients, finishing the proof.

3.4 Levi-Civita Connection

Local Torsion

Having established the torsion tensor

$$T_{\nu\mu}^\gamma = \Gamma_{\nu\mu}^\gamma - \Gamma_{\mu\nu}^\gamma,$$

let us now choose any point P on the manifold and examine what happens in a neighborhood local to P . Presuming the manifold is differentiable, it follows that the metric evaluated at a point corresponds to a ‘flat’ space:

$$\mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)} \Big|_P = g_{\mu\nu} \Big|_P = \eta_{\mu\nu}$$

Moreover, second derivatives of the basis vectors are zero, thus

$$\frac{\partial \mathbf{a}_{(\mu)}}{\partial q^\nu} \Big|_P = \Gamma_{\nu\mu}^\gamma \mathbf{a}_{(\gamma)} \Big|_P = 0,$$

which is to say the connection coefficients at P are zero to first order.

Restating the connection coefficient transformation law

$$\Gamma_{\nu' \mu'}^{\gamma'} = \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial q^\beta}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \Gamma_{\nu\beta}^\alpha + \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial^2 q^\alpha}{\partial q^{\nu'} \partial q^{\mu'}}$$

choose indices α, ν, β such that Γ is local to P . Then, in transformed coordinates, we find:

$$\Gamma_{\nu' \mu'}^{\gamma'} \Big|_P = \frac{\partial q^{\gamma'}}{\partial q^\alpha} \frac{\partial^2 q^\alpha}{\partial q^{\nu'} \partial q^{\mu'}}$$

Due to the symmetry in the partial derivatives above, there is symmetry in the lower indices on Γ :

$$\Gamma_{\nu' \mu'}^{\gamma'} \Big|_P = \Gamma_{\mu' \nu'}^{\gamma'} \Big|_P$$

This immediately means that the torsion at point P is zero.

Zero Torsion

The *Levi-Civita connection* has us presume that torsion is zero not only at point P , but also everywhere on the manifold. By this, the connection coefficients are symmetric in the two lower indices:

$$\Gamma_{\nu\mu}^\gamma = \Gamma_{\mu\nu}^\gamma$$

A more direct derivation of the above begins with

$$\frac{\partial \mathbf{a}_{(\mu)}}{\partial q^\nu} = \Gamma_{\nu\mu}^\gamma \mathbf{a}_{(\gamma)},$$

and then distribute $\mathbf{a}^{(\gamma)}$ into both sides to get a formula for Γ :

$$\Gamma_{\nu\mu}^\gamma = \mathbf{a}^{(\gamma)} \cdot \frac{\partial}{\partial q^\nu} \frac{\partial}{\partial q^\mu} \mathbf{S} = \mathbf{a}^{(\gamma)} \cdot \partial_\nu \mathbf{a}_{(\mu)}$$

Then, swapping the indices ν, μ should leave the result unchanged, and the same symmetry in the lower indices on Γ is implied.

Problem 4

If $A_{\mu\nu}$ is an antisymmetric type (0,2) tensor in the absence of torsion, show that all connection coefficients cancel out of $D [\mu A_{\nu\sigma}]$. Answer:

$$\begin{aligned} D [\mu A_{\nu\sigma}] &= \frac{1}{3!} \cdot \left\{ D_\mu A_{\nu\sigma} + D_\nu A_{\sigma\mu} + D_\sigma A_{\mu\nu} \right. \\ &\quad \left. - D_\mu A_{\sigma\nu} - D_\nu A_{\mu\sigma} - D_\sigma A_{\nu\mu} \right\} \\ &= \frac{1}{3} (D_\mu A_{\nu\sigma} + D_\nu A_{\sigma\mu} + D_\sigma A_{\mu\nu}) \end{aligned}$$

$$\begin{aligned} D [\mu A_{\nu\sigma}] &= \partial_\mu A_{\nu\sigma} - \Gamma_{\mu\nu}^\gamma A_{\gamma\sigma} - \Gamma_{\mu\sigma}^\gamma A_{\nu\gamma} \\ &\quad + \partial_\nu A_{\sigma\mu} - \Gamma_{\nu\sigma}^\gamma A_{\gamma\mu} - \Gamma_{\nu\mu}^\gamma A_{\sigma\gamma} \\ &\quad + \partial_\sigma A_{\mu\nu} - \Gamma_{\sigma\mu}^\gamma A_{\gamma\nu} - \Gamma_{\sigma\nu}^\gamma A_{\mu\gamma} \\ &= \partial [\mu A_{\nu\sigma}] \end{aligned}$$

Problem 5

Consider a given point p on the manifold, along with a nearby point p' . Write the first-order approximation to $\partial_\sigma \Gamma_{\mu\nu}^\gamma$. Answer:

$$\partial_\sigma \Gamma_{\mu\nu}^\gamma = \frac{\Gamma_{\mu\nu}^\gamma(q^\mu(p')) - \Gamma_{\mu\nu}^\gamma(q^\mu(p))}{q^\sigma(p') - q^\sigma(p)}$$

Metric Compatibility

To remove some abstraction between the covariant derivative and the connection coefficients, the Levi-Civita connection the foundational statement that the covariant derivative of the metric is identically zero:

$$D_\sigma g_{\mu\nu} = 0$$

The same is true in contravariant form, i.e. take the derivative of $g^{\mu\nu} g_{\nu\beta} = \delta_\beta^\mu$ to derive:

$$D_\sigma g^{\mu\nu} = 0$$

The zero-derivative of the metric is called *metric compatibility*. Quantitatively, we can apply the tensor field derivative to $g_{\mu\nu}$ using the formula

$$D_\sigma C_{\mu\nu} = \partial_\sigma C_{\mu\nu} - \Gamma_{\sigma\mu}^\gamma C_{\gamma\nu} - \Gamma_{\sigma\nu}^\gamma C_{\mu\gamma}$$

to derive the metric compatibility condition:

$$0 = \partial_\sigma g_{\mu\nu} - \Gamma_{\sigma\mu}^\gamma g_{\gamma\nu} - \Gamma_{\sigma\nu}^\gamma g_{\mu\gamma}$$

Two more copies of the compatibility condition arise by cycling the indices σ, μ, ν :

$$0 = \partial_\mu g_{\nu\sigma} - \Gamma_{\mu\nu}^\gamma g_{\gamma\sigma} - \Gamma_{\mu\sigma}^\gamma g_{\nu\gamma}$$

$$0 = \partial_\nu g_{\sigma\mu} - \Gamma_{\nu\sigma}^\gamma g_{\gamma\mu} - \Gamma_{\nu\mu}^\gamma g_{\sigma\gamma}$$

Next, subtract the latter two equations from the first, keeping in mind the symmetry of the metric and of the connection coefficients, to find

$$0 = \partial_\sigma g_{\mu\nu} - \partial_\mu g_{\nu\sigma} - \partial_\nu g_{\sigma\mu} + 2\Gamma_{\mu\nu}^\gamma g_{\sigma\gamma},$$

which has only one instance of Γ . Solving for this gives a formula for the connection coefficients in terms of the metric:

$$\Gamma_{\mu\nu}^\gamma = \frac{1}{2} g^{\gamma\alpha} (\partial_\mu g_{\nu\alpha} + \partial_\nu g_{\alpha\mu} - \partial_\alpha g_{\mu\nu})$$

Problem 6

Use the metric compatibility condition to calculate all nonzero connection coefficients for the Rindler metric characterized by

$$dS^2 = -c^2 d\tau^2 \left(1 + \frac{\xi \tilde{a}}{c^2}\right)^2 + d\xi^2.$$

Answer:

$$\Gamma_{\tau\tau}^\xi = \frac{\tilde{a}}{c^2} \left(1 + \frac{\xi \tilde{a}}{c^2}\right)$$

$$\Gamma_{\xi\tau}^\tau = \frac{\tilde{a}/c^2}{1 + \xi \tilde{a}/c^2}$$

Problem 7

Use the metric compatibility condition to calculate all nonzero connection coefficients for the Friedman-Robertson-Walker metric characterized by

$$dS^2 = -c^2 dt^2 + a^2(t) \delta_{jk} dq^j dq^k.$$

Answer:

$$\Gamma_{jk}^t = a \frac{da}{dt} \delta_{jk}$$

$$\Gamma_{tk}^j = \frac{1}{a} \frac{da}{dt} \delta_{jk}$$

3.5 Vector Calculus Operators**Del Operator**

While introducing the gradient, it's customary to reach for the Del notation via vector calculus:

$$\nabla f = \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^\mu}$$

In light of the covariant derivative however, it's helpful to reassign the ∇ symbol as:

$$\nabla = \mathbf{a}^{(\mu)} D_\mu$$

This does no harm to the gradient of the scalar function f . Operating on a scalar function, ∂_ν is the same as D_ν .

Divergence

Recall that for a vector field V , the covariant derivative is given by

$$D_\nu V = (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) \mathbf{a}_{(\mu)}.$$

Letting $\nu = \mu$ and distributing $\mathbf{a}^{(\mu)}$ into each side, we calculate the *divergence* of the field:

$$\mathbf{a}^{(\mu)} \cdot D_\mu V = (\partial_\mu V^\mu + \Gamma_{\mu\gamma}^\mu V^\gamma) = D_\mu V^\mu$$

In terms of the Del operator, the divergence is written:

$$\nabla \cdot V = \mathbf{a}^{(\mu)} \cdot D_\mu V$$

Laplacian

For a scalar field $f(q^\mu)$, the Laplacian operator $\nabla \cdot \nabla = \nabla^2 = \Delta$ involves two derivatives and must return a type $(0,0)$ tensor. In terms of the covariant derivative, this is simply:

$$\nabla^2 f = D^\mu D_\mu f = g^{\mu\nu} D_\nu D_\mu f$$

Levi-Civita Symbol

The cross product between two vectors is generalized to N dimensions using the *Levi-Civita* symbol $\epsilon_{ijk\dots}$. If the indices i, j, k are an even permutation of the sequence 1, 2, 3, then $\epsilon = 1$. For odd permutations, $\epsilon = -1$. For any two equal indices, $\epsilon = 0$. Note ϵ_{ijk} is *not* a tensor and ignores the up- or down-placement of indices. To generalize to higher dimensions, add more indices next to i, j, k .

For the standard three-dimensional cross product, one writes, for two vectors U, V :

$$(U \times V)^i = \epsilon_{ijk} U^j V^k$$

Curl

In terms of the Levi-Civita symbol, the i th component of the curl of a vector V reads

$$(\nabla \times V)^i = \epsilon_{ijk} D_j V^k,$$

or, for the full curl:

$$\nabla \times V = \left(\epsilon_{ijk} D_j V^k \right) a_{(i)}$$

Alternatively, one can begin with the operator

$$\nabla = a^{(\mu)} D_\mu,$$

and the curl is

$$\nabla \times V = a^{(\mu)} D_\mu \times a_{(\nu)} V^\nu,$$

yielding a curious formula:

$$\nabla \times V = \left(a^{(\nu)} \times a_{(\mu)} \right) D_\nu V^\mu$$

Jacobian

In terms of Levi-Civita-like symbols, the determinant of the metric, which is the square of the Jacobian, is given by

$$\det g = \frac{1}{n!} \epsilon^{i_1 i_2 \dots i_n} \epsilon^{j_1 j_2 \dots j_n} g_{i_1 j_1} g_{i_2 j_2} \dots g_{i_n j_n}.$$

For three dimensions, the formula calms down a little:

$$\det g = \frac{1}{6} \epsilon^{ijk} \epsilon^{pqr} g_{ip} g_{jq} g_{kr}$$

4 Flat Manifolds

4.1 Cartesian Coordinates

The Cartesian coordinate system in three dimensions is characterized by the position vector

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}} .$$

The corresponding differential line element is

$$d\mathbf{S} = dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}} + dz \hat{\mathbf{z}} ,$$

with differential interval

$$dS^2 = dx^2 + dy^2 + dz^2 .$$

In the language of tensors, assign the generalized coordinates $q^x = x$, $q^y = y$, $q^z = z$, and the above means

$$dS^2 = dq^\mu dq_\mu = g_{\mu\nu} dq^\nu dq_\mu ,$$

with the corresponding metric

$$g_{\mu\nu} = \eta_{\mu\nu} ,$$

which is the Minkowski metric without the 0-component:

$$\begin{aligned} \eta_{11} &= \eta_{22} = \eta_{33} = 1 \\ \eta_{\mu\nu} &= 0 \text{ if } \mu \neq \nu \end{aligned}$$

Cartesian coordinates correspond to a flat manifold with no surprises. The basis vectors $\mathbf{a}_{(\mu)}$ are fixed, and thus there are no connection coefficients Γ .

4.2 Cylindrical Coordinates

Starting with the Cartesian coordinate system, make the change of coordinates

$$\begin{aligned} x &= r \cos(\theta) \\ y &= r \sin(\theta) \end{aligned}$$

with

$$\begin{aligned} \hat{\mathbf{r}} &= \cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}} \\ \hat{\boldsymbol{\theta}} &= -\sin(\theta) \hat{\mathbf{x}} + \cos(\theta) \hat{\mathbf{y}} . \end{aligned}$$

With this, the position vector in cylindrical coordinates is

$$\mathbf{S} = r \hat{\mathbf{r}} + z \hat{\mathbf{z}} .$$

Line Element and Basis

From the position vector $\mathbf{S}$, the differential line element reads

$$d\mathbf{S} = dr \hat{\mathbf{r}} + r d\theta \hat{\boldsymbol{\theta}} + dz \hat{\mathbf{z}} ,$$

and right away we can write the differential interval:

$$dS^2 = dr^2 + r^2 d\theta^2 + dz^2$$

Choosing the generalized coordinates $q^r = r$, $q^\theta = \theta$, $q^z = z$, we pick out the basis vectors

$$\begin{aligned} \mathbf{a}_{(r)} &= \hat{\mathbf{r}} \\ \mathbf{a}_{(\theta)} &= r \hat{\boldsymbol{\theta}} \\ \mathbf{a}_{(z)} &= \hat{\mathbf{z}} . \end{aligned}$$

Note that the coordinates q^μ themselves aren't strictly limited to having units of length. The term $q^\theta = \theta$ is indeed dimensionless, however the basis vector $\mathbf{a}_{(\theta)}$ has units of length.

Using $\mathbf{a}_{(\mu)} \cdot \mathbf{a}^{(\nu)} = \delta_\mu^\nu$, the contravariant basis vectors are

$$\begin{aligned} \mathbf{a}^{(r)} &= \hat{\mathbf{r}} \\ \mathbf{a}^{(\theta)} &= r^{-1} \hat{\boldsymbol{\theta}} \\ \mathbf{a}^{(z)} &= \hat{\mathbf{z}} . \end{aligned}$$

Metric

From the basis vectors $\mathbf{a}_{(\mu)}$, all nonzero components of the metric $g_{\mu\nu}$ are straightforwardly calculated:

$$\begin{aligned} g_{rr} &= \mathbf{a}_{(r)} \cdot \mathbf{a}_{(r)} = 1 \\ g_{\theta\theta} &= \mathbf{a}_{(\theta)} \cdot \mathbf{a}_{(\theta)} = r^2 \\ g_{zz} &= \mathbf{a}_{(z)} \cdot \mathbf{a}_{(z)} = 1 \end{aligned}$$

Similar goes for $g^{\mu\nu}$:

$$\begin{aligned} g^{rr} &= \mathbf{a}^{(r)} \cdot \mathbf{a}^{(r)} = 1 \\ g^{\theta\theta} &= \mathbf{a}^{(\theta)} \cdot \mathbf{a}^{(\theta)} = r^{-2} \\ g^{zz} &= \mathbf{a}^{(z)} \cdot \mathbf{a}^{(z)} = 1 \end{aligned}$$

Alternatively, or perhaps more elegantly, the same metric can be calculated by transforming the Cartesian metric:

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu\nu}$$

In the above, the primed indices refer to parameters r, θ, z , whereas the unprimed indices are for x, y , and z . For example, one finds for $g_{\theta\theta}$:

$$\begin{aligned} g_{\theta\theta} &= \frac{\partial q^x}{\partial q^\theta} \frac{\partial q^x}{\partial q^\theta} + \frac{\partial q^y}{\partial q^\theta} \frac{\partial q^y}{\partial q^\theta} \\ &= \left(\sin^2(\theta) + \cos^2(\theta) \right) r^2 \\ &= r^2 \end{aligned}$$

Connection

Most of the connection coefficients for cylindrical coordinates are zero. Using either of

$$\Gamma_{\mu\nu}^\gamma = \mathbf{a}^{(\gamma)} \cdot \partial_\mu \mathbf{a}_{(\nu)},$$

$$\Gamma_{\mu\nu}^\gamma = \frac{1}{2} g^{\gamma\alpha} (\partial_\mu g_{\nu\alpha} + \partial_\nu g_{\mu\alpha} - \partial_\alpha g_{\mu\nu}),$$

one calculates the only nonzero Γ to be:

$$\begin{aligned} \Gamma_{r\theta}^\theta &= \frac{1}{r} \\ \Gamma_{\theta\theta}^r &= -r \end{aligned}$$

Gradient

For a scalar function $f = f(r, \theta, z)$, the gradient is:

$$\nabla f = \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^\mu} = \frac{\partial f}{\partial r} \hat{\mathbf{r}} + \frac{1}{r} \frac{\partial f}{\partial \theta} \hat{\boldsymbol{\theta}} + \frac{\partial f}{\partial z} \hat{\mathbf{z}}$$

Normalized Components

When dealing with vectors $\mathbf{V} = V^\mu \mathbf{a}_{(\mu)}$, notice that the basis vectors are not unit-normalized. Often it's convenient to work in such a system, in which case $\mathbf{V}$ is written

$$\mathbf{V} = v^\mu \hat{\mathbf{a}}_{(\mu)},$$

where the contravariant components V^μ are rescaled to v^μ according to

$$\begin{aligned} v^r &= V^r \\ v^\theta &= r V^\theta \\ v^z &= V^z. \end{aligned}$$

$$\begin{aligned} \nabla \cdot \mathbf{V} &= \mathbf{a}^{(\mu)} \cdot D_\mu \mathbf{V} = D_\mu V^\mu \\ &= \partial_r V^r + \left(\partial_\theta V^\theta + \Gamma_{\theta r}^\theta V^r \right) + \partial_z V^z \\ &= \frac{1}{r} \partial_r (r V^r) + \partial_\theta V^\theta + \partial_z V^z \\ &= \frac{1}{r} \partial_r (r v^r) + \frac{1}{r} \partial_\theta v^\theta + \partial_z v^z \end{aligned}$$

Laplacian

For a scalar function $f = f(r, \theta, z)$, the Laplacian is:

$$\begin{aligned} \nabla^2 f &= g^{\mu\nu} D_\nu D_\mu f = g^{\mu\nu} D_\nu \partial_\mu f \\ &= g^{\mu\nu} (\partial_\nu \partial_\mu f - \Gamma_{\nu\mu}^\gamma \partial_\gamma f) \\ &= g^{rr} \partial_r \partial_r f + g^{\theta\theta} (\partial_\theta \partial_\theta f + r \partial_r f) + g^{zz} \partial_z \partial_z f \\ &= \frac{1}{r} \partial_r (r \partial_r f) + \frac{1}{r^2} \partial_\theta \partial_\theta f + \partial_z \partial_z f \end{aligned}$$

Curl

$$\begin{aligned} \nabla \times \mathbf{V} &= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) D_\nu V^\mu \\ &= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^z &= \frac{1}{r} \left(\partial_r (r^2 V^\theta) - \partial_\theta V^r \right) \\ &= \frac{1}{r} \left(\partial_r (r v^\theta) - \partial_\theta v^r \right) \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^r &= \frac{1}{r} \partial_\theta V^z - r \partial_z V^\theta \\ &= \frac{1}{r} \partial_\theta v^z - \partial_z v^\theta \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^\theta &= \partial_z V^r - \partial_r V^z \\ &= \partial_z v^r - \partial_r v^z \end{aligned}$$

4.3 Spherical Coordinates

Starting again from

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}},$$

we go from Cartesian coordinates to spherical coordinates with the substitution

$$\begin{aligned}x &= r \sin(\theta) \cos(\phi) \\y &= r \sin(\theta) \sin(\phi) \\z &= r \cos(\theta)\end{aligned}$$

such that

$$\mathbf{S} = r \hat{\mathbf{r}}.$$

Normalized basis vectors are calculated via

$$\hat{\mathbf{r}} = \frac{\partial \mathbf{S} / \partial r}{|\partial \mathbf{S} / \partial r|}$$

and similar for θ and ϕ , resolving to:

$$\begin{aligned}\hat{\mathbf{r}} &= \sin(\theta) \cos(\phi) \hat{\mathbf{x}} + \sin(\theta) \sin(\phi) \hat{\mathbf{y}} + \cos(\theta) \hat{\mathbf{z}} \\ \hat{\boldsymbol{\theta}} &= \cos(\theta) \cos(\phi) \hat{\mathbf{x}} + \cos(\theta) \sin(\phi) \hat{\mathbf{y}} - \sin(\theta) \hat{\mathbf{z}} \\ \hat{\boldsymbol{\phi}} &= -\sin(\phi) \hat{\mathbf{x}} + \cos(\phi) \hat{\mathbf{y}}\end{aligned}$$

Line Element and Basis

The differential line element $d\mathbf{S}$ in spherical coordinates is calculated by direct substitution into

$$d\mathbf{S} = dx \hat{\mathbf{x}} + dy \hat{\mathbf{y}} + dz \hat{\mathbf{z}},$$

resulting in

$$d\mathbf{S} = dr \hat{\mathbf{r}} + r d\theta \hat{\boldsymbol{\theta}} + r \sin(\theta) d\phi \hat{\boldsymbol{\phi}},$$

and subsequently

$$dS^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2(\theta) d\phi^2.$$

Choosing the generalized coordinates $q^r = r$, $q^\theta = \theta$, $q^\phi = \phi$, we pick out the basis vectors

$$\begin{aligned}\mathbf{a}_{(r)} &= \hat{\mathbf{r}} \\ \mathbf{a}_{(\theta)} &= r \hat{\boldsymbol{\theta}} \\ \mathbf{a}_{(\phi)} &= r \sin(\theta) \hat{\boldsymbol{\phi}}.\end{aligned}$$

Using $\mathbf{a}_{(\mu)} \cdot \mathbf{a}^{(\nu)} = \delta_\mu^\nu$, the contravariant basis vectors are

$$\begin{aligned}\mathbf{a}^{(r)} &= \hat{\mathbf{r}} \\ \mathbf{a}^{(\theta)} &= \frac{1}{r} \hat{\boldsymbol{\theta}} \\ \mathbf{a}^{(\phi)} &= \frac{1}{r \sin(\theta)} \hat{\boldsymbol{\phi}}.\end{aligned}$$

Metric

From the basis vectors $\mathbf{a}_{(\mu)}$, all nonzero components of the metric $g_{\mu\nu}$ are straightforwardly calculated:

$$\begin{aligned}g_{rr} &= \mathbf{a}_{(r)} \cdot \mathbf{a}_{(r)} = 1 \\ g_{\theta\theta} &= \mathbf{a}_{(\theta)} \cdot \mathbf{a}_{(\theta)} = r^2 \\ g_{\phi\phi} &= \mathbf{a}_{(\phi)} \cdot \mathbf{a}_{(\phi)} = r^2 \sin^2(\theta)\end{aligned}$$

Similar goes for $g^{\mu\nu}$:

$$\begin{aligned}g^{rr} &= \mathbf{a}^{(r)} \cdot \mathbf{a}^{(r)} = 1 \\ g^{\theta\theta} &= \mathbf{a}^{(\theta)} \cdot \mathbf{a}^{(\theta)} = \frac{1}{r^2} \\ g^{\phi\phi} &= \mathbf{a}^{(\phi)} \cdot \mathbf{a}^{(\phi)} = \frac{1}{r^2 \sin^2(\theta)}\end{aligned}$$

Alternatively, the same metric can be calculated by transforming the Cartesian metric:

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu\nu}$$

In the above, the primed indices refer to parameters r , θ , ϕ , whereas the unprimed indices are for x , y , and z .

Connection

Using either of

$$\Gamma_{\mu\nu}^\gamma = \mathbf{a}^{(\gamma)} \cdot \partial_\mu \mathbf{a}_{(\nu)},$$

$$\Gamma_{\mu\nu}^\gamma = \frac{1}{2} g^{\gamma\alpha} (\partial_\mu g_{\nu\alpha} + \partial_\nu g_{\mu\alpha} - \partial_\alpha g_{\mu\nu}),$$

one calculates the nonzero Γ to be:

$$\begin{aligned}\Gamma_{r\theta}^\theta &= \Gamma_{r\phi}^\phi = \frac{1}{r} \\ \Gamma_{\theta\theta}^r &= -r \\ \Gamma_{\phi\phi}^r &= -r \sin^2(\theta) \\ \Gamma_{\phi\phi}^\theta &= -\sin(\theta) \cos(\theta) \\ \Gamma_{\phi\theta}^\phi &= \cot(\theta)\end{aligned}$$

Gradient

For a scalar function $f = f(r, \theta, \phi)$, the gradient is:

$$\nabla f = \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^\mu} = \frac{\partial f}{\partial r} \hat{\mathbf{r}} + \frac{1}{r} \frac{\partial f}{\partial \theta} \hat{\boldsymbol{\theta}} + \frac{1}{r \sin(\theta)} \frac{\partial f}{\partial \phi} \hat{\boldsymbol{\phi}}$$

Normalized Components

Reconciliation between the non-normalized vector $V = V^\mu \mathbf{a}_{(\mu)}$ and its normalized counterpart $\mathbf{V} = v^\mu \hat{\mathbf{a}}_{(\mu)}$ involves transferring the ‘extra’ coefficients from $\mathbf{a}_{(\mu)}$ to v^μ as:

$$\begin{aligned} v^r &= V^r \\ v^\theta &= r V^\theta \\ v^\phi &= r \sin(\theta) V^\phi \end{aligned}$$

Divergence

$$\begin{aligned} \nabla \cdot \mathbf{V} &= \mathbf{a}^{(\mu)} \cdot D_\mu \mathbf{V} = D_\mu V^\mu \\ &= \partial_r V^r + \partial_\theta V^\theta + \Gamma_{\theta r}^\theta V^r \\ &\quad + \partial_\phi V^\phi + \Gamma_{\phi r}^\phi V^r + \Gamma_{\phi \theta}^\phi V^\theta \\ &= \frac{\partial_r (r^2 v^r)}{r^2} + \frac{\partial_\theta (\sin(\theta) v^\theta)}{r \sin(\theta)} + \frac{\partial_\phi (v^\phi)}{r \sin(\theta)} \end{aligned}$$

Laplacian

For a scalar function $f = f(r, \theta, z)$, the Laplacian is:

$$\begin{aligned} \nabla^2 f &= g^{\mu\nu} D_\nu D_\mu f = g^{\mu\nu} D_\nu \partial_\mu f \\ &= g^{\mu\nu} (\partial_\nu \partial_\mu f - \Gamma_{\nu\mu}^\gamma \partial_\gamma f) \\ &= g^{rr} \partial_r \partial_r f + g^{\theta\theta} (\partial_\theta \partial_\theta f - \Gamma_{\theta\theta}^r \partial_r f) \\ &\quad + g^{\phi\phi} (\partial_\phi \partial_\phi f - \Gamma_{\phi\phi}^r \partial_r f - \Gamma_{\phi\phi}^\theta \partial_\theta f) \\ &= \frac{\partial_r (r^2 \partial_r f)}{r^2} + \frac{\partial_\theta (\sin(\theta) \partial_\theta f)}{r^2 \sin(\theta)} + \frac{\partial_\phi \partial_\phi f}{r^2 \sin^2(\theta)} \end{aligned}$$

Curl

$$\begin{aligned} \nabla \times \mathbf{V} &= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) D_\nu V^\mu \\ &= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^\phi &= r \partial_r V^\theta + 2 V^\theta - \frac{1}{r} \partial_\theta V^r \\ &= \frac{1}{r} \left(\partial_r (r v^\theta) - \partial_\theta v^r \right) \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^r &= \sin(\theta) \partial_\theta V^\phi + 2 \cos(\theta) V^\phi - \frac{\partial_\phi V^\theta}{\sin(\theta)} \\ &= \frac{1}{r \sin(\theta)} \left(\partial_\theta (\sin(\theta) v^\phi) - \partial_\phi v^\theta \right) \end{aligned}$$

$$\begin{aligned} (\nabla \times \mathbf{V})^\theta &= \frac{\partial_\phi V^r}{r \sin(\theta)} - 2 \sin(\theta) V^\phi \\ &\quad - r \sin(\theta) \partial_r V^\phi \\ &= \frac{1}{r} \left(\frac{\partial_\phi v^r}{\sin(\theta)} - \partial_r (r v^\phi) \right) \end{aligned}$$

4.4 Parabolic Coordinates 2D

Consider the two-dimensional Cartesian flat space with position vector

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}}$$

with the substitution

$$\begin{aligned} x &= uv \\ y &= \frac{1}{2} (u^2 - v^2), \end{aligned}$$

called *parabolic coordinates*.

Normalized basis vectors are calculated via

$$\begin{aligned} \hat{\mathbf{u}} &= \frac{\partial \mathbf{S} / \partial u}{|\partial \mathbf{S} / \partial u|} = \frac{v \hat{\mathbf{x}} + u \hat{\mathbf{y}}}{\sqrt{u^2 + v^2}} \\ \hat{\mathbf{v}} &= \frac{\partial \mathbf{S} / \partial v}{|\partial \mathbf{S} / \partial v|} = \frac{u \hat{\mathbf{x}} - v \hat{\mathbf{y}}}{\sqrt{u^2 + v^2}}, \end{aligned}$$

which can be inverted to say

$$\begin{aligned} \hat{\mathbf{x}} &= \frac{\partial \mathbf{S} / \partial u}{|\partial \mathbf{S} / \partial u|} = \frac{v \hat{\mathbf{u}} + u \hat{\mathbf{v}}}{\sqrt{u^2 + v^2}} \\ \hat{\mathbf{y}} &= \frac{\partial \mathbf{S} / \partial v}{|\partial \mathbf{S} / \partial v|} = \frac{u \hat{\mathbf{u}} - v \hat{\mathbf{v}}}{\sqrt{u^2 + v^2}}. \end{aligned}$$

All in terms of parabolic coordinates, the position vector becomes

$$\mathbf{S} = \frac{1}{2} \sqrt{u^2 + v^2} (u \hat{\mathbf{u}} + v \hat{\mathbf{v}}).$$

Problem 8

In terms of u, v , what is the radius of a circle centered at the origin? Answer:

$$r = \frac{1}{2} (u^2 + v^2)$$

Line Element and Basis

Prom the position vector $\mathbf{S}$ carefully work out the differential line element $d\mathbf{S}$:

$$d\mathbf{S} = \sqrt{u^2 + v^2} (du \hat{\mathbf{u}} + dv \hat{\mathbf{v}})$$

From the differential line element, pick out generalized coordinates $q^u = u$, $q^v = v$, along with basis vectors $\mathbf{a}_{(\mu)}$ are

$$\begin{aligned}\mathbf{a}_{(u)} &= \sqrt{u^2 + v^2} \hat{\mathbf{u}} \\ \mathbf{a}_{(v)} &= \sqrt{u^2 + v^2} \hat{\mathbf{v}},\end{aligned}$$

and correspondingly

$$\begin{aligned}\mathbf{a}^{(u)} &= \frac{1}{\sqrt{u^2 + v^2}} \hat{\mathbf{u}} \\ \mathbf{a}^{(v)} &= \frac{1}{\sqrt{u^2 + v^2}} \hat{\mathbf{v}}.\end{aligned}$$

Metric

The metric $g_{\mu\nu}$ is straightforwardly calculated via $\mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}$, however it's more interesting to start from the tensor transformation law

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu\nu},$$

working out to

$$g_{uu} = g_{vv} = u^2 + v^2,$$

and correspondingly

$$g^{uu} = g^{vv} = \frac{1}{u^2 + v^2}.$$

Connection

Using either of

$$\Gamma_{\mu\nu}^\gamma = \mathbf{a}^{(\gamma)} \cdot \partial_\mu \mathbf{a}_{(\nu)},$$

$$\Gamma_{\mu\nu}^\gamma = \frac{1}{2} g^{\gamma\alpha} (\partial_\mu g_{\nu\alpha} + \partial_\nu g_{\mu\alpha} - \partial_\alpha g_{\mu\nu}),$$

one calculates the nonzero Γ to be:

$$\begin{aligned}\Gamma_{uu}^u &= \Gamma_{uv}^v = \frac{u}{u^2 + v^2} \\ \Gamma_{vv}^v &= \Gamma_{vu}^u = \frac{v}{u^2 + v^2} \\ \Gamma_{vv}^u &= \frac{-u}{u^2 + v^2} \\ \Gamma_{uu}^v &= \frac{-v}{u^2 + v^2}\end{aligned}$$

Gradient

For a scalar function $f = f(u, v)$, the gradient is:

$$\nabla f = \mathbf{a}^{(\mu)} \frac{\partial f}{\partial q^\mu} = \frac{1}{\sqrt{u^2 + v^2}} \left(\frac{\partial f}{\partial u} \hat{\mathbf{u}} + \frac{\partial f}{\partial v} \hat{\mathbf{v}} \right)$$

Normalized Components

Reconciliation between the non-normalized vector $\mathbf{W} = W^\mu \mathbf{a}_{(\mu)}$ and its normalized counterpart $\mathbf{W} = w^\mu \hat{\mathbf{a}}_{(\mu)}$ involves transferring the 'extra' coefficients from $\mathbf{a}_{(\mu)}$ to w^μ as:

$$\begin{aligned}w^u &= \sqrt{u^2 + v^2} W^u \\ w^v &= \sqrt{u^2 + v^2} W^v\end{aligned}$$

Divergence

$$\begin{aligned}\nabla \cdot \mathbf{W} &= \mathbf{a}^{(\mu)} \cdot D_\mu \mathbf{W} = D_\mu W^\mu \\ &= \partial_u W^u + \partial_v W^v + \frac{2uW^u}{u^2 + v^2} + \frac{2vW^v}{u^2 + v^2} \\ &= \frac{\partial_u ((u^2 + v^2) W^u) + \partial_v ((u^2 + v^2) W^v)}{u^2 + v^2} \\ &= \frac{\partial_u (\sqrt{u^2 + v^2} w^u) + \partial_v (\sqrt{u^2 + v^2} w^v)}{u^2 + v^2}\end{aligned}$$

Laplacian

For a scalar function $f = f(u, v)$, the Laplacian is:

$$\begin{aligned}\nabla^2 f &= g^{\mu\nu} D_\nu D_\mu f = g^{\mu\nu} D_\nu \partial_\mu f \\ &= g^{uu} (\partial_{uu} f - \Gamma_{uu}^\gamma \partial_\gamma f) + g^{vv} (\partial_{vv} f - \Gamma_{vv}^\gamma \partial_\gamma f) \\ &= \frac{1}{u^2 + v^2} (\partial_{uu} f + \partial_{vv} f)\end{aligned}$$

Curl

Thankfully the curl has only one component.

$$\begin{aligned}
\nabla \times V &= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) D_\nu V^\mu \\
&= \left(\mathbf{a}^{(\nu)} \times \mathbf{a}_{(\mu)} \right) (\partial_\nu V^\mu + \Gamma_{\nu\gamma}^\mu V^\gamma) \\
&= \partial_u W^\nu - \partial_\nu W^u + \frac{2(uW^\nu - \nu W^u)}{u^2 + v^2} \\
&= \frac{\partial_u ((u^2 + v^2) W^\nu) - \partial_\nu ((u^2 + v^2) W^u)}{u^2 + v^2} \\
&= \frac{\partial_u (\sqrt{u^2 + v^2} w^\nu) - \partial_\nu (\sqrt{u^2 + v^2} w^u)}{u^2 + v^2}
\end{aligned}$$

4.5 Parabolic Coordinates 3D

Consider the two-dimensional Cartesian flat space with position vector

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}}$$

with the substitution

$$\begin{aligned}
x &= uv \cos(\phi) \\
y &= uv \sin(\phi) \\
z &= \frac{1}{2} (u^2 - v^2),
\end{aligned}$$

which are parabolic coordinates in three dimensions.

As it turns out, curves with constant u and constant v form confocal paraboloids characterized by:

$$\begin{aligned}
z_u &= \frac{1}{2} \left(\frac{x^2 + y^2}{v^2} - v^2 \right) \\
z_v &= \frac{1}{2} \left(u^2 - \frac{x^2 + y^2}{u^2} \right)
\end{aligned}$$

Line Element and Basis

From the position vector $\mathbf{S}$, define generalized coordinates $q^u = u$, $q^v = v$, $q^\phi = \phi$, work out the differential line element $d\mathbf{S}$ as

$$d\mathbf{S} = dq^u \mathbf{a}_{(u)} + dq^v \mathbf{a}_{(v)} + dq^\phi \mathbf{a}_{(\phi)},$$

where:

$$\begin{aligned}
\mathbf{a}_{(u)} &= v (\cos(\phi) \hat{\mathbf{x}} + \sin(\phi) \hat{\mathbf{y}}) + u \hat{\mathbf{z}} \\
\mathbf{a}_{(v)} &= u (\cos(\phi) \hat{\mathbf{x}} + \sin(\phi) \hat{\mathbf{y}}) - v \hat{\mathbf{z}} \\
\mathbf{a}_{(\phi)} &= uv (\sin(\phi) \hat{\mathbf{x}} + \cos(\phi) \hat{\mathbf{y}})
\end{aligned}$$

Metric

The metric $g_{\mu\nu}$ is calculated via $\mathbf{a}_{(\mu)} \cdot \mathbf{a}_{(\nu)}$ or using

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu\nu}.$$

In either case, one finds

$$\begin{aligned}
g_{uu} &= g_{vv} = u^2 + v^2 \\
g_{\phi\phi} &= u^2 v^2,
\end{aligned}$$

and furthermore:

$$\begin{aligned}
g^{uu} &= g^{vv} = \frac{1}{u^2 + v^2} \\
g^{\phi\phi} &= \frac{1}{u^2 v^2}
\end{aligned}$$

Connection

Connection coefficients Γ are largely unchanged from the two-dimensional case. New to the mix are:

$$\begin{aligned}
\Gamma_{\phi\phi}^u &= \frac{-uv^2}{u^2 + v^2} \\
\Gamma_{\phi\phi}^v &= \frac{-vu^2}{u^2 + v^2}
\end{aligned}$$

Laplacian

For a scalar function $f = f(u, v, \phi)$, the Laplacian is:

$$\begin{aligned}
\nabla^2 f &= g^{\mu\nu} D_\nu D_\mu f = g^{\mu\nu} D_\nu \partial_\mu f \\
&= g^{uu} (\partial_{uu} f - \Gamma_{uu}^\gamma \partial_\gamma f) + g^{vv} (\partial_{vv} f - \Gamma_{vv}^\gamma \partial_\gamma f) \\
&\quad + g^{\phi\phi} (\partial_{\phi\phi} f - \Gamma_{\phi\phi}^\gamma \partial_\gamma f) \\
&= \frac{1}{u^2 + v^2} \left(\frac{\partial_u (u \partial_u f)}{u} + \frac{\partial_v (v \partial_v f)}{v} \right) + \frac{\partial_{\phi\phi} f}{u^2 v^2}
\end{aligned}$$

4.6 Hyperspherical Coordinates

Four Spatial Dimensions

Consider a four-dimensional Cartesian flat space characterized by the position vector:

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}} + w \hat{\mathbf{w}},$$

and the metric extends to four dimensions accordingly:

$$\begin{aligned}
\eta_{11} &= \eta_{22} = \eta_{33} = \eta_{44} = 1 \\
\eta_{\mu\nu} &= 0 \text{ if } \mu \neq \nu
\end{aligned}$$

Hyperspherical Coordinates

Such a space can be mapped by *hyperspherical coordinates* according to the substitution:

$$\begin{aligned}x &= r \cos (\psi) \\y &= r \sin (\psi) \cos (\theta) \\z &= r \sin (\psi) \sin (\theta) \cos (\phi) \\w &= r \sin (\psi) \sin (\theta) \sin (\phi)\end{aligned}$$

Metric

Jumping to the metric calculation, the easiest route is to use the transformation law

$$g_{\mu' \nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu \nu} ,$$

leading to:

$$\begin{aligned}g_{rr} &= 1 \\g_{\psi\psi} &= r^2 \\g_{\theta\theta} &= r^2 \sin^2 (\psi) \\g_{\phi\phi} &= r^2 \sin^2 (\psi) \sin^2 (\theta)\end{aligned}$$

Chapter 28

Riemannian Geometry

1 Geodesic Equation

The notion of a ‘straight line’ in flat space is essentially always the same in the sense that one can imagine holding a ruler or straight edge to trace a line. This is admittedly difficult to imagine in higher dimensions, nonetheless the path between two points in flat space can always be characterized by

$$\Delta S = \sqrt{\Delta x_1^2 + \Delta x_2^2 + \Delta x_3^2 + \dots}.$$

Notion of Geodesic

The formula for a straight line breaks down on any manifold that is not flat, i.e. can’t be mapped by Cartesian coordinates. The idea we must dismiss is the rigidity of straight lines, however the idea we keep is that a line is the shortest path connection two points. In light of this, a straight line on a manifold is defined as the shortest possible path that can be traversed between any two points. Such a path is called a *geodesic*.

To get a sense of geodesics we consider the surface of a sphere. Using planet Earth for the sake of example, a geodesic is any line of longitude or the equator. Lines of latitude (except the equator) are not geodesics.

Another analogy we can borrow comes from kinematics. Since a particle’s motion deviates from linear only when acceleration is involved, it should stand that geodesics correspond to non-accelerated motion on the manifold.

1.1 Parallel Transport

Recall that the connection coefficients Γ arise from the fact that derivatives of the basis vectors $\mathbf{a}_{(\mu)}$ risk pointing off of

the manifold. This phenomenon can be reviewed manually by analysis called *parallel transport*.

To demonstrate, consider the two-dimensional polar coordinate system summarized by the position vector

$$\mathbf{r} = \cos(\theta) \hat{\mathbf{x}} + \sin(\theta) \hat{\mathbf{y}}.$$

This is merely the cylindrical coordinate system with the z -component flattened, thus the relevant metric components are

$$\begin{aligned} g_{rr} &= 1 \\ g_{\theta\theta} &= r^2, \end{aligned}$$

leading to the connection coefficients:

$$\begin{aligned} \Gamma_{r\theta}^{\theta} &= \frac{1}{r} \\ \Gamma_{\theta\theta}^r &= -r \end{aligned}$$

To proceed, situate a covariant vector $\mathbf{V} = V^\mu \mathbf{a}_{(\mu)}$ based at the position vector and calculate the norm:

$$\|\mathbf{V}\|^2 = g_{\mu\nu} V^\mu V^\nu = (V^r)^2 + r^2 (V^\theta)^2$$

Let ϕ represent the angle made between $\mathbf{V}$ and the position vector $\mathbf{r}$. In terms of ϕ , the components of $\mathbf{V}$ can be written:

$$\begin{aligned} V^r &= \|\mathbf{V}\| \cos(\phi) \\ V^\theta &= \|\mathbf{V}\| \frac{1}{r} \sin(\phi) \end{aligned}$$

Radial Parallel Transport

Now we’re ready to impose a parallel transport on $\mathbf{V}$. Keeping the length, and thus the norm of $\mathbf{V}$ the same, let the

location $\mathbf{r}$ in which $\mathbf{V}$ is based change via $r \rightarrow r + \Delta r$. Labeling the transported vector $\tilde{V}$, we write:

$$\begin{aligned}\tilde{V}^r &= \|V\| \cos(\phi) \\ \tilde{V}^\theta &= \|V\| \frac{1}{r + \Delta r} \sin(\phi)\end{aligned}$$

Eliminate $\sin(\phi)$ between the two θ -equations

$$\tilde{V}^\theta = V^\theta \frac{r}{r + \Delta r},$$

and expand for small Δr :

$$\tilde{V}^\theta \approx V^\theta - V^\theta \frac{\Delta r}{r}$$

Angular Parallel Transport

Instead of varying the radius, we can also impose parallel transport via $\theta \rightarrow \theta + \Delta\theta$. For this, we have:

$$\begin{aligned}\tilde{V}^r &= \|V\| \cos(\phi - \Delta\theta) \\ \tilde{V}^\theta &= \|V\| \frac{1}{r} \sin(\phi - \Delta\theta)\end{aligned}$$

Simplify the $\tilde{V}^r$ -equation to write, to first order in $\Delta\theta$:

$$\tilde{V}^r \approx \|V\| (\cos(\phi) + \sin(\phi) \Delta\theta)$$

Replace each ϕ -term in terms of the original vector components:

$$\tilde{V}^r = V^r + r V^\theta \Delta\theta$$

Connection

Taking the $\tilde{V}^r$ - and $\tilde{V}^\theta$ -equations introduce a temporary unknown symbol (?) in attempt to find a general pattern:

$$0 \approx \tilde{V}^\gamma - V^\gamma + (?)_{\mu\nu}^\gamma V^\nu dq^\mu$$

The above appears, as it should, as the first-order correction to the vector V given the parallel transport involving coordinates q^γ .

Meanwhile recall that the full differential of the vector V can be written

$$dV = \left(\frac{\partial V^\gamma}{\partial q^\mu} dq^\mu + \Gamma_{\mu\nu}^\gamma V^\nu dq^\mu \right) \mathbf{a}_{(\gamma)}.$$

Comparing the two equations, one can't help but notice the unknown symbol (?) plays the same role as the connection coefficients Γ . For a quick check, notice that the factors $1/r$, $-r$ in the respective $\tilde{V}^\theta$, V^r equations are equal to the two connection coefficients for the polar system.

To summarize, we come up with a tensor equation for the first-order change to a vector V under parallel transport:

$$0 \approx \tilde{V}^\gamma - V^\gamma + \Gamma_{\mu\nu}^\gamma V^\nu dq^\mu$$

1.2 Geodesic Equation

To remind where we are, let S denote the position vector on an N -dimensional manifold parameterized by generalized coordinates $q^1, q^2, q^3, \dots, q^N$.

Stating the result first, the geodesic equation is

$$0 = \frac{d^2 q^\gamma}{d\tau^2} + \Gamma_{\mu\nu}^\gamma \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau},$$

which is worth memorizing. By letting U^γ equal the proper velocity, i.e. $U^\gamma = dq^\gamma/d\tau$, the geodesic equation is also written

$$0 = \frac{dU^\gamma}{d\tau} + \Gamma_{\mu\nu}^\gamma U^\mu U^\nu.$$

Derivation by Parallel Transport

A simple derivation of the geodesic equation comes from parallel transport. With U in place of V , start with

$$0 \approx \tilde{U}^\gamma - U^\gamma + \Gamma_{\mu\nu}^\gamma U^\nu dq^\mu.$$

Divide through by a change in time to get

$$0 \approx \frac{\tilde{U}^\gamma - U^\gamma}{\Delta\tau} + \Gamma_{\mu\nu}^\gamma U^\nu \frac{dq^\mu}{\Delta\tau}.$$

In the small $\Delta\tau$ limit, the above converges to the geodesic equation.

Derivation by Covariant Derivative

A stronger foundation for the geodesic equation starts with the generalized coordinates q^γ being parameterized by $q^\gamma(\lambda)$. With respect to λ , the corresponding velocity is written V^γ .

The derivative of V^γ with respect to λ goes like:

$$\begin{aligned}D_\lambda V^\gamma &= \frac{dq^\mu}{d\lambda} D_\mu V^\gamma \\ &= \frac{dq^\mu}{d\lambda} \partial_\mu V^\gamma + \Gamma_{\mu\nu}^\gamma V^\nu \frac{dq^\mu}{d\lambda} \\ &= \frac{d^2 q^\gamma}{d\lambda^2} + \Gamma_{\mu\nu}^\gamma \frac{dq^\mu}{d\lambda} \frac{dq^\nu}{d\lambda}\end{aligned}$$

From the above, one sees that the geodesic equation is recovered if λ is replaced by τ , in which case the vector V becomes the proper velocity. We uncover a tight expression for the geodesic equation using these substitutions:

$$D_\tau U^\mu = 0$$

Affine Parameter

It's worth checking 'how close' the quantity $D_\lambda V^\gamma$ is to the geodesic equation without jumping right to $\lambda \rightarrow \tau$. By the chain rule, we find

$$\begin{aligned} D_\lambda V^\gamma &= \frac{d\tau}{d\lambda} \frac{d}{d\tau} \left(\frac{d\tau}{d\lambda} \frac{dq^\gamma}{d\tau} \right) + \left(\frac{d\tau}{d\lambda} \right)^2 \Gamma_{\mu\nu}^\gamma \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} \\ &= \frac{dq^\gamma}{d\lambda} \frac{d}{d\tau} \left(\frac{1}{d\lambda/d\tau} \right) + \left(\frac{d\tau}{d\lambda} \right)^2 \cancel{D_\tau U^\gamma} \\ &= -V^\gamma \frac{(d^2\lambda/d\tau^2)}{(d\lambda/d\tau)^2}. \end{aligned}$$

In order to get zero on the right, it's only required that λ have no second derivative with respect to τ . This is the fingerprint of an affine parameter:

$$d\lambda \propto d\tau \propto dS$$

Derivation by Variations

Define the *action* as the integral with respect to τ between two proper times τ_1, τ_2 on the manifold. Scale the action by the invariant c such that the action has units of arc length:

$$S = \int_{\tau_1}^{\tau_2} c \, d\tau = \int_{\tau_1}^{\tau_2} \sqrt{-g_{\mu\nu} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau}} \, d\tau$$

To proceed, introduce the variation $S \rightarrow S + \delta S$ by the small change of coordinates

$$q^\mu \rightarrow q^\mu + \delta q^\mu,$$

which must also be reflected in the metric:

$$g_{\mu\nu} \rightarrow g_{\mu\nu} + \delta g_{\mu\nu} = g_{\mu\nu} + \delta q^\alpha \frac{\partial}{\partial q^\alpha} g_{\mu\nu}$$

Per standard prescription in variational calculus, all variations are set to zero at the endpoints $\tau_{1,2}$.

Given the above variations, the quantity under the square root in the action integral takes the form $\sqrt{c^2 - A}$, where

$$\begin{aligned} A &= \delta q^\alpha \frac{\partial g_{\mu\nu}}{\partial q^\alpha} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} + g_{\mu\nu} \frac{d\delta q^\mu}{d\tau} \frac{dq^\nu}{d\tau} \\ &\quad + g_{\mu\nu} \frac{dq^\mu}{d\tau} \frac{d\delta q^\nu}{d\tau} \end{aligned}$$

Then, since c^2 is much greater than A , we can write the first-order approximation:

$$\sqrt{c^2 - A} \approx c - \frac{A}{2c}$$

This is a handy result, as the factor of c allows us to cancel the bulk of the action integral, namely the S term. The remainder of the action integral is written

$$\delta S = \delta S_1 + \delta S_2 + \delta S_3,$$

where:

$$\begin{aligned} \delta S_1 &= \frac{-1}{2c} \int_{\tau_1}^{\tau_2} \delta q^\alpha \frac{\partial g_{\mu\nu}}{\partial q^\alpha} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} \, d\tau \\ \delta S_2 &= \frac{-1}{2c} \int_{\tau_1}^{\tau_2} g_{\mu\nu} \frac{d\delta q^\mu}{d\tau} \frac{dq^\nu}{d\tau} \, d\tau \\ \delta S_3 &= \frac{-1}{2c} \int_{\tau_1}^{\tau_2} g_{\mu\nu} \frac{dq^\mu}{d\tau} \frac{d\delta q^\nu}{d\tau} \, d\tau \end{aligned}$$

Use integration by parts on the δS_3 integral by letting:

$$\begin{aligned} u &= g_{\mu\nu} \frac{dq^\mu}{d\tau} \\ du &= \frac{dg_{\mu\nu}}{d\tau} \frac{dq^\mu}{d\tau} + g_{\mu\nu} \frac{d^2 q^\mu}{d\tau^2} \\ dv &= \frac{d\delta q^\nu}{d\tau} \, d\tau \\ v &= \delta q^\nu \end{aligned}$$

The boundary term $uv|$ vanishes because δq^ν is zero at the endpoints. The δS_3 integral becomes:

$$\delta S_3 = \frac{1}{2c} \int_{\tau_1}^{\tau_2} \delta q^\alpha \left(\frac{\partial g_{\mu\alpha}}{\partial q^\nu} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} + g_{\mu\alpha} \frac{d^2 q^\mu}{d\tau^2} \right) d\tau$$

A similar calculation is done for the δS_2 integral, after changing some indices:

$$\delta S_2 = \frac{1}{2c} \int_{\tau_1}^{\tau_2} \delta q^\alpha \left(\frac{\partial g_{\nu\alpha}}{\partial q^\mu} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} + g_{\alpha\mu} \frac{d^2 q^\mu}{d\tau^2} \right) d\tau$$

Assembling the sum δS results in two chunks:

$$\begin{aligned} \delta S &= \frac{1}{c} \int_{\tau_1}^{\tau_2} \delta q^\alpha g_{\mu\alpha} \frac{d^2 q^\mu}{d\tau^2} \, d\tau \\ &\quad + \frac{1}{c} \int_{\tau_1}^{\tau_2} \delta q^\alpha \left(\frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} \right) \\ &\quad \times \frac{1}{2} \left(\frac{\partial g_{\nu\alpha}}{\partial q^\mu} + \frac{\partial g_{\alpha\mu}}{\partial q^\nu} - \frac{\partial g_{\mu\nu}}{\partial q^\alpha} \right) d\tau \end{aligned}$$

Next, we invoke the argument that the variation δS must vanish entirely for the 'correct' path of motion, therefore the integrands must sum to zero:

$$0 = g_{\mu\alpha} \frac{d^2 q^\mu}{d\tau^2} + \frac{1}{2} \left(\frac{\partial g_{\nu\alpha}}{\partial q^\mu} + \frac{\partial g_{\alpha\mu}}{\partial q^\nu} - \frac{\partial g_{\mu\nu}}{\partial q^\alpha} \right) \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau}$$

Contract with $g^{\gamma\alpha}$ to clean up:

$$0 = \frac{d^2 q^\gamma}{d\tau^2} + \frac{1}{2} g^{\gamma\alpha} \left(\frac{\partial g_{\nu\alpha}}{\partial q^\mu} + \frac{\partial g_{\alpha\mu}}{\partial q^\nu} - \frac{\partial g_{\mu\nu}}{\partial q^\alpha} \right) \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau}$$

The second term contains the exact definition of $\Gamma_{\mu\nu}^\gamma$ if you haven't spotted it already, and the whole calculation results in the geodesic equation:

$$0 = \frac{d^2 q^\gamma}{d\tau^2} + \Gamma_{\mu\nu}^\gamma \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau}$$

1.3 Plane Geodesics

Parallel Transport of Unit Vector

Consider a vector field V^μ that has unit length and points in the positive $\hat{x}$ direction everywhere. It should make sense that such a vector has zero covariant derivative in the Cartesian system, but it's worth checking if the same is true in plane polar coordinates given by:

$$x = r \cos(\theta)$$

$$y = r \sin(\theta)$$

Using the tensor transformation law, one straightforwardly calculates

$$\begin{aligned} V^r &= \frac{\partial q^r}{\partial q^x} V^x + \frac{\partial q^r}{\partial q^y} V^y = \cos(\theta) \\ V^\theta &= \frac{\partial q^\theta}{\partial q^x} V^x + \frac{\partial q^\theta}{\partial q^y} V^y = \frac{-\sin(\theta)}{r} \end{aligned}$$

For thoroughness, we need to calculate $D_r V^r$, $D_r V^\theta$, $D_\theta V^r$, $D_\theta V^\theta$. By brute force, we find each resolving to zero:

$$\begin{aligned} D_r V^r &= \partial_r V^r + \Gamma_{r\gamma}^r V^\gamma = 0 \\ D_r V^\theta &= \partial_r V^\theta + \Gamma_{r\gamma}^\theta V^\gamma \\ &= \partial_r \left(\frac{-\sin(\theta)}{r} \right) - \frac{\sin(\theta)}{r^2} = 0 \\ D_\theta V^r &= \partial_\theta V^r + \Gamma_{\theta\gamma}^r V^\gamma \\ &= -\sin(\theta) + \sin(\theta) = 0 \\ D_\theta V^\theta &= \partial_\theta V^\theta + \Gamma_{\theta\gamma}^\theta V^\gamma \\ &= \frac{-\cos(\theta)}{r} + \frac{\cos(\theta)}{r} = 0 \end{aligned}$$

Straight Lines in two Dimensions

Geodesics in the Cartesian plane are parameterized by

$$x(\lambda) = a\lambda + x_0$$

$$y(\lambda) = b\lambda + y_0,$$

where λ is a dimensionless parameter with a , b , x_0 , y_0 constant.

Switching to polar coordinates, it's straightforward to show that the same line is given by

$$\begin{aligned} r(\lambda) &= \sqrt{(a\lambda + x_0)^2 + (b\lambda + y_0)^2} \\ \theta(\lambda) &= \tan^{-1} \left(\frac{b\lambda + y_0}{a\lambda + x_0} \right). \end{aligned}$$

Meanwhile, the geodesic equation spawns two entities for plane polar coordinates. Using the notation on hand, we write

$$\begin{aligned} 0 &= \frac{d^2 r}{d\lambda^2} - r \left(\frac{d\theta}{d\lambda} \right)^2 \\ 0 &= \frac{d^2 \theta}{d\lambda^2} + \frac{2}{r} \frac{dr}{d\lambda} \frac{d\theta}{d\lambda}, \end{aligned}$$

or using prime notation for λ -derivatives:

$$\begin{aligned} 0 &= r'' - r (\theta')^2 \\ 0 &= \theta'' + \frac{2}{r} r' \theta' \end{aligned}$$

The task now is to show that the r - and θ -equations satisfy the above. Along the way, one finds:

$$\begin{aligned} r' &= \frac{a(a\lambda + x_0) + b(b\lambda + y_0)}{r} \\ \theta' &= \frac{bx_0 - ay_0}{r^2} \\ r'' &= \frac{a^2 + b^2}{r} - \frac{(r')^2}{r} \\ \theta'' &= \frac{-2r'\theta'}{r} \end{aligned}$$

The final θ'' -equation is identical to the θ -instance of the geodesic equation. As for the radial instance, we have

$$\begin{aligned} r'' - r (\theta')^2 &= \\ &= \frac{r^2 (a^2 + b^2) - (a(a\lambda + x_0) + b(b\lambda + y_0))^2}{r^3} \\ &= -\frac{a^2 y_0^2 + b^2 x_0^2 - 2abx_0 y_0}{r^3} = 0 \end{aligned}$$

as expected.

1.4 Spherical Shell Geodesics

Equatorial Geodesic

On a spherical shell of radius R , the only path with constant θ that qualifies as a geodesic is the equator. To prove this, write the geodesic equation

$$\frac{d^2 q^\theta}{d\tau^2} + \Gamma_{\mu\nu}^\theta \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} = 0,$$

simplifying to

$$\frac{d^2 q^\theta}{d\tau^2} - \sin(\theta) \cos(\theta) \left(\frac{dq^\phi}{d\tau} \right)^2 = 0.$$

The second term is zero only if $\theta = \pi/2$, corresponding to the equator.

To handle the first term, write the differential interval

$$dS^2 = R^2 d\theta^2 + R^2 \sin^2(\theta) d\phi^2,$$

and use the chain rule to reason

$$c^2 = R^2 \left(\frac{d\theta}{d\tau} \right)^2 + R^2 \left(\sin(\theta) \frac{d\phi}{d\tau} \right)^2.$$

where c is a constant speed. From this, we see that setting $d\phi/d\tau = 0$ means $d\theta/d\tau$ is a constant, and thus its second derivative is zero.

It's worth having a formula for the arc length for an embedded curve, geodesic or not, for constant $\theta = \theta_0$. For this we use the differential interval to write

$$S_{\theta_0} = R \sin(\theta_0) \int_{\phi_1}^{\phi_2} d\phi.$$

On the equator we have $\theta_0 = \pi/2$ and $0 \leq \phi \leq 2\pi$, and then $S_{\pi/2} = 2\pi R$.

Longitudinal Geodesics

On a spherical shell, a line of constant longitude has constant ϕ , and all lines of longitude are geodesics. For this case, write the geodesic equation

$$\frac{d^2 q^\phi}{d\tau^2} + \Gamma_{\mu\nu}^\phi \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} = 0,$$

simplifying to

$$\frac{d^2 q^\phi}{d\tau^2} + \cot(\theta) \frac{dq^\phi}{d\tau} \frac{dq^\theta}{d\tau} = 0.$$

For constant ϕ , the above holds trivially.

The arc length of a geodesic of constant $\phi = \phi_0$ is the integral

$$S_{\phi_0} = R \int_{\theta_1}^{\theta_2} d\theta = R\Delta\theta.$$

Fixed Latitude Parallel Transport

Consider a spherical surface that embeds a circle C at constant latitude $\theta = \theta_0$. At $\phi = 0$, a vector V has components $V^\theta = 0, V^\phi = 1$.

Allowing ϕ to vary, we can calculate the components $V^\theta(\phi), V^\phi(\phi)$ as V is parallel-transported around C . For this we choose ϕ to be an affine parameter and write the covariant derivative of V as

$$D_\phi V^\gamma = \frac{dq^\mu}{d\phi} \partial_\mu V^\gamma + \Gamma_{\mu\nu}^\gamma V^\mu \frac{dq^\nu}{d\phi},$$

which must resolve to zero for parallel transport.

For the $\gamma = \theta$ branch, we find $\mu = \nu = \phi$ and thus

$$0 = \frac{dV^\theta}{d\phi} - \sin(\theta) \cos(\theta) V^\phi.$$

For the $\gamma = \phi$ branch, we have $\mu = \theta, \nu = \phi$ and thus

$$0 = \frac{dV^\phi}{d\phi} + \cot(\theta) V^\theta$$

Differentiate each equation with respect to ϕ

$$0 = \frac{d^2 V^\theta}{d\phi^2} - \sin(\theta) \cos(\theta) \frac{dV^\phi}{d\phi}$$

$$0 = \frac{d^2 V^\phi}{d\phi^2} + \cot(\theta) \frac{dV^\theta}{d\phi},$$

and tease out a pair of differential equations

$$0 = \frac{d^2 V^\theta}{d\phi^2} + k^2 V^\theta$$

$$0 = \frac{d^2 V^\phi}{d\phi^2} + k^2 V^\phi,$$

where $k = \cos(\theta)$. Solving this system in light of the initial conditions, we have

$$V^\theta(\phi) = A \sin(k\phi)$$

$$V^\phi(\phi) = \cos(k\phi),$$

where A must be determined from the covariant derivative equation(s). Substituting into either, find $A = \sin(\theta)$ to finish the problem.

1.5 Minkowski Space Geodesics

Recall that the 1+3-dimensional Minkowski space is mapped by spacetime coordinates

$$q^\mu = (ct, \mathbf{x}) ,$$

whose derivative with respect to proper time τ is the four-velocity

$$U^\mu = (\gamma c, \gamma \mathbf{v}) ,$$

where γ is the so-called gamma factor. The normalization condition on the four-velocity is:

$$-c^2 = \eta_{\mu\nu} U^\mu U^\nu$$

Geodesics in Minkowski Space are characterized by

$$\frac{d^2 q^\mu}{d\tau^2} = 0 ,$$

corresponding to non-accelerated motion. Integrating with respect to proper time, we write an equation for straight-line motion:

$$q^\mu(\tau) = q_0^\mu + U_0^\mu \tau$$

Problem 1

Show that a path with constant spatial coordinates is a geodesic of Friedman-Robertson-Walker space.

Null Geodesics

The special case $d\tau = 0$ corresponds to a *null geodesic*, quantified by

$$0 = \eta_{\mu\nu} U^\mu U^\nu .$$

Keep in mind that this means $dS = 0$, or in terms of space-time coordinates:

$$0 = -c^2 dt^2 + d\mathbf{x} \cdot d\mathbf{x}$$

From this, we see that the speed of a body moving on a null geodesic can only be $v = c$, corresponding to a photon.

1.6 Connection by Variation

The geodesic equation can be used with the calculus of variations to directly calculate connection coefficients. To demonstrate, recall that the Friedman-Robertson-Walker space is characterized by

$$dS^2 = -c^2 dt^2 + a^2(t) \delta_{jk} dq^j dq^k ,$$

where $a(t)$ characterizes the acceleration and the indices j, k cover the three spatial dimensions.

For this, we note that the geodesic equation minimizes the integral

$$I = \frac{1}{2} \int g_{\mu\nu} \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} d\tau ,$$

which becomes

$$I = \frac{1}{2} \int \left(-c^2 \left(\frac{dt}{d\tau} \right)^2 + a^2(t) \delta_{\alpha\beta} \frac{dq^\alpha}{d\tau} \frac{dq^\beta}{d\tau} \right) d\tau$$

Choosing the variation

$$t \rightarrow t + \Delta t ,$$

all spatial dependence remains the same as I is written:

$$I + \Delta I = \frac{1}{2} \int d\tau \left(-c^2 \left(\frac{d(t + \Delta t)}{d\tau} \right)^2 + a^2(t + \Delta t) \delta_{\alpha\beta} \frac{dq^\alpha}{d\tau} \frac{dq^\beta}{d\tau} \right)$$

Retaining only terms to first order in Δt after simplifying, the above simplifies to

$$\Delta I = \int \Delta t \left(c^2 \frac{d^2 t}{d\tau^2} + a \frac{da}{dt} \delta_{\alpha\beta} \frac{dq^\alpha}{d\tau} \frac{dq^\beta}{d\tau} \right) d\tau .$$

For the variation ΔI to be zero, we must have

$$0 = c^2 \frac{d^2 t}{d\tau^2} + a \frac{da}{dt} \delta_{\alpha\beta} \frac{dq^\alpha}{d\tau} \frac{dq^\beta}{d\tau} .$$

The above is the time component of the geodesic equation

$$0 = c^2 \frac{d^2 t}{d\tau^2} + \Gamma_{\alpha\beta}^t \frac{dq^\alpha}{d\tau} \frac{dq^\beta}{d\tau} .$$

comparing the two, one must conclude

$$\Gamma_{jk}^t = a \frac{da}{dt} \delta_{jk} ,$$

as expected.

Problem 2

For Friedman-Robertson-Walker space, try the variation $q^k \rightarrow q^k + \Delta q^k$, where k excludes the time component, to derive

$$\Gamma_{tk}^j = \frac{1}{a} \frac{da}{dt} \delta_{jk} .$$

2 Curvature

2.1 Notion of Curvature

A *curved space* is any manifold where the rules of Euclidean geometry don't work.

Embedded Triangle

Consider small intelligent creatures confined to the surface of a large sphere. Unable to see the total surface, such creatures may draw and study 'straight lines' - which we know to be geodesics on a two-sphere - but the creatures don't better. Attempting to draw a triangle on the sphere however, the creatures can detect there's something strange about their confining surface.

In particular, when laying a triangle onto a sphere, one discovers that the sum of interior angles can be *greater* than $\pi/2$. Such a manifold is said to exhibit *positive curvature*.

Correspondingly, one can conceive of cases, as on the surface of a saddle, that the sum of the embedded triangle's interior angles would be lesser than $\pi/2$. Such a manifold has *negative curvature*.

Manual Parallel Transport

A test for curvature on an arbitrary surface involves parallel transport. Suppose you're standing at a given location and are handed an arrow pointing in a fixed direction. You receive instructions to walk a closed path on the surface *without* twisting or swiveling the arrow. That is, you're allowed to step forward, back, left, right, but never to swivel or pivot so the arrow changes direction.

After completing such a walk on a flat manifold, the arrow will have pointed in one and the same direction throughout travel. On a curved manifold however, results will vary.

To illustrate, consider the following circuit on the surface of a sphere:

1. Starting at the north pole, walk to the equator.
2. Strafe any distance along the equator, such as one quarter of its total length.
3. Walk back to the north pole.

Without pivoting the arrow, one will notice on a spherical surface that the direction of the arrow is not pointing in the same direction as before the walk began. This is enough to conclude that the surface you're on has nonzero curvature.

2.2 Riemann Curvature Tensor

The task now is to find a mathematical recipe for detecting curvature on a manifold.

Parallel Transport Analysis

Take an arbitrary manifold $\mathcal{S}$ and consider the parallel transport of a vector $V_{(j)}^\mu$, where $j = 0 \dots 3$ denotes the four corners of a small parallelogram on the manifold.

Proceed by writing the geodesic equation in the form

$$0 = dV^\gamma + \Gamma_{\mu\nu}^\gamma dq^\mu V^\nu,$$

and write a first-order equation for the components of V at any point:

$$\begin{aligned} V_{(j')}^\gamma &= V_{(j)}^\gamma + \int_j^{j'} \frac{dV^\gamma}{dq^\mu} dq^\mu \\ &= V_{(j)}^\gamma + \int_j^{j'} (-\Gamma_{\mu\nu}^\gamma V^\nu) dq^\mu \end{aligned}$$

Next, implement the above on the parallelogram (a_1, b_1) , (a_2, b_1) , (a_2, b_2) , (a_1, b_2) , where $a_2 - a_1 = \delta a$ and $b_2 - b_1 = \delta b$:

$$V_{(1)}^\gamma = V_{(0)}^\gamma + \int_{a_1}^{a_2} (-\Gamma_{\mu\nu}^\gamma (b_1) V_{(b_1)}^\nu) dq^\mu$$

$$V_{(2)}^\gamma = V_{(1)}^\gamma + \int_{b_1}^{b_2} (-\Gamma_{\mu\nu}^\gamma (a_2) V_{(a_2)}^\nu) dq^\mu$$

$$V_{(3)}^\gamma = V_{(2)}^\gamma + \int_{a_2}^{a_1} (-\Gamma_{\mu\nu}^\gamma (b_2) V_{(b_2)}^\nu) dq^\mu$$

$$V_{(0')}^\gamma = V_{(3)}^\gamma + \int_{b_2}^{b_1} (-\Gamma_{\mu\nu}^\gamma (a_1) V_{(a_1)}^\nu) dq^\mu$$

Now define the difference $\delta V^\gamma = V_{(0')}^\gamma - V_{(0)}^\gamma$, which is the sum of the four above integrals, to find:

$$\begin{aligned} \delta V^\gamma &= \int_{a_1}^{a_2} (\Gamma_{\mu\nu}^\gamma (b_2) V_{(b_2)}^\nu - \Gamma_{\mu\nu}^\gamma (b_1) V_{(b_1)}^\nu) dq^\mu \\ &\quad + \int_{b_1}^{b_2} (\Gamma_{\mu\nu}^\gamma (a_1) V_{(a_1)}^\nu - \Gamma_{\mu\nu}^\gamma (a_2) V_{(a_2)}^\nu) dq^\mu \end{aligned}$$

Each parenthesized term is suggestive of a first-order derivative, motivating us to write

$$\begin{aligned} \delta V^\gamma &= \int_{a_1}^{a_2} \delta b^\alpha \frac{\partial}{\partial q^\alpha} (\Gamma_{\mu\nu}^\gamma V^\nu) dq^\mu \\ &\quad - \int_{b_1}^{b_2} \delta a^\mu \frac{\partial}{\partial q^\mu} (\Gamma_{\alpha\nu}^\gamma V^\nu) dq^\alpha, \end{aligned}$$

after swapping dummy indices in the latter term.

Staying to first order, the integrals can be replaced by products:

$$\delta V^\gamma = \delta a^\mu \delta b^\alpha \left(\frac{\partial}{\partial q^\alpha} (\Gamma_{\mu\nu}^\gamma V^\nu) - \frac{\partial}{\partial q^\mu} (\Gamma_{\alpha\nu}^\gamma V^\nu) \right)$$

Simplifying while replacing any derivatives of V using the geodesic equation as written above yields

$$\begin{aligned} \delta V^\gamma &= \delta a^\mu \delta b^\alpha V^\sigma (\partial_\alpha \Gamma_{\mu\sigma}^\gamma - \Gamma_{\mu\nu}^\gamma \Gamma_{\alpha\sigma}^\nu) \\ &\quad + \delta a^\mu \delta b^\alpha V^\sigma (-\partial_\mu \Gamma_{\alpha\sigma}^\gamma + \Gamma_{\alpha\nu}^\gamma \Gamma_{\mu\sigma}^\nu) . \end{aligned}$$

Analyzing this result, first recall that $\delta V^\gamma = 0$ when the space is flat. For curved space however, we cannot assume $\delta V^\gamma = 0$ and thus the parenthesized sum is nonzero. The deciding factor is called the *Riemann curvature tensor*:

$$R_{\sigma\mu\nu}^\gamma = \partial_\mu \Gamma_{\nu\sigma}^\gamma - \partial_\nu \Gamma_{\mu\sigma}^\gamma + \Gamma_{\mu\alpha}^\gamma \Gamma_{\nu\sigma}^\alpha - \Gamma_{\nu\alpha}^\gamma \Gamma_{\mu\sigma}^\alpha$$

Geodesic Deviation Analysis

On the manifold, consider a family of geodesics characterized by $q^\mu(\tau, s)$, where τ is an affine parameter and s is a perpendicular arc length parameter. To remind, the geodesic equation reads

$$\frac{d^2 q^\gamma}{d\tau^2} + \Gamma_{\mu\nu}^\gamma \frac{dq^\mu}{d\tau} \frac{dq^\nu}{d\tau} = 0 .$$

For small deviations S^γ from the geodesic, let us make the replacement

$$q^\gamma \rightarrow q^\gamma + S^\gamma ,$$

where S^γ is perpendicular to the four-velocity at any point. Correspondingly, the connection coefficients become, to first order:

$$\Gamma_{\mu\nu}^\gamma \rightarrow \Gamma_{\mu\nu}^\gamma + S^\alpha \partial_\alpha \Gamma_{\mu\nu}^\gamma$$

With these, the geodesic equation yields

$$\frac{d^2 S^\gamma}{d\tau^2} + U^\mu U^\sigma (2\Gamma_{\mu\alpha}^\gamma \partial_\sigma S^\alpha + S^\alpha \partial_\alpha \Gamma_{\mu\sigma}^\gamma) = 0 .$$

The second derivative term is simplified via the geodesic equation and the chain rule

$$\begin{aligned} \frac{d^2 S^\gamma}{d\tau^2} &= \frac{d}{d\tau} (U^\mu \partial_\mu S^\gamma) \\ &= U^\mu U^\sigma (\partial_\mu \partial_\sigma S^\gamma - \Gamma_{\mu\sigma}^\alpha \partial_\alpha S^\gamma) , \end{aligned}$$

availing the identity:

$$\partial_\mu \partial_\sigma S^\gamma - \Gamma_{\mu\sigma}^\alpha \partial_\alpha S^\gamma = -2\Gamma_{\mu\alpha}^\gamma \partial_\sigma S^\alpha - S^\alpha \partial_\alpha \Gamma_{\mu\sigma}^\gamma$$

To proceed, define the *relative velocity* and *relative acceleration*

$$\begin{aligned} V^\gamma &= \frac{d}{d\tau} S^\gamma = U^\sigma D_\sigma S^\gamma \\ A^\gamma &= \frac{d}{d\tau} V^\gamma , \end{aligned}$$

and substitute V into A to get

$$A^\gamma = (U^\mu D_\mu U^\sigma) D_\sigma S^\gamma + U^\mu U^\sigma D_\mu D_\sigma S^\gamma .$$

The parenthesized term contains the geodesic equation and is identically zero. Unpacking the second term gives

$$A^\gamma = U^\mu U^\sigma (\partial_\mu D_\sigma S^\gamma + \Gamma_{\mu\alpha}^\gamma D_\sigma S^\alpha - \Gamma_{\mu\sigma}^\alpha D_\alpha S^\gamma) ,$$

and then simplify while making use of the above identity to find

$$A^\gamma = U^\mu U^\sigma S^\nu R_{\sigma\mu\nu}^\gamma ,$$

which is to say the relative acceleration is proportional to the Riemann curvature tensor. Note that the above is not an equation of motion. Rather, A^γ tracks adjacent geodesics as a bundle.

Problem 3

Show that the relative acceleration is equivalently written:

$$A^\gamma = U^\mu U^\sigma (D_\mu \partial_\sigma S^\gamma + (D_\mu \Gamma_{\sigma\alpha}^\gamma) S^\alpha + \Gamma_{\sigma\alpha}^\gamma D_\mu S^\alpha)$$

Use the same above identity to arrive at

$$A^\gamma = U^\mu U^\sigma S^\nu (D_\mu \Gamma_{\sigma\nu}^\gamma + \Gamma_{\sigma\alpha}^\gamma \Gamma_{\mu\nu}^\alpha - \partial_\nu \Gamma_{\mu\sigma}^\gamma) .$$

Next use

$$D_\mu \Gamma_{\sigma\nu}^\gamma = \partial_\mu \Gamma_{\sigma\nu}^\gamma + \Gamma_{\mu\alpha}^\gamma \Gamma_{\sigma\nu}^\alpha - \Gamma_{\mu\sigma}^\alpha \Gamma_{\alpha\nu}^\gamma - \Gamma_{\mu\nu}^\alpha \Gamma_{\sigma\alpha}^\gamma$$

to reach the final form for A^γ .

Second Derivative Analysis

Given a vector V on the manifold, consider the operation

$$\Delta V_{\nu\mu\sigma} = (D_\nu D_\mu - D_\mu D_\nu) V_\sigma .$$

Evaluating the first term, we find

$$\begin{aligned} D_\nu D_\mu V_\sigma &= \partial_\nu D_\mu V_\sigma - \Gamma_{\nu\mu}^\alpha D_\alpha V_\sigma - \Gamma_{\nu\sigma}^\alpha D_\mu V_\alpha \\ &= \partial_\nu (\partial_\mu V_\sigma - \Gamma_{\mu\sigma}^\alpha V_\alpha) \\ &\quad - \Gamma_{\nu\mu}^\alpha (\partial_\alpha V_\sigma - \Gamma_{\alpha\sigma}^\gamma V_\gamma) \\ &\quad - \Gamma_{\nu\sigma}^\alpha (\partial_\mu V_\alpha - \Gamma_{\mu\alpha}^\gamma V_\gamma) , \end{aligned}$$

and the second term swaps the indices $\mu \leftrightarrow \nu$. Taking the difference results in

$$\begin{aligned}\Delta V_{\nu\mu\sigma} &= \partial_\mu (\Gamma_{\nu\sigma}^\gamma V_\gamma) - \partial_\nu (\Gamma_{\mu\sigma}^\gamma V_\gamma) \\ &\quad - (\Gamma_{\nu\sigma}^\gamma \partial_\mu V_\gamma - \Gamma_{\nu\sigma}^\alpha \Gamma_{\mu\alpha}^\gamma V_\gamma) \\ &\quad + (\Gamma_{\mu\sigma}^\gamma \partial_\nu V_\gamma - \Gamma_{\mu\sigma}^\alpha \Gamma_{\nu\alpha}^\gamma V_\gamma) ,\end{aligned}$$

simplifying to

$$\Delta V_{\nu\mu\sigma} = R_{\sigma\mu\nu}^\gamma V_\sigma .$$

General Properties

As a type-4 tensor, the Riemann curvature tensor

$$R_{\sigma\mu\nu}^\gamma = \partial_\mu \Gamma_{\nu\sigma}^\gamma - \partial_\nu \Gamma_{\mu\sigma}^\gamma + \Gamma_{\mu\alpha}^\gamma \Gamma_{\nu\sigma}^\alpha - \Gamma_{\nu\alpha}^\gamma \Gamma_{\mu\sigma}^\alpha$$

has N^4 total components on an N -dimensional manifold. Due to certain symmetries though, the number of independent components is significantly reduced. For the covariant form $R_{\alpha\beta\gamma\delta}$, you are encouraged to verify the following:

- $R_{\alpha\beta\gamma\delta}$ is antisymmetric in the last two indices:

$$R_{\alpha\beta\gamma\delta} = -R_{\alpha\beta\delta\gamma}$$

- $R_{\alpha\beta\gamma\delta}$ is antisymmetric in the first two indices:

$$R_{\alpha\beta\gamma\delta} = -R_{\beta\alpha\gamma\delta}$$

- $R_{\alpha\beta\gamma\delta}$ is symmetric when swapping pairs of indices:

$$R_{\alpha\beta\gamma\delta} = R_{\gamma\delta\alpha\beta}$$

- $R_{\alpha\beta\gamma\delta}$ obeys the cyclic identity:

$$R_{\alpha\beta\gamma\delta} + R_{\alpha\gamma\delta\beta} + R_{\alpha\delta\beta\gamma} = 0$$

- Using the above symmetries, one can show that the Riemann curvature tensor in d dimensions has

$$N = \frac{1}{12} d^2 (d^2 - 1)$$

unique components.

- Derivatives of $R_{\alpha\beta\gamma\delta}$ obey the *Bianchi identity*:

$$D_\alpha R_{\beta\gamma\delta\epsilon} + D_\beta R_{\gamma\alpha\delta\epsilon} + D_\gamma R_{\alpha\beta\delta\epsilon} = 0$$

- On a one-dimensional manifold, all components of $R_{\alpha\beta\gamma\delta}$ are zero.

- Curvature in two dimensions is characterized by one independent component R_{0101} .

Spherical Coordinates

Recall that the spherical coordinate system has connection coefficients given by

$$\begin{aligned}\Gamma_{r\theta}^\theta &= \Gamma_{r\phi}^\phi = \frac{1}{r} \\ \Gamma_{\theta\theta}^r &= -r \\ \Gamma_{\phi\phi}^r &= -r \sin^2(\theta) \\ \Gamma_{\phi\phi}^\theta &= -\sin(\theta) \cos(\theta) \\ \Gamma_{\phi\theta}^\phi &= \cot(\theta) .\end{aligned}$$

As given, the spherical coordinate system covers all of three-dimensional Cartesian space, which is surely a flat space. It should follow that all components of the Riemann curvature tensor are zero, for instance

$$\begin{aligned}R_{\phi\theta\phi}^\theta &= \partial_\theta \Gamma_{\phi\phi}^\theta - \cancel{\partial_\phi \Gamma_{\theta\phi}^\theta} + \Gamma_{\theta\alpha}^\theta \Gamma_{\phi\phi}^\alpha - \Gamma_{\phi\alpha}^\theta \Gamma_{\theta\phi}^\alpha \\ &= -\cos^2(\theta) + \sin^2(\theta) - \sin^2(\theta) + \cos^2(\theta) \\ &= 0 ,\end{aligned}$$

or

$$\begin{aligned}R_{\theta\phi\theta}^\phi &= \cancel{\partial_\phi \Gamma_{\theta\theta}^\phi} - \partial_\theta \Gamma_{\phi\theta}^\phi + \Gamma_{\phi\alpha}^\phi \Gamma_{\theta\theta}^\alpha - \Gamma_{\theta\alpha}^\phi \Gamma_{\phi\theta}^\alpha \\ &= -\partial_\theta (\cot(\theta)) - \frac{r}{r} - \cot^2(\theta) \\ &= 0\end{aligned}$$

On the other hand, suppose we fix the spherical radius to a constant $r = r_0$. This corresponds to a curved space, and the component calculated above becomes

$$\begin{aligned}R_{\phi\theta\phi}^\theta &= \partial_\theta \Gamma_{\phi\phi}^\theta - \cancel{\partial_\phi \Gamma_{\theta\phi}^\theta} + \Gamma_{\theta\alpha}^\theta \Gamma_{\phi\phi}^\alpha - \Gamma_{\phi\alpha}^\theta \Gamma_{\theta\phi}^\alpha \\ &= -\cos^2(\theta) + \sin^2(\theta) + \cos^2(\theta) \\ &= \sin^2(\theta) ,\end{aligned}$$

and similarly

$$\begin{aligned}R_{\theta\phi\theta}^\phi &= \cancel{\partial_\phi \Gamma_{\theta\theta}^\phi} - \partial_\theta \Gamma_{\phi\theta}^\phi + \Gamma_{\phi\alpha}^\phi \Gamma_{\theta\theta}^\alpha - \Gamma_{\theta\alpha}^\phi \Gamma_{\phi\theta}^\alpha \\ &= -\partial_\theta (\cot(\theta)) - \cot^2(\theta) \\ &= 1 .\end{aligned}$$

2.3 Ricci Tensor and Ricci Scalar

Ricci Tensor

Consider the Riemann curvature tensor in the form $R_{\mu\beta\nu}^{\alpha}$. A contraction on the first and third indices, i.e. letting $\beta = \alpha$, results in the type $(0, 2)$ *Ricci tensor*:

$$R_{\mu\nu} = R_{\mu\alpha\nu}^{\alpha} = g^{\alpha\sigma} R_{\alpha\mu\sigma\nu}$$

Explicitly, one has:

$$R_{\mu\nu} = R_{\mu\alpha\nu}^{\alpha} = \partial_{\alpha}\Gamma_{\nu\mu}^{\alpha} - \partial_{\nu}\Gamma_{\alpha\mu}^{\alpha} + \Gamma_{\alpha\beta}^{\alpha}\Gamma_{\nu\mu}^{\beta} - \Gamma_{\nu\beta}^{\alpha}\Gamma_{\alpha\mu}^{\beta}$$

Ricci Scalar

Converting the Ricci tensor to type $(1, 1)$ and contracting the remaining index results in the *Ricci scalar*:

$$g^{\mu\nu} R_{\mu\nu} = R_{\mu}^{\mu} = R$$

(The so-called Ricci curvature is related to other types of curvature in differential geometry, namely the mean curvature or Gaussian curvature. In two dimensions, it turns out that R is twice the Gaussian curvature.)

Spherical Shell

Applying the Ricci apparatus to a spherical shell of fixed radius r_0 , prepare by writing the components of the Ricci tensor:

$$\begin{aligned} R_{\theta\theta} &= R_{\theta\alpha\theta}^{\alpha} = R_{\theta\phi\theta}^{\phi} = 1 \\ R_{\phi\phi} &= R_{\phi\alpha\phi}^{\alpha} = R_{\phi\theta\phi}^{\theta} = \sin^2(\theta) \\ R_{\theta\phi} &= R_{\phi\theta} = 0 \end{aligned}$$

Then, the Ricci scalar is:

$$\begin{aligned} R &= g^{\mu\nu} R_{\mu\nu} \\ &= g^{\theta\theta} R_{\theta\theta} + g^{\phi\phi} R_{\phi\phi} \\ &= \frac{1}{r_0^2} + \frac{\sin^2(\theta)}{r_0^2 \sin^2(\theta)} = \frac{2}{r_0^2} \end{aligned}$$

Note that the result would be zero if r were allowed to be variable.

FRW Space

Let us calculate the Ricci tensor for the Friedman-Robertson-Walker space characterized by

$$dS^2 = -c^2 dt^2 + a^2(t) \delta_{jk} dq^j dq^k.$$

Turning the crank through the Riemann curvature tensor, one finds

$$\begin{aligned} R_{tt} &= -\frac{\partial}{\partial t} \Gamma_{tk}^k - \Gamma_{tj}^k \Gamma_{tk}^j \\ &= -\frac{\partial}{\partial t} \left(\frac{1}{a} \frac{da}{dt} \right) - \left(\frac{1}{a} \frac{da}{dt} \right)^2 = -\frac{3}{a} \frac{d^2 a}{dt^2}, \end{aligned}$$

along with

$$\begin{aligned} R_{tk} &= 0 \\ R_{jk} &= 3\delta_{jk} a \frac{d^2 a}{dt^2}. \end{aligned}$$

Problem 4

Show that if $a(t) = \exp Ht$, then the FRW space Ricci tensor is proportional to the metric.

2.4 Einstein Tensor

In terms of the Ricci curvature tensor, Ricci tensor, and Ricci scalar, one may wonder, as Einstein did, just which combination of these has zero covariant derivative. Denoting the result $G^{\mu\nu}$, we seek a type-2 tensor that satisfies

$$\begin{aligned} D_{\nu} G^{\nu\mu} &= 0 \\ D_{\nu} G_{\mu}^{\nu} &= 0 \end{aligned}$$

To begin, take the Bianchi identity and raise the last index using $g^{\sigma\epsilon}$ to write

$$D_{\alpha} R_{\mu\nu\delta}^{\sigma} + D_{\mu} R_{\nu\alpha\delta}^{\sigma} + D_{\nu} R_{\alpha\mu\delta}^{\sigma} = 0.$$

Raise the (new) last index using $g^{\alpha\delta}$ to then get

$$D_{\alpha} R_{\mu\nu}^{\sigma\alpha} + D_{\mu} R_{\nu\alpha}^{\sigma\alpha} + D_{\nu} R_{\alpha\mu}^{\sigma\alpha} = 0,$$

inducing a contraction on α . Let $\sigma = \nu$ to have another contraction:

$$D_{\alpha} R_{\mu\nu}^{\nu\alpha} + D_{\mu} R_{\nu\alpha}^{\nu\alpha} + D_{\nu} R_{\alpha\mu}^{\nu\alpha} = 0$$

The second term is the divergence of the Ricci scalar. Since the Riemann curvature tensor is antisymmetric in the first two and the last two indices, the first and third terms are equal, and we have

$$D_{\nu} R_{\mu}^{\nu} - \frac{1}{2} D_{\mu} R = 0,$$

or

$$D_{\nu} \left(R_{\mu}^{\nu} - \frac{1}{2} g_{\mu}^{\nu} R \right) = 0.$$

The parenthesized term is the type $(1, 1)$ *Einstein tensor*. Tidying up the indices, we conclude:

$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R$$

Perturbed Flat Metric

A manifold that is essentially flat while exhibiting slight curvature admits a metric

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} ,$$

where the components of h are much smaller than the components of the flat metric η . The contravariant instance of the metric is written

$$g^{\mu\nu} = \eta^{\mu\nu} + H^{\mu\nu} ,$$

where H is some function of h .

To solve for H , insert both instances of the metric into the Kronecker delta:

$$\begin{aligned} \delta_{\sigma}^{\mu} &= g^{\mu\alpha} g_{\alpha\sigma} \\ \cancel{\delta}_{\sigma}^{\mu} &= \cancel{\delta}_{\sigma}^{\mu} + g^{\mu\alpha} h_{\alpha\sigma} + H^{\mu\alpha} \eta_{\alpha\sigma} \\ H^{\mu\nu} &= -g^{\mu\alpha} h_{\alpha}^{\nu} \end{aligned}$$

Then, the contravariant metric is written

$$g^{\mu\nu} = \eta^{\mu\nu} - g^{\mu\alpha} h_{\alpha}^{\nu} ,$$

which of course contains itself in the form $g^{\mu\alpha}$, thus $g^{\mu\nu}$ is recursively calculated:

$$\begin{aligned} g^{\mu\nu} &= \eta^{\mu\nu} - (\eta^{\mu\alpha} + H^{\mu\alpha}) h_{\alpha}^{\nu} \\ &= \eta^{\mu\nu} - h^{\mu\nu} + g^{\mu\beta} h_{\beta}^{\alpha} h_{\alpha}^{\nu} \\ &= \eta^{\mu\nu} - h^{\mu\nu} + h^{\mu\alpha} h_{\alpha}^{\nu} - g^{\mu\gamma} h_{\gamma}^{\beta} h_{\beta}^{\alpha} h_{\alpha}^{\nu} \\ &= \eta^{\mu\nu} - h^{\mu\nu} + h^{\mu\alpha} h_{\alpha}^{\nu} - h^{\mu\beta} h_{\beta}^{\alpha} h_{\alpha}^{\nu} \\ &\quad + g^{\mu\delta} h_{\delta}^{\gamma} h_{\gamma}^{\beta} h_{\beta}^{\alpha} h_{\alpha}^{\nu} + \dots \end{aligned}$$

The same result can be derived by expanding g and H as a power series in the dummy parameter λ :

$$\begin{aligned} g^{\mu\alpha} &= \eta^{\mu\alpha} + \lambda H_{(1)}^{\mu\alpha} + \lambda^2 H_{(2)}^{\mu\alpha} + \lambda^3 H_{(3)}^{\mu\alpha} + \dots \\ g_{\alpha\sigma} &= \eta_{\alpha\sigma} + \lambda h_{\alpha\sigma} \end{aligned}$$

Then, the Kronecker delta can be written

$$\begin{aligned} \cancel{\delta}_{\sigma}^{\mu} &= \cancel{\delta}_{\sigma}^{\mu} + \lambda \left(h_{\sigma}^{\mu} + \eta_{\nu\sigma} H_{(1)}^{\mu\nu} \right) \\ &\quad + \lambda^2 \left(h_{\nu\sigma} H_{(1)}^{\mu\nu} + \eta_{\nu\sigma} H_{(2)}^{\mu\nu} \right) \\ &\quad + \lambda^2 \left(h_{\nu\sigma} H_{(2)}^{\mu\nu} + \eta_{\nu\sigma} H_{(3)}^{\mu\nu} \right) + \dots , \end{aligned}$$

where each parenthesized quantity is equal to zero. From this we find

$$\begin{aligned} H_{(1)}^{\mu\nu} &= -\eta^{\alpha\nu} h_{\alpha}^{\mu} = -h^{\mu\nu} \\ H_{(2)}^{\mu\nu} &= -H_{(1)}^{\mu\alpha} h_{\alpha}^{\nu} = h^{\mu\alpha} h_{\alpha}^{\nu} \\ H_{(3)}^{\mu\nu} &= -H_{(2)}^{\mu\alpha} h_{\alpha}^{\nu} = -h^{\mu\beta} h_{\beta}^{\alpha} h_{\alpha}^{\nu} , \end{aligned}$$

and we see the same terms attained by the recursive method.

Acceleration from Curvature

Consider a perturbed Minkowski space admitting a metric

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} .$$

In the low-velocity non-relativistic limit, the proper velocity of a particle at position q^{μ} is approximately

$$U^{\mu} \approx (c, 0, 0, 0) ,$$

which embeds the approximation $dt \approx d\tau$.

Meanwhile, the geodesic equation tells us

$$\frac{d^2 q^{\mu}}{d\tau^2} + \Gamma_{00}^{\mu} U^0 U^0 = 0 ,$$

where the connection coefficient is, to first order in h :

$$\Gamma_{00}^{\mu} = -\frac{1}{2} g^{\mu\nu} \partial_{\nu} g_{00} \approx -\frac{1}{2} \eta^{\mu\nu} \partial_{\nu} h_{00}$$

Evidently, we have

$$\frac{d^2 q^{\mu}}{dt^2} = \frac{c^2}{2} \eta^{\mu\nu} \partial_{\nu} h_{00} ,$$

where the $\mu = 0$ channel tells us h_{00} has no time derivative. In vector form, the spatial components obey:

$$\frac{d^2 \mathbf{q}}{dt^2} = \frac{c^2}{2} \nabla h_{00}$$

Keep in mind that we haven't mentioned any external forces, yet we have the presence of acceleration due to some kind of scalar potential of the form

$$h_{00} = -\frac{2}{c^2} V(\mathbf{q}) .$$

(If this smells like gravity, you would be correct!)

Feeling around for other formulas that illuminate h_{00} , calculate the 00-component of the Ricci tensor to write

$$R_{0\beta 0}^{\beta} = \partial_{\beta} \Gamma_{00}^{\beta} - \partial_0 \Gamma_{\beta 0}^{\beta} + \Gamma_{\beta\alpha}^{\beta} \Gamma_{00}^{\alpha} - \Gamma_{0\alpha}^{\beta} \Gamma_{\beta 0}^{\alpha} ,$$

where to first order, only the first term survives. Then, we find:

$$\begin{aligned} R_{00} &= -\frac{1}{2} \eta^{\mu\nu} \partial_{\mu} \partial_{\nu} h_{00} \\ &= -\frac{1}{2} \partial^{\mu} \partial_{\mu} h_{00} = -\frac{1}{2} \nabla^2 h_{00} \end{aligned}$$

3 Curved Manifolds

3.1 Two-Sphere

Spherical Surface

On the surface of a sphere of constant radius r_0 , the differential line element is

$$dS^2 = r_0^2 \left(d\theta^2 + \sin^2(\theta) d\phi^2 \right).$$

Summarizing some results found previously, we have the metric

$$\begin{aligned} g_{\theta\theta} &= r_0^2 \\ g_{\phi\phi} &= r_0^2 \sin^2(\theta), \end{aligned}$$

along with the connection coefficients

$$\begin{aligned} \Gamma_{\phi\phi}^\theta &= -\sin(\theta) \cos(\theta) \\ \Gamma_{\theta\theta}^\phi &= \cot(\theta), \end{aligned}$$

and components of the Riemann curvature tensor

$$\begin{aligned} R_{\phi\theta\phi}^\theta &= \sin^2(\theta) \\ R_{\theta\phi\theta}^\phi &= 1. \end{aligned}$$

From these, the corresponding Ricci scalar is

$$R = \frac{2}{r_0^2}.$$

Non-Round Sphere

A non-round spherical surface can be expressed by replacing

$$\begin{aligned} \sin(\theta) &\rightarrow f(\theta) \\ \cos(\theta) &\rightarrow g(\theta), \end{aligned}$$

so then

$$\begin{aligned} x &= r_0 f(\theta) \cos(\phi) \\ y &= r_0 f(\theta) \sin(\phi) \\ z &= r_0 g(\theta), \end{aligned}$$

which maintains axial symmetry along $\hat{z}$.

The differential line element becomes:

$$dS^2 = r_0^2 \left(\left(\frac{df}{d\theta} \right)^2 + \left(\frac{dg}{d\theta} \right)^2 \right) d\theta^2 + r_0^2 (f(\theta))^2 d\phi^2$$

If we suppose that the functions f, g obey a trig-like identity

$$\left(\frac{df}{d\theta} \right)^2 + \left(\frac{dg}{d\theta} \right)^2 = 1,$$

then dS^2 simplifies to

$$dS^2 = r_0^2 \left(d\theta^2 + (f(\theta))^2 d\phi^2 \right).$$

By inspecting the above, note that curves of constant ϕ obey the same arc length equation as the round sphere. For curves of constant $\theta = \theta_0$, the arc length becomes

$$S_{\theta_0} = r_0 f(\theta_0) \int_{\phi_1}^{\phi_2} d\phi.$$

The equator is described by

$$S_{\pi/2} = 2\pi r_0 f(\pi/2).$$

Problem 5

Earth is not a perfectly round sphere. The polar radius, measured to be $r_{\text{pol}} = 6357$ km, is slightly less than the equatorial radius $r_{\text{equ}} = 6378$ km. Supposing the surface is modeled using

$$f(\theta) = \sin(\theta) \left(1 + \epsilon \sin^2(\theta) \right),$$

what values of r_0 and ϵ best represent the data? Answer:

$$\begin{aligned} 2\pi r_{\text{equ}} &= 2\pi r_0 \sin(\pi/2) \left(1 + \epsilon \sin^2(\pi/2) \right) \\ 2\pi r_{\text{pol}} &= 2\pi r_0 \\ \epsilon &= \frac{r_{\text{equ}}}{r_{\text{pol}}} - 1 = \frac{21}{6357} \approx 0.003303 \end{aligned}$$

3.2 Three-Sphere

In flat hyperspherical coordinates, we can fix the radius by setting $r = 1$, resulting in a curved *three-sphere*. It's possible to choose a parameterization such that the differential interval is

$$dS^2 = d\theta^2 + \sin^2(\theta) d\psi^2 + \cos^2(\theta) d\phi^2.$$

The corresponding metric reads

$$\begin{aligned} g_{\theta\theta} &= 1 \\ g_{\psi\psi} &= \sin^2(\theta) \\ g_{\phi\phi} &= \cos^2(\theta), \end{aligned}$$

and the nonzero connection coefficients are

$$\begin{aligned}\Gamma_{\psi\psi}^{\theta} &= -\sin(\theta) \cos(\theta) \\ \Gamma_{\phi\phi}^{\theta} &= \sin(\theta) \cos(\theta) \\ \Gamma_{\theta\psi}^{\psi} &= \cot(\theta) \\ \Gamma_{\theta\phi}^{\phi} &= -\tan(\theta) .\end{aligned}$$

Using our standard apparatus, it's straightforward to show that

$$\begin{aligned}R_{\theta\theta} &= 2 \\ R_{\psi\psi} &= 2 \sin^2(\theta) \\ R_{\phi\phi} &= 2 \cos^2(\theta) ,\end{aligned}$$

or more compactly, $R_{\mu\nu} = 2g_{\mu\nu}$. The Ricci scalar is simply

$$R = g^{\mu\nu} R_{\mu\nu} = 2(g^{\mu\nu} g_{\mu\nu}) = 2(3) = 6 .$$

3.3 Two-Sheeted Hyperboloid

A hyperboloid of two sheets can be studied by considering a three-dimensional flat space parameterized by a radius and two angles

$$\begin{aligned}x &= r \sinh(\psi) \cos(\phi) \\ y &= r \sinh(\psi) \sin(\phi) \\ z &= \pm r \cosh(\psi) ,\end{aligned}$$

where

$$\mathbf{S} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}} .$$

Note that $x_{,\mu}^{\mu}$ is not equal to r^2 . Instead, one finds

$$\sqrt{x^2 + y^2 + z^2} = rH ,$$

where

$$H = \sqrt{\cosh^2(\psi) + \sinh^2(\psi)} ,$$

and meanwhile

$$x^2 + y^2 - z^2 = -r^2 .$$

Normalized basis vectors are calculated via

$$\hat{\mathbf{a}}_{(\mu)} = \frac{\partial \mathbf{S} / \partial q^{\mu}}{|\partial \mathbf{S} / \partial q^{\mu}|} ,$$

which work out to

$$\begin{aligned}\hat{\mathbf{r}} &= \frac{\sinh(\psi) (\cos(\phi) \hat{\mathbf{x}} + \sin(\phi) \hat{\mathbf{y}}) + \cosh(\psi) \hat{\mathbf{z}}}{H} \\ \hat{\boldsymbol{\psi}} &= \frac{\cosh(\psi) (\cos(\phi) \hat{\mathbf{x}} + \sin(\phi) \hat{\mathbf{y}}) + \sinh(\psi) \hat{\mathbf{z}}}{H} \\ \hat{\boldsymbol{\phi}} &= -\sin(\phi) \hat{\mathbf{x}} + \cos(\phi) \hat{\mathbf{y}} ,\end{aligned}$$

and furthermore

$$\mathbf{S} = rH \hat{\mathbf{r}} .$$

The differential line element is

$$d\mathbf{S} = (drH + r dH) \hat{\mathbf{r}} + rH d\hat{\mathbf{r}} ,$$

where

$$d\hat{\mathbf{r}} = \left(-\frac{dH}{H}\right) \hat{\mathbf{r}} + d\psi \hat{\boldsymbol{\psi}} + \frac{\sinh(\psi) d\phi}{H} \hat{\boldsymbol{\phi}} .$$

Simplifying, we find

$$d\mathbf{S} = (H dr) \hat{\mathbf{r}} + (rH d\psi) \hat{\boldsymbol{\psi}} + (r \sinh(\psi) d\phi) \hat{\boldsymbol{\phi}} ,$$

implying the differential interval

$$dS^2 = H^2 dr^2 + r^2 H^2 d\psi^2 + r^2 \sinh^2(\psi) d\phi^2 .$$

From the differential interval, read the metric $g_{\mu\nu}$ as

$$\begin{aligned}g_{rr} &= H^2 \\ g_{\psi\psi} &= r^2 H^2 \\ g_{\phi\phi} &= r^2 \sinh^2(\psi) .\end{aligned}$$

Problem 6

Recover the differential interval by the brute force calculation

$$dS^2 = dx^2 + dy^2 + dz^2 .$$

Problem 7

Recover the components of the metric using the tensor transformation law

$$g_{\mu'\nu'} = \frac{\partial q^{\mu}}{\partial q^{\mu'}} \frac{\partial q^{\nu}}{\partial q^{\nu'}} \eta_{\mu\nu} .$$

All of the above applies to a flat space mapped by hyperbolic coordinates. Now we break the flatness by restricting $dr = 0$ such that r is a constant r_0 . Noting that

$$\frac{dH}{d\psi} = \frac{2 \sinh(\psi) \cosh(\psi)}{H} ,$$

the nonzero connection coefficients are

$$\begin{aligned}\Gamma_{\psi\psi}^{\psi} &= \frac{1}{2} g^{\psi\psi} \partial_{\psi} g_{\psi\psi} = \frac{2 \sinh(\psi) \cosh(\psi)}{H^2} \\ \Gamma_{\phi\phi}^{\psi} &= \frac{1}{2} g^{\psi\psi} (\partial_{\phi} g_{\phi\psi} + \partial_{\phi} g_{\psi\phi} - \partial_{\psi} g_{\phi\phi}) \\ &= \frac{-\sinh(\psi) \cosh(\psi)}{H^2}\end{aligned}$$

$$\begin{aligned}\Gamma_{\phi\psi}^{\phi} &= \frac{1}{2}g^{\phi\phi}(\partial_{\phi}g_{\psi\psi} + \partial_{\psi}g_{\phi\phi} - \partial_{\phi}g_{\phi\psi}) \\ &= \coth(\psi) .\end{aligned}$$

From the connection coefficients we calculate the relevant components of the Riemann curvature tensor. Using

$$\frac{\partial}{\partial\psi}\left(\frac{\sinh(\psi)\cosh(\psi)}{H^2}\right) = \frac{1}{H^4},$$

we find:

$$\begin{aligned}R_{\phi\psi\phi\psi} &= g_{\phi\phi}R_{\psi\phi\psi}^{\phi} \\ &= g_{\phi\phi}\left(0 - \partial_{\psi}\Gamma_{\phi\psi}^{\phi} + \Gamma_{\phi\psi}^{\phi}\Gamma_{\psi\psi}^{\psi} - \Gamma_{\psi\phi}^{\phi}\Gamma_{\phi\psi}^{\phi}\right) \\ &= r_0^2\sinh^2(\psi)\left(\frac{1}{H^2}\right)\end{aligned}$$

$$\begin{aligned}R_{\psi\phi\psi\phi} &= g_{\psi\psi}R_{\phi\psi\phi}^{\psi} \\ &= g_{\psi\psi}\left(\partial_{\psi}\Gamma_{\phi\phi}^{\psi} - 0 + \Gamma_{\psi\psi}^{\psi}\Gamma_{\phi\phi}^{\psi} - \Gamma_{\phi\phi}^{\psi}\Gamma_{\psi\phi}^{\phi}\right) \\ &= r_0^2H^2\left(\frac{\sinh^2(\psi)}{H^4}\right)\end{aligned}$$

Note that the Riemann curvature tensor components are equal in accordance with the antisymmetry in the first and last pairs of indices.

Components of the Ricci tensor are calculated via

$$\begin{aligned}R_{\psi\psi} &= g^{\phi\phi}R_{\phi\psi\phi\psi} = \frac{1}{H^2} \\ R_{\phi\phi} &= g^{\psi\psi}R_{\psi\phi\psi\phi} = \frac{\sinh^2(\psi)}{H^4},\end{aligned}$$

which are equivalent to the parenthesized terms above. The Ricci Scalar is finally:

$$R = g^{\psi\psi}R_{\psi\psi} + g^{\phi\phi}R_{\phi\phi} = \frac{2}{r_0^2H^4}$$

The hyperboloid is sphere-like near the vertex and the curvature reflects this accordingly. For large ψ , the hyperboloid approximates a flat space (the wall of a cone).

4 Symmetry and Invariance

4.1 Killing's Equation

Continuing the analysis of manifolds, let S denote the position vector on an N -dimensional manifold parameterized by generalized coordinates $q^1, q^2, q^3, \dots, q^N$. Some key insights into the structure of the manifold can be attained by considering how the metric $g_{\mu\nu}$ changes with respect to translations along vector field lines.

Let K^μ denote a vector field embedded on the manifold. Starting from a point p_0 , take an infinitesimal step along a contour of K^μ to another point p_1 such that

$$q^\mu(p_1) = q^\mu(p_0) + \epsilon K^\mu,$$

where ϵ is a small parameter.

The new location $q^\mu(p_1)$ is treated as a change in coordinates from the point $\tilde{q}$ to q on the manifold such that

$$\tilde{q}^\mu(p_0) = q^\mu(p_1),$$

which is to say p_1 is the new base point in the shifted system.

Given the change of coordinates, the metric changes by an amount $\Delta g_{\mu\nu}$ given as

$$\Delta g_{\mu\nu} = g_{\mu\nu}(q(p_1)) - \tilde{g}_{\mu\nu}(\tilde{q}(p_0)).$$

Unpacking each term on the right to a first-order approximation, the first term

$$g_{\mu\nu}(q(p_1)) \approx g_{\mu\nu}(q(p_0)) + \epsilon K^\gamma \partial_\gamma g_{\mu\nu}(q(p_0)),$$

where the derivative of the metric is given by the compatibility condition

$$\partial_\gamma g_{\mu\nu} = \Gamma_{\gamma\mu}^\alpha g_{\alpha\nu} + \Gamma_{\gamma\nu}^\beta g_{\mu\beta}.$$

The second term $\tilde{g}_{\mu\nu}(\tilde{q}(p_0))$ requires the tensor transformation law

$$\tilde{g}_{\mu\nu}(\tilde{q}(p_0)) = \frac{\partial q^\alpha}{\partial \tilde{q}^\mu} \frac{\partial q^\beta}{\partial \tilde{q}^\nu} g_{\alpha\beta}(q(p_0)),$$

where $q^\mu(p_0)$ in terms of $\tilde{q}^\mu(p_0)$ is written

$$q^\mu(p_0) = \tilde{q}^\mu(p_0) - \epsilon K^\mu,$$

from which we gain

$$\frac{\partial q^\mu}{\partial \tilde{q}^\nu} = \delta_\nu^\mu - \epsilon \partial_\nu K^\mu.$$

Thus at the point p_0 , we find:

$$\begin{aligned} \tilde{g}_{\mu\nu} &\approx \left(\delta_\mu^\alpha - \epsilon \partial_\mu K^\alpha \right) \left(\delta_\nu^\beta - \epsilon \partial_\nu K^\beta \right) g_{\alpha\beta} \\ &\approx g_{\mu\nu} - \epsilon \left(g_{\alpha\nu} \partial_\mu K^\alpha + g_{\mu\beta} \partial_\nu K^\beta \right) + O(\epsilon^2) \end{aligned}$$

Reassembling $\Delta g_{\mu\nu}$, the metric $g_{\mu\nu}$ cancels out and we're left with the first order relationship

$$\begin{aligned} \Delta g_{\mu\nu} &= \epsilon \left(K^\gamma \Gamma_{\gamma\mu}^\alpha g_{\alpha\nu} + K^\gamma \Gamma_{\gamma\nu}^\beta g_{\mu\beta} \right) \\ &\quad - \epsilon \left(g_{\alpha\nu} \partial_\mu K^\alpha + g_{\mu\beta} \partial_\nu K^\beta \right) \\ &= \epsilon g_{\mu\beta} \left(\partial_\nu K^\beta + \Gamma_{\gamma\nu}^\beta K^\gamma \right) \\ &\quad - \epsilon g_{\alpha\nu} \left(\partial_\mu K^\alpha + \Gamma_{\gamma\mu}^\alpha K^\gamma \right) \\ &= \epsilon \left(g_{\mu\beta} D_\nu K^\beta + g_{\alpha\nu} D_\mu K^\alpha \right), \end{aligned}$$

which by the metric compatibility condition simplifies to

$$\Delta g_{\mu\nu} = \epsilon (D_\nu K_\mu + D_\mu K_\nu).$$

Metric Isometry

Under the coordinate transformation given, the condition $\Delta g_{\mu\nu} = 0$ is called *metric isometry*. That is, the metric does not change along the contour of the vector field K^μ .

Killing's Equation

In order to have metric isometry, we can only have

$$D_\nu K_\mu + D_\mu K_\nu = 0,$$

known as *Killing's Equation*, named after Wilhelm Killing. The vector field K^μ is called a *Killing vector*.

Properties

It's possible to show that the maximum number of Killing vectors on a manifold of dimension d is equal to

$$N = d + \frac{1}{2}d(d-1) = \frac{1}{2}d(d+1),$$

which means there are d translational symmetries and $d(d-1)/2$ rotations.

If time is a dimension on the manifold, any constant-velocity shift is the familiar Lorentz boost, and the collection of three boosts, three rotations, and four translations comprise the Poincare group.

Furthermore, it can be shown that curved manifolds carrying maximal symmetry have constant Ricci curvature R such that the Riemann curvature tensor is written

$$R_{\alpha\beta\gamma\delta} = \frac{R}{d(d-1)} (g_{\alpha\gamma}g_{\beta\delta} - g_{\beta\gamma}g_{\alpha\delta}) .$$

4.2 Invariant Quantities

Consider a scalar quantity

$$Q = K_\mu U^\mu ,$$

where K_μ is a vector field satisfying Killing's equation and U^μ is the proper velocity, i.e. tangen vector to the position on a manifold. Taking the full derivative of Q , one finds

$$\begin{aligned} \frac{dQ}{d\tau} &= \frac{dq^\gamma}{d\tau} D_\gamma (K_\mu U^\mu) \\ &= U^\gamma U^\mu D_\gamma K_\mu + K_\mu (U^\gamma D_\gamma U^\mu) , \end{aligned}$$

where the parenthesized term contains the geodesic equation and is thus zero. Due to the symmetry in the γ, μ indices, the above is also written

$$\frac{dQ}{d\tau} = \frac{1}{2} U^\gamma U^\mu (D_\gamma K_\mu + D_\mu K_\gamma) ,$$

which contains Killing's equation and is thus zero, meaning Q is invariant. From this, we reason that any Killing vector K implies an invariant Q on the manifold.

Minkowski Space Invariant

By replacing

$$\begin{aligned} K_\mu &\rightarrow q_\mu \\ U^\mu &\rightarrow P^\mu , \end{aligned}$$

where q^μ is the position four-vector and P_μ is the four-momentum, invariance of the Minkowski space quantity

$$q_\mu P^\mu = -Et + \mathbf{p} \cdot \mathbf{x}$$

emerges trivially from the above analysis.

Cartesian Space Isometry

Killing's equation reduces to

$$\partial_\nu K_\mu + \partial_\mu K_\nu = 0$$

in Cartesian space, as the covariant derivative simplifies to the partial derivative.

For position contours $q^\mu(\lambda)$, the velocity vector U^μ has constant components and automatically satisfies Killing's equation. For this reason, we call the q^μ contour a *direction of symmetry* in Cartesian space with U^μ as the Killing vector.

Similar can be shown for rotations about the origin. Demonstrating on the two-dimensional Cartesian plane, a rotated position vector is given by $V = (-y, x)$. All instances of Killing's equation are trivially zero for this case, except one which requires an extra step to resolve to zero:

$$\partial_x V_y + \partial_y V_x = 1 + (-1) = 0$$

Illustrating the same idea in polar coordinates, rotations in θ with constant r are characterized by $V^\mu = (0, 1)$ or $V_\mu = (0, r^2)$ using the metric $g_{\theta\theta} = r^2$. The nontrivial instance of Killing's equation becomes

$$\begin{aligned} D_r V_\theta + D_\theta V_r &= \partial_r (r^2) - \Gamma_{r\theta}^\alpha V_\alpha + \partial_\theta V_r - \Gamma_{\theta r}^\alpha V_\alpha \\ &= 2r - \frac{r^2}{r} - \frac{r^2}{r} = 0 . \end{aligned}$$

4.3 Conformal Killing's Equation

A variation on Killing's equation dispenses with metric isometry and instead replaces zero with something proportional to the metric such that

$$D_\nu K_\mu + D_\mu K_\nu = f(q^\lambda) g_{\mu\nu} ,$$

where $f(q^\lambda)$ is an arbitrary function.

The previously-invariant quantity Q has a nonzero derivative in light of the conformal Killing's equation:

$$\frac{dQ}{d\tau} = \frac{1}{2} U^\mu U^\nu f(q^\lambda) g_{\mu\nu} = \frac{1}{2} (U^\mu U_\mu) f(q^\lambda)$$

If U^μ traces a null geodesic, then $U^\mu U_\mu = 0$ and Q remains invariant.

Contours in Minkowski space are given by $q^\mu = (ct, \mathbf{x})$. It happens that q^μ satisfies the conformal Killing's equation:

$$\begin{aligned} D_\nu q_\mu + D_\mu q_\nu &= \partial_\nu q_\mu + \partial_\mu q_\nu \\ &= \eta_{\mu\alpha} \frac{\partial}{\partial q^\nu} q^\alpha + \eta_{\beta\nu} \frac{\partial}{\partial q^\mu} q^\beta \\ &= \eta_{\mu\alpha} \delta_\nu^\alpha + \eta_{\beta\nu} \delta_\mu^\beta \\ &= 2\eta_{\mu\nu} \end{aligned}$$

Two-Sphere

The three-dimensional Cartesian space has Killing vectors

$$\begin{aligned} \mathbf{a}_{(x)} &= (1, 0, 0) \\ \mathbf{a}_{(y)} &= (0, 1, 0) \\ \mathbf{a}_{(z)} &= (0, 0, 1) , \end{aligned}$$

which rests on the standard basis vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$.

Meanwhile, consider an embedded two-sphere of unit radius given by

$$\begin{aligned} x &= \sin(\theta) \cos(\phi) \\ y &= \sin(\theta) \sin(\phi) \\ z &= \cos(\theta) \end{aligned}$$

such that

$$dS^2 = d\theta^2 + \sin^2(\theta) d\phi^2 .$$

Between the two coordinate systems, a vector V^μ transforms via

$$V^{\mu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} V^\mu ,$$

or:

$$\begin{aligned} \mathbf{a}_{(x)} &= \cos(\theta) \cos(\phi) \mathbf{a}_{(\theta)} - \frac{\sin(\phi)}{\sin(\theta)} \mathbf{a}_{(\phi)} \\ \mathbf{a}_{(y)} &= \cos(\theta) \sin(\phi) \mathbf{a}_{(\theta)} + \frac{\cos(\phi)}{\sin(\theta)} \mathbf{a}_{(\phi)} \\ \mathbf{a}_{(z)} &= -\sin(\theta) \mathbf{a}_{(\theta)} , \end{aligned}$$

where the replacements $V^\theta = g^{\theta\theta} V_\theta$, $V^\phi = g^{\phi\phi} V_\phi$ are made on the right.

From the above, pick out three conformal Killing vectors:

$$\begin{aligned} X_\theta &= \cos(\theta) \cos(\phi) \\ X_\phi &= -\frac{\sin(\phi)}{\sin(\theta)} \\ Y_\theta &= \cos(\theta) \sin(\phi) \\ Y_\phi &= \frac{\cos(\phi)}{\sin(\theta)} \\ Z_\theta &= -\sin(\theta) \\ Z_\phi &= 0 \end{aligned}$$

Problem 8

Show that each vector X_μ, Y_μ, Z_μ satisfies the conformal Killing's equation.

4.4 Poincare Metric

Part V

Classical Physics

Chapter 29

Newtonian Mechanics

1 Foundation

Newton's laws of motion and universal gravitation, along with the foundations of calculus, were devised around 1687. The modern approach to these subjects differs drastically from that of the seventeenth century, particularly when it comes to *energy*. Newton was satisfied using (what we now know as) force and momentum vectors to make his conclusions. Things like potential energy and conservation of energy were popularized in later centuries by Joule, Helmholtz, and others. Noether informed us in 1918 that any differentiable symmetry of the action of a physical system leads to a conserved quantity.

In the following we use the full arsenal of the 21st century to develop Newtonian mechanics.

1.1 Position, Velocity, Acceleration

Newtonian mechanics assumes a single universal reference frame in which all dynamics take place. In this scheme, it makes sense to define an origin from which all positions $\mathbf{r}(t)$ at all times t are represented as vectors. For a pair of particles or events, we could write:

$$\begin{aligned}\text{position: } & \mathbf{r}(t) \\ \text{displacement (time): } & \mathbf{r}(t_2) - \mathbf{r}(t_1) \\ \text{displacement (space): } & \mathbf{r}_2(t) - \mathbf{r}_1(t) \\ \text{time interval: } & t_2 - t_1 = \Delta t\end{aligned}$$

The average velocity of a particle is defined as the displacement (in time) divided by the time interval. In the differential limit, this becomes the instantaneous velocity $\mathbf{v}(t)$:

$$\mathbf{v}(t) = \frac{d}{dt} \mathbf{r}(t)$$

The average acceleration of a particle is defined as the velocity divided by the time interval. In the differential limit, this becomes the instantaneous acceleration $\mathbf{a}(t)$:

$$\mathbf{a}(t) = \frac{d}{dt} \mathbf{v}(t) = \frac{d^2}{dt^2} \mathbf{r}(t)$$

For a system of N total particles, we label each particle with a subscript i, j , etc., depending on context. For instance, the i th particle has position vector $\mathbf{r}_i$, velocity vector $\mathbf{v}_i$, mass m_i , etc.

1.2 Newton's Laws

The axioms of Newtonian mechanics are the famous Newton's laws, of which there are three:

1. **Law of Inertia:** An object at rest stays at rest, and an object in motion stays in motion with the same speed and in the same direction unless acted upon by an unbalanced force. This means if the total force $\mathbf{F}$ on an object is zero, the velocity doesn't change:

$$\mathbf{F} = 0 \rightarrow \mathbf{v} = \text{constant}$$

2. **Law of Force and Acceleration:** The acceleration of an object is directly proportional to the net force $\mathbf{F}$ acting on the object, is in the same direction as the net force, and is inversely proportional to the mass m of the object:

$$\mathbf{F} = m\mathbf{a}$$

For a system of N particles, Newton's second law reads:

$$m_i \mathbf{a}_i = \sum_{j=1}^N \mathbf{F}_{ij}$$

3. **Law of Action-Reaction:** For every action, there is an equal and opposite reaction. When one object exerts a force on another object, the second object exerts an equal force back on the first object in the opposite direction:

$$\mathbf{F}_{ij} = -\mathbf{F}_{ji}$$

Force as Gradient

In order for Newton's second law to be of use, we need to declare several things about the force vector $\mathbf{F}$, more precisely the interaction between two particles $\mathbf{F}_{ij}$:

1. The force $\mathbf{F}_{ij}$ depends only on the positions $\mathbf{r}_i, \mathbf{r}_j$:

$$\mathbf{F}_{ij} = \mathbf{F}_{ij}(\mathbf{r}_i, \mathbf{r}_j)$$

2. The force vector is defined as the spatial derivative, i.e. the gradient of a scalar function depending only on $\mathbf{r}_i, \mathbf{r}_j$:

$$\mathbf{F}_{ij} = -\frac{\partial}{\partial \mathbf{r}_i} U_{ij}(\mathbf{r}_i, \mathbf{r}_j)$$

3. The $\mathbf{r}$ -dependence in $U_{ij}(\mathbf{r}_i, \mathbf{r}_j)$ occurs as the magnitude of the displacement between $\mathbf{r}_i, \mathbf{r}_j$:

$$U_{ij}(\mathbf{r}_i, \mathbf{r}_j) = U_{ji}(\mathbf{r}_1, \mathbf{r}_2) = U_{ij}(|\mathbf{r}_1 - \mathbf{r}_2|)$$

From the above, we reason that the strength of the interaction between two particles is, as far as space is concerned, only a function of the separation between the particles. This should mean the force vector $\mathbf{F}_{ij}$ is equivalent to a scalar function f_{ij} multiplied by the distance $\mathbf{r}_i - \mathbf{r}_j$ between the particles:

$$\mathbf{F}_{ij} = f_{ij}(|\mathbf{r}_i - \mathbf{r}_j|)(\mathbf{r}_i - \mathbf{r}_j)$$

Equivalently, we have

$$\mathbf{F}_{ij} = f_{ij}(|\mathbf{r}_i - \mathbf{r}_j|) r \hat{\mathbf{r}}_{ij},$$

telling us

$$-\frac{\partial}{\partial \mathbf{r}_i} U_{ij}(r) = f_{ij}(|\mathbf{r}_i - \mathbf{r}_j|) r,$$

or

$$f_{ij}(r) = f_{ji}(r) = \frac{-1}{r} \frac{d}{dr} (U_{ij}(r)).$$

Swapping particle i with particle j , we see $f_{ij}(r)$ remains unchanged but the unit vector $\hat{\mathbf{r}}$ reverses sign. This constitutes a proof of Newton's third law.

In light of the interpretation of force as the gradient of a scalar function, we compile an 'updated' Newton's second law, which contains the other two laws, as:

$$m_i \mathbf{a}_i = \sum_{j=1}^N -\frac{\partial}{\partial \mathbf{r}_i} U_{ij}(|\mathbf{r}_i - \mathbf{r}_j|)$$

1.3 Reference Frame

Galilean Invariance

Newton's laws exhibit the remarkable property that all physics remains the same when introducing a shift of coordinates

$$\mathbf{r}'_i = \mathbf{r}_i - \Delta \mathbf{r},$$

or more strongly

$$\mathbf{r}'_i = \mathbf{r}_i - \mathbf{v}_0 \Delta t,$$

where $\mathbf{v}_0$ is a constant velocity vector. Taking a single time derivative gives an analogous statement for a shift of velocity:

$$\mathbf{v}'_i = \mathbf{v}_i - \mathbf{v}_0$$

The pattern stops before acceleration. Due to the same shift $\Delta \mathbf{r} = \mathbf{v}_0 \Delta t$, the acceleration remains the same:

$$\mathbf{a}'_i = \mathbf{a}_i$$

Problem 1

Prove the notion of Galilean invariance by evaluating:

$$m_i \mathbf{a}'_i = \sum_{j=1}^N -\frac{\partial}{\partial \mathbf{r}'_i} U_{ij}(|\mathbf{r}'_i - \mathbf{r}'_j|)$$

Hint: Simplify until all prime symbols are gone.

Center of Mass

The center of mass of a system of N particles denoted $\mathbf{R}(t)$, is defined as:

$$\mathbf{R}(t) = \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + m_3 \mathbf{r}_3 + \cdots}{m_1 + m_2 + m_3 + \cdots}$$

The denominator is the total mass M of the system

$$M = m_1 + m_2 + m_3 + \cdots = \sum_{j=1}^N m_j,$$

and the quantity $\mathbf{R}(t)$ is tightly expressed in sigma notation:

$$\mathbf{R}(t) = \frac{1}{M} \sum_{j=1}^N m_j \mathbf{r}_j$$

1.4 Momentum

For a system of N interacting particles, it should make intuitive sense that the center of mass coordinate $\mathbf{R}(t)$ does not accelerate or rotate on its own. We establish this formally with the notation of momentum.

Linear Momentum

Take the sum of all products $\mathbf{p}_i = m_i \mathbf{v}_i$ in the system to yield the total linear momentum

$$\mathbf{P}(t) = \sum_{i=1}^N \mathbf{p}_i = \sum_{i=1}^N m_i \mathbf{v}_i ,$$

which is related to the center of mass via

$$\mathbf{P}(t) = M \frac{d}{dt} \mathbf{R}(t) .$$

The total linear momentum vector is a constant of motion. To prove this, take a time derivative to write

$$\begin{aligned} \frac{d}{dt} \mathbf{P}(t) &= \sum_{i=1}^N m_i \frac{d}{dt} \mathbf{v}_i \\ &= \sum_{i=1}^N \sum_{j=1}^N \mathbf{F}_{ij} , \end{aligned}$$

which seems like we did nothing.

The trick now is to swap the indices $i \leftrightarrow j$. The order of the sums stays the same, but the force term reverses the indices:

$$\frac{d}{dt} \mathbf{P}(t) = \sum_{i=1}^N \sum_{j=1}^N \mathbf{F}_{ji}$$

Then, use Newton's third law to restore the original $\mathbf{F}_{ij}$ with a negative sign:

$$\frac{d}{dt} \mathbf{P}(t) = - \sum_{i=1}^N \sum_{j=1}^N \mathbf{F}_{ij}$$

At this point we're saying $d\mathbf{P}(t)/dt$ equals the negative of itself, which can only mean $\mathbf{P}(t)$ is constant in time.

Angular Momentum

The total angular momentum of a system of particles is given by the sum

$$\mathbf{L}(t) = \sum_{i=1}^N \mathbf{r}_i \times m_i \mathbf{v}_i ,$$

which is also constant as we'll prove.

For the time derivative of $\mathbf{L}(t)$, we have

$$\frac{d}{dt} \mathbf{L}(t) = \sum_{i=1}^N \frac{d}{dt} \mathbf{r}_i \times m_i \mathbf{v}_i + \sum_{i=1}^N \mathbf{r}_i \times m_i \frac{d}{dt} \mathbf{v}_i .$$

The first term contains $\mathbf{v}_i \times \mathbf{v}_i$, which is zero by definition of the cross product. Simplifying what's left gives

$$\frac{d}{dt} \mathbf{L}(t) = \sum_{i=1}^N \left(\mathbf{r}_i \times \sum_{j=1}^N \mathbf{F}_{ij} \right) = \sum_{i,j} \mathbf{r}_i \times \mathbf{F}_{ij} ,$$

where substituting

$$\mathbf{F}_{ij} = f_{ij}(r) (\mathbf{r}_i - \mathbf{r}_j)$$

means

$$\frac{d}{dt} \mathbf{L}(t) = - \sum_{ij} f_{ij}(r) \mathbf{r}_i \times \mathbf{r}_j .$$

Problem 2

By swapping indices $i \leftrightarrow j$ and exploiting the cross product, prove:

$$\frac{d}{dt} \mathbf{L}(t) = 0$$

1.5 Symmetry

Similar to Galilean invariance, the equations of Newtonian mechanics exhibit several inherent symmetries, particularly with respect to rotations in space and translations in time. In detail, this means solutions to

$$m_i \mathbf{a}_i = \sum_{j=1}^N - \frac{\partial}{\partial \mathbf{r}_i} U_{ij} (|\mathbf{r}_i - \mathbf{r}_j|)$$

remain intact under the following changes:

1. **Coordinate Translation:** A translation is a constant shift in all position vectors such that

$$\mathbf{r}'_i = \mathbf{r}_i - \mathbf{r}_0 .$$

2. **Time Translation:** A shift of the time coordinate $t' = t - t_0$ implies all position vectors take the form

$$\mathbf{r}'_i(t) = \mathbf{r}_i(t - t_0) .$$

3. **Time Reversal:** A reversal of the time coordinate $t' = -t$ implies all position vectors take the form

$$\mathbf{r}'_i(t) = \mathbf{r}_i(-t) .$$

4. **Coordinate Rotation:** A rotation of all coordinates about the origin is represented by

$$(r'_i)_j = \sum_{k=1}^3 R_{jk} (r_i)_k ,$$

where R_{jk} is the rotation matrix in three dimensions.

For thoroughness we will work each of the above cases. Coordinate translations entail $\mathbf{r}'_i = \mathbf{r}_i - \mathbf{r}_0$, along with $\mathbf{v}'_i = \mathbf{v}_i$ and $\mathbf{a}'_i = \mathbf{a}_i$. We also note that

$$\mathbf{r}'_i - \mathbf{r}'_j = \mathbf{r}_i - \mathbf{r}_j ,$$

and furthermore

$$\frac{\partial}{\partial \mathbf{r}'_i} = \frac{\partial}{\partial \mathbf{r}_i} .$$

Starting with

$$m_i \mathbf{a}'_i = \sum_{j=1}^N -\frac{\partial}{\partial \mathbf{r}'_i} U_{ij} (|\mathbf{r}'_i - \mathbf{r}'_j|)$$

and replacing all primed variables via the above yields the unprimed version. (This repeats the proof of Galilean invariance.)

For time translations, start with $t' = t - t_0$ and take derivatives of

$$\mathbf{r}'_i(t) = \mathbf{r}_i(t - t_0)$$

to find

$$\mathbf{a}'_i(t) = \mathbf{a}_i(t - t_0) .$$

From here it suffices to build the primed version of Newton's second law from the left. Doing so, we have:

$$\begin{aligned} m_i \mathbf{a}'_i(t) &= m_i \mathbf{a}_i(t - t_0) \\ &= \sum_{j=1}^N \mathbf{F}_{ij}(\mathbf{r}_i(t - t_0), \mathbf{r}_j(t - t_0)) \\ &= \sum_{j=1}^N \mathbf{F}_{ij}(\mathbf{r}'_i(t), \mathbf{r}'_j(t)) \end{aligned}$$

Problem 3

Show that Newton's second law is maintained under time reversal symmetry $t \rightarrow -t$.

Proving that Newton's second law remains intact under coordinate rotations is a bit more technical. We first acknowledge the rotation matrix in three dimensions obeys

$$\sum_{k=1}^3 R_{ik} R_{jk} = \sum_{k=1}^3 R_{ki} R_{kj} = \delta_{ij} ,$$

which is to say the transpose is the same as the inverse

$$R^T = R^{-1} ,$$

so if $\mathbf{r}' = R\mathbf{r}$, then $\mathbf{r} = R^T \mathbf{r}'$.

We still need to see what happens to the gradient operator under coordinate rotations. For this we write using the chain rule

$$\frac{\partial}{\partial (r'_i)_j} = \sum_{k=1}^3 \frac{\partial (r_i)_k}{\partial (r'_i)_j} \frac{\partial}{\partial (r_i)_k} = \sum_{k=1}^3 R_{jk} \frac{\partial}{\partial (r_i)_k} ,$$

or

$$\frac{\partial}{\partial \mathbf{r}'_i} = R \frac{\partial}{\partial \mathbf{r}_i} .$$

The proof is now easily worked out:

$$\begin{aligned} m_i \mathbf{a}'_i &= R m_i \mathbf{a}_i = \sum_{j=1}^N -R \frac{\partial}{\partial \mathbf{r}_i} U_{ij} (|\mathbf{r}_i - \mathbf{r}_j|) \\ &= \sum_{j=1}^N -\frac{\partial}{\partial \mathbf{r}'_i} U_{ij} (|\mathbf{r}'_i - \mathbf{r}'_j|) \end{aligned}$$

1.6 Energy

In a system of N particles, consider the force and displacement of the j th particle as it moves from the initial position $\mathbf{r}_i$ to the final position $\mathbf{r}_f$.

Kinetic Energy

The plan is to project the force on the j th particle onto its displacement along the arc of motion, called a *work integral*:

$$W_j = m_j \int_{\mathbf{r}_i}^{\mathbf{r}_f} \mathbf{a}_j \cdot d\mathbf{r}_j$$

The work W has the same dimension as $U(r)$, namely energy units such as Joules.

We can evaluate the work integral without knowing the motion of the particle or the force upon it. First use the chain rule to change the integration domain from spacial to temporal:

$$W_j = m_j \int_{t_i}^{t_f} \mathbf{a}_j \cdot \mathbf{v}_j dt$$

Integrating by parts, let

$$\begin{aligned} U &= m_j \mathbf{v}_j \\ dU &= m_j \mathbf{a}_j dt \\ d\mathbf{V} &= \mathbf{a}_j dt \\ \mathbf{V} &= \mathbf{v}_j \end{aligned}$$

and according to the recipe,

$$W_j = (\mathbf{U} \cdot \mathbf{V}) \Big|_{t_i}^{t_f} - \int_{t_i}^{t_f} \mathbf{V} \cdot d\mathbf{U}.$$

Evaluating, we find

$$W_j = m_j (\mathbf{v}_j \cdot \mathbf{v}_j) \Big|_{t_i}^{t_f} - m_j \int_{t_i}^{t_f} \mathbf{a}_j \cdot \mathbf{v}_j dt,$$

where the remaining integral is just W_j again. Thus the energy added to the j th particle throughout the motion is

$$W_j = \frac{1}{2} m_j v_j^2 \Big|_{t_i}^{t_f},$$

which we interpret as the kinetic energy, denoted T .

At a given instant in any particle's motion, the instantaneous kinetic energy is a function of the instantaneous velocity $\mathbf{v}_j(t)$, precisely given by

$$T_j(\mathbf{v}_j) = \frac{1}{2} m_j (\mathbf{v}_j \cdot \mathbf{v}_j) = \frac{1}{2} m_j v_j^2,$$

where m is the mass of the particle.

Total Energy

For a system of N particles, the total kinetic energy is easy to write down:

$$T = \sum_{i=1}^N T_i = \frac{1}{2} \sum_{i=1}^N m_i v_i^2$$

The total potential energy involves a double sum due to the two indices on U_{ij} . For starters let's try

$$U \propto \sum_{i=1}^N \sum_{j=1}^N U_{ij},$$

but notice this double-counts the total energy. We could slip in a factor of 1/2 and proceed with

$$U = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N U_{ij} = \frac{1}{2} \sum_{i,j} U_{ij},$$

or more efficiently, the configuration

$$U = \sum_{j=1}^N \sum_{i=1}^j U_{ij} = \sum_{i < j} U_{ij}$$

does the same job.

The sum of the kinetic energy and the potential energy is the total energy, denoted E :

$$E = T + U = \frac{1}{2} \sum_{i=1}^N m_i v_i^2 + \sum_{i < j} U_{ij} (|\mathbf{r}_1 - \mathbf{r}_2|)$$

Conservation of Energy

The total energy E in a system of N interacting particles is a constant. Establish this by first calculating

$$\frac{dT}{dt} = \frac{1}{2} \sum_{i=1}^N m_i (2\mathbf{v}_i \cdot \mathbf{a}_i) = \sum_{i,j} \mathbf{v}_i \cdot \mathbf{F}_{ij}.$$

To proceed we must make the right-most term into something symmetric. Impose the swap of indices $i \leftrightarrow j$ and sum the two equations to write

$$\frac{dT}{dt} = \frac{1}{2} \sum_{i,j} (\mathbf{v}_i \cdot \mathbf{F}_{ij} + \mathbf{v}_j \cdot \mathbf{F}_{ji}).$$

Replace each force vector in terms of the gradient of the potential energy, exploiting the symmetry $U_{ij} = U_{ji}$:

$$\frac{dT}{dt} = \frac{-1}{2} \sum_{i,j} \left(\mathbf{v}_i \cdot \left(\frac{\partial}{\partial \mathbf{r}_i} U_{ij} \right) + \mathbf{v}_j \cdot \left(\frac{\partial}{\partial \mathbf{r}_j} U_{ij} \right) \right)$$

Notice this is exactly $-dU/dt$ by the chain rule:

$$\frac{dU}{dt} = \frac{d}{dt} \left(\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N U_{ij} \right) = -\frac{dT}{dt}$$

Putting both results together we find

$$\frac{dE}{dt} = \frac{d}{dt} (T + U) = 0,$$

thus the total energy is conserved.

Example: Collision of Spheres

Consider two spheres of radius $R_{1,2}$ and mass $m_{1,2}$, each moving with uniform velocity $\mathbf{v}_{1,2}$. With this setup, the system has a total energy T (scalar) and linear momentum $\mathbf{P}$ (vector):

$$T = \frac{1}{2}m_1 v_1^2 + \frac{1}{2}m_2 v_2^2$$

$$\mathbf{P} = m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2$$

If the spheres are to make contact via an elastic collision without exchanging mass, each sphere emerges with a new velocity vector $\mathbf{u}_{1,2}$ obeying

$$T = \frac{1}{2}m_1 u_1^2 + \frac{1}{2}m_2 u_2^2$$

$$\mathbf{P} = m_1 \mathbf{u}_1 + m_2 \mathbf{u}_2 ,$$

which is to say energy and momentum are conserved throughout the collision. The task is to solve for $\mathbf{u}_1, \mathbf{u}_2$.

To make the problem easier, we can define a momentum exchange vector $\mathbf{q}$ such that

$$m_1 \mathbf{u}_1 = m_1 \mathbf{v}_1 - \mathbf{q}$$

$$m_2 \mathbf{u}_2 = m_2 \mathbf{v}_2 + \mathbf{q} .$$

This pair of equations can recover the answers $\mathbf{u}_{1,2}$ from $\mathbf{q}$, so the whole problem becomes finding $\mathbf{q}$.

The two spheres exchange momentum at the point of contact, thus $\mathbf{q}$ is normal to each sphere's surface at that point. By the same token, $\mathbf{q}$ is parallel to the vector connecting the center of each sphere. Thus each vector on hand relates by:

$$\mathbf{q} = |\mathbf{q}| \hat{\mathbf{n}} = |\mathbf{q}| \left(\frac{\mathbf{r}_1 - \mathbf{r}_2}{|\mathbf{r}_1 - \mathbf{r}_2|} \right)$$

Since the positions $\mathbf{r}_{1,2}$ are given, the direction of $\mathbf{q}$ is already clear, and the task reduces to finding $|\mathbf{q}|$.

To proceed, square each momentum exchange equation to write

$$\frac{1}{2}m_1 u_1^2 = \frac{1}{2}m_1 v_1^2 - \mathbf{v}_1 \cdot \mathbf{q} + \frac{q^2}{2m_1}$$

$$\frac{1}{2}m_2 u_2^2 = \frac{1}{2}m_2 v_2^2 + \mathbf{v}_2 \cdot \mathbf{q} + \frac{q^2}{2m_2}$$

Take the sum of these, and notice all kinetic energy terms cancel, leaving

$$0 = \mathbf{q} \cdot (\mathbf{v}_1 - \mathbf{v}_2) - q^2 \left(\frac{1}{m_1} + \frac{1}{m_2} \right) .$$

Using $\mathbf{q} = |\mathbf{q}| \hat{\mathbf{n}}$, finally solve for $|\mathbf{q}|$:

$$|\mathbf{q}| = \left(\frac{2m_1 m_2}{m_1 + m_2} \right) \hat{\mathbf{n}} \cdot (\mathbf{v}_1 - \mathbf{v}_2)$$

Chapter 30

Central Forces

1 Planetary Motion

Early Progress

The ‘modern’ understanding of planetary motion arguably began with Johannes Kepler (1571 - 1630), whose career predates the invention of calculus and Newton’s laws of motion by decades. Already familiar with the Heliocentric model of the solar system, Kepler studied meticulously-recorded charts of night sky measurements recorded by Tycho Brahe (1546 - 1601).

Paying attention to the positions of observable planets in the night sky, Kepler astonishingly figured out that planetary orbits were elliptical in shape with the sun at a focus. This became known as Kepler’s first law, which survives to this day among two other laws written by Kepler.

Aware of Kepler’s first law, Newton proposed the existence of a law of mutual Earth-sun attraction that gives rise to elliptical planetary orbits. In the modern vector notation, he began with something like

$$\mathbf{F} = F(r) \hat{\mathbf{r}},$$

and the quest was to find whatever $F(r)$ is.

Using the calculus of his own invention, Newton found the answer to be a unified force depending on the masses involved and the inverse square of the distance separating them. We know this as Newton’s law of universal gravitation.

The plan here is to develop the equations of planetary motion using a similar approach, at least in spirit, to Newton.

Shell Theorem

One assumption we’ll make early on, which happens to be *true*, and will be proven with triple integration, is *any object can be considered as a point mass located at the object’s*

center of mass. For instance, if we need to calculate the gravitational attraction between two asteroids, the shape of each does not matter. Only the center-to-center distance and the mass of each body is important.

Newton’s Second Law

The one-dimensional version of Newton’s second law

$$m \frac{d^2}{dt^2} x(t) = -\frac{d}{dx} U(x)$$

generalizes to more dimensions where the force and acceleration become vectors:

$$m \frac{d^2 \mathbf{r}}{dt^2} = m \frac{d\mathbf{v}}{dt} = m\mathbf{a} = \mathbf{F}$$

I avoided saying exactly how $-dU/dx$ becomes $\mathbf{F}$. Note that in one dimension,

$$F = -\frac{dU}{dx}$$

is true by definition, but the three dimensional version of this requires a vector derivative operator. The exact details aren’t needed in order to proceed.

Newton’s Third Law

The classic phrase, *for every action, there is an equal and opposite reaction*, is Newton’s third law. It means that the force from object 1 onto object 2 is exactly opposite of the force from object 2 onto object 1. This is concisely stated via vectors:

$$\mathbf{F}_{12} = -\mathbf{F}_{21}$$

1.1 Two-Body Problem

Consider two bodies in space, one of mass m_1 at position $\mathbf{r}_1(t)$, and the other of mass m_2 at position $\mathbf{r}_2(t)$. The force imposed onto body 1 by body 2 is given by

$$m_1 \frac{d^2}{dt^2} \mathbf{r}_1(t) = m_1 \frac{d}{dt} \mathbf{v}_1(t) = \mathbf{F}_{12},$$

and the force imposed onto particle 2 by particle 1 is given by

$$m_2 \frac{d^2}{dt^2} \mathbf{r}_2(t) = m_2 \frac{d}{dt} \mathbf{v}_2(t) = \mathbf{F}_{21}.$$

This setup is called the *two-body problem*.

Center of Mass

In the two-body system, the *center of mass* is defined as a point in space $\mathbf{R}(t)$ such that

$$\mathbf{R}(t) = \frac{m_1 \mathbf{r}_1(t) + m_2 \mathbf{r}_2(t)}{m_1 + m_2}.$$

The time derivative of the center of mass gives a quantity called the *center of velocity*:

$$\mathbf{V}(t) = \frac{d}{dt} \mathbf{R}(t) = \frac{m_1 \mathbf{v}_1(t) + m_2 \mathbf{v}_2(t)}{m_1 + m_2}.$$

Taking the time derivative of the center of velocity gives something interesting:

$$\begin{aligned} \frac{d^2}{dt^2} \mathbf{R}(t) &= \frac{m_1 (d\mathbf{v}_1(t)/dt) + m_2 (d\mathbf{v}_2(t)/dt)}{m_1 + m_2} \\ &= \frac{\mathbf{F}_{12} + \mathbf{F}_{21}}{m_1 + m_2} = \frac{\mathbf{F}_{12} - \mathbf{F}_{12}}{m_1 + m_2} = 0 \end{aligned}$$

Evidently, the second derivative of the center of mass is precisely zero because $\mathbf{F}_{12} = -\mathbf{F}_{21}$, regardless of how the forces act. This means that two bodies, while free to move individually, are not accelerating anywhere as a group. Moreover, this result proves that the center of velocity $\mathbf{V}$ is a constant $\mathbf{V}_0$.

Relative Displacement

If the distance separating the two bodies is r , define a vector

$$\mathbf{r}(t) = \mathbf{r}_1(t) - \mathbf{r}_2(t)$$

with $|\mathbf{r}| = r$, capturing the relative displacement between the two.

Listing this with the center of mass $\mathbf{R}(t)$, we have a system of two equations that can be solved for $\mathbf{r}_1(t)$, $\mathbf{r}_2(t)$

separately: (We know everything is a function of t by now, so drop the extra notation.)

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{R} + \frac{m_2}{m_1 + m_2} \mathbf{r} \\ \mathbf{r}_2 &= \mathbf{R} - \frac{m_1}{m_1 + m_2} \mathbf{r} \end{aligned}$$

Reduced Mass

From the equations above, multiply through by m_1 , m_2 , respectively, and take two time derivatives:

$$\begin{aligned} m_1 \frac{d^2 \mathbf{r}_1}{dt^2} &= m_1 \cancel{\frac{d^2 \mathbf{R}}{dt^2}} + \frac{m_1 m_2}{m_1 + m_2} \frac{d^2 \mathbf{r}}{dt^2} \\ m_2 \frac{d^2 \mathbf{r}_2}{dt^2} &= m_2 \cancel{\frac{d^2 \mathbf{R}}{dt^2}} - \frac{m_1 m_2}{m_1 + m_2} \frac{d^2 \mathbf{r}}{dt^2} \end{aligned}$$

These results say the same thing, as the left sides are $\mathbf{F}_{12}$, $\mathbf{F}_{21}$, respectively, and the right sides differ by the proper negative sign.

Evidently, we have

$$\mathbf{F}_{12} = \frac{m_1 m_2}{m_1 + m_2} \frac{d^2 \mathbf{r}}{dt^2}.$$

That is, there is only one force equation to worry about, and thus one position to worry about if we work with the relative displacement vector $\mathbf{r}$ rather than two explicit position vectors $\mathbf{r}_{1,2}$.

The price we pay is the mass term became a mess. This group of symbols is called the *reduced mass*:

$$m_* = \frac{m_1 m_2}{m_1 + m_2}$$

Representing the effective mass of the total system as m_* , the two-body problem is summarized in one equation:

$$\mathbf{F}_{12} = m_* \frac{d^2 \mathbf{r}}{dt^2} = m_* \mathbf{a}$$

A handy identity involving the reduced mass, somewhat reminiscent of resistors in parallel, goes as:

$$\frac{1}{m_*} = \frac{1}{m_1} + \frac{1}{m_2}$$

Linear Momentum

The linear momentum $\mathbf{p} = m_* \mathbf{v}$ is not constant in the two-body system. To see this, take a derivative and simplify

using Newton's third law:

$$\begin{aligned}\frac{d\mathbf{p}}{dt} &= m_* \left(\frac{d\mathbf{v}_1}{dt} - \frac{d\mathbf{v}_2}{dt} \right) = m_* \left(\frac{\mathbf{F}_{12}}{m_1} - \frac{\mathbf{F}_{21}}{m_2} \right) \\ \frac{d\mathbf{p}}{dt} &= \mathbf{F}_{12}\end{aligned}$$

1.2 Angular Momentum

Alongside the notion of forces, we'll need to put the ideas of angular momentum to use. In particular, we can show that the angular momentum of the two-body system is constant, and find what it is.

By definition, the angular momentum $\mathbf{L}$ of the two-body system reads

$$\mathbf{L} = m_* \mathbf{r} \times \mathbf{v},$$

where $\mathbf{r}$ is the relative displacement vector, and $\mathbf{v}$ is its time derivative. Now calculate the time derivative of $\mathbf{L}$:

$$\begin{aligned}\frac{d}{dt} \mathbf{L} &= m_* \frac{d}{dt} (\mathbf{r} \times \mathbf{v}) \\ &= m_* \left(\mathbf{v} \times \mathbf{v} + \mathbf{r} \times \frac{d\mathbf{v}}{dt} \right) \\ &= \mathbf{r} \times \mathbf{F}\end{aligned}$$

For the remaining cross product to vanish, we go back to Newton's original assumption that

$$\mathbf{F} = F(r) \hat{\mathbf{r}},$$

which means the force vector and the displacement vector are parallel. Using this, we see that the derivative of $\mathbf{L}$ resolves to zero.

Without knowing the exact motion of the two-body system, we can still write a formula for the angular momentum. For some $r(t)$, $\theta(t)$, we have, in polar coordinates:

$$\begin{aligned}\mathbf{r} &= r \hat{\mathbf{r}} \\ \mathbf{v} &= \frac{dr}{dt} \hat{\mathbf{r}} + r \frac{d\theta}{dt} \hat{\boldsymbol{\theta}}\end{aligned}$$

Remembering $\hat{\mathbf{r}} \times \hat{\mathbf{r}}$ is zero, we then have

$$\mathbf{L} = m_* r^2 \frac{d\theta}{dt} (\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}).$$

The angular momentum is a constant vector that points perpendicular to the plane of motion. We take its magnitude

$$L = m_* r^2 \frac{d\theta}{dt}$$

as a constant of motion in the two-body system.

It's easy to show that the position vector and the angular momentum vector are always perpendicular. Starting with the definition of $\mathbf{L}$, project $\mathbf{r}$ into both sides:

$$\mathbf{r} \cdot \mathbf{L} = m_* \mathbf{r} \cdot (\mathbf{r} \times \mathbf{v}),$$

and then make use of the triple product:

$$\mathbf{r} \cdot \mathbf{L} = m_* \mathbf{v} \cdot (\mathbf{r} \times \mathbf{r}) = 0$$

1.3 Inverse-Square Acceleration

We've made it this far without knowing the magnitude gravitational force $F(r)$, although we have harmlessly assumed that gravity acts in a straight line. Here we will derive the proper gravitational force by using Kepler's first law as a starting point.

In detail, Kepler noticed that the orbit of any planet around the sun takes an elliptical form described by

$$r(\theta) = \frac{r_0}{1 + e \cos(\theta)},$$

where e is the *eccentricity* of the orbit, and r_0 is a positive characteristic length. Notice that $r(\theta)$ as written places the origin (the sun) at the *right* focus of the ellipse. Reverse the sign on the cosine term for the sun at the left focus.

To really get started, take the time derivative of the (constant) angular momentum of the two-body system:

$$0 = \frac{dL}{dt} = m_* r \left(2 \frac{dr}{dt} \frac{d\theta}{dt} + r \frac{d^2\theta}{dt^2} \right)$$

Perhaps you recognize the parenthesized term as being identically the $\hat{\boldsymbol{\theta}}$ -component of the acceleration vector in polar coordinates. In terms of L , the acceleration vector is

$$\mathbf{a} = \left(\frac{d^2r}{dt^2} - \frac{L^2}{m_*^2 r^3} \right) \hat{\mathbf{r}} + \frac{1}{m_* r} \left(\frac{dL}{dt} \right) \hat{\boldsymbol{\theta}}.$$

We need the polar form of the ellipse to calculate d^2r/dt^2 . For this, we find, after simplifying,

$$\frac{dr}{dt} = \frac{d\theta}{dt} \frac{d}{d\theta} \left(\frac{r_0}{1 + e \cos(\theta)} \right) = \frac{L}{m_* r_0} e \sin(\theta),$$

and keep going to the second derivative:

$$\frac{d^2r}{dt^2} = \frac{L^2}{m_*^2 r^2} e \cos(\theta) = \frac{L^2}{m_*^2 r^2} \left(\frac{1}{r} - \frac{1}{r_0} \right)$$

The full acceleration vector then reads

$$\mathbf{a} = \frac{L^2}{m_*^2 r^2} \left(\frac{1}{r} - \frac{1}{r_0} - \frac{1}{r} \right) \hat{\mathbf{r}},$$

which simplifies nicely:

$$\mathbf{a} = \frac{-L^2}{m_*^2 r_0} \frac{\hat{\mathbf{r}}}{r^2}$$

This finally reveals the nature of $F(r)$. The r -dependence is present as $-1/r^2$, hence the name inverse-square acceleration.

Going back to the equations that led to the reduced mass, i.e.

$$\begin{aligned} m_1 \frac{d^2 \mathbf{r}_1}{dt^2} &= m_* \frac{d^2 \mathbf{r}}{dt^2} \\ m_2 \frac{d^2 \mathbf{r}_2}{dt^2} &= -m_* \frac{d^2 \mathbf{r}}{dt^2}, \end{aligned}$$

we can solve for the absolute acceleration of each body:

$$\begin{aligned} \mathbf{a}_1 &= \frac{m_*}{m_1} \mathbf{a} \\ \mathbf{a}_2 &= \frac{-m_*}{m_2} \mathbf{a} \end{aligned}$$

Eliminate $\mathbf{a}$ between the two equations to recover Newton's third law:

$$m_1 \mathbf{a}_1 + m_2 \mathbf{a}_2 = 0$$

1.4 Universal Gravitation

Enough ground work has been done to push toward Newton's universal law of gravitation.

Recall the absolute acceleration of each body $\mathbf{a}_1$, $\mathbf{a}_2$, and replace the reduced mass m_* and acceleration $\mathbf{a}$ with expanded forms:

$$\begin{aligned} \mathbf{a}_1 &= \left(\frac{m_2}{m_1 + m_2} \right) \frac{-L^2}{m_*^2 r_0} \frac{\hat{\mathbf{r}}}{r^2} \\ \mathbf{a}_2 &= \left(\frac{-m_1}{m_1 + m_2} \right) \frac{-L^2}{m_*^2 r_0} \frac{\hat{\mathbf{r}}}{r^2} \end{aligned}$$

Multiply each equation through by m_1 , m_2 , respectively to turn accelerations into forces:

$$\begin{aligned} \mathbf{F}_{12} &= \left(\frac{m_1 m_2}{m_1 + m_2} \right) \frac{-L^2}{m_*^2 r_0} \frac{\hat{\mathbf{r}}}{r^2} \\ \mathbf{F}_{21} &= \left(\frac{-m_1 m_2}{m_1 + m_2} \right) \frac{-L^2}{m_*^2 r_0} \frac{\hat{\mathbf{r}}}{r^2} \end{aligned}$$

Newton decided to introduce a new proportionality constant G , named after 'gravity', to wrangle all of the constants inherent to the two-body, system such that

$$G = \left(\frac{1}{m_1 + m_2} \right) \frac{L^2}{m_*^2 r_0}.$$

Of course, both force equations are saying the same thing due to Newton's third law, thus we write Newton's law of universal gravitation:

$$\mathbf{F}_{12} = -G \frac{m_1 m_2}{r^2} \hat{\mathbf{r}}$$

Note that the force vector bears the 12-subscript and not the other way around. The subscript is often omitted because the unit vector $\hat{\mathbf{r}}$ has an implied 12-subscript that goes back to the definition of $\mathbf{r}$.

While this calculation was set up in the context of planetary motion, note that the gravitational force is in fact *universal*, which is to say that every pair of particles in the universe obeys the same law.

Problem 1

Show that

$$G = \frac{L^2}{m_* m_1 m_2 r_0}.$$

1.5 Equations of Motion

With the law of universal gravitation on hand, we should be able to run the analysis in reverse by starting with $\mathbf{F}_{12}$ and finishing with the shape of the ellipse, along with all other allowed possibilities.

Acceleration

Use

$$L = m_* r^2 \frac{d\theta}{dt}$$

to eliminate $1/r^2$ in the force vector:

$$\mathbf{F}_{12} = -G m_1 m_2 \frac{m_*}{L} \frac{d\theta}{dt} \hat{\mathbf{r}}$$

Also replace $\mathbf{F}_{12}$ to keep simplifying

$$m_1 \mathbf{a}_1 = m_* \mathbf{a} = -G m_1 m_2 \frac{m_*}{L} \frac{d\theta}{dt} \hat{\mathbf{r}},$$

and solve for the relative acceleration:

$$\mathbf{a} = -G \frac{m_1 m_2}{L} \frac{d\theta}{dt} \hat{\mathbf{r}}$$

Velocity

To proceed, replace the acceleration vector as the derivative of the relative velocity by $\mathbf{a} = d\mathbf{v}/dt$. Also replace $\hat{\mathbf{r}}$ via $-\hat{\mathbf{r}} = d\hat{\boldsymbol{\theta}}/d\theta$ to get

$$\frac{d\mathbf{v}}{dt} = G \frac{m_1 m_2}{L} \frac{d\theta}{dt} \frac{d\hat{\boldsymbol{\theta}}}{d\theta},$$

simplifying with the chain rule to:

$$d\mathbf{v} = G \frac{m_1 m_2}{L} d\hat{\boldsymbol{\theta}}$$

Integrate both sides of the above to get a vector equation for the velocity

$$\mathbf{v}(t) = G \frac{m_1 m_2}{L} \hat{\boldsymbol{\theta}}(t) + \mathbf{v}_0,$$

where $\mathbf{v}_0$ is the integration constant. Letting $\theta = 0$ correspond with the positive x -axis, it must be that $\mathbf{v}_0 = v_0 \hat{\mathbf{y}}$.

Position

To goal is get hold of a position equation $r(\theta)$. To get closer, calculate the full angular momentum vector:

$$\begin{aligned} \mathbf{L} &= m_* \mathbf{r} \times \mathbf{v} \\ &= m_* \mathbf{r} \times \left(G \frac{m_1 m_2}{L} \hat{\boldsymbol{\theta}} + v_0 \hat{\mathbf{y}} \right) \\ &= m_* G \frac{m_1 m_2}{L} r (\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}) + m_* v_0 r (\hat{\mathbf{r}} \times \hat{\mathbf{y}}) \end{aligned}$$

To handle the cross products, note that

$$\begin{aligned} |\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}| &= 1 \\ |\hat{\mathbf{r}} \times \hat{\mathbf{y}}| &= |\cos(\theta)|, \end{aligned}$$

and we can work with just magnitudes:

$$L = m_* G \frac{m_1 m_2}{L} r + m_* v_0 r \cos(\theta)$$

To simplify, use

$$G = \frac{L^2}{m_* m_1 m_2 r_0},$$

and work to isolate r , arriving at

$$r_0 = r \left(1 + r_0 \frac{m_* v_0}{L} \cos(\theta) \right),$$

and finally get

$$r(\theta) = \frac{r_0}{1 + (m_* r_0 v_0 / L) \cos(\theta)}.$$

With $r(\theta)$ known, the position vector is straightforwardly written:

$$\mathbf{r} = r(\theta) \hat{\mathbf{r}}$$

Eccentricity

Comparing the above to the general form of a conic section in polar coordinates, we pick out the eccentricity to be

$$e = \frac{m_* r_0 v_0}{L}.$$

Circular orbits arise from the special case $v_0 = 0$. Another special case is $e = 1$ for a parabolic trajectory. For all $e < 1$, the orbit is strictly an ellipse. For $e > 1$, the path (also technically an orbit) is hyperbolic.

This surely nails the case shut for Kepler's first law. All results reinforce the fact that planetary orbits occur on ellipses with the sun at a focus.

The eccentricity can be expressed by a variety of combinations of terms. For a version without L , one can find

$$e = \frac{\sqrt{r_0} v_0}{\sqrt{G(m_1 + m_2)}},$$

or, if you need to get rid of r_0 :

$$e = \frac{v_0 L}{G m_1 m_2}$$

In terms of the eccentricity, the equations of motion can be simplified. For the position, we simply have

$$r(\theta) = \frac{r_0}{1 + e \cos(\theta)}.$$

For the velocity and acceleration, shuffle the constants around to establish

$$\frac{G m_1 m_2}{L} = \frac{v_0}{e},$$

which is only defined for non-circular orbits. With this, we have:

$$\begin{aligned} \mathbf{v} &= v_0 \left(\frac{\hat{\boldsymbol{\theta}}}{e} + \hat{\mathbf{y}} \right) \\ \mathbf{a} &= \frac{-v_0}{e} \frac{d\theta}{dt} \hat{\mathbf{r}} \end{aligned}$$

1.6 Runge-Lorenz Vector

The two-body problem exhibits conservation of angular momentum via the constant vector $\mathbf{L}$. There is, in fact, another constant vector of motion lurking about called the *Runge-Lorenz* vector

$$\mathbf{Z} = \mathbf{v} \times \mathbf{L} - G m_1 m_2 \hat{\mathbf{r}}.$$

Constant of Motion

Take a time derivative to prove $\mathbf{Z}$ is constant:

$$\begin{aligned}\frac{d}{dt}\mathbf{Z} &= \frac{d}{dt}(\mathbf{v} \times \mathbf{L}) - Gm_1m_2 \frac{d\hat{\mathbf{r}}}{dt} \\ &= \frac{d\mathbf{v}}{dt} \times \mathbf{L} + \mathbf{v} \times \frac{d\mathbf{L}}{dt} - Gm_1m_2 \frac{d\hat{\mathbf{r}}}{dt}\end{aligned}$$

Keep simplifying with

$$\frac{d\mathbf{v}}{dt} = \frac{1}{m_*}\mathbf{F} = -G \frac{m_1m_2}{m_*r^2} \hat{\mathbf{r}},$$

and also with $\mathbf{L} = m_*\mathbf{r} \times \mathbf{v}$, so we have

$$\frac{d}{dt}\mathbf{Z} = Gm_1m_2 \left(-\frac{\hat{\mathbf{r}} \times (\mathbf{r} \times \mathbf{v})}{r^2} - \frac{d\hat{\mathbf{r}}}{dt} \right).$$

Replace $\mathbf{v}$ with its polar expression and note that

$$\mathbf{r} \times \mathbf{v} = \mathbf{r} \times \left(\frac{dr}{dt} \hat{\mathbf{r}} + r \frac{d\theta}{dt} \hat{\boldsymbol{\theta}} \right) = r^2 \frac{d\theta}{dt} (\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}),$$

and furthermore, using the BAC-CAB formula:

$$\hat{\mathbf{r}} \times (\mathbf{r} \times \mathbf{v}) = r^2 \frac{d\theta}{dt} \hat{\mathbf{r}} \times (\hat{\mathbf{r}} \times \hat{\boldsymbol{\theta}}) = -r^2 \frac{d\theta}{dt} \hat{\boldsymbol{\theta}}$$

Summarizing, we find

$$\frac{d}{dt}\mathbf{Z} = Gm_1m_2 \left(\frac{d\theta}{dt} \hat{\boldsymbol{\theta}} - \frac{d\hat{\mathbf{r}}}{dt} \right) = 0$$

as proposed.

Perigee

With $\mathbf{Z}$ known to be constant, we're free to evaluate it at any point along the trajectory. Choose a point $\mathbf{r}_p = r_p \hat{\mathbf{x}}$ that has $\mathbf{v}_p \cdot \mathbf{r}_p = 0$, called a *perigee*:

$$\begin{aligned}\mathbf{Z} &= \mathbf{v}_p \times \mathbf{L} - Gm_1m_2 \hat{\mathbf{x}} \\ &= \mathbf{v}_p \times (m_*r_p \hat{\mathbf{x}} \times \mathbf{v}_p) - Gm_1m_2 \hat{\mathbf{x}} \\ &= (m_*r_p v_p^2 - Gm_1m_2) \hat{\mathbf{x}}\end{aligned}$$

At the perigee, the velocity $\mathbf{v}_p$ is momentarily equal to $r_p d\theta/dt$, which we'll call

$$v_p = r_p \omega_p.$$

In the same notation, the angular momentum is

$$L = m_*r_p^2\omega_p = m_*r_p v_p,$$

and the vector $\mathbf{Z}$ becomes

$$\mathbf{Z} = \left(\frac{L^2}{m_*r_p} - Gm_1m_2 \right) \hat{\mathbf{x}}.$$

We can keep simplifying. Replace L^2 with the expression involving G :

$$\mathbf{Z} = Gm_1m_2 \left(\frac{r_0}{r_p} - 1 \right) \hat{\mathbf{x}}.$$

The ratio r_0/r_p can be calculated by setting $\theta = 0$ in the polar equation $r(\theta)$ for a conic section:

$$r_p = \frac{r_0}{1 + (r_0 m_* v_0 / L)} = \frac{r_0}{1 + e}$$

Finally, the simplest form for $\mathbf{Z}$ is:

$$\mathbf{Z} = Gm_1m_2 e \hat{\mathbf{x}}$$

What $\mathbf{Z}$ tells us, apart from containing all information about the trajectory, is that all gravitational trajectories contain at least one perigee, defining the x-axis of the coordinate system about which the motion is symmetric.

Apogee

The perigee is also known as the nearest distance attained between the two bodies. For an elliptical orbit or hyperbolic orbit, the perigee is given by $\theta = 0$:

$$r_{\text{perigee}} = \frac{r_0}{1 + e}$$

For elliptical orbits, there is also the notion of *apogee*, which is the furthest distance attained between the two bodies. Set $\theta = \pi$ to find

$$r_{\text{apogee}} = \frac{r_0}{1 - e}$$

Problem 2

Take derivatives of

$$r(\theta) = \frac{r_0}{1 + e \cos(\theta)}$$

to verify the locations of the perigee and apogee.

Problem 3

Show that:

$$e = \left| \frac{r_p - r_a}{r_p + r_a} \right|$$

Conic Trajectory

The Runge-Lorenz vector

$$\mathbf{Z} = \mathbf{v} \times \mathbf{L} - Gm_1m_2 \hat{\mathbf{r}},$$

together with its particular expression

$$\mathbf{Z} = Gm_1m_2 e \hat{\mathbf{x}}$$

can be used together to quickly recover the polar equation for conic sections by projecting the position vector across the equation and simplifying:

$$\begin{aligned} \mathbf{r} \cdot \mathbf{Z} &= \mathbf{r} \cdot (\mathbf{v} \times \mathbf{L}) - Gm_1m_2 \mathbf{r} \cdot \hat{\mathbf{r}} \\ rZ \cos(\theta) &= \mathbf{L} \cdot (\mathbf{r} \times \mathbf{v}) - Gm_1m_2 r \\ Gm_1m_2 r e \cos(\theta) &= \frac{L^2}{m_*} - Gm_1m_2 r \end{aligned}$$

Now solve for $r(\theta)$ and simplify more:

$$\begin{aligned} r(\theta) &= \left(\frac{L^2}{Gm_1m_2m_*} \right) \frac{1}{1 + e \cos(\theta)} \\ &= \frac{r_0}{1 + e \cos(\theta)} \end{aligned}$$

Relation to Ellipse

An ellipse is classified by two perpendicular lengths we know as the semi-major and semi-minor axes, denoted a , b , respectively. By studying the ellipse, it's straightforward to show that

$$a = \frac{r_0}{1 - e^2},$$

along with

$$b = \frac{r_0}{\sqrt{1 - e^2}},$$

so that

$$\frac{b}{a} = \sqrt{1 - e^2}$$

and

$$r_0 = \frac{b^2}{a}.$$

The a -equation can be derived by taking the difference between $r(0)$ and $r(\pi)$, i.e. the distance between the perigee and apogee. This pair of points defines the distance $2a$.

The b -equation can be derived by finding r_* , θ_* that correspond to $y = b$, the highest point on the ellipse:

$$\begin{aligned} 0 &= \frac{d}{d\theta} (y(\theta)) = \frac{d}{d\theta} (r(\theta) \sin(\theta)) \Big|_{r_*, \theta_*} \\ &= \left(\frac{r_0 e \sin^2(\theta)}{(1 + e \cos(\theta))^2} + \frac{r_0 \cos(\theta)}{1 + e \cos(\theta)} \right) \Big|_{r_*, \theta_*} \\ &= \frac{r_*^2}{r_0} (e + \cos(\theta_*)) \end{aligned}$$

Evidently, we have

$$\cos(\theta_*) = -e.$$

Taking this with

$$\begin{aligned} b &= r_* \sin(\theta_*) \\ r_* &= \sqrt{e^2 a^2 + b^2} \end{aligned}$$

is enough to finish the job. Note that similar relationships can be drawn for hyperbolic orbits.

Problem 4

Show that $\mathbf{r} \cdot \mathbf{v} = 0$ is true only at the apogee and perigee.

Dimensionless Runge-Lorenz

The Runge-Lorenz vector can be made into a dimensionless vector ϵ by dividing Gm_1m_2 across the whole equation

$$\epsilon = \frac{\mathbf{v} \times \mathbf{L}}{Gm_1m_2} - \hat{\mathbf{r}},$$

where by the properties of $\mathbf{Z}$, we also know

$$\epsilon = e \hat{\mathbf{x}}.$$

With this setup, write

$$\hat{\mathbf{r}} + e \hat{\mathbf{x}} = \frac{\mathbf{v} \times \mathbf{L}}{Gm_1m_2},$$

and then project $\mathbf{r}$ into each side to recover the equation of a conic section:

$$r(1 + e \cos(\theta)) = \frac{\mathbf{r} \cdot (\mathbf{v} \times \mathbf{L})}{Gm_1m_2} = r_0$$

1.7 Kepler's Laws

We spent a good effort developing the nature of gravitational orbits, and it would be difficult to imagine doing this

without all of the modern advantages, particularly calculus and vectors. Somehow, Kepler was able to find enough pattern in sixteenth-century astronomical data to work out three correct laws of planetary motion. The data itself was recorded by astronomer Tycho Brahe over a span of at least thirty years.

Law of Ellipses (1609)

The orbit of each planet is an ellipse, with the sun at a focus.

This law we know very well by now, as did Newton. For the sun at the right focus (reverse the sign for the left focus), a planetary orbit looks like

$$r(t) = \frac{r_0}{1 + e \cos(\theta(t))},$$

where e is the eccentricity.

Law of Equal Areas (1609)

A line drawn between the sun and the planet sweeps out equal areas in equal times.

This is an amazing thing to notice from looking at charts of numbers. It turns out that this law is actually stating the conservation of angular momentum, although Kepler wouldn't have known so.

To derive the law in familiar language, recall the setup for the area integral in polar coordinates, particularly

$$A = \frac{1}{2} \int_{\theta_0}^{\theta_1} r^2 d\theta.$$

In differential form, this same notion reads

$$dA = \frac{1}{2} r^2 d\theta.$$

Or, by the chain rule, we can also write

$$\frac{dA}{dt} = \frac{1}{2} r^2 \frac{d\theta}{dt}.$$

Notice, though, that $r^2 d\theta/dt$ is also present in the angular momentum

$$L = m_* r^2 \frac{d\theta}{dt},$$

which can only mean

$$\frac{dA}{dt} = \frac{L}{2m_*},$$

thus dA/dt is constant. This is the literal mathematical statement of 'equal areas swept in equal times'.

Problem 5

For a body moving on a path $r = f(\theta)$ obeying Kepler's second law, show that the acceleration is:

$$\mathbf{a} = \frac{L^2}{m_* r^3} \left(\frac{f''(\theta)}{f(\theta)} - 2 \left(\frac{f'(\theta)}{f(\theta)} \right)^2 - 1 \right) \hat{\mathbf{r}}$$

Harmonic Law (1618)

The square of the period of a planet is directly proportional to the cube of the semi-major axis of the orbit.

Years after his first two discoveries, Kepler discerned yet another relationship for linking the time scale of the orbit to its length scale. While Kepler only knew of the proportionality between the period T and the semi-major axis a , we can do better by finding the associated constant.

Integrate the area equation for a full period of the orbit:

$$A = \frac{1}{2} \int_0^{2\pi} r^2 d\theta = \frac{L}{2m_*} \int_0^T dt = \frac{L}{2m_*} T$$

The area is simply πab , so we find

$$T = \pi ab \frac{2m_*}{L}.$$

Replace b using $b = a\sqrt{1 - e^2}$, and eliminate L using

$$G = \frac{L^2}{m_* m_1 m_2 r_0}.$$

To deal with the r_0 term, recall $a = r_0/(1 - e^2)$ and reason that

$$T = \frac{2\pi a^{3/2}}{\sqrt{G(m_1 + m_2)}}.$$

1.8 Energy Considerations

With some fine details of planetary motion finished, it's worth pointing out that the notion of 'energy' was not used at all. To develop some of this now, recall that in one dimension, the force relates to the potential energy by

$$F = -\frac{d}{dx} U(x).$$

Planetary motion, on the other hand, requires three dimensions to express the force, or two dimensions if we already know the plane of the motion. This is why the force is a vector:

$$\mathbf{F} = -\frac{Gm_1 m_2}{r^2} \hat{\mathbf{r}}$$

Notice, though, that the force is dependent on one spacial quantity, the length, which to say the force is effectively one-dimensional.

Gravitational Potential Energy

Since the gravitational force acts in strictly the radial direction, it stands to reason that the gravitational potential energy $U(r)$ relates to the force by:

$$\mathbf{F}(\mathbf{r}) = -\frac{d}{dr}(U(r)) \hat{\mathbf{r}}$$

This is just like the one-dimensional Newton's law $F = -dU/dx$, except the force is a vector, balanced by $\hat{\mathbf{r}}$ on the right.

To solve for $U(r)$, project $\hat{\mathbf{r}}$ into both sides of the above to get

$$\frac{Gm_1m_2}{r^2} = \frac{d}{dr}(U(r)) ,$$

solved by:

$$U(r) = -\frac{Gm_1m_2}{r}$$

This is the total gravitational potential energy stored between the two masses m_1, m_2 .

For a more formal definition, turn Newton's second law into a definite integral in the variable $d\mathbf{r}$ to get

$$\int_{r_0}^{r_1} \mathbf{F}(r) \cdot d\mathbf{r} = - \int_{r_0}^{r_1} \frac{d}{dr} U(r) \hat{\mathbf{r}} \cdot d\mathbf{r} ,$$

where the integral on the right is redundant to the derivative, leaving $U(r)$ evaluated at the endpoints:

$$\int_{r_0}^{r_1} \mathbf{F}(r) \cdot d\mathbf{r} = -(U(r_1) - U(r_0))$$

Set r_0 to infinity to recover the previous form.

Kinetic Energy

Containing two objects in total, the kinetic energy T of the two-body system is

$$T = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 .$$

What we need, however, is to express the kinetic energy in terms of the relative velocity

$$\mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2 .$$

Working out the algebra for this is left as an exercise, but the effort results in

$$T = \frac{1}{2}m_*v^2 + \frac{1}{2}(m_1 + m_2)V_0^2 ,$$

where V_0 is the (constant) center of velocity of the whole system. It's harmless to set this term to zero.

Conservation of Energy

The total energy of the two-body system is the sum of the kinetic and the potential contributions:

$$E = T + U = \frac{1}{2}m_*v^2 - \frac{Gm_1m_2}{r}$$

As it turns out, the energy of the system is constant.

To prove this, begin with Newton's second law

$$\mathbf{F} = -\frac{Gm_1m_2}{r^2} \hat{\mathbf{r}} ,$$

and project the velocity vector into each side:

$$\mathbf{v} \cdot \mathbf{F} = -\frac{Gm_1m_2}{r^2} (\mathbf{v} \cdot \hat{\mathbf{r}})$$

Replace $\mathbf{F}$ on the left and $\mathbf{v}$ on the right

$$m_* \left(\mathbf{v} \cdot \frac{d\mathbf{v}}{dt} \right) = -\frac{Gm_1m_2}{r^2} \left(\frac{dr}{dt} \hat{\mathbf{r}} + r \frac{d\theta}{dt} \hat{\boldsymbol{\theta}} \right) \cdot \hat{\mathbf{r}} ,$$

which simplifies to

$$\frac{1}{2}m_* \frac{d}{dt} (\mathbf{v} \cdot \mathbf{v}) = -\frac{Gm_1m_2}{r^2} \frac{dr}{dt} .$$

(Note we didn't really need the polar expression for the velocity. The r -component of the velocity is always dr/dt .)

The right side can be undone with the chain rule:

$$\frac{d}{dt} \left(\frac{1}{2}m_*v^2 \right) = \frac{d}{dt} \left(\frac{Gm_1m_2}{r} \right)$$

Finally, we have found

$$\frac{d}{dt} (T + U) = 0 ,$$

as expected.

Apocalypse Problem

Problem 6

If a planet were suddenly stopped in its orbit, supposed circular, show that it would fall into the sun in a time which is $\sqrt{2}/8$ times the period of the planet's revolution.

Answer: Begin with the total energy of the system

$$-\frac{Gm_1m_2}{a} = \frac{1}{2}m_* \left(\frac{dr}{dt} \right)^2 - \frac{Gm_1m_2}{r(t)} ,$$

where a is the radius of the orbit. Solve for dr/dt to get

$$\frac{dr}{dt} = \sqrt{\frac{2G(m_1 + m_2)}{a}} \sqrt{\frac{a}{r} - 1},$$

which can be separated into two equal integrals:

$$\int_a^0 \frac{dr}{\sqrt{a/r - 1}} = \sqrt{\frac{2G(m_1 + m_2)}{a}} \int_0^{t_*} dt,$$

where t_* is the answer we're after.

To solve the r -integral, choose the peculiar substitution

$$\begin{aligned} r &= a \cos^2(\theta) \\ dr &= -2a \cos(\theta) \sin(\theta) d\theta, \end{aligned}$$

and the above reduces to

$$2a \int_{\pi/2}^{\pi} \cos^2(\theta) d\theta = t_* \sqrt{\frac{2G(m_1 + m_2)}{a}}.$$

The remaining θ -integral resolves to $\pi/4$. Solving for t_* gives

$$t_* = \frac{\sqrt{2}}{8} \left(\frac{2\pi a^{3/2}}{\sqrt{G(m_1 + m_2)}} \right) = \frac{\sqrt{2}}{8} T,$$

as stated. This is about 0.1768 years, or just over two months, supposing there are twelve months per year on that planet.

Inverse Cube Attraction

Problem 7

A particle released from rest a distance D from the origin is attracted by the force

$$F = -\frac{mk^2}{x^3}.$$

Show that the time required to fall to the origin is D^2/k .

Hint:

$$v = \frac{dx}{dt} = k \sqrt{\frac{1}{x^2} - \frac{1}{D^2}}$$

1.9 Solid Sphere

We've taken on assumption (correctly) the shell theorem, which says a gravitational body with finite size can be treated as a point located at its center of mass.

With the shell theorem, we can calculate the gravitational force inside a uniform sphere of mass M and radius R at any distance $r < R$ from the center. A uniform sphere has the same density throughout, which we'll call λ :

$$\lambda = \frac{M}{4\pi R^3/3}$$

Force Inside Solid Sphere

At a distance $r < R$ from the center, according to the theorem, all of the sphere's mass that is located further from the center than r can be ignored. Only the sphere's mass obeying $r < R$ contributes to the force at distance r . This portion is called the *enclosed mass*. The enclosed mass is written $m(r)$, given by

$$m(r) = \lambda \frac{4}{3} \pi r^3.$$

If the test particle has mass m_0 , the magnitude of the force on the test particle is

$$F(r) = -\frac{Gm_0 m(r)}{r^2} = -\frac{Gm_0 M r}{R^3}.$$

Due to the r^3 factor that enters the numerator, the usual r^{-2} factor is replaced by r . The gravitational force inside a sphere grows linearly with distance until $r = R$.

As a vector, the force inside the solid sphere reads

$$\mathbf{F}(r) = -\frac{Gm_0 M}{R^3} \mathbf{r}.$$

Energy Inside Solid Sphere

The gravitational potential energy inside a solid sphere is not $U \propto -1/r$. To find the proper answer, first define

$$\lim_{r \rightarrow \infty} U(r) = 0$$

which assumes there is no energy when infinitely far from the solid sphere, assumed centered at the origin.

Starting from infinity, let a test particle of mass m_0 approach the solid sphere, eventually penetrating the its surface, stopping at r_1 . The energy spent during approach is broken into two integrals:

$$U(r_1) = -\int_{\infty}^R \mathbf{F}_{\text{out}} \cdot d\mathbf{r} - \int_R^{r_1} \mathbf{F}_{\text{in}} \cdot d\mathbf{r},$$

where $\mathbf{F}_{\text{out}}$, $\mathbf{F}_{\text{in}}$ are the forces felt by m_0 outside and inside the sphere, respectively.

Carrying out the integrals and simplifying, one finds

$$U(r_1) = \frac{Gm_0 M}{2} \left(\frac{r_1^2 - 3R^2}{R^3} \right).$$

Note that the special point $r_1 = R$ corresponds to being on the sphere's surface, and the potential energy takes a familiar form

$$U(R) = -\frac{Gm_0 M}{R}.$$

Self Energy

The *self energy* of a solid object is the energy required to assemble the object from parts initially at infinity. Since the gravitational force is attractive, we can expect the self energy due to gravity to be negative.

Demonstrating on a uniform solid sphere of radius R and mass M , write

$$dU = -\frac{Gm}{r} dm,$$

where dU is the energy added when a particle of mass dm is added to the existing sphere of mass $m = m(r)$. The particle settles at radius r , and each particle does so until $r = R$.

Since the sphere has uniform density λ , we further write

$$dU = -\frac{Gm\lambda v}{r} \lambda dv,$$

where $v = v(r)$ is the volume of the sphere. Turning the crank gives:

$$\begin{aligned} dU &= -G\lambda^2 \frac{4\pi r^3/3}{r} 4\pi r^2 dr \\ U &= -G(4\pi\lambda)^2 \frac{1}{3} \int_0^R r^4 dr \\ U &= -\frac{3}{5} \frac{GM^2}{R} \end{aligned}$$

Variable Density

If the gravitational force inside a solid sphere is independent of position, determine the density $\lambda(r)$ of the sphere.

While there exist more rigorous ways to solve this problem, assume the density takes the form

$$\lambda(r) = Ar^n,$$

where n is an integer and A is a unit-balancing constant. With this, all mass contained within a radius $r < R$ is given by

$$m(r) = A \frac{4\pi}{3} r^{3+n}.$$

Then, the magnitude of the force on the particle is

$$F = -\frac{4\pi Gm_0}{3} r^{1+n}.$$

If the force is to be independent of r , we must have $n = -1$, or $\lambda(r) = A/r$.

1.10 Gravity Near Earth

Students of classical physics find out early that the force due to gravity near Earth's surface is a vector pointing straight down

$$\mathbf{F}_g = -mg \hat{\mathbf{y}},$$

with corresponding potential energy

$$U(y) = mgy,$$

where y is the height above the surface (or a location near it), and g is the local gravitation constant:

$$g = \frac{9.8 \text{ m}}{\text{s}^2}$$

On the other hand, we just went through all the pains of showing that the gravitational force is

$$\mathbf{F}(\mathbf{r}) = -\frac{Gm_1 m_2}{r^2} \hat{\mathbf{r}}$$

with potential energy

$$U(r) = -\frac{Gm_1 m_2}{r}.$$

Clearly, these two pictures must be reconciled. To do so, let r be replaced by the quantity $R + y$, where R is a constant distance we'll take to be the radius of the Earth, and y is the effective height, approximately from sea level. What we assume throughout is that $y \ll R$.

Without loss of generality, we can assume all displacements are one dimensional and thus $\hat{\mathbf{r}} = \hat{\mathbf{y}}$. This identifies m_1 for the mass of the Earth, and m_2 for the mass of a test projectile.

With these restrictions, the force and energy become:

$$\begin{aligned} \mathbf{F}(y) &= -\frac{Gm_1 m_2}{(R+y)^2} \hat{\mathbf{y}} \\ U(y) &= -\frac{Gm_1 m_2}{(R+y)} \end{aligned}$$

Next apply binomial expansion to each denominator, particularly:

$$\begin{aligned} (R+y)^{-2} &\approx \frac{1}{R^2} - \frac{2y}{R^3} + \frac{3y^2}{R^4} - \dots \\ (R+y)^{-1} &\approx \frac{1}{R} - \frac{y}{R^2} + \frac{y^2}{R^3} - \dots \end{aligned}$$

To first order, the above equations become

$$\begin{aligned} \mathbf{F}(y) &\approx -\frac{Gm_1 m_2}{R^2} \left(1 - \frac{2y}{R}\right) \hat{\mathbf{y}} \\ U(y) &\approx -\frac{Gm_1 m_2}{R} \left(1 - \frac{y}{R}\right). \end{aligned}$$

We want the force equation to be constant, thus we see the quantity $2y/R$ must be negligible so the effective force at the surface is

$$\mathbf{F}_g = -\frac{Gm_1m_2}{R^2} \hat{\mathbf{y}}.$$

The acceleration term is identically g :

$$g = \frac{Gm_{\text{Earth}}}{R_{\text{Earth}}^2}.$$

For the potential energy, we have

$$U(y) = U_0 + mgy,$$

where U_0 is the potential energy at $y = 0$, often defined to be zero, and the unscripted mass m is that of a test particle (not the Earth).

Note that the first-order potential term is maintained despite y/R being a very small number. The reason for this not only to recover the form mgy , but also that the first derivative must equal a constant, which is what we asked of the force.

Problem 8

A person's weight is F_0 at sea level. To first order in y , what is the person's weight 4 km above sea level? Answer: $F_{4\text{ km}} = F_0 (1 - 2(4\text{ km})/R)$

Escape Velocity

In a two-body system with gravity being the only force present, suppose we imparted an initial carefully-chosen *escape velocity* v_e along the line between the bodies such that the kinetic energy goes to zero as the separation becomes infinite.

As a two-body problem, we can apply conservation of energy to write

$$\frac{1}{2}m_*v_0^2 - \frac{Gm_1m_2}{d} = \frac{1}{2}m_*v^2 - \frac{Gm_1m_2}{r} = 0,$$

where d is the initial separation between the bodies. The total energy is zero by definition.

From the energy statement, we can easily solve for the escape velocity from a starting separation d :

$$v_e = \sqrt{\frac{2Gm_1m_2}{m_*d}} = \sqrt{\frac{2G(m_1 + m_2)}{d}}$$

For the case of a small body escaping Earth, the above becomes

$$v_e \approx \sqrt{2gR_{\text{Earth}}}.$$

Hole Through Earth

An object of mass m is dropped through a straight tube connecting two points on Earth's surface. Neglecting rotational effects and friction, what happens to the object?

Unless the hole is drilled through the center, which we will not assume, the gravitational force on the object resolves into two components - one component parallel to the tube, the other component perpendicular. The equilibrium position of the object is in the center of the tube where the parallel component of the force is zero.

Let x be the parallel displacement from equilibrium, and let $\mathbf{r}$ be the position vector from the center. The force on the object parallel to the tube is:

$$F_x = \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{x}} = -\frac{GMm}{R^3} (\mathbf{r} \cdot \mathbf{x}),$$

or

$$m \frac{d^2x}{dt^2} = -\frac{GMm}{R^3}x = -\frac{gm}{R}x.$$

The above is the differential equation

$$\frac{d^2x}{dt^2} = -\omega^2x$$

for a simple harmonic oscillator solved by

$$x(t) = A \cos(\omega t + \phi),$$

where A is the length of the tube and ϕ chooses the initial position of the object. The angular frequency ω is given by

$$\omega = \sqrt{\frac{g}{R}}.$$

1.11 Energy and Orbit

Parabolic Orbit

Suppose now that a two-body system has zero total energy

$$E = 0.$$

but the motion is not strictly along the line connecting the two bodies. In this special case, the system is *always* at escape velocity. This does not mean the escape velocity is constant. The distance d is playing the role of r in the v_e equation.

To develop this, recall that the velocity for a parabolic orbit can be written

$$\mathbf{v} = v_0 (\hat{\boldsymbol{\theta}} + \hat{\mathbf{y}}),$$

which means

$$v^2 = \mathbf{v} \cdot \mathbf{v} = 2v_0^2 \frac{r_0}{r}.$$

Using the escape velocity in place of v allows us to write

$$\frac{2G(m_1 + m_2)}{r} = 2v_0^2 \frac{r_0}{r},$$

or

$$v_0^2 = \frac{G(m_1 + m_2)}{r_0}.$$

Elliptical Orbit

Elliptical orbits are called *bound* orbits, and have negative total energy:

$$E < 0$$

Interestingly, if we take a parabolic orbit with $E = 0$ and subtract a little energy from the total (by some external means), then the parabola becomes an ellipse by having the second focus come in from infinity.

We ought to be able to prove the total energy is negative for an elliptical orbit. Start with the total energy

$$E = \frac{1}{2}m_*v^2 - \frac{Gm_1m_2}{r},$$

and substitute v^2 using

$$\mathbf{v} = \frac{v_0}{e} (\hat{\boldsymbol{\theta}} + e \hat{\mathbf{y}}),$$

which excludes the case of circles. Proceeding carefully, find

$$\begin{aligned} v^2 &= \frac{v_0^2}{e^2} \left(\frac{2r_0}{r} - 1 + e^2 \right) \\ &= \frac{2Gm_1m_2}{m_*r} - \frac{Gm_1m_2}{m_*r_0} (1 - e^2), \end{aligned}$$

so the kinetic term is

$$E_{\text{kin}} = \frac{Gm_1m_2}{r} - \frac{Gm_1m_2}{2r_0} (1 - e^2).$$

The total energy sums the potential plus the kinetic, which happens to contain equal and opposite $1/r$ -like terms, leaving just the constant:

$$E = \frac{-Gm_1m_2}{2r_0} (1 - e^2) = \frac{-Gm_1m_2}{2a},$$

in terms of v_0 ,

$$E = -\frac{1}{2}m_*v_0^2 \left(\frac{1 - e^2}{e^2} \right).$$

Hyperbolic Orbit

Hyperbolic orbits are called *unbound* orbits, and have positive total energy:

$$E > 0$$

The analysis of this situation follows exactly like the elliptical case. For the total energy, you can see $e > 1$ simply flips the sign to make

$$E = \frac{Gm_1m_2}{2a} = \frac{1}{2}m_*v_0^2 \left(\frac{e^2 - 1}{e^2} \right).$$

Circular Orbit

For circular orbits, we need to go back to the velocity equation

$$\mathbf{v} = \frac{Gm_1m_2}{L} \hat{\boldsymbol{\theta}},$$

which has no v_0 -term.

The angular momentum is

$$L = m_*R^2 \frac{d\theta}{dt} = m_*a^2 \frac{2\pi}{T},$$

where T is the period of the orbit and R is the radius. Simplifying gives

$$L = m_*\sqrt{G(m_1 + m_2)R},$$

and then the square of the velocity is:

$$v^2 = \frac{G(m_1 + m_2)}{R}$$

The time derivative of $\mathbf{v}$ gives a familiar equation for the acceleration

$$\mathbf{a} = -\frac{Gm_1m_2}{L} \frac{d\theta}{dt} \hat{\mathbf{r}},$$

which for circular orbits simplifies to

$$\mathbf{a} = \frac{-v^2}{R} \hat{\mathbf{r}},$$

as expected for circular motion in general.

The energy of a circular orbit is

$$E = \frac{1}{2} \frac{Gm_1m_2}{R} - \frac{Gm_1m_2}{R} = \frac{-Gm_1m_2}{2R},$$

thus the kinetic energy is half the potential energy, and the total is negative.

Eccentricity and Orbit

Begin with the Runge-Lorenz vector and replace $\mathbf{L}$ using its definition:

$$\mathbf{Z} = \mathbf{v} \times (m_* \mathbf{r} \times \mathbf{v}) - Gm_1 m_2 \hat{\mathbf{r}},$$

and square the whole equation:

$$\begin{aligned} \mathbf{Z} \cdot \mathbf{Z} &= |\mathbf{v} \times (m_* \mathbf{r} \times \hat{\mathbf{v}})|^2 \\ &\quad - 2Gm_1 m_2 \mathbf{v} \times (m_* \mathbf{r} \times \hat{\mathbf{v}}) \cdot \hat{\mathbf{r}} + G^2 m_1^2 m_2^2 \end{aligned}$$

For the first term on the right, notice $\mathbf{v}$ is perpendicular to $\mathbf{r} \times \mathbf{v}$, so

$$|\mathbf{v} \times (m_* \mathbf{r} \times \hat{\mathbf{v}})| = m_* r v^2 |\sin(\phi)|,$$

where ϕ is the angle between $\mathbf{r}$ and $\mathbf{v}$.

For the second term, the scalar triple product can be rewritten

$$\mathbf{v} \times (m_* \mathbf{r} \times \hat{\mathbf{v}}) \cdot \hat{\mathbf{r}} = m_* (\mathbf{r} \times \mathbf{v}) \cdot (\hat{\mathbf{r}} \times \mathbf{v}).$$

The remaining vectors are parallel and the whole quantity simplifies to

$$\mathbf{v} \times (m_* \mathbf{r} \times \hat{\mathbf{v}}) \cdot \hat{\mathbf{r}} = m_* r v^2 \sin^2(\phi).$$

Rewriting $\mathbf{Z} \cdot \mathbf{Z}$ with this in mind, we have

$$\begin{aligned} Z^2 &= (m_* r v^2)^2 \sin^2(\phi) \\ &\quad - 2Gm_1 m_2 m_* r v^2 \sin^2(\phi) + G^2 m_1^2 m_2^2, \end{aligned}$$

or

$$\frac{Z^2}{G^2 m_1^2 m_2^2} = 1 + \sin^2(\phi) (q^2 - 2q),$$

where

$$q = \frac{m_* r v^2}{G m_1 m_2}.$$

simplifying this further gives

$$\frac{Z^2}{G^2 m_1^2 m_2^2} = \cos^2(\phi) + \sin^2(\phi) (1 - q)^2$$

Finally, note that the left side is actually the square of the eccentricity, giving, after restoring q :

$$e^2 = \cos^2(\phi) + \sin^2(\phi) \left(1 - \frac{r v^2}{G(m_1 + m_2)}\right)^2$$

This is an enlightening result. For $\phi = 0$ the motion is purely radial and uninteresting. For all other cases, we see

the combination of variables being suspiciously like to the escape velocity. Swapping this in gives

$$e^2 = \cos^2(\phi) + \sin^2(\phi) \left(1 - \frac{2v^2}{v_e^2}\right)^2$$

We see if $v = v_e$, then the eccentricity is precisely one, which is consistent with what we know of parabolic orbits. Similarly we see the cases $v < v_e$ and $v > v_e$ give $e < 1$ and $e > 1$ respectively, which is the signature of elliptic and hyperbolic orbits. A circular orbit has $\phi = \pi/2$.

1.12 Shell Theorem

Newton's law of gravitation tells us that every particle in the universe is trying to pull every other particle toward itself with a force proportional to the masses involved and inversely proportional to the square of the separation, and this is duly used to calculate the force onto planets, moons, satellites, and so on.

Using triple integration and spherical coordinates, something Newton didn't have, we finally address an assumption made early in gravitational analysis, namely *why* we're allowed to represent voluminous objects as single points located at the center of mass. This is called the shell theorem, and entails two important proofs.

Outside a Sphere

Consider a solid sphere of radius R , total mass M , and uniform density λ . Also let there be a test particle of mass m somewhere in space. Without loss of generality, place the test particle on the z -axis at the point $\mathbf{D} = D \hat{\mathbf{z}}$. The length D is the distance from the test particle to the center of the sphere.

In order to 'properly' calculate the gravitational attraction between the test mass and the sphere, a volume integral over the entire sphere must be calculated. Choose any element of volume dV inside the sphere at location $\mathbf{r}$, which is located distance r from the center, at an angle θ from the z -axis.

Let vector $\mathbf{q}$ denote the line connecting $\mathbf{D}$ to $\mathbf{r}$ such that

$$\mathbf{r} + \mathbf{q} = D \hat{\mathbf{z}},$$

and also let α be the angle between $\hat{\mathbf{z}}$ and $\hat{\mathbf{q}}$. From the law of cosines, we can say:

$$q^2 = r^2 + D^2 - 2rD \cos(\theta)$$

$$r^2 = q^2 + D^2 - 2qD \cos(\alpha)$$

The total force on the test particle is the vector $\mathbf{F}$. However, due to the ϕ -symmetry of this picture, only the z -component of the force will have a net effect on the particle. All xy -components cancel equally and oppositely:

$$\mathbf{F} = \int_{\mathcal{D}} d\mathbf{F} \cdot \hat{\mathbf{z}} = \int \int \int_{\text{volume}} dF \cos(\alpha)$$

The differential force is

$$dF = \frac{-Gm}{q^2} dm,$$

where dm is the mass of the differential volume element influencing the test particle. The mass term can be replaced using the density

$$\frac{dm}{dV} = \frac{M}{4\pi R^3/3} = \lambda,$$

where it is appropriate to replace dV with the volume element in spherical coordinates.

The force integral now is

$$F = -Gm\lambda \int_0^{2\pi} \int_0^\pi \int_0^R \frac{\cos(\alpha)}{q^2} r^2 \sin(\theta) dr d\theta d\phi,$$

which, after substituting and simplifying a bit, becomes:

$$F = -Gm\lambda \frac{2\pi}{2D} \int_0^\pi \int_0^R \left(\frac{1}{q} + \frac{D^2 - r^2}{q^3} \right) r^2 \sin(\theta) dr d\theta$$

Perform implicit differentiation on the q^2 equation to find, remembering r and θ are independent,

$$q dq = r D \sin(\theta) d\theta,$$

and rewrite the integral with the intent of integrating over r last. Make sure you know why the limits are now changed:

$$F = -Gm\lambda \frac{\pi}{D^2} \int_0^R \int_{(D-r)}^{(D+r)} \left(1 + \frac{D^2 - r^2}{q^2} \right) r dq dr$$

The whole q -integral treats r as a constant and resolves to $4r$, so

$$F = -Gm\lambda \frac{\pi}{D^2} \int_0^R 4r^2 dr,$$

and the r -integral is elementary. Simplifying everything gives

$$F = -Gm \left(\frac{3M}{4\pi R^3} \right) \frac{\pi}{D^2} \frac{4}{3} R^3 = \frac{-GMm}{D^2}.$$

Conveniently, the force acts as if *all* of its mass were concentrated at the center. This result is also true in general, where the notion of ‘center’ means center of mass, not necessarily the center of the volume.

Inside a Shell

Another interesting question that arises in the course of studying gravity is, what does it feel like inside a hollow uniform shell? To pursue this question, suppose we have a thin spherical shell of radius R and thickness $2a$ that is much less than R , and the test particle is inside anywhere within the shell.

This setup borrows all of the geometry from the previous setup, except this time we have $D < R$, which is the important part. Setting up the same integral and doing the same simplifications, we can jump to

$$F = -Gm\lambda \frac{\pi}{D^2} \int_{R-a}^{R+a} \int_{(r-D)}^{(D+r)} \left(1 + \frac{D^2 - r^2}{q^2} \right) r dq dr.$$

Most notably, the lower integration in the q -integral is swapped to accommodate $D < R$. This causes the q -integral to resolve to zero, and we find

$$F = 0$$

inside the shell.

2 Central Potential

The whole apparatus for studying planetary motion can be grown from the gravitational potential energy

$$U(r) = -\frac{Gm_1m_2}{r}$$

of a two-body system. The plan now is to develop the theory while assuming as little as possible about $U(r)$.

Two-Body Analysis

As a two-body system, we still deal with the center of mass

$$\mathbf{R}(t) = \frac{m_1\mathbf{r}_1(t) + m_2\mathbf{r}_2(t)}{m_1 + m_2},$$

whose time derivative is the center of velocity $\mathbf{V}(t)$. Two time derivatives of $\mathbf{R}(t)$ yields the center of acceleration, which is always zero by Newton's third law:

$$\frac{d^2}{dt^2}\mathbf{R}(t) = 0$$

In terms of $\mathbf{R}$, the absolute position of each body is

$$\begin{aligned}\mathbf{r}_1 &= \mathbf{R} + \frac{m_2}{m_1 + m_2} \mathbf{r} \\ \mathbf{r}_2 &= \mathbf{R} - \frac{m_1}{m_1 + m_2} \mathbf{r},\end{aligned}$$

and we define the relative displacement

$$\mathbf{r}(t) = \mathbf{r}_1(t) - \mathbf{r}_2(t)$$

to deal with one variable instead of two. The time derivative of the relative displacement is the relative velocity $\mathbf{v}(t)$.

In terms of $\mathbf{r}(t)$, Newton's second law takes special form

$$\mathbf{F}_{12} = m_* \frac{d^2\mathbf{r}}{dt^2},$$

where

$$m_* = \frac{m_1m_2}{m_1 + m_2}$$

is the reduced mass.

Energy Conservation

The total energy in the two-body system is the sum of a kinetic term and a potential term. For the kinetic energy we have

$$T = \frac{1}{2}m_*v^2 + \frac{1}{2}(m_1 + m_2)V_0^2,$$

where we take $V_0 = 0$ without loss of generality. For the potential energy we're stuck with just $U(r)$.

For the total energy, we write

$$E = T + U = \frac{1}{2}m_*v^2 + U(r).$$

Take a single time derivative to find E to be constant

$$\begin{aligned}\frac{dE}{dt} &= \frac{1}{2}m_*\frac{d}{dt}(v^2) + \frac{d}{dt}U(r) \\ &= \frac{1}{2}m_*2\mathbf{v} \cdot \frac{d\mathbf{v}}{dt} + \mathbf{v} \cdot \frac{d}{d\mathbf{r}}(U(r)) \\ &= \mathbf{v} \cdot \left(m_*\frac{d\mathbf{v}}{dt} + \frac{d}{d\mathbf{r}}(U(r)) \right),\end{aligned}$$

because the parenthesized term is identically Newton's second law and resolves to zero.

Angular Momentum

The angular momentum

$$\mathbf{L} = m_*\mathbf{r} \times \mathbf{v}$$

doesn't depend on $U(r)$ at all, thus we recycle the constant of motion from planetary motion analysis:

$$L = m_*r^2\omega,$$

where $\omega = d\theta/dt$. Since L is constant, we know the motion to be planar.

Polar Coordinates

In the polar coordinate system, the relative position, velocity, and acceleration vectors read

$$\begin{aligned}\mathbf{r} &= r \hat{\mathbf{r}} \\ \mathbf{v} &= \frac{dr}{dt} \hat{\mathbf{r}} + r\omega \hat{\boldsymbol{\theta}} \\ \mathbf{a} &= \left(\frac{d^2r}{dt^2} - r\omega^2 \right) \hat{\mathbf{r}} + \left(r\frac{d\omega}{dt} + 2\frac{dr}{dt}\omega \right) \hat{\boldsymbol{\theta}}.\end{aligned}$$

The $\hat{\boldsymbol{\theta}}$ -term in the acceleration is proportional to dL/dt and vanishes entirely, leaving a purely radial acceleration vector

$$\mathbf{a} = \left(\frac{d^2r}{dt^2} - \frac{L^2}{m_*^2r^3} \right) \hat{\mathbf{r}}.$$

2.1 Effective Potential

In terms of the angular momentum L , the velocity vector can be written

$$\mathbf{v} = \frac{dr}{dt} \hat{\mathbf{r}} + \frac{L}{m_* r} \hat{\boldsymbol{\theta}},$$

meaning

$$v^2 = \left(\frac{dr}{dt}\right)^2 + \frac{L^2}{m_*^2 r^2}.$$

Feed this v^2 into the energy conservation statement:

$$E = \frac{1}{2} m_* \left(\frac{dr}{dt}\right)^2 + \frac{L^2}{2m_* r^2} + U(r)$$

The latter two terms constitute the *effective potential energy*

$$U_{\text{eff}}(r) = U(r) + \frac{L^2}{2m_* r^2},$$

also known as the *centrifugal potential energy*, the gradient of which is the centrifugal force.

In terms of the effective potential, the total energy goes back down to two terms, one with time dependence and one with spatial dependence:

$$E = \frac{1}{2} m_* \left(\frac{dr}{dt}\right)^2 + U_{\text{eff}}(r)$$

Problem 9

For the elliptical orbit of a planet, show for a given radius r_0 that:

$$|U_{\text{eff}}(r_0)| = \frac{-Gm_1 m_2}{2r_0}$$

2.2 One-Dimensional Systems

In terms of the effective potential energy, the total energy is reduced to a one-dimensional system in the variable r . Taking this literally, let us study the generic one-dimensional system

$$E = \frac{1}{2} m \left(\frac{dx}{dt}\right)^2 + U(x)$$

to remove complications from planar motion.

Solving the above for the velocity, one writes

$$\frac{dx}{dt} = \pm \sqrt{\frac{2}{m} (E - U(x))}$$

to establish the equation of motion

$$t = \pm \sqrt{\frac{m}{2}} \int_{x_i}^{x_f} \frac{dx}{\sqrt{E - U(x)}}.$$

Time-Reversal Symmetry

The $\pm$ symbol in the equation of motion indicates *time-reversal symmetry* of the problem. Typically in one-dimensional systems, the *solution* to the equation of motion exhibits such symmetry, a stronger constraint than what we have. Supposing $x(t)$ is a solution to the equation of motion, time-reversal symmetry implies that $x_1(t) = x(t_0 - t)$ is also a solution that differs from original $x(t)$ by an integration constant.

For most configurations, there exists at least one turning point t_* at which the velocity goes to zero. We exploit time-reversal symmetry to write an exact time-reversed-and-shifted equation

$$x_1(t) = x(2t_* - t).$$

Next, we note from function- and derivative matching that

$$x_1(t_*) = x(t_*)$$

$$\frac{d}{dt} x_1(x(t_*)) = -\frac{d}{dt} x(x(t_*)) = 0,$$

and so on for higher derivatives. We may then drop the 1-subscript to get

$$x(t) = x(2t_* - t).$$

Shifting the above by t_* , the symmetric equation

$$x(t_* + t) = x(t_* - t)$$

emerges. In one dimension, the essence of time-reversal symmetry means that equations of motion are symmetric about turning points t_* .

Trapped Particle

Potential energy functions $U(x)$ that exhibit at least one local minimum can ‘trap’ a particle into an oscillatory pattern. Supposing x_i and x_f correspond to turning points in the motion, the oscillatory period is given by

$$T = \sqrt{2m} \int_{x_i}^{x_f} \frac{dx}{\sqrt{E - U(x)}}.$$

The quantity $E - U(x)$ is always positive except at the turning points, at which the speed of the particle is instantaneously zero.

Harmonic Oscillations

In the vicinity of a local energy minimum at x_* , the first- and second-derivatives of $U(x_*)$ are

$$\begin{aligned}\frac{d}{dx}U(x_*) &= 0 \\ \frac{d^2}{dx^2}U(x_*) &= \lambda > 0,\end{aligned}$$

which allows $U(x)$ to be approximated by Taylor series:

$$U(x) \approx U(x_*) + \frac{1}{2}\lambda(x - x_*)^2$$

Applying Newton's second law, the corresponding equation of motion is

$$\frac{d^2}{dx^2}x(t) = -\frac{\lambda}{m}(x - x_*),$$

whose solution is known as the *harmonic oscillator*

$$x(t) = x_* + A \sin\left(\sqrt{\frac{\lambda}{m}}t - \phi_0\right).$$

The amplitude of oscillation is A , and the initial phase is contained in ϕ_0 .

Unstable Equilibrium

An equilibrium point x_* exists at any local maximum of $U(x)$, however motion around such a point is unstable (non-oscillatory). To show this, reverse the sign on λ to arrive at the differential equation

$$\frac{d^2}{dx^2}x(t) = \frac{\lambda}{m}(x - x_*),$$

generally solved by

$$x(t) = x_* + Ae^{\lambda t} + Be^{-\lambda t}.$$

That is, the particle is pulled away from x_* and rides $U(x)$ downhill.

2.3 Planar Orbits

Returning to the case of planar orbits, we can extrapolate all one-dimensional results to two dimensions by replacing $x \rightarrow r$ and accounting for $\theta(t)$ as a dynamic variable.

Equations of Motion

In terms of the total energy E , the time evolution of the two-body system is given by

$$t(r) = \pm \sqrt{\frac{m_*}{2}} \int_{r_i}^{r_f} \frac{dr}{\sqrt{E - U_{\text{eff}}(r)}}.$$

An equation for $\theta(t)$ is attained by integrating the angular momentum $L = m_* r^2 \omega$ to write

$$\theta(t) = \theta_0 + \frac{L}{m_*} \int_0^t \frac{dt'}{(r(t'))^2}.$$

Switching to the r domain, the above is also written

$$\theta(r) = \pm \frac{L}{\sqrt{2m_*}} \int_{r_i}^{r_f} \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}.$$

Apogee and Perigee

Supposing there exists a time t_* at which the radius r reaches a turning point (i.e. $dr/dt = 0$), the corresponding point (r_*, θ_*) in the plane is called the *apogee* if r is at a maximum, and the *perigee* if r is at a minimum.

Solutions to the θ -equation occur in four explicit branches:

Apogee, $\theta > \theta_*$

$$\theta = \theta_0 + \frac{L}{\sqrt{2m_*}} \int_{r_*}^r \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}$$

Apogee, $\theta < \theta_*$

$$\theta = \theta_0 - \frac{L}{\sqrt{2m_*}} \int_{r_*}^r \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}$$

Perigee, $\theta > \theta_*$

$$\theta = \theta_0 + \frac{L}{\sqrt{2m_*}} \int_r^{r_*} \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}$$

Perigee, $\theta < \theta_*$

$$\theta = \theta_0 - \frac{L}{\sqrt{2m_*}} \int_r^{r_*} \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}$$

Evident in the above is the time-refersal symmetry about θ_* , namely

$$r(\theta_* - \theta) = r(\theta_* + \theta).$$

Bounded Orbits

When the potential energy $U(x)$ contains a local minimum, a particle with a sufficiently low energy may become ‘trapped’ in the so-called *potential well*. Looking at the evolution of θ between two extreme points (perigee to apogee or vice-versa), we have

$$\theta = \frac{L}{\sqrt{2m_*}} \int_{r_p}^{r_a} \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}} ,$$

which is some number that is not generally a rational fraction of π . That is, we find that orbits are bounded aren’t necessarily repeated shapes, but may have a any (or an infinite) number of apogees and perigees. We will soon find there are two exceptions to this, where if the energy U has certain dependence on r , closed orbits are possible.

Circular Orbit

The circular orbit is characterized by $dr/dt = 0$, or equivalently $r(t) = r_0$. The equation of motion for $\theta(t)$ becomes trivial

$$\theta(t) = \theta_0 + \frac{Lt}{m_* r_0^2} ,$$

and the time T_0 required for $\theta(t) - \theta_0 = 2\pi$ corresponds to the period of the circular orbit:

$$T_0 = \frac{2\pi m_* r_0^2}{L}$$

As for the energy of a circular orbit, we have

$$E = \frac{1}{2} m_* \left(\frac{dr}{dt} \right)^2 + \frac{L^2}{2m_* r^2} + U(r) ,$$

which remains constant. Taking an r -derivative at r_0 yields

$$0 = -\frac{L^2}{m_* r_0^3} + \frac{d}{dr} (U(r)) \Big|_{r_0} ,$$

or, in shorthand:

$$U'(r_0) = \frac{L^2}{m_* r_0^3}$$

Eliminate L from the above to write

$$U'(r_0) = m_* r_0 \omega^2 = \frac{m_* v_0^2}{r_0} ,$$

which is precisely the force ‘felt’ by an object constrained to uniform circular motion. (The force vector itself points to the center.)

We can eliminate L once more to express the period T of the circular orbit in terms of $U'(r_0)$:

$$T_0 = 2\pi \sqrt{\frac{m_* r_0}{U'(r_0)}}$$

Problem 10

Check that the circular orbit is described by

$$\frac{d}{dr} (U_{\text{eff}}(r)) \Big|_{r_0} = 0 .$$

Problem 11

Show that Kepler’s law of equal areas hold for any central force, including straight-line motion.

3 Power Law Potential

Finally we must choose a particular form for the potential energy $U(r)$, thus we'll pose a *central power law potential*

$$U(r) = -\frac{\Lambda}{r^\alpha},$$

where Λ is an arbitrary constant (positive or negative), and $\alpha = 2$ reproduces the case for planetary motion. In the general case, we use

$$U_{\text{eff}}(r) = -\frac{\Lambda}{r^\alpha} + \frac{L^2}{2m_* r^2}$$

for the effective potential energy.

3.1 Circular Orbit

For orbits that have a nearly-circular radius r_0 , we may approximate the effective potential energy via Taylor expansion in the vicinity $r \approx r_0$:

$$U_{\text{eff}}(r) = U_{\text{eff}}(r_0) + \left. \frac{d}{dr} (U_{\text{eff}}(r)) \right|_{r_0} (r - r_0) + \left. \frac{d^2}{dr^2} (U_{\text{eff}}(r)) \right|_{r_0} \frac{(r - r_0)^2}{2!} + \dots$$

Stability

The first-order derivative term is identically zero by nature of the circular orbit. The second-order term must be done by brute force:

$$\begin{aligned} \left. \frac{d^2}{dr^2} U_{\text{eff}}(r) \right|_{r_0} &= \left. \frac{d}{dr} \left(\frac{\Lambda\alpha}{r^{\alpha+1}} - \frac{L^2}{m_* r^3} \right) \right|_{r_0} \\ &= -\frac{\Lambda\alpha(\alpha+1)}{r_0^{\alpha+2}} + \frac{3L^2}{m_* r_0^4} \end{aligned}$$

Now, from the first-order equation we learn

$$\frac{\Lambda\alpha}{r_0^{\alpha+1}} = \frac{L^2}{m_* r_0^3},$$

where eliminating L^2 in the second-order term now gives

$$\begin{aligned} \left. \frac{d^2}{dr^2} U_{\text{eff}}(r) \right|_{r_0} &= -\frac{\Lambda\alpha(\alpha+1)}{r_0^{\alpha+2}} + \frac{3\Lambda\alpha}{r_0^{\alpha+2}} \\ &= \frac{\Lambda\alpha(2-\alpha)}{r_0^{\alpha+2}} \\ &= \frac{L^2}{m_* r_0^4} (2-\alpha). \end{aligned}$$

We can also write the angular frequency ω_0 in terms of the angular momentum via

$$\omega_0 = \frac{L}{m_* r_0^2},$$

and thus

$$\left. \frac{d^2}{dr^2} U_{\text{eff}}(r) \right|_{r_0} = m_* \omega_0^2 (2-\alpha)$$

Rewriting the Taylor expansion, we now have

$$U_{\text{eff}}(r) \approx U_{\text{eff}}(r_0) + m_* \omega_0^2 (2-\alpha) \frac{(r - r_0)^2}{2!}.$$

For all $\alpha > 2$, near-circular orbits are unstable, meaning particles with high enough energy will slip away to $r \rightarrow \infty$, whereas particles with sufficiently low energy or sufficiently low radius will inevitably collapse to $r = 0$. For $\alpha < 2$, the system corresponds to a one-dimensional harmonic oscillator in r , thus near-circular orbits are stable. The angular frequency in the r -variable is given by

$$\omega_r = \omega_0 \sqrt{2-\alpha},$$

implying that periodic closed orbits occur when $\sqrt{2-\alpha}$ is a rational number. Conveniently we'll see that the Coulomb and gravitational potentials ($\alpha = 1$), along with the harmonic oscillator ($\alpha = -2$) each produce closed orbits not limited to circles. The next closed orbit corresponds to $\alpha = -7$.

Period

In terms of ω_0 , we can write the period T_0 of a circular orbit. Start with the definition $\omega T = 2\pi$, we have

$$T_0 = \frac{2\pi}{\omega_0} = \frac{2\pi m_* r_0^2}{L}$$

Harmonic Potential

The *harmonic potential* is defined by

$$U(r) = \Lambda r^2.$$

As a central force, all motion is confined to a plane and thus we separate into components as

$$U(r) = \Lambda (x^2 + y^2).$$

Using $F_x = -\partial U/\partial x$ and similar for the y-component, the above implies a pair of independent one-dimensional differential equations

$$\begin{aligned}\frac{d^2}{dt^2}x(t) &= -\omega^2 x(t) \\ \frac{d^2}{dt^2}y(t) &= -\omega^2 y(t),\end{aligned}$$

where $\omega = \sqrt{2\Lambda/m_*}$.

General solutions to the above are trigonometric, namely

$$\begin{aligned}x(t) &= A_x \cos(\omega t - \phi_x) \\ y(t) &= A_y \cos(\omega t - \phi_y),\end{aligned}$$

where $A_{x,y}$ and $\phi_{x,y}$ are determined from initial conditions.

We can do away with the ϕ_x -term by placing the particle at A_x at $t = 0$ and defining the x -axis to pass through that point. Then, the y -component of the position must be zero, telling us $\phi_y = \pi/2$. Finally, we find a closed equation for elliptical orbits with the origin at the center:

$$\begin{aligned}x(t) &= A_x \cos(\omega t) \\ y(t) &= A_y \sin(\omega t)\end{aligned}$$

The envelops of positions draws an ellipse:

$$\frac{x^2}{A_x^2} + \frac{y^2}{A_y^2} = 1$$

3.2 Dimensionless Variables

For a power law potential $U(r) = -\Lambda/r^\alpha$, we can find a circular orbit characterized by $r = r_0$ that minimizes $U_{\text{eff}}(r)$ at r_0 . The period of such an orbit is T_0 .

Let us now replace quantities of radius, energy, and time units with dimensionless variables by the following substitutions:

$$\begin{aligned}\rho(t) &= \frac{r(t)}{r_0} \\ \mathcal{E} &= \frac{E}{|U_{\text{eff}}(r_0)|} \\ \tau &= \frac{t}{T_0} = \frac{Lt}{2\pi m_* r_0^2}\end{aligned}$$

We bring Λ into the mix by recalling

$$\frac{L^2}{m_* r_0^2} = \frac{\Lambda \alpha}{r_0^\alpha}$$

for circular orbits, and it further follows that

$$|U_{\text{eff}}(r_0)| = \frac{L^2}{2m_* r_0^2} \left| \frac{2-\alpha}{\alpha} \right|.$$

Problem 12

For a planet on an elliptical orbit with semi-major axis a , use

$$E = \frac{-Gm_1 m_2}{2a}$$

to show that the eccentricity is given by

$$e = \sqrt{1 + \mathcal{E}}.$$

Problem 13

For a planet on an elliptical orbit with semi-major axis a and semi-minor axis b , show that

$$\frac{b}{a} = \sqrt{-\mathcal{E}}.$$

If the orbit is hyperbolic, show instead that

$$\frac{b}{a} = \sqrt{\mathcal{E}}.$$

Equations of Motion

The equations of motion

$$\begin{aligned}t(r) &= \pm \sqrt{\frac{m_*}{2}} \int_{r_i}^{r_f} \frac{dr}{\sqrt{E - U_{\text{eff}}(r)}} \\ \theta(r) &= \pm \frac{L}{\sqrt{2m_*}} \int_{r_i}^{r_f} \frac{dr/r^2}{\sqrt{E - U_{\text{eff}}(r)}}\end{aligned}$$

must be recast in dimensionless variables.

Staying in the special case $\Lambda > 0$, $\alpha < 2$, proceed by simplifying the radical term first:

$$\begin{aligned}E - U_{\text{eff}}(r) &= E + \frac{\Lambda}{r^\alpha} - \frac{L^2}{2m_* r^2} \\ &= \frac{\mathcal{E}\Lambda\alpha}{2r_0^\alpha} \left(\frac{2-\alpha}{\alpha} \right) + \frac{\Lambda}{r_0^\alpha \rho^\alpha} - \frac{\Lambda\alpha}{2r_0^\alpha \rho^2} \\ &= \left(\frac{\Lambda\alpha}{2r_0^\alpha} \right) \left((2-\alpha)\mathcal{E}/\alpha + (2/\alpha)/\rho^\alpha - 1/\rho^2 \right)\end{aligned}$$

Substituting carefully, one finds

$$\begin{aligned}\tau &= \pm \frac{1}{2\pi} \int_{\rho_i}^{\rho_f} \frac{d\rho}{\sqrt{(2-\alpha)\mathcal{E}/\alpha + (2/\alpha)/\rho^\alpha - 1/\rho^2}} \\ \theta &= \pm \int_{\rho_i}^{\rho_f} \frac{d\rho/\rho^2}{\sqrt{(2-\alpha)\mathcal{E}/\alpha + (2/\alpha)/\rho^\alpha - 1/\rho^2}}.\end{aligned}$$

Note that solutions to

$$0 = (2 - \alpha) \mathcal{E}/\alpha + (2/\alpha) / \rho^\alpha - 1/\rho^2$$

indicate all apogees and perigees in the motion.

3.3 Inverse Square Attraction

The attractive power law potential with $\Lambda > 0$ and $\alpha = 1$ corresponds to the gravitational force and the attractive static electric force. Such a potential naturally hosts a circular orbit with:

$$\begin{aligned} r_0 &= \frac{L^2}{\Lambda m_*} \\ T_0 &= \frac{2\pi m_* r_0^2}{L} \\ |U_{\text{eff}}| &= \frac{L^2}{2m_* r_0^2} = \frac{\Lambda^2 m_*}{2L^2} \end{aligned}$$

Spatial Dynamics

The equations of motion simplify significantly with $\alpha = 1$. For the θ -equation, we have

$$\theta = \pm \int_{\rho_i}^{\rho_f} \frac{d\rho/\rho^2}{\sqrt{\mathcal{E} + 2/\rho - 1/\rho^2}} .$$

Let $\xi = 1/\rho$ to find

$$\theta = \pm \int \frac{-d\xi}{\sqrt{\mathcal{E} - (\xi - 1)^2 + 1}} ,$$

and then let $\beta = \xi - 1$ to get

$$\theta = \pm \int \frac{-d\beta}{\sqrt{\mathcal{E} - \beta^2 + 1}} .$$

Factor $\sqrt{1 + \mathcal{E}}$ from the denominator and also let $\gamma = \beta/\sqrt{1 + \mathcal{E}}$:

$$\begin{aligned} \theta &= \pm \int \frac{-d\beta}{\sqrt{1 + \mathcal{E}} \sqrt{1 - \beta^2/(1 + \mathcal{E})}} \\ &= \pm \int \frac{-d\gamma}{\sqrt{1 - \gamma^2}} \end{aligned}$$

Next, let $\gamma = \cos(\psi)$ to find

$$\theta = \pm \int \frac{\sin(\psi) d\psi}{\sin(\psi)} = \pm \int d\psi .$$

The remaining integral has a trivial solution

$$\theta = \theta_0 \pm \psi ,$$

and undoing all substitutions gives

$$\theta = \theta_0 \pm \arccos\left(\frac{1/\rho - 1}{\sqrt{1 + \mathcal{E}}}\right) ,$$

where the integration constant θ_0 is an ignorable rotation in the plane. Continue solving for ρ to get the equation of a conic section:

$$\rho = \frac{1}{1 + \sqrt{1 + \mathcal{E}} \cos(\theta)}$$

Eccentricity

The combination $\sqrt{1 + \mathcal{E}}$ is none other than the eccentricity of the orbit:

$$e = \sqrt{1 + \mathcal{E}}$$

From what we know of conic sections, recall that $e = 1$ makes a parabola, $e < 1$ makes an ellipse, and $e > 1$ makes a hyperbola. In terms of the dimensionless energy $\mathcal{E}$, this also means:

$$\begin{aligned} \mathcal{E} = 0 &\rightarrow \text{parabola} \\ \mathcal{E} < 0 &\rightarrow \text{ellipse} \\ \mathcal{E} > 0 &\rightarrow \text{hyperbola} \end{aligned}$$

Conserved Quantities

As a special case of the two-body central potential system, we're aware that the inverse-square attraction supports conservation of energy E and conservation of angular momentum L .

Also conserved is the Runge-Lorenz vector

$$\mathbf{Z} = \mathbf{v} \times \mathbf{L} - \Lambda \hat{\mathbf{r}} ,$$

which fixes the orientation of the total orbit in its plane of motion. Take a time derivative to quickly prove $\mathbf{Z}$ is constant:

$$\begin{aligned} \frac{d}{dt} \mathbf{Z} &= \frac{d\mathbf{v}}{dt} \times \mathbf{L} + \mathbf{v} \times \frac{d\mathbf{L}}{dt} - \Lambda \frac{d\hat{\mathbf{r}}}{dt} \\ &= -\frac{1}{m_*} \frac{\Lambda}{r^2} \hat{\mathbf{r}} \times (\mathbf{r} \times m_* \mathbf{v}) - \Lambda \frac{d}{dt} \left(\frac{\mathbf{r}}{r} \right) \\ &= \Lambda \left(\frac{-(\hat{\mathbf{r}}(\hat{\mathbf{r}} \cdot \mathbf{v}) - \mathbf{v}(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}))}{r} - \frac{\mathbf{v}}{r} + \frac{\hat{\mathbf{r}}}{r} (\hat{\mathbf{r}} \cdot \mathbf{v}) \right) \\ &= 0 \end{aligned}$$

Being constant, we're free to evaluate $\mathbf{Z}$ anywhere on the orbit. Choosing a perigee at $\mathbf{r}_p = r_p \hat{\mathbf{x}}$ where $\mathbf{v}_p \cdot \mathbf{r}_p = 0$, we find:

$$\begin{aligned}\mathbf{Z} &= \mathbf{v}_p \times (\mathbf{r}_p \times m_* \mathbf{v}_p) - \Lambda \hat{\mathbf{r}}_p \\ &= m_* \mathbf{r}_p (\mathbf{v}_p \cdot \mathbf{v}_p) - m_* \mathbf{v}_p (\mathbf{v}_p \cdot \mathbf{r}_p) - \Lambda \hat{\mathbf{r}}_p \\ &= \left(\frac{L^2}{m_* r_p} - \Lambda \right) \hat{\mathbf{r}}_p = \Lambda \left(\frac{r_0}{r_p} - 1 \right) \hat{\mathbf{x}} \\ &= \Lambda \left(\frac{1}{\rho_p} - 1 \right) \hat{\mathbf{x}} = \Lambda \left(1 + \sqrt{1 + \mathcal{E}} - 1 \right) \hat{\mathbf{x}} \\ \mathbf{Z} &= \Lambda e \hat{\mathbf{x}}\end{aligned}$$

Conic Trajectory

Problem 14

Derive the dimensionless Runge-Lorenz vector equation

$$\hat{\mathbf{r}} + e \hat{\mathbf{x}} = \frac{\mathbf{v} \times \mathbf{L}}{\Lambda},$$

and then project $\mathbf{r}$ into each side to recover the equation of a conic section, namely

$$r = \frac{r_0}{1 + e \cos(\theta)}.$$

Temporal Dynamics

The integral for the dimensionless time τ is straightforwardly solved with $\alpha = 1$. Begin with

$$\tau = \pm \frac{1}{2\pi} \int_{\rho_i}^{\rho_f} \frac{d\rho}{\sqrt{\mathcal{E} + 2/\rho - 1/\rho^2}}$$

and assume $\mathcal{E} \neq 0$. Simplify a bit to get

$$\tau = \pm \frac{\sqrt{\mathcal{E}}}{2\pi e} \int_{\rho_i}^{\rho_f} \frac{\rho d\rho}{\sqrt{\mathcal{E} \left(\rho\sqrt{\mathcal{E}} + 1/\sqrt{\mathcal{E}} \right)^2 / e^2 - 1}}.$$

Let $\xi = \sqrt{\mathcal{E}} \left(\rho\sqrt{\mathcal{E}} + 1/\sqrt{\mathcal{E}} \right) / e$ so that $d\xi = d\rho \mathcal{E}/e$ and

$$\rho d\rho = e \left(\frac{\xi e - 1}{\mathcal{E}^2} \right) d\xi,$$

giving

$$\tau = \pm \frac{1}{2\pi \mathcal{E}^{3/2}} \int \frac{(\xi e - 1)}{\sqrt{\xi^2 - 1}} d\xi.$$

So far we've been a bit loose with the sign on $\mathcal{E}$. For $\mathcal{E} > 0$, everything stays as-is. However for $\mathcal{E} < 0$, we have to

propagate $\mathcal{E} \rightarrow -|\mathcal{E}|$ through the calculation. Maintaining both channels, we have:

$$\begin{aligned}\tau_{\mathcal{E}>0} &= \pm \frac{1}{2\pi |\mathcal{E}|^{3/2}} \int \frac{(\xi e - 1)}{\sqrt{\xi^2 - 1}} d\xi \\ \tau_{\mathcal{E}<0} &= \pm \frac{1}{2\pi |\mathcal{E}|^{3/2}} \int \frac{(\xi e - 1)}{\sqrt{1 - \xi^2}} d\xi\end{aligned}$$

Let

$$\begin{aligned}\xi_{\mathcal{E}>0} &= \cosh(\psi) \\ \xi_{\mathcal{E}<0} &= \cos(\psi)\end{aligned}$$

and each integral becomes trivial, setting arbitrary integration constants to zero:

$$\begin{aligned}\tau_{\mathcal{E}>0} &= \pm \frac{1}{2\pi |\mathcal{E}|^{3/2}} (e \sinh(\psi) - \psi) \\ \tau_{\mathcal{E}<0} &= \pm \frac{1}{2\pi |\mathcal{E}|^{3/2}} (\psi - e \sin(\psi))\end{aligned}$$

The period of an elliptical orbit corresponds to one full cycle in ψ , i.e. $0 \leq \psi < 2\pi$. For this we have

$$\tau_{\text{period}} = 2(\tau_{\pi} - \tau_0) = |\mathcal{E}|^{-3/2}.$$

Problem 15

Use

$$\tau = \frac{t}{T_0} = \frac{Lt}{2\pi m_* r_0^2}$$

and let $\tau = |\mathcal{E}|^{-3/2}$. Solve for $t = T$ to recover Kepler's third law for elliptical orbits.

Radial Component

Using the ρ -substitution from the above integral, namely

$$\begin{aligned}\rho_{\mathcal{E}>0} &= \frac{e\xi - 1}{\mathcal{E}} \\ \rho_{\mathcal{E}<0} &= \frac{e\xi - 1}{-|\mathcal{E}|},\end{aligned}$$

we get solutions for $\rho(\psi)$:

$$\begin{aligned}\rho_{\mathcal{E}>0} &= \frac{e \cosh(\psi) - 1}{\mathcal{E}} \\ \rho_{\mathcal{E}<0} &= \frac{1 - e \cos(\psi)}{|\mathcal{E}|}\end{aligned}$$

The perigee corresponds to $\psi = 0$. For $\mathcal{E} < 0$, the apogee is at $\psi = \pi$.

Zero-Energy Case

Returning to the equations of motion for τ, θ with $\mathcal{E} = 0$ and $\alpha = 1$, we have

$$\tau = \pm \frac{1}{2\pi} \int_{\rho_i}^{\rho_f} \frac{\rho \, d\rho}{\sqrt{2\rho - 1}}$$

$$\theta = \pm \int_{\rho_i}^{\rho_f} \frac{d\rho/\rho}{\sqrt{2\rho - 1}},$$

resolving to, discarding the integration constants:

$$\tau = \pm \frac{1}{6\pi} \sqrt{2\rho - 1} (\rho + 1)$$

$$\theta = \pm 2 \arctan \left(\sqrt{2\rho - 1} \right)$$

3.4 Inverse Square Repulsion

Consider the central potential given by

$$U = \frac{\Lambda}{r},$$

as one finds with the repulsive Coulomb force. This is the same as the inverse square attractive case with the modification $\Lambda \rightarrow -\Lambda$, and the recipe for the equations of motion is essentially the same. Some key results are:

$$\rho = \frac{1}{-1 + \sqrt{1 + \mathcal{E}} \cos(\theta)}$$

$$\tau = \pm \frac{1}{2\pi |\mathcal{E}|^{3/2}} (e \sinh(\psi) + \psi)$$

$$\rho = \frac{1}{\mathcal{E}} (e \cosh(\psi) + 1)$$

Chapter 31

Thermodynamics

1 Heat and Energy

Thermodynamic System

A *thermodynamic system* is any collection of parts that permits energy transfer between its members. A closed thermodynamic system allows no energy or material transfer across its boundary. Conversely, an open thermodynamic system does allow such interactions with its environment.

Temperature

Temperature is a scale used for quantifying the amount of ‘hotness’ in a system. We’ll soon build up the vocabulary to create a sophisticated definition of temperature, but it suffices for now to say:

Temperature is a quantity that eventually becomes the same for two systems after sustained contact.

Typical temperature scale units, namely Fahrenheit and Centigrade, aren’t defined in fundamental terms of length, space, or time.

Problem 1

At atmospheric pressure (also known as *standard pressure*), liquid water boils at 212°F and freezes at 32°F. On the centigrade scale, these temperatures are 100°C and 0°C, respectively. Calculate the one temperature at which the Fahrenheit and centigrade scales, assumed linear with temperature, are in agreement. Answer: $-40^\circ\text{F} = -40^\circ\text{C}$

Zeroth Law

Sustained contact between two systems can lead to a state called *thermal equilibrium*, achieved when there is zero

net energy flowing between each system. This idea is also contained in the *zeroth law of thermodynamics*:

If system A and system B are each in equilibrium with system C, then A and B are in equilibrium with each other.

Quasistatic Approximation

The *quasistatic approximation* is the limit when thermodynamic processes occur ‘slowly’ enough such the whole substance maintains instantaneous equilibrium throughout any process.

1.1 Matter and Temperature

Thermal Expansion

The volume of an object is subject to its temperature. Adding heat usually causes the volume to increase, an effect called *thermal expansion*.

Taking a one-dimensional example, consider a thin metal rod of length L_0 and initial temperature T_0 . An external heat source changes the rod’s temperature by ΔT , and the length is observed to change by ΔL . Restoring the temperature to T_0 , the length restores to L_0 . It follows that a linear ‘law’ for thermal expansion can be written as

$$L = L_0 + \alpha L_0 \Delta T, \quad (31.1)$$

where the parameter α is the *linear thermal expansion coefficient*, measuring typically around 10^{-5} per degree centigrade. The temperature change ΔT may be positive or negative, and the rod expands or contracts, respectively.

Thermal expansion occurs in more than one spatial dimensions. For two dimensions, consider a plate of area

$L_0 \times W_0$. With a change in temperature, the area of the plate becomes

$$A = L_0 W_0 (1 + \alpha \Delta T)^2 \approx L_0 W_0 (1 + 2\alpha \Delta T) . \quad (31.2)$$

For small temperature displacements, $\alpha^2 \Delta T^2$ is a reasonably small contribution to thermal expansion and is ignored to first-order approximation. Thus the area scales linearly with temperature, with effective expansion factor 2α .

For three dimensions, the effective expansion constant is $\beta = 3\alpha$. If the volume of the object is V_0 , this means

$$V \approx V_0 (1 + \beta \Delta T) . \quad (31.3)$$

Problem 2

A circular piece with diameter d is removed from a room-temperature sheet of aluminum. If the sheet is heated in an oven, how does the area of the hole change with temperature? Does the hole increase or decrease in area? Answer: $\Delta A = \pi d^2 \alpha \Delta T / 2$, increasing.

Problem 3

On a cool 4°C morning, a driver fills his aluminum gasoline tank to the full capacity of 106.0 L. That evening, he checks the fuel level and finds 103.4 L remaining in the tank. The coefficient of volume expansion for gasoline is $9.5 \times 10^{-4} / ^\circ\text{C}$, and for aluminum is $7.25 \times 10^{-5} / ^\circ\text{C}$. What was the maximum temperature reached by the fuel tank during the day? Answer: 32.70°C

Problem 4

Consider a metal rod having a thermal expansion coefficient that varies with temperature such that $\alpha(T) = A + BT + CT^2$. Derive a formula for the length of the material as a function of temperature change. Answer:

$$\Delta L = L_0 \left(A\Delta T + B\Delta T^2/2 + C\Delta T^3/3 \right)$$

Thermal Stress

An object confined in space may not be able to physically undergo thermal expansion (or contraction), in which case the object undergoes *thermal stress*.

Suppose a thin metal rod of fixed length L_0 is held between two clamps at temperature T_0 with no initial forces on the ends of the rod. If the temperature changes, the rod ‘wants’ to obey the thermal expansion Equation (31.1), written here as

$$\frac{\Delta L}{L_0} = \alpha \Delta T ,$$

where α is the linear thermal expansion coefficient.

Meanwhile, the Young’s modulus for the rod is defined as $Y = (F/A)/(\Delta L/L_0)$, so we write

$$\frac{\Delta L}{L_0} = \frac{F}{AY} ,$$

where F is the linear tension in the rod (positive or negative) and A is the cross-sectional area. Since ΔL must be zero, we therefore have

$$\frac{F}{A} = -Y\alpha\Delta T . \quad (31.4)$$

1.2 Thermal Energy

The phenomenon of ‘heat’ occurs as a consequence of some type of energy expense: mechanical, chemical, electrical, nuclear, etc. We define *thermal energy*, denoted Q , as the broad subset of energies and potentials that couple strongly to temperature. Like all other energies in physics, thermal energy must be conserved.

It was known to Sir James Joule (1818-1889) and his contemporaries that the amount of work required to raise the temperature of water is directly proportional to the temperature change. Energy was measured in calories, where one *calorie* (c) is the amount of energy required to raise the temperature of one gram of water by one degree centigrade. (One food calorie, denoted with a capital C, is 1000 ordinary calories.)

Experiments conducted by Joule found that one calorie equals 4.186 energy units in the SI system, also known as Joules. Contemporary to this is the Btu, or *British thermal unit*, being $252 \text{ cal} = 1055 \text{ J}$. The Btu is defined as the amount of energy required to raise one pound of water by one degree Fahrenheit.

Specific Heat

A small thermal energy quantity dQ transferred to a substance of mass m (without changing the phase) is linearly proportional to the change in temperature. This means we write

$$c = \frac{1}{m} \frac{dQ}{dT} , \quad (31.5)$$

where the proportionality constant c is the *specific heat* of the substance. Equation (31.5) can often be integrated in the variable T , giving a more familiar statement

$$Q = mc\Delta T . \quad (31.6)$$

A typical solid metal has c around 300 J per kilogram of material per degree centigrade. In the same units, liquid

water has $c = 4187$, frozen water has $c = 2108$, and water vapor has $c = 1996$.

Heat Capacity

The product of mass and specific heat yields the *heat capacity*:

$$C = mc \quad (31.7)$$

Like specific heat, the heat capacity is proportional to Joules per degree centigrade, but the notion of ‘per kilogram of material’ has been multiplied away. Heat capacity refers to the whole object.

1.3 Phases of Matter

Phase is a term synonymous with ‘state of matter’, used for classifying the overall spatial arrangement and dynamics of the particles in a substance. Common phases of matter are solid, liquid, and gas - but the list doesn’t stop there. More exotic phases such as plasmas and liquid crystals occur in nature (and the laboratory).

The phase of a given substance is chiefly determined by (i) the surrounding temperature, and (ii) the surrounding pressure. This is visualized by using a *phase diagram*, which comes in two flavors: *PT* and *PV*, standing for a pressure-temperature plot or pressure-volume plot, respectively.

PT Diagram

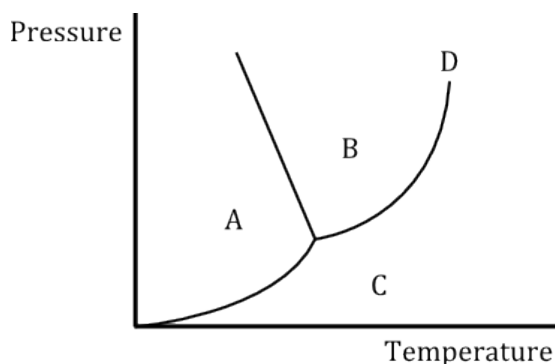


Figure 31.1. *PT* phase diagram.

Figure 31.1 illustrates a typical *PT* diagram for a typical substance (water for example). A small enough ‘parcel’ of a substance occupies just one point in the phase diagram at a given time, and the parcel’s ‘trajectory’ in a phase diagram is continuous over time, but not necessarily smooth.

Boundaries between regions in the phase diagram are associated with abrupt changes in the phase:

- Region *A* has low temperature and high pressure, corresponding to the solid state.
- Region *B* has moderate temperature and moderate pressure, corresponding to the liquid state.
- Region *C* has high temperature and low pressure, corresponding to the gaseous state.
- Region *D* indicates the supercritical state, a phase mixture of liquid and gas.
- Solid lines are boundaries that separate two phases.
- The junction of regions *A*, *B*, *C* is called the *triple point*.

PV Diagram

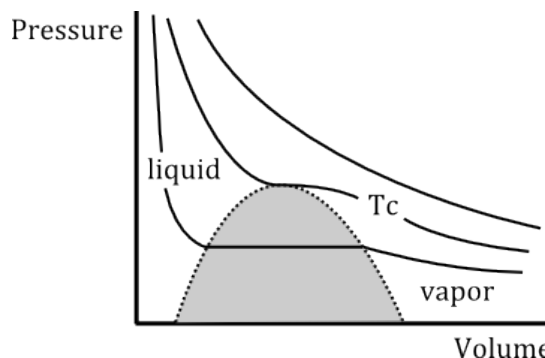


Figure 31.2. *PV* phase diagram.

A variation of the *PT* diagram is the *PV* diagram, where the volume replaces temperature on the horizontal axis. Figure 31.2 illustrates a typical *PV* diagram for another typical substance, excluding the solid state. The contour curves are called *isotherms*, having constant temperature. The shaded area under the dotted arch represents the liquid-vapor equilibrium phase. The isotherm T_c is the *critical temperature*.

Mixtures and Dalton’s Law

A mixture of gases is itself a gas and may be treated as a unit. For instance, if a pressurized vessel holds a mixture of helium and nitrogen, then the gases are fully mixed, sharing the same volume and same temperature.

Each element also makes its own contribution to the pressure, called a *partial pressure*, where the total pressure P is $P = P_{He} + P_{Ni}$ for the example on hand. This is known

as Dalton's law. In general, the law of partial pressures reads

$$P = \sum_j P_j, \quad (31.8)$$

where index j sums over the species in the mixture.

Latent Heat

Changing the temperature or pressure of a substance causes some shift in the phase diagram. Crossing the boundary between any two regions (see Figure 31.1) corresponds to a *phase change*. At such a boundary, molecular order of the substance changes radically, which involves an energy change at the *chemical* level.

Now comes an important point about latent heat. For a substance *on* a phase boundary at a given pressure, all energy added or removed from the substance contributes to phase change with the temperature remaining constant.

To illustrate, consider a glass containing an ice + water mixture having initial temperature 0°C , the freezing point of water. Placed in a room-temperature environment, the mixture steadily absorbs heat from the surroundings. Does the temperature of the ice + water mixture increase? *No*, not for a while. Rather, any new energy added to the system contributes to melting the ice. Only when the ice is melted will the system's temperature begin to increase. The same experiment works in reverse: if instead the ice + water mixture were placed in a frigid -100°C environment, the temperature would remain 0°C until all liquid water phase changes to ice. Only after this point would the overall temperature begin to plunge. (Of course, this all assumes that water is a perfect conductor of heat.)

During a phase change, the thermal energy of a substance is not governed by $Q = mc\Delta T$ because ΔT is zero during the change. Instead, the change in energy is called *latent heat*, which is proportional to the mass of material being changed:

$$Q = mL_\alpha \quad (31.9)$$

The amount of mass that changes phase per unit energy is a constant denoted L_α , where subscript α denotes which phase boundary is being crossed.

Subscript $\alpha = v$ stands for 'vaporization', corresponding to the liquid-gas boundary, and $\alpha = f$ stands for 'fusion', corresponding to the solid-liquid boundary. Water at atmospheric pressure has $L_v = 2.256 \times 10^6 \text{ J/kg}$ and $L_f = 3.34 \times 10^5 \text{ J/kg}$.

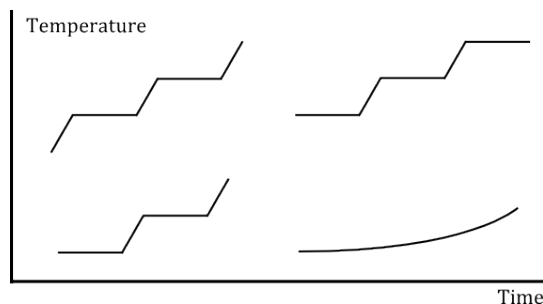


Figure 31.3. Temperature curves for tea kettle problem.

Problem 5

A tea kettle is filled with ice at -10°C and placed on a burner that is steadily getting hotter. Over a long enough time, the kettle empties due to evaporation. Of the four curves drawn in Figure 31.3, which best represents the temperature inside the kettle as a function of time? Use the quasistatic approximation. Answer: Top-left

Problem 6

How much ice at 0°C must be added to a liter of water at 80°C so as to end up with all liquid at 20°C ? Answer:

$$\frac{m_{\text{ice}}}{m_0} = \frac{80^\circ\text{C} - 20^\circ\text{C}}{20^\circ\text{C} + (L_f/c)_{\text{water}}}$$

Problem 7

A 15 g ice cube at 0°C is mixed with 2 g of steam at 100°C in an isolated container. What is the final state of the system?

Answer: (i) Since $m_{\text{steam}}L_v - m_{\text{ice}}L_f < 0$, all steam condenses, releasing 4512 J, which melts 13.51 g of ice, leaving 1.49 g in the solid state. (ii) Next, the 2 g of hot water must come to equilibrium with the melted ice, giving $0 = 2(x - 100^\circ\text{C}) + 13.51x$, where x solves to 12.9°C . (iii) Finally, let 15.51 g of water at 12.9°C come to equilibrium with the remaining ice, giving $0 = 15.51(y - 12.9^\circ\text{C}) + 1.49L_f/c$. Solving for y , the final system has 17 g of water at about 5°C .

Problem 8

A 20 g ice cube at -10°C is mixed with 2 g of steam at 100°C in an isolated container. What is the final state of the system?

Answer: (i) Energy required to warm ice to 0°C : 422 J. (ii) Amount of steam condensed to accomplish this: 0.187 g. (iii) Energy required to melt ice: 6680 J. (iv) Energy released by remaining steam condensation: 4090 J. (v) Energy released from cooling 2 g of water from 100°C to 0°C : 837 J. (vi) Energy available to melt ice: 4927 J. (vii) Mass of ice

melted: 14.8 g. (viii) Final state has about 5 g of ice and 17 g of water at 0°C.

1.4 Heat Transfer

Conduction

Conduction is heat transfer due to microscopic motions of particles, where more rapidly-moving particles exchange kinetic energy by collision with ‘colder’ particles. A material’s affinity to conduct heat in this way is the *thermal conductivity*, k , measured in Watts per meter per degree centigrade. (Incidentally, most materials that easily conduct electric currents also have high thermal conductivity.)

Consider two objects H and C that are maintained at constant temperatures T_H and T_C , respectively. The ‘hotter’ object H is connected to the ‘colder’ object C by a conductor having cross-sectional area A and thermal conductivity k .

From ordinary classical mechanics, we can write the rate of energy flow dQ/dt across the conductor:

$$\frac{dQ}{dt} = -\frac{d}{dt} \int \nabla U \cdot d\mathbf{x} = - \int \nabla U \cdot \frac{d\mathbf{x}}{dt}$$

The gradient ∇U term is proportional to (i) the temperature difference $\Delta T = T_C - T_H$, and (ii) the conductor cross section A . The $d\mathbf{x}/dt$ term represents the rate and direction of energy flow across A , proportional to the thermal conductivity k . The differential equation for heat conduction is therefore

$$\frac{dQ}{dt} = -kA\Delta T, \quad (31.10)$$

where ΔT is known as the *temperature gradient*. The negative sign reminds that heat flows from hot to cold.

R-Value

The R -value of a material is equal to the thickness (along the direction of heat flow) divided by the conductivity:

$$R = \frac{L}{k}$$

One inch of wood has $R \approx 1$. Typical New England houses have $R \approx 19$ for the walls and $R \approx 30$ for the roof. The R -value is an additive quantity, i.e. $R_{\text{eff}} = R_1 + R_2 + \dots$

Problem 9

The walls of a house are insulated with 2.0 cm of Styrofoam ($k_{\text{SF}} = 0.01 \text{ W/mK}$), and 15.0 cm of fiberglass

($k_{\text{FG}} = 0.04 \text{ W/mK}$). The Styrofoam is on the exterior of the house. The outside temperature is 0°C, and the interior of the house is maintained at 20°C. (i) Calculate the temperature on the SF-FG interface. (ii) What is the rate of heat transfer per m^2 through both layers of insulation? (iii) Which provides greater net insulation, an additional 1.0 cm of Styrofoam, or an additional 5.0 cm of fiberglass? Answer: 6.96°C, 3.5 W, fiberglass.

Problem 10

Inside a conductive cylinder of length L and radius R , heat flows radially from the axis $r = 0$ toward the wall at R with $T(0) > T(R)$. Calculate the thermal energy flow rate between any two radii r_1 and r_2 in cylinder. Answer:

$$dQ/dt = k2\pi RL (T_1 - T_2) / \ln(r_2/r_1)$$

Problem 11

If a tank of water initially at 0°C is left in very cold conditions (assume -10°C), a sheet of ice forms on the water surface and grows downward over time. Supposing all heat transfer takes place through the ice, and not through the container walls: (i) Determine the thickness Z of the ice as a function of time t . (ii) Calculate the thickness of the ice sheet that will form in one day. (iii) If the tank is 50 cm deep, how many days does it take to freeze all of the water? (iv) If the tank is 10 m deep, how many days does it take to freeze all of the water? Answer: (i) $Z = (\Delta T 2k / (\rho L_f))^{1/2} \sqrt{t}$, (ii) 9.49 cm, (iii) 2.40×10^6 sec, (iv) 6.40×10^6 days

Convection

Convection is heat transfer due to macroscopic displacement of a fluid. The displacement can be spontaneous, as when steam carries heat away from a hot cup of tea - or forced, as done inside a refrigerator compressor.

For example, consider a pot of water initially at 10°C that sits on a kitchen stove burner maintained at 110°C.

1. Conduction: Heat is introduced to the liquid by contact with the bottom of its metal container, in contact with the heat source.
2. Convection cells: Fluid parcels nudge their way to the surface, spreading horizontally, and submerging when displaced by warmer fluid.
3. Turbulence: The bottom of the fluid becomes much warmer than the surface, and the convection cells burst into mushroom-cloud shapes.

4. Boiling: Discernible patterns vanish, and the fluid expels its heat by ejecting the most energetic particles.
5. Steam: Steam rapidly expands (if not constrained), doing work on the atmosphere.

Steps (2) through (5) each involve convection.

Electromagnetic Radiation

Electromagnetic radiation is energy transfer that uses photons as a carrier. All objects continuously absorb and expel energy through photon exchange with the environment. The net rate of heat flow is given by Stefan's law,

$$\frac{dQ}{dt} = Ae\sigma (T_{\text{obj}}^4 - T_{\text{env}}^4), \quad (31.11)$$

where A is the object's exposed surface area, e is the emissivity ($0 \leq e \leq 1$) of the material, and σ is the *Stefan-Boltzmann* constant, $5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$. Note the temperature is counted in Kelvin units, which scales proportionally to the centigrade scale but equals zero at $T = -273.15^\circ\text{C}$.

A hot electric stove coil emits intense infrared radiation with wavelengths ranging from 10^3 nm to 10^6 nm . (Your skin easily detects these rays.) With increasing power, the range of emitted wavelengths widens, and eventually the coil visibly glows red with $\lambda = 700 \text{ nm}$. (Your eyes easily detect these.) With even more increasing power, the stove coil becomes 'white hot', emitting a very wide spectrum of wavelengths.

A hypothetical object useful for discussing radiation is the *blackbody*, which has $e = 1$, absorbing all incident radiation. Contrastly, a *reflector* absorbs no radiation.

Problem 12

Earth's upper atmosphere receives roughly $1.50 \times 10^3 \text{ W/m}^2$ of energy from the sun by thermal radiation. The distance from the earth to the sun is $1.50 \times 10^{11} \text{ m}$, and the radius of the sun $6.96 \times 10^8 \text{ m}$. Use conservation of flux to calculate the surface temperature of the sun. Answer: About 6000 Kelvin.

Problem 13

Wandering in the Desert: Consider an unfortunate person walking at 5 km/h on a hot day in the desert, wearing only a bathing suit. The person's skin temperature tends to rise due to four mechanisms:

- Energy is generated by metabolic reactions in the body at a rate of 280 W, and almost all of this energy is converted to heat that flows to the skin.

- Heat is delivered to the skin by convection from the outside air at a rate equal to $k' A_{\text{skin}} (T_{\text{air}} - T_{\text{skin}})$, where k' is $54 \text{ J/}^\circ\text{C m}^2\text{h}$, the exposed skin area A_{skin} is 1.5 m^2 , and the air temperature T_{air} is 47°C .
- The skin absorbs radiant energy from the sun at a rate of 1400 W/m^2 .
- The skin absorbs radiant energy from the environment, which has temperature 47°C .

Assume the emissivity of the skin is $e = 1$ and the skin temperature is 36°C . (i) Determine the net heat flow into the person due to the four mechanisms. (ii) At what rate (in liters per hour) must perspiration evaporate from the person's skin to maintain a constant skin temperature? (Assume sweat is made of water which has $L_v = 2.42 \times 10^6 \text{ J/kg}$.) (iii) Suppose instead that the person is protected by light colored clothing with $e = 0$ such that the skin exposed is 0.45 m^2 . (The convective heat exchange is not affected by the clothing.) What is the rate of perspiration now required? Answer: (i) 280 W, 0.248 W, $2.10 \times 10^3 \text{ W}$, 893 W, (ii) 4.87 L/h, (iii) 1.75 L/h

1.5 Atmospheric Qualities

Vapor Pressure

The *vapor pressure* of a substance is the pressure at which the vapor phase is in equilibrium with the solid or liquid phase at a given temperature. For example, water and air in a sealed container will begin mixing by evaporation and condensation (of water molecules in and out of the air). Equilibrium is achieved when the partial pressure P_W of water in the air is equal to the vapor pressure P_0 .

The following table maps out the vapor pressure of water as a function of temperature:

Temperature ($^\circ\text{C}$)	Vapor Pressure ($\text{Pa} \times 10^3$)
10	1.23
12	1.40
14	1.60
16	1.81
18	2.06
20	2.34
22	2.65
24	2.99
26	3.36
28	3.78
30	4.25
40	7.34

Note one Pascal (Pa) is equal to one Newton per square meter.

Humidity

The *relative humidity* is defined as the ratio

$$H = \frac{P_W}{P_0},$$

always expressed as a percentage and never exceeding 100%. Because Earth's atmosphere is considered open instead of closed, H is typically much less than unity. On a rainy day, H becomes close to unity. In either case, P_0 is much less than the atmospheric pressure P_A , which is near 10^5 Pa.

Problem 14

At 24°C and at 50.0% relative humidity, what is the partial pressure of water vapor in the atmosphere? Answer: $P_W = 1.50 \times 10^3$ Pa

Problem 15

At 24°C and at 50.0% relative humidity, what is the mass density of the water vapor in the air? (Hint: $M = 18 \times 10^{-3}$ kg/mol.) Answer: $\rho = 0.0109$ kg/m³

Dew Point

Consider an ice-cold glass sitting in a room. After a short time, water droplets form on the outside of the glass. To explain this, observe first that the air surrounding the glass, which itself contains water molecules, is cooled by conduction. The vapor pressure of the surrounding molecules therefore decreases (see table above), causing the relative humidity surrounding the glass to increase. When the relative humidity reaches 100%, the corresponding temperature is called the *dew point*, denoted T_{dew} .

At the dew point temperature, *further cooling cannot result in increased humidity*, which is already at unity, so the density of water molecules surrounding the glass must decrease. The excess molecules stick to the glass to form droplets. Note that the dew point is always less than the surrounding room temperature. The higher the relative humidity, the closer the dew point is to the ambient temperature.

Problem 16

Suppose the air temperature in a laboratory is 24°C , and a physics student cools a half-filled drink by adding ice. He finds that water begins to condense on the glass when its temperature reaches 12°C . What is the relative humidity in the room? Answer: 46.8%

Clouds and Fog

Clouds and fog are easily explained in the terms defined above. Simply put, condensation (100% relative humidity) occurs at the altitude where the air temperature equals the dew point.

Problem 17

Consider a spring day when the air temperature on the ground is 24°C and the relative humidity is 50%. Assume that the partial pressure of water does not change with elevation, and that the air temperature decreases with altitude at a rate of $0.06^\circ\text{C}/100$ m. (i) At what altitude will clouds form? (ii) If the relative humidity on the ground is higher, will the clouds form at a higher or lower altitude? (iii) If the temperature gradient is less, will the clouds form at higher or lower altitude? Answer: (i) 1833 m, (ii) lower, (iii) higher

1.6 Ideal Gas

The chain of discoveries leading to the understanding of gases spans across several centuries.

Boyle's Law

In the seventeenth century, Robert Boyle (1627 - 1691) performed numerous studies on the properties of gases using the materials of his day (simple glassware and metals). The most famous fruit of his effort, called *Boyle's Law*, states that *the pressure of a fixed amount of gas maintained at constant temperature is inversely proportional to the volume of the gas*:

$$P \propto 1/V$$

Charles's Law

A similar discovery about gases that came out of France in the 1780s, now known as *Charles's Law*, stating that *the volume of a fixed amount of gas maintained at constant pressure is directly proportional to the absolute temperature of the gas*:

$$V \propto T$$

Lord William Kelvin

The linear relationship between volume V and temperature T found by Charles has an interesting feature that he probably didn't notice. All experimental data was indicating that any plot that displays $V(T)$ for *any* substance at *any* pressure has an x -intercept at -273.15°C . Scottish physicist Kelvin realized the significance of this in 1848, and seized the opportunity to name a new temperature scale after himself.

The *Kelvin temperature scale* resembles the centigrade system in every way with one exception: the bottom of the scale, 0 *Kelvin*, corresponds to -273.15°C , also known as *absolute zero*. As it turns out, absolute zero is the limit low temperature for any object in the universe.

Avogadro's Law

Italian scientist Amedeo Avogadro made his contribution in 1811, hypothesizing correctly that *the volume of a gas at constant temperature and pressure is proportional to the number of particles N composing the gas*:

$$V \propto N$$

Ideal Gas Equation

Compiling the above achievements into a single equation of state, we arrive at the *ideal gas equation*:

$$PV = NK_B T \quad (31.12)$$

Equation (31.12) describes the macroscopic behavior of all N particles in a gas at pressure P and temperature T (in Kelvin) occupying volume V . The constant K_B is the *Boltzmann constant*, and has the experimental value $1.381 \times 10^{-23} \text{ J/K}$.

Sometimes in physics (and more often in chemistry) the quantity NK_B is written instead as nR , where n is the number of moles of the gas, and R is the *ideal gas constant*, measured as $8.315 \text{ J mol}^{-1} \text{ K}^{-1}$ (Joules per mole-Kelvin).

Ideal Gas Model

The ideal gas law is an approximate (but usually accurate) description of a realistic gas. As an equation of state, (31.12) holds if the following assumptions can be safely made:

- The volume of the container holding the gas is much greater than the total volume of the constituent gas particles.
- The separation between gas particles is much greater than the diameter of a given particle.
- A given particle's position and direction of travel (not its speed) is completely random.
- Particles do not interact among themselves, with the exception of perfectly elastic collisions.
- The container walls exchange momentum with the gas particles isotropically and don't deform.
- At a given instant, P , V , N , and T are uniform throughout the gas (quasistatic approximation).

Typically, the subscript B is typically dropped from the Boltzmann constant; thus we simply write K for now on.

Problem 18

A rigid box of capacity 0.5 m^3 is initially open, but is then sealed, trapping air inside at 20°C at atmospheric pressure. The box is then heated until the pressure inside becomes 3 times that of the atmosphere. Calculate the temperature of the trapped air (in Kelvin) after heating. How many particles are in the box? Answer: $T = 879.45 \text{ K}$ and $N = 1.25 \times 10^{25}$.

2 Kinematic Gas Theory

Consider a sealed rigid vessel of volume V containing N non-interacting particles, not necessarily ideal gas. Such a gas has constant pressure and temperature at equilibrium.

2.1 Relative Velocity

Of all N particles in the gas, choose any two of them, where particle '1' has mass m_1 and moves with velocity $\mathbf{v}_1$. Similar can be said for particle '2'. Borrowing from two-body analysis, we can write the kinetic energy of the pair as

$$T = \frac{1}{2} m^* v_{\text{rel}}^2 + \frac{1}{2} (m_1 + m_2) v_0^2,$$

where v_0 is the velocity of the center of mass, v_{rel} is the relative velocity between the particles, and m^* is the reduced mass. If m_1, m_2 are equal to the same mass m then m^* reduces to $m/2$.

Starting with the relative velocity

$$\mathbf{v}_{\text{rel}} = \mathbf{v}_2 - \mathbf{v}_1,$$

square both sides and take the average of each term to write

$$\langle v_{\text{rel}} \rangle = \sqrt{\langle v_1 \rangle^2 + \langle v_2 \rangle^2 - 2 \langle \mathbf{v}_1 \cdot \mathbf{v}_2 \rangle}.$$

If we assume all particles are identical, then the averages $\langle v_1 \rangle, \langle v_2 \rangle$ are equal to the same velocity $\langle v \rangle$. Furthermore, we have that $\langle \mathbf{v}_1 \cdot \mathbf{v}_2 \rangle$ resolves to zero, because all velocity vectors are random and uncorrelated. From this we conclude

$$\langle v_{\text{rel}} \rangle = \sqrt{2} \langle v \rangle,$$

which is to say the average relative velocity between each particle is about 1.41 times the average absolute velocity of the particles in the gas.

Mean Free Path

Assuming a particle in the gas are spherical with radius r , the size of the particle is characterized by the cross-sectional area σ given by

$$\sigma = \pi (2r)^2.$$

In the time Δt between any two collisions, a particle carves out a Gaussian 'tube' of volume

$$V = \sigma \langle v_{\text{rel}} \rangle \Delta t,$$

where $\langle v_{\text{rel}} \rangle$ is the relative speed between the particle and its target (another particle).

We can also write the collision rate $f = \Delta N / \Delta t$. At the least, f is proportional to the density of the gas N/V , and is also proportional to the product $\sigma \langle v_{\text{rel}} \rangle$. Thus we write

$$f = \frac{\Delta N}{\Delta t} = \sigma \langle v_{\text{rel}} \rangle \frac{N}{V}.$$

Borrowing from wave mechanics, a frequency f is associated with a wavelength λ by a velocity term: $\lambda f = \langle v \rangle$, and works out to be

$$\lambda = \frac{\langle v \rangle}{f} = \frac{\langle v \rangle}{\sigma \langle v_{\text{rel}} \rangle (N/V)} = \frac{1}{\sigma \sqrt{2}} \frac{V}{N}. \quad (31.13)$$

The distance λ is the *mean free path* for particles of radius r in the gas. Notice the inverse proportionality to the density and the cross-sectional area. The smaller the particles and the less dense they are, the longer they go before a collision.

Problem 19

Show that the mean free path for ideal gas molecules is

$$\lambda = \frac{1}{\sigma \sqrt{2}} \frac{KT}{P}.$$

2.2 Pressure

We advance on the kinematic gas problem by calculating the pressure in two not so very different ways. For a setup we continue considering a sealed container with volume V containing gas at temperature T . Starting from so little, the only freebie we get is that the particle number density is uniform and equal to N/V . Going forward we'll refer to N/V as density, but remember there are no mass units involved.

Heuristic Pressure Calculation

Inside the vessel, consider a cylindrical Gaussian cylinder of length L , held so the longitudinal z -axis is perpendicular and adjacent to the inner surface of the vessel with intersection area A , assumed circular. At any given instant, particles contained in the membrane are moving in all directions.

Question: How many of the particles ΔN enclosed in the membrane will collide with the container wall in time interval Δt ? Answer: Statistically half of the enclosed particles collide with the wall (the other half are heading away). A particle will collide with the wall in a time interval Δt if it is heading toward the wall and within a distance $L = |v_z| \cdot \Delta t$ of the wall, where v_x is the x -component particle's absolute

speed, $\sqrt{\langle v^2 \rangle}$. The collision rate is also proportional to the overall density of the gas, N/V . Therefore we may write

$$\Delta N = (1/2) (A |v_z| \Delta t) (N/V) .$$

On any collision with the container wall, there are two contributions to the momentum exchange: (i) the wall first absorbs momentum p_z to momentarily stop the gas particle, and (ii) the wall then imparts momentum $-p_z$ to the particle, with a total exchange of $\Delta p = 2mv_z$ per collision.

By the impulse-momentum theorem $F dt = m dv$, the definition of pressure gives

$$P = \frac{F}{A} = \Delta N \frac{1}{A} \frac{\Delta p}{\Delta t} ,$$

and after inserting what we know about ΔN and Δp , we have

$$P = \frac{N}{V} m \langle v_z^2 \rangle .$$

Assuming the gas to be isotropic and living in three dimensions, it follows that

$$\langle v^2 \rangle = \langle v_x^2 \rangle + \langle v_y^2 \rangle + \langle v_z^2 \rangle = 3 \langle v_z^2 \rangle ,$$

bringing us to the result

$$P = \frac{1}{3} \frac{N}{V} m \langle v^2 \rangle . \quad (31.14)$$

Problem 20

In a 60 sec interval, 700 hailstones strike a 0.5 m^2 glass window at an angle of 45° with respect to the normal to the window surface. Each hailstone has mass of 4.0 g and speed 8.0 m/s. If the collisions are purely elastic, Calculate the average force and pressure on the window.

Formal Pressure Calculation

Inside the vessel, a given particle has any speed between zero and infinity in any direction. Introducing $f(v)$ as a generalized velocity distribution, we write the normalization condition as

$$1 = \int_{\text{all } v} f(v) d^d v ,$$

where d is the number of dimensions in velocity space.

Performing the prior analysis using more vector language, the same Gaussian cylindrical membrane of length L inside the vessel touches the inner z -wall with common area element $-dA \hat{z}$, where $dA = dx dy$. For particles with speed v inside the Gaussian cylinder, the length is represented by the vector $\mathbf{L} = (dt)v$. The number of particles ΔN destined

to collide with the wall is proportional to the macroscopic density N/V times the volume $dA \cdot \mathbf{L} = dA dt v_z$. Upon collision, the particle reverses the z -component of its momentum, with $dp_z = 2p_z$. Particles with negative p_z are heading away from the wall and don't collide.

Collecting these observations, write the 'pressure operator' as

$$[P_v] = \Delta N \frac{dF}{dA} = \Delta N \frac{dp_z/dt}{dA} = \frac{N}{V} v_z 2p_z ,$$

and the total pressure is given by

$$P = \int_v [P_v] f(v) d^3 v ,$$

expanding to:

$$P = \int_{v_x=-\infty}^{v_x=\infty} \int_{v_y=-\infty}^{v_y=\infty} \int_{v_z=0}^{v_z=\infty} \times \frac{N}{V} v_z 2p_z f(v) dv_x dv_y dv_z$$

Note the careful choice of limits on the v_z variable: no particles are coming from outside the container.

Since $f(v)$ is an even function by symmetry arguments, it does no harm to integrate over the interval $-\infty < v_z < \infty$ and divide by a factor of 2. The integral simplifies to

$$P = \frac{N}{V} \int_v (v_z p_z) f(v) d^3 v = \frac{N}{V} \langle v_z p_z \rangle ,$$

where the statistical definition of average has been used. Finally, note again from symmetry that $3 \langle v_z p_z \rangle = \langle v p \rangle$, and the pressure takes a form reminiscent of Equation (31.14):

$$P = \frac{1}{3} \frac{N}{V} \langle v p \rangle$$

2.3 Effusion

Effusion takes place as gas particles are allowed to escape through a small aperture in the enclosing vessel. While the substance is losing particles, energy, and so on - we still work within the quasistatic approximation to gain insight into the phenomenon.

Effusion Rate

Let us calculate the rate R (measured in s^{-1}) at which particles that strike a small patch of area A from inside the vessel. This is analogous to the pressure calculation, however now we ignore momentum exchange. For this, we begin with

$$R = \int_v [R_v] f(v) d^3 v$$

such that

$$[R_v] = A \frac{N}{V} v_z .$$

Expressed in 3D spherical coordinates, the effusion rate is

$$R = \int_{\phi=0}^{\phi=2\pi} \int_{\theta=0}^{\theta=\pi/2} \int_{v=0}^{v=\infty} A \frac{N}{V} v \cos(\theta) f(v) v^2 dv \sin(\theta) d\theta d\phi ,$$

where the z -axis is perpendicular to the container wall on the patch A . The gas exists only above the $z = 0$ plane, as indicated by the choice of limits on the variable θ . In this coordinate system, it follows that the normalization condition is

$$\int_0^{\infty} f(v) 4\pi v^2 dv = 1 .$$

Using the statistical definition of average velocity, the instantaneous effusion rate simplifies down to

$$R = \frac{A N}{4 V} \langle v \rangle . \quad (31.15)$$

Effusion Energy

With the effusion rate known, it's possible to frame the calculation for the average energy per particle carried away by effusion. Begin with the definition of the average energy

$$\langle E \rangle = \int_v [E_z] f(v) d^3v ,$$

and use

$$[E_z] = \left(\frac{1}{R} \right) \left(A \frac{N}{V} v_z \right) \left(\frac{mv_z^2}{2} \right) .$$

Note that the θ -integral resolves via

$$\int_0^{\pi/2} \cos^3(\theta) \sin(\theta) d\theta = \frac{1}{4} ,$$

and simplify the $\langle E \rangle$ expression to deduce

$$\langle E \rangle = \frac{m \langle v^3 \rangle}{4 \langle v \rangle} . \quad (31.16)$$

Problem 21

Show that the average effusion energy per ideal gas particle is $\langle E \rangle = KT$. Explain why the total energy carried away per effused particle is $2KT$. Compare with $3NKT/2$ to explain why the temperature of the vessel holding an effusing gas decreases so rapidly.

2.4 Maxwell Speed Distribution

Our recent calculations for pressure and effusion are generally applicable and have not assumed the ideal gas model. To gain traction on the ideal gas, we borrow from a more general study of Boltzmann factors to write the *Maxwell(-Boltzmann) speed distribution*:

$$D(v) dv = \left(\frac{m}{2\pi KT} \right)^{3/2} 4\pi v^2 e^{-mv^2/2KT} dv \quad (31.17)$$

The chance of finding an ideal gas particle with speed within the interval $[v_1, v_2]$ equals the area under $D(v)$ curve spanning that interval, or

$$P(v_1 < v < v_2) = \int_{v_1}^{v_2} D(v) dv .$$

The curve ‘opens up’ parabolically near speed $v = 0$, has a finite peak, and dies off exponentially for very large v . $D(v)$ is already normalized such that $P = 1$ for the interval $[0, \infty]$.

Note also there is one dependent variable in $D(v)$, but it is indeed a three-dimensional equation: the $4\pi v^2$ term is byproduct of spherical coordinates in ‘velocity space’.

Average Speed

With $D(v)$ on hand, we use Gaussian integrals to calculate the average speed $\langle v \rangle$ for the ideal gas:

$$v_{\text{ave}} = \langle v \rangle = \int_0^{\infty} v D(v) dv = \sqrt{\frac{8KT}{\pi m}} \quad (31.18)$$

Problem 22

Use Gaussian integrals to prove the following:

$$\begin{aligned} \langle v^2 \rangle &= \frac{3KT}{m} \\ \langle v^3 \rangle &= \frac{4KT \langle v \rangle}{m} \\ \langle v^4 \rangle &= \frac{5KT \langle v^2 \rangle}{m} \end{aligned}$$

Maximum Speed

A single derivative of Equation (31.17) determines the peak v_{max} :

$$v_{\text{max}} = \sqrt{\frac{2KT}{m}} \quad (31.19)$$

RMS Speed

The *root mean square* speed is defined as $\sqrt{\langle v^2 \rangle}$, or

$$v_{\text{rms}} = \sqrt{\frac{3KT}{m}}. \quad (31.20)$$

Writing each result side-by-side, we see:

$$v_{\text{max}} < v_{\text{ave}} < v_{\text{rms}},$$

or

$$\sqrt{\frac{2KT}{m}} < \sqrt{\frac{8KT}{\pi m}} < \sqrt{\frac{3KT}{m}}$$

The numerical coefficients are approximately 1.414, 1.596, and 1.732 respectively.

Problem 23

Ideal gas molecules in a container have the following measured speed distribution:

Speed (m/s)	Percentage
220	10%
250	10%
500	15%
650	30%
900	20%
1300	15%

Calculate the average speed, the rms speed, and the most probable speed. If the system is ideal gas of molecular mass $M = 50 \times 10^{-3}$ kg/mol, what is the temperature? Answer: 690 m/s, 768 m/s, 650 m/s, 1180 K

Problem 24

The escape speed for a particle to leave the gravitational influence of a massive body is given by $(2GM/R)^{1/2}$. The temperature near the top of Jupiter's multicolored cloud layer is about 140 K. The temperature of Earth's atmosphere at 20 km is 220 K. Calculate the RMS speed of H_2 molecules in each of those environments. Give the answer as a fraction of the escape speed from the respective planet. Answer: 14.7% for Earth and 2.16% for Jupiter, explaining the rarity of H_2 in Earth's atmosphere. Jupiter's atmosphere consists of 89% H_2 plus other gases.

Problem 25

Ceres is an asteroid with a mass equal to 0.014 times the mass of the Moon, has density 2400 kg/m^3 , and surface temperature around 200 K. Can this object support an O_2 atmosphere? Answer: $v_{\text{rms}}/v_{\text{esc}} = 72.9\%$ thus O_2 will largely leave the asteroid.

2.5 Equipartition of Energy

Internal Energy

Assuming gas particles do not mutually interact at a distance, the quantity $m \langle v^2 \rangle / 2$ represents the average internal energy per particle in the system. According to Equation (31.14), we find the general result

$$\langle U \rangle = \frac{m \langle v^2 \rangle}{2} = \frac{3PV}{2N},$$

or, scaling by all N particles,

$$U = \frac{3}{2}PV, \quad (31.21)$$

where for an ideal gas, the internal energy resolves to

$$U = \frac{3}{2}NKT.$$

Focusing on the factor of 3 momentarily, recall that it arises from symmetry by generalizing to three dimensions from one. Let us further generalize the idea of 'dimension' to mean 'degree of freedom'. That is, point-like gas particles in a box have three translational degrees of freedom. When molecules are *not* point-like, their motions may include rotation and vibration. These modes of motion (unavailable to points or spheres) make new contributions to the internal energy.

For an ideal gas comprised of non-point-like molecules, we will eventually prove the internal energy obeys

$$U = \frac{f}{2}NKT, \quad (31.22)$$

where f is the number of degrees of freedom available to a given molecule. Note that each degree of freedom must be *quadratic* with respect to its dependent variable. Kinetic energy is $mv^2/2$, and similarly rotational and vibrational energies are $I\omega^2/2$ and $k\Delta x^2/2$, respectively.

Problem 26

Insert the ideal gas Equation (31.12) into the internal energy Equation (31.21) to recover the formula for v_{rms} .

Problem 27

The temperature of 7.00 mol of helium gas, considered ideal, is increased by 2.00 K. What is the change in internal energy?

Equipartition Theorem

Equation (31.22) is a corollary of a more general statement called the *equipartition theorem*, stating:

The internal energy per molecule gains a factor of $KT/2$ for each independent quadratic degree of freedom.

This theorem can be *derived* in full generality using statistical mechanics (also accounting for non-quadratic potentials), but is beyond the scope of this study.

Diatomic Molecules

Consider a generalized ideal gas made of N dumbbell-shaped H_2 molecules. In accordance with (31.22), we count the quadratic degrees of freedom per molecule as follows:

- Three translational modes, so f is at least 3.
- Two degrees of rotational freedom. The diatomic molecule may spin like a baton in two orthogonal modes, however rotations about the axis joining the atoms do not change during collision. So far, $f = 3 + 2$.
- Two contributions from the single vibrational mode (a combination of both kinetic and potential energies), increasing f again by 2.

We expect therefore that the energy of a diatomic gas is

$$U_{\text{dia}} = (7/2) NKT .$$

However, $f = 7$ is only observed at high temperatures. Experiments performed on H_2 gas have shown that only translational modes occur at temperatures under 200 K. Below this, the rotational and vibrational energies are ‘frozen out’ and don’t contribute to the internal energy. By 400 K, the two rotational modes are fully active. Vibrational modes finally show themselves around 4000 K. Evidently, diatomic molecules have f appearing as 3, 5, or 7, depending on T .

Problem 28

Use the equipartition theorem to find the total rotational energy of the molecules in 4.00 mol of diatomic gas at 310 K.

2.6 Heat Capacity

The heat capacity is the factor that couples energy change to temperature change for a given substance (not news):

$$C = \frac{dQ}{dT} \quad (31.23)$$

It is generally wrong to assume that *all* added energy contributes to the kinetic energy of the gas, as some effort is ‘wasted’ on volume expansion of the system. We therefore frame heat capacity calculations somewhere between two extremes: constant volume and constant pressure, where C gains a respective V - or a P -subscript.

Ideal Gas at Constant Volume

Consider a sample of ideal gas inside a sealed, infinitely rigid container of fixed volume V . It follows that all added heat contributes to the kinetic energy, as only the pressure and temperature are variable. Using equations (31.22) and (31.23), we find

$$C_V = \frac{f}{2} N K . \quad (31.24)$$

For one mole of monatomic gas with point-like constituents with $f = 3$ (such as helium or argon), the prediction for C_V evaluates to

$$C_{V_{\text{mon}}} = \frac{3}{2} R = 12.47 \frac{\text{J}}{\text{mol} \cdot \text{K}} ,$$

in very good agreement with experiments. Due to complications that are later solved by quantum mechanics, Equation (31.24) is only accurate for some gases.

Ideal Gas at Constant Pressure

For systems allowed to change volume while maintained at constant pressure, the added energy ΔQ splits according to

$$\Delta Q = \Delta U_{\text{int}} + P\Delta V ,$$

where the increase in translational, rotational, and vibrational motions of the molecules is all contained in ΔU_{int} , and $P\Delta V$ is the energy needed to shove the environment out of the way as the gas expands. Taking a T -derivative of the above yields

$$C_P = C_V + N K , \quad (31.25)$$

where $P\Delta V/\Delta T$ has been replaced using the ideal gas equation.

Gamma Factor

The dimensionless ratio C_P/C_V is called the *gamma factor*, which comes out to

$$\gamma = \frac{C_P}{C_V} = 1 + \frac{2}{f} . \quad (31.26)$$

Problem 29

Methane (CH_4) is a 3-dimensional molecule - a tetrahedron with carbon in the center. Determine the best approximations for C_V and γ . Answer: $24.9 \text{ J mol}^{-1}\text{K}^{-1}$ and 1.33

Elemental Solids

The equipartition theorem allows easy calculation of the heat capacity of many solids. For our purposes, a solid is regarded as a semi-frozen ideal gas, which we take as a periodic arrangement of molecules with no contributions to thermal energy from translational or rotational motions. The structure is held together by quadratic potentials, thus such solids have six degrees of freedom per molecule: three components of kinetic energy due to vibration and three corresponding quadratic potentials.

According to (31.24), one mole of ‘frozen gas’ with $f = 6$ has

$$C_{V\text{solids}} = \frac{6}{2}R = 24.9 \frac{\text{J}}{\text{mol} \cdot \text{K}} ,$$

which is obeyed by *all* elemental solids at high enough temperature, known as the *rule of Dulong and Petit*.

2.7 Van der Waals Model

The *van der Waals* model is an extension of the ideal gas law that accounts for the particle size and the interactions

between particles. To write a generalized ideal gas law, two experimental constants enter the mix - blandly named a and b .

The starting point is the ideal gas model, with equation of state $PV = NKT$. Supposing each of N particles in a sample (gas or liquid) has volume b , we make the replacement

$$V \rightarrow V - Nb ,$$

where the variable V is understood as the volume available for the sample to occupy. The particles themselves do not contribute to V .

Next we assume particles interact through the attractive *Leonnard Jones potential*. Without increasing temperature, particle interactions lead to an increase in number density N/V and pressure P , as evident by considering a particle on the surface layer of the sample. On the surface layer, the net force on a particle is inward, proportional in strength to N/V . The total force inward is proportional to the surface area of V , contributing another factor N/V . It follows that the correction to pressure is

$$P \rightarrow P + a (N/V)^2 ,$$

where a is a constant of dimension $\text{kg m}^5 \text{s}^{-2}$.

The van der Waals equation is

$$\left(P + \frac{aN^2}{V^2} \right) (V - Nb) = NKT , \quad (31.27)$$

and is accurate for many types of gases and fluids. Water vapor has $a = 0.5507 \text{ J m}^3 \text{ mol}^{-2}$ and $b = 3.04 \times 10^{-5} \text{ m}^3/\text{mol}$. In the same units, N_2 gas has $a = 0.1361$ and $b = 3.85 \times 10^{-5}$, and H_2 gas has $a = 0.0247$ and $b = 2.65 \times 10^{-5}$.

3 First Law

3.1 First Law of Thermodynamics

Thermodynamic systems are subject to energy conservation as any other mechanical system would. If U is the internal energy of a substance at pressure P and volume V , any thermal energy dQ added (or removed) must follow

$$dU = dQ - PdV, \quad (31.28)$$

known as the *first law of thermodynamics*.

To make sense of the sign of the PdV term, imagine adding heat $+dQ$ to a gas sample using a candle. If the experiment is prepared such that the temperature does not change, making $dU = 0$, then the gas necessarily expands its volume. The PdV term must appear negative to balance out dQ .

3.2 Thermodynamic Processes

A *thermodynamic process* is any change in the state variables characterizing a thermodynamic system. In many cases, only some variables may change, such as temperature and pressure, while other variables remain fixed, such as volume and number of particles. In the following we classify some of the common thermodynamic processes.

Isothermal Process

An *isothermal* process is one that leaves the system's temperature unchanged, which generally means there is no change in the system's internal energy. The first law with $dU = 0$ reads $0 = dQ - PdV$.

Isochoric Process

A process that leaves the system's volume unchanged (easier said than done) is classified as *isochoric*. The dV term in the first law is zero by construction, so all heat added to the system contributes to internal energy via $dU = dQ - 0$.

Isobaric Process

A process that leaves the system's pressure unchanged (easier said and easily done) is classified as *isobaric*. For example, the macroscopic work done on a gas is not an integral $\int PdV$, but simply the product $P\Delta V$.

Adiabatic Process

Adiabatic processes are those that have zero net heat transfer into or out of the system. The first law of thermodynamics quantifies this by $dU = 0 - PdV$, where any internal energy change dU is balanced out by response in pressure and volume.

Reversible Processes

Many processes in classical mechanics (oscillations, orbits, etc.) are *reversible*, where the system may backtrack along its path in configuration space by reversing time $t \rightarrow -t$ without violating any laws of physics or probabilistic expectations. Only *some* thermodynamic processes are reversible, and only if conditions are right, with one example being isothermal phase change between water and ice at 0°C in a sealed container.

In reality, thermodynamic phenomena are naturally *irreversible*, such as when a bomb explodes, or when scrambling an egg. Witnessing either of these in reverse would be extremely unlikely. To restore such a system to its original state, an external re-organizing force would have to intervene.

3.3 Ideal Gas Processes

We now examine the consequences of thermodynamic processes for a controlled sample of ideal gas, always obeying the ideal gas Equation (31.12).

Isothermal

For isothermal expansion in ideal gas, the total energy entering the gas equals the total work done on the environment, given by the integral

$$W = \int PdV = \int_{V_i}^{V_f} \frac{NKT}{V} dV,$$

simplifying to

$$W = NKT \ln \left(\frac{V_f}{V_i} \right). \quad (31.29)$$

That is, all added heat contributes to volume expansion.

Isochoric

An isochoric process permits no volume change, which we studied when writing the heat capacity of ideal gas at constant volume. The first law tells us $dQ = dU$, where

meanwhile we know $dQ/dT = C_V = fNK/2$. It follows that the work done during an isochoric process is

$$W = C_V \Delta T = \frac{f}{2} NK \Delta T. \quad (31.30)$$

Isobaric

For an ideal gas, we found $C_P = C_V + NK$, leading to

$$W = C_P \Delta T = \left(\frac{f}{2} + 1 \right) NK \Delta T. \quad (31.31)$$

Adiabatic

An adiabatic process usually involves a rapid expansion or compression of a gas that is ‘too fast’ for external heat transfer, making $dQ = 0 = dQ - PdV$. Inserting the ideal gas internal energy relation (31.22) into the first law, we write a differential equation

$$\frac{f}{2} dT = -T \frac{dV}{V}, \quad (31.32)$$

implying:

$$\begin{aligned} PV^\gamma &= \text{constant} \\ TV^{\gamma-1} &= \text{konstant} \end{aligned} \quad (31.33)$$

Problem 30

For a purely kinetic ideal gas, show that $VU^{3/2}$ is constant.

The work entering an ideal gas during an adiabatic process is

$$W = \int P dV = NK \int \frac{T}{V} dV = NK \frac{f}{2} (T_i - T_f),$$

simplifying to

$$W = -\frac{PV}{T} \frac{f}{2} \Delta T. \quad (31.34)$$

Note the term (PV/T) may be evaluated at *any* point during the adiabatic process, so long as the combination of P , V , and T are known simultaneously. Using relation (31.26), we may alternatively write

$$W = \frac{1}{1-\gamma} (V_f P_f - V_i P_i). \quad (31.35)$$

The variable γ is always > 1 by construction, thus (31.35) has a buried negative sign. If the *overall* sign of W is positive, the work is done *by* the gas. If negative, the work is done *on* the gas.

Free Expansion of Ideal Gas

Consider a sealed, perfectly-insulating vessel with ideal gas contained within half of its total volume. The other half of the vessel is vacuum, and a thin membrane separates the two compartments. The membrane is then punctured and the gas is allowed to freely expand into the vacuum until the two compartments are in equilibrium. (The system is not quasistatic during expansion.)

Such free expansion is an adiabatic process, as no heat enters or leaves the vessel through the insulated walls. The system does zero work, as no force is required to expand into vacuum. The temperature change is zero according to Equation (31.34), thus the internal energy change is also zero. In summary, we see

$$Q = W = \Delta T = \Delta U = 0,$$

showing the energy state of the gas is unchanged. Of course, the gas will never accidentally find itself crammed into one side of the box again; work *would* be required to compress the system to its original state. Free expansion of ideal gas is therefore irreversible.

Problem 31

Take a P -derivative of Equation (31.33) to derive

$$\frac{dT}{dP} = \frac{2}{f+2} \frac{T}{P}. \quad (31.36)$$

Problem 32

Consider an ideal gas of N particles with initial temperature T_i and initial volume V_i that is compressed to final volume V_f . Which process requires more energy compress the gas, isothermal or adiabatic?

Problem 33

At height z in the Earth’s atmosphere, (i) show that a horizontal thin layer of air with (massless) volume density $\rho(z)$ and pressure $P(z)$ obeys the equation $dP/dz = -mg\rho(z)$ at mechanical equilibrium, where m is the average mass of molecules in the air. (ii) Consider a bubble of gas that moves upward fast by expanding or shrinking adiabatically. Show that the temperature gradient experienced by the bubble is given by

$$\frac{dT}{dz} = -mg\rho(z) \frac{2}{f+2} \frac{T}{P}.$$

(iii) At low altitudes above sea level, the pressure decreases by about 1.2 kPa for every 100 m. For higher altitudes within the troposphere, show that

$$P(z) = P_0 e^{-mgz/KT_0},$$

where P_0 and T_0 are measured at sea level ($P_0 \approx 101325$ Pa, $T_0 \approx 288.15$ K).

Problem 34

A scuba diver is swimming a depth of 25 m where the pressure is 3.5 atm. The air she exhales forms bubbles 8.0 mm in radius, which rise to the surface where the pressure is 1 atm. Assume the bubbles remain at the uniform 300 K temperature of the surrounding water. How much work is done by each bubble as it expands to the rising surface? Answer: 0.953 J

Problem 35

When a quantity of monatomic ideal gas expands at a constant pressure of 4.00×10^4 Pa, the volume of the gas increases from 2.00×10^{-3} m³ to 8.00×10^{-3} m³. (i) What is the change in internal energy of the gas? (ii) Does heat flow into out of the gas? (iii) What is the magnitude of the heat flow? Answer: 360 J, 600 J into the gas.

3.4 Thermodynamic Cycles

A *thermodynamic cycle* is a chain of processes that bring an open system's state variables to a previous configuration. It follows that the net internal energy change ΔU is zero for a complete thermodynamic cycle, meaning that any heat added to a system must be removed somewhere in the same cycle.

During any infinitesimal volume change, work done is equal to PdV . Integrated over a thermodynamic cycle, the total work done equals the area enclosed by the loop on the PV diagram. If the loop traces out in a clockwise direction, the work output of the system is positive. When going counter-clockwise, the system absorbs work over a complete cycle.

Heat Engine

A *heat engine* is a device that utilizes a thermodynamic cycle in repetition to perform 'useful' work. The *efficiency* of a heat engine is defined as the ratio of output work over input heat energy. Using the first law we write

$$e = \frac{W}{Q_H} = \frac{Q - \Delta U}{Q_H} = \frac{Q}{Q_H}. \quad (31.37)$$

Through a complete thermodynamic cycle, the net heat flow Q equals $Q_H + Q_C$. The efficiency is therefore

$$e = \frac{Q_H + Q_C}{Q_H} = 1 - \left| \frac{Q_C}{Q_H} \right|,$$

using the fact that Q_C is negative for a typical heat engine.

Problem 36

Two moles of an ideal diatomic gas are taken around the cycle abc as shown in Figure 31.4. Data: $P_a = 1.00 \times 10^5$ Pa, $P_c = 1.40 \times 10^5$ Pa, $V_a = 0.049$ m³. If the path from b to c is isothermal, calculate the work done by the gas. Answer: -2300 J

Problem 37

Two moles of an ideal diatomic gas are taken around the cycle abc as shown in Figure 31.4. Data: $P_a = 1.00 \times 10^5$ Pa, $P_c = 1.40 \times 10^5$ Pa, $V_a = 0.100$ m³. If the path from b to c is adiabatic, calculate the volume of the gas at point c . Answer: 0.13 m³

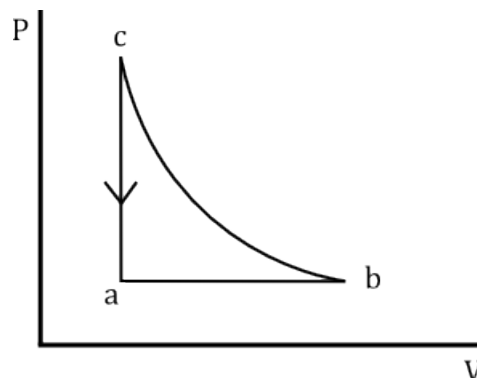


Figure 31.4. PV diagram showing thermodynamic cycle.

Problem 38

A cylinder with a piston contains 0.150 mol of nitrogen at 1.80×10^5 Pa at 300 K. The nitrogen may be treated as an ideal diatomic gas whose molecules can rotate, but not vibrate. The gas is first compressed isobarically to half its original volume. It then expands adiabatically back to its original volume, and is finally heated isochorically to its original pressure. (i) Do the arrows on the corresponding PV diagram point clockwise or counterclockwise? Draw it. (ii) Compute the temperatures at the beginning and end of the adiabatic expansion. (iii) Compute the minimum pressure. (iv) Compute the work done by nitrogen during one thermodynamic cycle. Answer: (i) counterclockwise, (ii) 150 K, 114 K, (iii) 6.82×10^4 Pa, (iv) -75 J

Otto Cycle

Many automobiles operate on the gasoline-burning *Otto cycle*. Inside Otto engines, a fuel-air mixture (approximately ideal gas) is compressed inside a piston and then ignited with a spark, resulting in rapid volume expansion to 'blast' the piston outward, applying work that ultimately reaches the

wheels. The spent fuel-air mixture is ejected from the piston and the process repeats. Engines in general have a different number of steps required to complete their thermodynamic cycles, with popular conventions being ‘two-stroke’ and ‘four-stroke’.

The Otto cycle consists of four thermodynamic processes illustrated in Figure 31.5. We begin analyzing an Otto engine piston as it is fully expanded at volume $V_{\max}$ and filled with a fuel-air mixture (approximately ideal gas), ready to ignite. In this ‘cool’ state, the piston has temperature is T_0 pressure P_0 .

1. Compression stroke (adiabatic): An external force (work left over from a previous cycle) rapidly compresses the piston to volume $V_{\min}$. The pressure of the fuel-air mixture thereby increases to P_1 , and the temperature increases to T_1 . Because no significant heat is added to the cylinder during the compression stroke, it is an adiabatic process.
2. Fuel ignition (isochoric): With the piston fully compressed, a spark ignites the fuel-air mixture, quickly releasing a great deal of kinetic energy Q_H into the gas. The pressure jumps quickly to P_2 , and likewise the temperature to T_2 . The volume remains constant at $V_{\min}$. The energy added follows the relation:

$$Q_H = nC_V (T_2 - T_1) ,$$

where n is the number of moles of fuel-air mixture in the piston.

3. Power stroke (adiabatic): The cylinder’s volume expands to $V_{\max}$ following ignition, and whatever device is attached to the moving part of the piston receives organized work W (driving the car forward, for instance). The pressure drops to P_+ during this process, but not quite down to P_0 . The spent fuel-air mixture is still ‘hot’ at this point, having temperature T_+ , slightly greater than T_0 .
4. Exhaust stroke (isochoric): The cylinder expels the spent fuel-air mixture of energy Q_H to the environment and takes in a new fuel-air packet, returning the pressure to P_0 . The piston effectively remains at $V_{\max}$ (after intake of new fuel), and is ready again for a compression stroke. The (negative) energy added follows the relation

$$Q_C = nC_V (T_0 - T_+) .$$

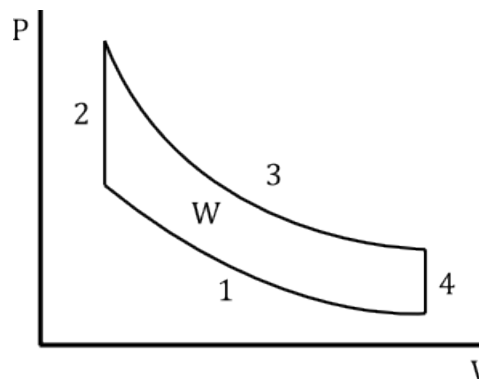


Figure 31.5. Otto cycle.

The efficiency of the Otto engine is straightforward to calculate by Equation (31.37), giving

$$e = 1 + \frac{Q_C}{Q_H} = 1 + \frac{T_0 - T_+}{T_2 - T_1} ,$$

but we can do better. The Otto cycle’s two adiabatic transitions obeys Equation (31.33), allowing all T -variables to cancel. Meanwhile, define the piston *compression ratio* $r = V_{\max}/V_{\min}$ so the formula for e boils down to

$$e = 1 - \frac{1}{r^{\gamma-1}} . \quad (31.38)$$

Problem 39

How much of Q_C may be used to bolster to the power stroke of an engine? Answer: None, not even theoretically.

Problem 40

A gasoline engine takes in 8.00×10^3 J of heat and delivers 1.60×10^3 J of work per cycle. The heat is obtained from gasoline with a heat combustion of 4.6×10^4 J/g. (i) What is thermal efficiency? (ii) How much heat is discarded in each cycle? (iii) What mass of fuel is burned in each cycle? (iv) If the engine goes through 70.0 cycles per second, what is the power output in watts? (v) What is the power in horsepower? Answer: 0.212, 6.70×10^3 J, 1.85×10^{-4} kg, 1.62×10^5 W, 217 hp.

Problem 41

A Toyota Celica GT has a four cylinder Otto-cycle engine with a compression ratio of $r = 9.50$. The diameter of each cylinder, called the *bore* of the engine is 87.1 mm. The distance that the piston moves during the compression, called the *stroke* of the engine is 90.9 mm. The initial (and minimal) pressure of the air - fuel mixture (when the volume is maximal) is 8.50×10^4 Pa, and the initial temperature is 300 K (same as the environment). Assume that 200 J of heat is added to each cylinder in each cycle by burning the fuel (gasoline), which has $C_V = 20.5 \text{ Jmol}^{-1}\text{K}^{-1}$ and $\gamma = 1.40$.

(i) Calculate the work done in each cycle. (ii) Calculate the thermal energy released when the gas is cooled to the temperature of the outside air. (iii) Calculate the minimal volume of the air - fuel mixture. (iv) Calculate P , V , and T just before ignition. Answer: 119 J, 81.3 J, $6.05 \times 10^{-4} \text{ m}^3$, $1.99 \times 10^6 \text{ Pa}$, $6.37 \times 10^{-5} \text{ m}^3$, 738 K

Diesel Cycle

Some automobiles and nearly all road shipping vehicles (excluding hybrid drives) operate on the *diesel cycle*. Like the Otto cycle, the diesel cycle entails four thermodynamic processes. Inside a diesel engine piston however, fuel-air mixture (approximated as ideal gas) is compressed until it self-ignites without the help of a spark, causing the power stroke. The ignition takes place in a constant-pressure condition, and the volume jumps almost instantly, much unlike the Otto cycle's ignition scenario.

Problem 42

Figure 31.6 is the PV diagram for the diesel cycle. The compression stroke, the instant of fuel ignition, power stroke, and exhaust stroke are labeled 1, 2, 3, 4, respectively. Assuming steps 1 and 3 are adiabatic, show that the efficiency of the diesel engine is given by

$$e = 1 - \frac{1}{\gamma} \frac{(c/b)^{-\gamma} - (c/a)^{-\gamma}}{(c/b)^{-1} - (c/a)^{-1}},$$

where volumes a , b , and c are indicated on the horizontal axis and γ is the ratio C_P/C_V .

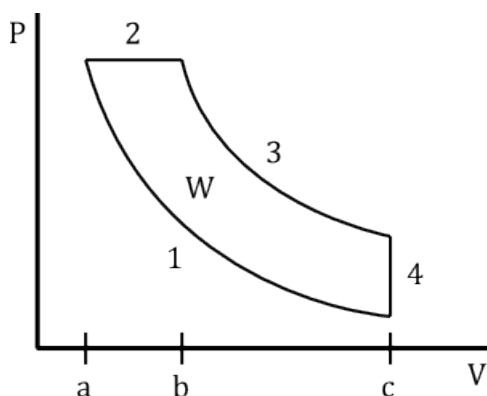


Figure 31.6. Diesel cycle.

Refrigerators

A refrigerator operates as a heat engine in reverse: work is required to displace heat from the cold reservoir to the hot

reservoir. The working substance that circulates inside a refrigerator is a fixed amount of water, freon, or ammonia (no chemical exhaust). The operation is outlined as follows:

1. Starting in the gaseous phase, the working substance is forced through a compressor, emerging with reduced volume and increased temperature and pressure.
2. The substance goes through a long heat-conducting tube called the *condenser*, which is in thermal contact with the environment (hence the refrigerator needs to ventilate). In the condenser, the working substance loses heat $Q_H < 0$ and changes phase to liquid.
3. The working substance then encounters a very small 'bottleneck' called the *throttle* that hinders most of the circulation. On the input side of the throttle, the temperature and pressure of the working substance are relatively high. The output side is at much lower temperature and pressure.
4. As a cold liquid, the working substance passes through the refrigerator body, warming again to the gaseous phase, absorbing energy $Q_C > 0$.

Like the heat engine, the refrigerator obeys the first law

$$Q_H + Q_C = Q_{in} = \Delta U + W,$$

where $\Delta U = 0$ and $W < 0$. The merit of a refrigerator is not its efficiency, but instead the *coefficient of performance*, defined as the ratio of heat removed from the refrigerator body over the work required to do so. That is,

$$K = \frac{|Q_C|}{|W|} = \frac{|Q_C|}{|Q_H| - |Q_C|}. \quad (31.39)$$

Problem 43

An ideal air conditioner takes heat from a room at 68.5°F and transfers it to the environment, which is at 93.1°F . For each joule of electrical energy required to operate the air conditioner, how many joules of thermal energy are removed from the room?

Problem 44

Liquid refrigerant at a pressure of $1.34 \times 10^5 \text{ Pa}$ leaves the expansion valve of a refrigerator at -22°C . It then flows through the vaporization coils inside the refrigerator and leaves as vapor at the same pressure at a temperature of -18°C , the same temperature as the inside of the refrigerator. The boiling point of the refrigerant at this pressure is -18°C , the heat of vaporization is $1.60 \times 10^5 \text{ J/kg}$, and the specific heat capacity of the vapor is $485 \text{ J/kg}^\circ\text{K}$. The coefficient of performance of the refrigerator is $K = 2.52$. If 8.00 kg

of refrigerant flows through the refrigerator each hour, find the electric power that must be supplied to the refrigerator. Answer: 123 W

Problem 45

A heat pump designed for southern climates extracts heat from the outside air, and delivers air at 20°C to the inside of the house. Compute the coefficient of performance of the heat pump (i) in the south, where the outside temperature is 5°C; (ii) in the north, where the outside temperature is -10°C. (iii) Two identical houses, one in the north and one in the south, are heated by this pump, and maintain temperatures of 20°C. Considering heat loss through the walls, windows, and roof, what is the ratio of the electrical power required to heat the two houses? Answer: (i) 18.5, (ii) 8.77, (iii) 4.0

Carnot Cycle

Sadi Carnot (1796-1832) designed an engine with maximal efficiency by using only reversible isothermal and reversible adiabatic processes. All heat transfer during the Carnot cycle occurs on isotherms; there is no heat transfer across finite temperature differences. No machine using the Carnot cycle would be 'fast' enough for industrial standards, but Carnot engines are still the most efficient type.

The gaseous working substance within a Carnot engine piston never leaves the piston, and undergoes no chemical change. Carnot engines rely on a literal 'hot' reservoir from which to draw energy, and a similar 'cold' reservoir to deposit energy. These two reservoirs may in practice be as simple as hot and cold containers of water. The four stages of the Carnot cycle go as follows:

1. Slow isothermal: A piston in the 'compressed' state enclosing a gas of volume $V_{\min}$ is attached to a reservoir at high temperature T_H and absorbs heat Q_H isothermally. The volume swells to V_1 , and the reservoir is removed. Recall from Equation (31.29) that

$$Q_H = NKT_H \ln \left(\frac{V_1}{V_{\min}} \right).$$

2. Fast adiabatic: The gas expands adiabatically, lowering its temperature from T_H until it reaches T_C . The volume swells once more to reach a maximum $V_{\max}$.
3. Slow isothermal: The system moves to contact the cold reservoir of temperature T_C , where isothermal

compression takes the volume to V_2 (with $V_2 < V_1$). The discarded energy is

$$Q_C = -NKT_C \ln \left(\frac{V_{\max}}{V_2} \right).$$

4. Fast adiabatic: The system moves away from the cold reservoir and the gas is compressed adiabatically, raising its temperature from T_C until it reaches T_H . The volume returns to $V_{\min}$.

The efficiency of the Carnot engine, according to definition (31.37), is

$$e = \frac{Q_H + Q_C}{Q_H} = 1 - \frac{T_C \ln(V_{\max}/V_2)}{T_H \ln(V_1/V_{\min})},$$

where Equation (31.33) for adiabatic processes tells us $V_{\max}/V_2 = V_1/V_{\min}$, and thus

$$\frac{Q_C}{Q_H} = -\frac{T_C}{T_H}. \quad (31.40)$$

Astonishingly, the efficiency depends on neither the construction of an engine nor the properties of the working substance, only the reservoir temperatures:

$$e = 1 - \frac{T_C}{T_H} \quad (31.41)$$

Problem 46

Consider a Carnot engine operating between temperatures T_H and T_C , where T_C is above the ambient temperature T_0 . A second engine operates between the temperatures T_C and T_0 . Show that the overall efficiency of this system is equal to the efficiency of a single Carnot engine operating between T_H and T_0 .

Problem 47

A Carnot engine operating between two thermal reservoirs has an efficiency of e . When it is run in reverse, it becomes a Carnot refrigerator. Calculate the coefficient of performance. Answer: $(1 - e)/e$

Reversible Loops

By riding isothermal and adiabatic curves, Carnot successfully identified a closed loop in the PV diagram that is entirely reversible. It should follow that we may consider any reversible path as a stack of Carnot cycles. To capture this idea, write Equation (31.40) in the form

$$\frac{Q_C}{T_C} + \frac{Q_H}{T_H} = 0$$

where a path made of many isothermal and adiabatic segments must obey

$$\sum \frac{Q}{T} = 0.$$

Heat transfer on shared isotherms cancel, and the sum refers only to the boundary of the PV -curve. Taking a finer

resolution, the sum generalizes to a path integral

$$\oint \frac{dQ}{T} = 0. \quad (31.42)$$

Evidently, there is *something* that links quantity dQ/T to the reversibility of thermodynamic processes.

4 Second Law

4.1 Entropy

The *entropy* of system is a quantity representing its ‘reversibility’. First surfacing in Equation (31.42), let us define the entropy as

$$S = \int \frac{dQ}{T}. \quad (31.43)$$

For reversible processes, the entropy can be found by direct integration of the above. For two states linked by an irreversible process, it suffices to calculate the entropy over some other path (in the PV diagram) consisting of reversible process that would connect those two states.

Adding to our list of terms for thermodynamic processes, any process that is both adiabatic and quasistatic is called *isentropic*, which means there is no entropy change in the process.

Problem 48

When 1.00 kg of water at 0°C is frozen to make ice, what is the entropy change? Answer: -1200 J/K

Problem 49

An object of mass m_1 , specific heat c_1 , and temperature T_1 is placed in contact with a second object mass m_2 , specific heat c_2 , and temperature $T_2 > T_1$. As a result, object 1 heats to temperature T and object 2 cools to T' . (i) Show that energy conservation requires that $m_1 c_1 (T - T_1) = m_2 c_2 (T_2 - T')$. (ii) Show that the entropy increase of the system is

$$\Delta S = m_1 c_1 \ln(T/T_1) + m_2 c_2 \ln(T'/T_2).$$

Isothermal Expansion of Ideal Gas

For isothermal expansion of ideal gas, we wrote the energy entering the system as Equation (31.29), implying $\Delta Q = NKT \ln(V_f/V_i)$. By the definition of entropy, it follows that

$$S = NK \ln\left(\frac{V_f}{V_i}\right). \quad (31.44)$$

Isochoric Expansion of Ideal Gas

For isochoric expansion of ideal gas, we wrote the energy entering the system as Equation (31.30), implying $dQ/dT = C_V$. By the definition of entropy, it follows that

$$S = \int \frac{dQ}{dT} \frac{dT}{T} = C_V \ln\left(\frac{T_f}{T_i}\right). \quad (31.45)$$

Isobaric Expansion of Ideal Gas

For isobaric expansion of ideal gas, we wrote the energy entering the system as Equation (31.31), implying $dQ/dT = C_P$. By the definition of entropy, it follows that

$$S = \int \frac{dQ}{dT} \frac{dT}{T} = C_P \ln\left(\frac{T_f}{T_i}\right). \quad (31.46)$$

Adiabatic Process in Ideal Gas

Starting with the definition of entropy, an adiabatic process in ideal gas obeys

$$\begin{aligned} S &= \int \frac{dQ}{T} = \int \frac{dU + PdV}{T} \\ &= NK \int \left(\frac{f}{2} \frac{dT}{T} + \frac{dV}{V} \right) = 0, \end{aligned}$$

which evaluates to zero because the terms inside the integral are equal and opposite by Equation (31.32).

Free Expansion of Ideal Gas

In discussing the free expansion of ideal gas, we found that no energy changes whatsoever, i.e., $\Delta U = \Delta Q = W = 0$. However, we also found that ideal gas expansion is irreversible, so the entropy shouldn't be zero.

To calculate the entropy of free expansion of ideal gas, imagine the effort required to isothermally compress the gas to its original volume. This is precisely the same condition that gave us Equation (31.44), thus the entropy is

$$S = NK \ln\left(\frac{V_f}{V_i}\right).$$

A change in entropy implies a change in the ‘quality’ of the energy in a system. This means ideal gas freely-expanding through a hole, as opposed to escaping into a vacuum, could instead be used to do something that keeps energy ‘organized’, such as to wind up a spring connected to a turbine.

Entropy of Light

A ‘gas’ of photons, also referred to as blackbody radiation, is subject to entropy. Recall that Stefan’s law (31.11) states that the rate of electromagnetic energy transfer is given by

$$\frac{dQ}{dt} = A\sigma (T_{\text{obj}}^4 - T_{\text{env}}^4),$$

where $e = 1$ and for simplicity we'll take T_{env} to be constant. The total energy in a volume V of photon gas is

$$Q = \frac{\sigma}{c} (T_{\text{obj}}^4 - T_{\text{env}}^4) \int d^3x = \frac{V\sigma}{c} (T_{\text{obj}}^4 - T_{\text{env}}^4),$$

and it follows that $dQ = (V\sigma/c)4T_{\text{obj}}^3 dT$. Thus the entropy S as a function of T_{obj} is

$$S = \int \frac{dQ}{T} = \frac{4V\sigma}{c} \int T^2 dT = \frac{4V\sigma}{3c} (T_{\text{obj}}^3 - T_0^3).$$

If we approximate $T_{\text{env}} = 0$ and $T_0 = 0$, the entropy becomes

$$S = \frac{4}{3} \frac{Q}{T}.$$

4.2 Second Law of Thermodynamics

The *second law of thermodynamics* is a restriction on the entropy of closed systems. The following statements more-or-less equally capture the second law:

Entropy statement

The entropy of a closed system can never decrease. The entropy of an open system can be forced to decrease, but the net entropy of the system + surroundings necessarily increases.

Engine statement

It is impossible for any system to undergo a cyclic process in which it absorbs heat at a single temperature and converts the heat completely into mechanical work.

Refrigerator statement

It is impossible for any process to have as its sole result the transfer of heat from a cooler to a hotter body.

4.3 Systems in Contact

Systems placed in contact will exchange energy, temperature, particles, and so on - depending on the intimacy of contact. Such conjoined systems always tend toward some kind of equilibrium state, and we take advantage of this to re-understand the notion temperature, pressure, etc.

Thermal Equilibrium

Consider two thermodynamic systems A and B that are in separate, rigid, sealed containers. With $T_A > T_B$ initially, the systems are then placed in contact to allow internal energy exchange by conduction or radiation, with the total $U = U_A + U_B$ remaining constant.

A heat packet $dQ = -dU_A = dU_B$ exchanged between systems changes the entropy of both:

$$dS_A = -\frac{|dQ|}{T_A}$$

$$dS_B = \frac{|dQ|}{T_B}$$

To accommodate the temperature terms, the entropy loss of the hotter system A is *smaller* in magnitude than the entropy gain of the cooler system B . The sum $dS_A + dS_B$ is necessarily positive, thus the energy exchange is irreversible. *Thermal equilibrium* is achieved when the entropy stops increasing, corresponding to $T_A = T_B = T$, or

$$\frac{1}{T} = \frac{dS_A}{dU_A} = \frac{dS_B}{dU_B} = \frac{1}{T}.$$

Definition of Temperature

Given the the role of entropy in thermal equilibrium, we seize this opportunity to actually *define* the notion of temperature as an entropy maximum with respect to internal energy (for fixed V and N):

$$T = \left(\frac{\partial S}{\partial U} \right)_{V,N}^{-1} \quad (31.47)$$

Mechanical Equilibrium

Supposing the partition between systems A and B were made flexible but not porous, the systems tend toward *mechanical equilibrium*, where maximum entropy corresponds to equal pressures. Supposing $T_A = T_B$ with $V_A > V_B$ initially, the heat packet exchanged between systems is purely mechanical as $dQ = -P_A dV_A = P_B dV_B$.

Using the same arguments as above, we may replace instances of dQ with PdV terms to get:

$$\frac{P}{T} = \frac{dS_A}{dV_A} = \frac{dS_B}{dV_B} = \frac{P}{T}$$

Alas, we're in position to relate pressure to entropy as we did the temperature. The equation

$$P = T \left(\frac{\partial S}{\partial V} \right)_{U,N} \quad (31.48)$$

describes systems with fixed internal energy and fixed number of particles. Note that the above is merely a tool for calculation, not a refined definition of pressure.

Diffusive Equilibrium

We now let systems A and B exchange both energy and particles through a non-flexing porous membrane. Placed in contact, the systems achieve *diffusive equilibrium* in addition to thermal equilibrium.

The heat dQ exchanged between systems is proportional to the number of particles dN that move across the membrane. The proportionality factor between energy and particles is called the *chemical potential*, denoted μ , as

$$dQ = \mu dN .$$

implying

$$\frac{-\mu}{T} = \frac{dS_A}{dN_A} = \frac{dS_B}{dN_B} = \frac{-\mu}{T} .$$

We thus define the *chemical potential* in terms of the entropy:

$$\mu = -T \left(\frac{\partial S}{\partial N} \right)_{U,V} \quad (31.49)$$

The minus sign in Equation (31.49) reminds us that particles flow from regions of higher μ to lower. When there are multiple particle species present in a system, μ adopts a subscript j for each species, and the quantity μdN becomes $\sum \mu_j dN_j$.

5 Thermodynamic Variables

5.1 Extensive and Intensive

Most thermodynamic variables belong to one of two classifications. A quantity may be *extensive*, where the scale of the variable goes with the size of the system, or it may be *intensive*, where the scale of the variable doesn't care about the scale of the system. To illustrate, consider a thermodynamic system with the laundry list of state variables. If the system is duplicated and attached to itself, we ask which variables scale appropriately, versus which variables stay the same? You should conclude:

Extensive: M, V, N, U, S

Intensive: P, T, μ, ρ

Properties of intensive and extensive variables:

- An extensive quantity multiplied by an intensive quantity yields an extensive quantity.
- The sum of two extensive quantities is also extensive.
- A ratio of extensive quantities yields an intensive quantity.
- The product of two extensive quantities is rare in calculations - check twice if you encounter this.
- The sum of an extensive quantity and an intensive quantity should never arise.

5.2 First Thermodynamic Identity

Recall the first law of thermodynamics as stated in Equation (31.28), namely

$$dU = dQ - PdV .$$

Our study of thermal and diffusive equilibrium has shown dQ coming in two flavors as $dQ = TdS$ and $dQ = \mu dN$, respectively. It follows that the first law has a special form

$$dU = TdS - PdV + \mu dN , \quad (31.50)$$

which we'll call the *first thermodynamic identity*.

Entropy of the Universe

The whole universe consists of any given system plus its total environment such that $S_{\text{uni}} = S_{\text{sys}} + S_{\text{env}}$. It follows that Equation (31.50) lets us write:

$$dS_{\text{uni}} = dS_{\text{sys}} + \left(\frac{1}{T} (dU + PdV - \mu dN) \right)_{\text{env}} \quad (31.51)$$

5.3 Free Energies

The meaning of the term *thermal energy* or *thermodynamic potential* can vary depending on the conditions surrounding a system. The distinction is characterized by the *free energy* of the system.

Enthalpy

In an environment with fixed pressure P , the total energy required to create a system with internal energy U and volume V is called the *enthalpy*, defined as

$$H = U + PV . \quad (31.52)$$

Enthalpy is the internal energy of a system plus the work required to shove the environment out of the way in order to make room for the system. Conversely, H is the total energy recovered if you could annihilate the system.

Problem 50

Show that the enthalpy of an ideal gas is $H = 5NKT/2$.

The differential enthalpy is $dH = dU + PdV + VdP$, where substituting Equation (31.50) gives a new identity

$$dH = TdS + VdP + \mu dN . \quad (31.53)$$

The enthalpy is generally not extremized as the system tends toward equilibrium. In the special case that P and S are constant, the enthalpy is minimized at equilibrium.

Helmholtz Free Energy

In an environment with fixed temperature T , one who creates a system with constant V and constant N shall borrow energy from the environment by an amount equal to TS , where S is the entropy of the system. The net work required is therefore

$$A = U - TS , \quad (31.54)$$

called the *Helmholtz Free Energy*. This is the energy recovered if the system is destroyed in the same conditions.

The differential Helmholtz free energy is $dA = dU - TdS - SdT$, where substituting Equation (31.50) gives another identity

$$dA = -SdT - PdV + \mu dN . \quad (31.55)$$

By assuming the system has fixed T, V, N , we assume the same about the environment. In accordance with Equation (31.51), we have that a such a system plus its environment will follow

$$dS_{\text{uni}} = dS_{\text{sys}} + \left(\frac{1}{T} (dU + PdV - \mu dN) \right)_{\text{env}} .$$

The energy exchanged between the two are equal and opposite, i.e. $dU_{\text{sys}} = -dU_{\text{env}}$. All in terms of system variables, the above becomes

$$dS_{\text{uni}} = \frac{-1}{T} (dU - TdS) .$$

On the right side of the above, the parenthesized expression is the differential Helmholtz free energy for constant T , i.e.

$$dS_{\text{uni}} = \frac{-1}{T} (dA)_{T,V,N} .$$

Thus, we find that such a system will do whatever it can to minimize the Helmholtz free energy.

Gibbs Free Energy

In an environment with fixed temperature T and pressure P , creating a system requires the *Gibbs free energy*

$$G = U + PV - TS , \quad (31.56)$$

which means you supply the internal energy U plus the work PV to shove the environment out of the way. The environment supplies TS for you, hence the minus sign. In the same conditions, G is the energy recovered if the system is annihilated.

Taking the differential Gibbs free energy and substituting (31.50) gives a corresponding identity:

$$dG = -SdT + VdP + \mu dN \quad (31.57)$$

At constant temperature and pressure, the dT - and dP -terms are zero, leaving us with $dG = \mu dN$, which integrates to

$$G = \mu N . \quad (31.58)$$

Notice this readily generalizes to a gas with multiple species of particles:

$$G = \sum_j \mu_j N_j$$

You might wish that a similar relation connects A to N by starting with $\mu = (\partial A / \partial N)_{T,V}$. However, adding particles while maintaining fixed T and V causes the density to slowly increase, which directly changes μ .

Assume the system and the environment have fixed T, P, N . In accordance with Equation (31.51), we have that a such a system plus its environment will follow

$$dS_{\text{uni}} = dS_{\text{sys}} + \left(\frac{1}{T} (dU + PdV - \mu dN) \right)_{\text{env}} .$$

The energy and volume exchanged between the two are equal and opposite, i.e. $dU_{\text{sys}} = -dU_{\text{env}}$, $dV_{\text{sys}} = -dV_{\text{env}}$. All in terms of system variables, the above becomes

$$dS_{\text{uni}} = \frac{-1}{T} (dU + PdV - TdS) .$$

On the right side of the above, the parenthesized expression is the differential Gibbs free energy for constant T and P , i.e.

$$dS_{\text{uni}} = \frac{-1}{T} (dG)_{T,P,N} .$$

Such a system will do whatever it can to minimize the Gibbs free energy. Note if T is allowed to vary, the system instead minimizes the enthalpy.

Summary

The above is summarized as follows:

- Fixed U, V :
Maximize entropy (S).
- Fixed P :
Minimize enthalpy ($H = U + PV$).
- Fixed T, V :
Minimize Helmholtz ($A = U - TS$).
- Fixed T, P :
Minimize Gibbs ($G = U + PV - TS$).

Grand Free Energy

A grouping of terms useful for describing irreversible processes in open systems is the *grand free energy*, defined as

$$\Phi = U - TS - \mu N . \quad (31.59)$$

The quantity (31.59) is the energy recovered by reducing the system to zero size; putting all particles and energy back into the reservoir. Using Equation (31.58), it's straightforwardly shown that

$$\Phi = -PV . \quad (31.60)$$

Suppose a system of fixed V is allowed to exchange both U and N with its environment. Assume the chemical potential μ of the system is already in equilibrium with

the environment, as is the temperature T . The differential version of Φ reads

$$d\Phi = dU - TdS - \mu dN ,$$

which is the right hand side of Equation (31.51). It immediately follows that

$$dS_{\text{uni}} = -\frac{1}{T} (d\Phi)_{T,V,\mu} ,$$

meaning a system with fixed V in thermal and diffusive equilibrium with the environment will minimize the grand free energy.

Energy Checksum

A handy consistency check on the internal energy U can be derived in terms of TS , PV and μN , avoiding terms like H , A , G , and Φ . Start with the first thermodynamic identity

$$dU = TdS - PdV + \mu dN ,$$

and ask: which potential could be added to dU in order to make the right side a total derivative? Denoting such a term Ξ , we write

$$d(U + \Xi) = d(TS) - d(PV) + d(\mu N) ,$$

where clearly,

$$d\Xi = SdT - VdP + d\mu N .$$

Notice though that this is (minus) the differential version of $\Phi + PV$, which is identically zero, meaning that Ξ is exactly zero. Integrating the above gives

$$U = TS - PV + \mu N ,$$

the familiar formula for the Gibbs free energy.

Gibbs-Duhem Equation

The arrangement of thermodynamic state variables that gives zero on either side of the equation gives a special result relating differential chemical potentials. Begin with the Grand free energy as given by (31.59), and add the PV term onto the left to get zero on the right:

$$\Phi + PV = U - TS - \sum_i \mu_i N_i + PV = 0$$

Note this discussion applies to a multi-species system, as indicated by the sum over particles with their corresponding chemical potentials. Taking the differential version, we find

$$d(\Phi + PV) = -SdT + VdP - \sum_i d\mu_i N_i = 0 ,$$

where for conditions of constant temperature and constant pressure, we arrive at the *Gibbs-Duhem* equation:

$$\sum_i d\mu_i N_i = 0 \quad (31.61)$$

5.4 Legendre Transform

We have found that the energy state of a system is beholden to its (sometimes controlled) relationship with the environment. For example, at fixed temperature, volume, and number of particles, the best representation is the Helmholtz free energy $A(T, V, N)$. Here we address *why* to ever bother with different potentials - that is, why not simply write $U = U(T, V, N)$ and deal with internal energy all the time?

Staying in the Helmholtz regime, to assign U as a function of T is patently incorrect, as temperature is already defined by Equation (31.47), namely

$$T = \left(\frac{\partial S}{\partial U} \right)^{-1}_{V,N} .$$

It is appropriate however to express the internal energy in terms of the entropy instead as $U = U(S, V, N)$. To proceed, consider a given point P on a $U(S)$ curve, and determine the slope $\partial U / \partial S = T$ at that point. Draw a straight line through point P having slope T , and extend the line until it hits the U -axis. The U -intercept is precisely the Helmholtz free energy A . That is, we have drawn the line

$$U(S, V, N) = A(T, V, N) + S \left(\frac{\partial U}{\partial S} \right)_{V,N} ,$$

which is equivalent to the usual formula $A = U - TS$. Subtracting TS from U in this way is called the *Legendre transform*. A similar procedure extends to each of the thermodynamic potentials.

5.5 State Variables as Derivatives

Thermodynamic identities (31.50), (31.53), (31.55), and (31.57) can be arranged to express certain state variables. For completeness, these are:

$$\begin{aligned} dU &= TdS - PdV + \mu dN \\ dH &= TdS + VdP + \mu dN \\ dA &= -SdT - PdV + \mu dN \\ dG &= -SdT + VdP + \mu dN \end{aligned}$$

From these, one readily derives the following:

Derivatives of Entropy

$$\begin{aligned}\frac{1}{T} &= \left(\frac{\partial S}{\partial U} \right)_{V,N} \\ \frac{P}{T} &= \left(\frac{\partial S}{\partial V} \right)_{U,N} \\ -\frac{\mu}{T} &= \left(\frac{\partial S}{\partial N} \right)_{U,V}\end{aligned}$$

Derivatives of Helmholtz

$$\begin{aligned}-S &= \left(\frac{\partial A}{\partial T} \right)_{V,N} \\ -P &= \left(\frac{\partial A}{\partial V} \right)_{T,N} \\ \mu &= \left(\frac{\partial A}{\partial N} \right)_{T,V}\end{aligned}$$

Derivatives of Gibbs

$$\begin{aligned}-S &= \left(\frac{\partial G}{\partial T} \right)_{P,N} \\ V &= \left(\frac{\partial G}{\partial P} \right)_{T,N} \\ \mu &= \left(\frac{\partial G}{\partial N} \right)_{T,P}\end{aligned}$$

Heat Capacity

$$\begin{aligned}C_V &= \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{3}{2} NK \\ C_P &= \left(\frac{\partial H}{\partial T} \right)_{P,N} = \frac{5}{2} NK\end{aligned}$$

Thermal Expansion

As a partial derivative, the coefficient of thermal expansion, namely $\beta = 3\alpha$ from Equation (31.3) may be written as

$$\beta = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P .$$

A similar property of matter is the isothermal compressibility κ_T , defined as

$$\kappa_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T .$$

Problem 51

Assume some equation of state (not necessarily the ideal gas equation) connects a system's pressure P , temperature T , and volume V . For the infinitesimal change $P \rightarrow P + dP$, $T \rightarrow T + dT$ with V fixed, (i) show that $(\partial V / \partial P)_T dP + (\partial V / \partial T)_P dT = 0$. (ii) Next show that

$$\frac{dP}{dT} = \left(\frac{\partial P}{\partial T} \right)_V = -\frac{(\partial V / \partial T)_P}{(\partial V / \partial P)_T} .$$

Problem 52

Near room temperature, liquid water has a thermal expansion coefficient $\beta = 2.5 \times 10^{-4} \text{ K}^{-1}$ and isothermal compressibility $\kappa_T = 4.5 \times 10^{-10} \text{ Pa}^{-1}$. Estimate the pressure that must be exerted on a drop of water in order to prevent it from expanding while increasing from 20°C to 30°C . Use this result to explain why it is easier to measure C_P rather than C_V . Answer: 55.8 atm

Problem 53

For a van der Waals fluid obeying Equation (31.27) at fixed T and N , use the thermodynamic identity (31.57) to show that the Gibbs free energy is:

$$\begin{aligned}G &= -NKT \ln(V - Nb) \\ &\quad + \frac{N^2 b K T}{V - Nb} - \frac{2aN^2}{V} + C(T)\end{aligned}\quad (31.62)$$

5.6 Maxwell Relations

A trick from calculus allows a deeper relationship between derivatives and state variables. Consider a function f of two variables x and y . It's easy to see that the mixed second derivatives of f are identical: $\partial_{xy} f = \partial_{yx} f$. Applying this to thermodynamics, Take for example the differential Helmholtz free energy, dA , with a fixed number of particles. By Equation (31.55), we have

$$dA = -SdT - PdV .$$

From the rules of partial derivatives, dA also reads

$$dA = \left(\frac{\partial A}{\partial T} \right)_V dT + \left(\frac{\partial A}{\partial V} \right)_T dV ,$$

where from calculus, we must have

$$\left(\frac{\partial}{\partial V} \left(\frac{\partial A}{\partial T} \right)_V \right)_T = \left(\frac{\partial}{\partial T} \left(\frac{\partial A}{\partial V} \right)_T \right)_V$$

Substituting state variables for the inner derivatives, but leaving the outer ones, we arrive at a *Maxwell relation*:

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V \quad (31.63)$$

In the same spirit, we can do the same for dU , dH , and dG , where the number of particles is fixed in all cases. Respectively, the results are:

$$\left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial P}{\partial S}\right)_V \quad (31.64)$$

$$\left(\frac{\partial T}{\partial P}\right)_S = \left(\frac{\partial V}{\partial S}\right)_P \quad (31.65)$$

$$\left(\frac{\partial S}{\partial P}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_P \quad (31.66)$$

6 Chemical Physics

6.1 Thermochemistry Tables

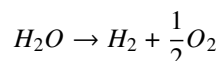
In response to change, a thermodynamic system's state variables will adjust in accordance with environmental con-

ditions. At 'standard' room temperature (298 K) and at 'standard' atmospheric pressure (1 atm), also known as *STP* conditions, the enthalpy H , entropy S , and Gibbs free energy G of common substances behave as listed in the following *thermochemistry table*, also called an entropy table:

Symbol	Species	Enthalpy ΔH [kJ mol ⁻¹]	Entropy S [J mol ⁻¹ K ⁻¹]	Gibbs free energy ΔG [kJ mol ⁻¹]
H_2O	Liquid Water	-285.83	69.95	-237.15
H_2O	Water Vapor	-241.83	188.84	-228.59
H_2	Hydrogen Gas	0	130.7	0
O_2	Oxygen Gas	0	205.07	0
O	Monatomic Oxygen	249.170	161.055	231.731
N	Nitrogen Gas	472.704	153.298	455.563
NH_3	Ammonia	-45.90	192.77	-16.37

Electrolysis of Water

Consider a 1 mol sample of liquid water at standard temperature and pressure conditions. When electric current is applied through the sample, the reaction



takes place. Using the data above, we may calculate the energy required to drive the reaction.

Rummaging through the various energy quantities studied previously, observe that the enthalpy $H = U + PV$ matches the scenario. From thermochemistry data, we find that the enthalpy decreases by $\Delta H = -286$ kJ. (Note: of the 286 kJ that enters the system, an amount of work equal to $1 \text{ atm} \times \Delta V = 4$ kJ is spent on shoving aside the atmosphere to make room for the new gases, with 282 kJ remaining.)

The entropy of the reaction is

$$\begin{aligned}\Delta S &= S_{H_2 + \frac{1}{2}O_2} - S_{H_2O} \\ &= \left(131 + \frac{205}{2} - 70\right) \text{ J K}^{-1} \\ &= 163 \text{ J K}^{-1},\end{aligned}$$

thus the environment contributes (positive) energy

$$T\Delta S = (298 \text{ K}) (163 \text{ J K}^{-1}) = 49 \text{ kJ}.$$

Finally, the change in Gibbs free energy $G = H - TS$ works out as $\Delta G = \Delta H - T\Delta S = -286 \text{ kJ} + 49 \text{ kJ} = -237 \text{ kJ}$, which 'predicts' the top-right value in the thermochemistry table. Evidently, the device driving electrolysis must only provide 237 kJ per mole of water.

6.2 Phase Change

At a boundary on a PT diagram, such as the gas-liquid boundary where the substance is stable in either form, the Gibbs free energy is the same on each side of the boundary:

$$G_{\text{gas}} = G_{\text{liquid}}$$

Using Equation (31.57) and noting that $dG_{\text{gas}} = dG_{\text{liquid}}$, we write for each side of the boundary (assuming N constant):

$$(-SdT + VdP)_{\text{gas}} = (-SdT + VdP)_{\text{liquid}}$$

Solving for dP/dT , one finds

$$\frac{dP}{dT} = \frac{S_{\text{gas}} - S_{\text{liquid}}}{V_{\text{gas}} - V_{\text{liquid}}} = \frac{\Delta S}{\Delta V}.$$

Next, recall that the latent heat of phase transformation at constant temperature is written L , which we interpret as equal to $T\Delta S$. From this we write

$$\frac{dP}{dT} = \frac{L}{T\Delta V},$$

also known as the *Clausius-Clapeyron relation*.

Vapor Pressure

In the case that L is constant, and also that $V_{\text{gas}} \gg V_{\text{liquid}}$ such that $\Delta V \approx V_{\text{gas}} = V$, we further assume that the evaporated gas is ideal gas. In this case, replace V to write

$$\frac{dP}{P} = \frac{L}{NK} \frac{dT}{T^2},$$

solved by

$$P = \psi e^{-L/NKT},$$

where ψ is a constant. Often, we set N to correspond to one mole of substance, and the final answer is written

$$P = \psi e^{-L/RT},$$

known as the *vapor pressure* equation.

6.3 Entropy of Mixing

Consider two systems such gases, equally-dense liquids, certain solids, or any other *ideal mixtures* A and B that are in thermal contact but in diffusive isolation by a partition. When the partition is removed, the change in entropy is strictly due to mixing, thus the analysis is highly analogous to ideal gas expansion. For each system respectively, start from Equation (31.44) to write

$$\begin{aligned}\Delta S_A &= N_A K \ln \left(1 + \frac{V_B}{V_A} \right) \\ \Delta S_B &= N_B K \ln \left(1 + \frac{V_A}{V_B} \right),\end{aligned}$$

and let y equal the volume and number ratio of species B to species A .

The sum $\Delta S_A + \Delta S_B$ is the *entropy of mixing* formula:

$$\Delta S_{\text{mix}} = N_A K [(1-y) \ln(1-y) - y \ln y] \quad (31.67)$$

Letting χ_α equal the ratio the volume and number ratio of either species over the total, the above formula is equivalent to:

$$\Delta S_{\text{mix}} = -(N_A + N_B) K (\chi_A \ln \chi_A + \chi_B \ln \chi_B) \quad (31.68)$$

Problem 54

Show that the entropy of mixing for two equal portions $N = N_A = N_B$ of ideal gas resolves to

$$\Delta S = 2NK \ln 2.$$

Free Energy of a Mixture

Consider (the same) two systems (gases, equally-dense liquids, certain solids) A and B that are in thermal contact but in diffusive isolation by a partition. Before the partition is removed, the Gibbs free energy is

$$G = \mu_A N_A + \mu_B N_B,$$

where letting $N = N_A + N_B$ and $x = N_B/N$, we have

$$G = \mu_A N (1-x) + \mu_B N x = (1-x) G_A^0 + x G_B^0,$$

where $G_A^0 = \mu_A N$ and $G_B^0 = \mu_B N$ denote the unmixed energies. When the partition is removed, the total Gibbs free energy is simply

$$G_{\text{mix}} = (1-x) G_A^0 + x G_B^0 + T \Delta S_{\text{mix}}. \quad (31.69)$$

Dilute Solutions

Consider a system that initially consists of N_A particles of species A . According to Equation (31.58), the Gibbs free energy reads

$$G = N_A \mu_0(T, P)$$

at fixed temperature and pressure. The volume of the system may be considered as the sum of N_A volume elements.

A single particle of species B added to the system will ‘freely expand’ (isothermally) into the existing volume in analogy to Equation (31.44), contributing $-KT \ln(N_A)$ to the Gibbs free energy. The chemical binding energy is represented by some unknown function $f(T, P)$, so we write the energy change as

$$dG = f(T, P) - KT \ln(N_A).$$

Adding a second particle of species B is not a matter of taking $dG \rightarrow 2dG$ due to the indistinguishability of particles. It follows that for two particles of species B , the total correction to the Gibbs free energy is

$$dG = 2f(T, P) - 2KT \ln N_A + KT \ln 2.$$

Generalizing to N_B particles while maintaining the dilute limit $N_B \ll N_A$ particles is straightforward. The $KT \ln 2$ term generalizes to $KT \ln(N_B!)$, which by Stirling’s approximation becomes $KT(N_B \ln N_B - N_B)$. Finally, we have the Gibbs free energy of a dilute solution,

$$\begin{aligned}G &= N_A \mu_0(T, P) + N_B f(T, P) - N_B KT \ln N_A \\ &\quad + N_B KT \ln N_B - N_B KT. \quad (31.70)\end{aligned}$$

The chemical potentials of species A and B are readily calculated from the Gibbs free energy. These are:

$$\begin{aligned}\mu_A &= \left(\frac{\partial G}{\partial N_A} \right)_{T,P,N_A} \\ &= \mu_0(T, P) - \frac{N_B K T}{N_A}\end{aligned}\quad (31.71)$$

$$\begin{aligned}\mu_B &= \left(\frac{\partial G}{\partial N_B} \right)_{T,P,N_B} \\ &= f(T, P) + K T \ln \left(\frac{N_B}{N_A} \right)\end{aligned}\quad (31.72)$$

6.4 Osmotic Pressure

Consider a system of total volume V and number of particles solvent N_A . Next, divide the system with a semipermeable membrane and introduce some number of solute particles N_B on one side of the membrane. The membrane is tuned to allow the passage of solute particles, but not solvent particles. It follows that the pressure is weakly discontinuous across the membrane, and so too is the chemical potential on either side of the membrane, leading us to write

$$\mu_0(T, P_2) \approx \mu_0(T, P_1) + (P_2 - P_1) \frac{\partial \mu_0}{\partial P}.$$

Using Equation (31.71) for the chemical potential of the solvent particles, arrive at

$$(P_2 - P_1) \frac{\partial \mu_0}{\partial P} = \frac{N_B K T}{N_A}.$$

Of course, the derivative $\partial \mu_0 / \partial P$ is simply V / N_A at constant T and N , and we arrive at

$$(P_2 - P_1) = \frac{N_B K T}{V}, \quad (31.73)$$

a result known as the *van't Hoff* formula. The pressure difference $P_2 - P_1$ is known as the *osmotic pressure*.

6.5 Fugacity

As a dilute solution tends to equilibrium, the pressure required for the system to become diffuse and behave like ideal gas is called the *fugacity*. Begin by considering the entropy of freely-expanding ideal gas

$$S = N K \ln \left(\frac{V_f}{V_i} \right),$$

which also obeys $P_i V_i = P_f V_f$. Next, insert S into the definition (31.49) of chemical potential to get

$$\mu = \mu_0 + K T \ln \left(\frac{P_f}{P_i} \right),$$

where the μ_0 term has been imposed into the equation to handle dilute solutions. The ideal gas has $\mu_0 = 0$.

The above result generalizes to multiple dilute solutions. Let the ‘initial’ pressure P_i equal the sum P of all partial pressures, so drop the subscript. The j -th partial substance thus obeys

$$\mu_j = \mu_0 + K T \ln \left(\frac{P_j}{P} \right), \quad (31.74)$$

which generalizes again to non-ideal gases by introducing a van der Waals constant b as

$$\mu_j = \mu_0 + K T \ln \left(\frac{P_j}{P} \right) + b P,$$

which is equivalent to adding a dimensionless term ϕ into the ln-term of Equation (31.74). Note that we have deployed a simplified van der Waals equation $P(V - Nb) = NKT$.

Problem 55

Show that:

$$bP = K T \ln \phi$$

The dimensionless quantity ϕ is called the *fugacity coefficient*, and the bare term ‘fugacity’ typically refers to the modified pressure

$$f = \phi P. \quad (31.75)$$

Fugacity comes into play when describing non-ideal gases near the ideal gas limit. Paraphrasing the differential version of Equation (31.74) for one chemical species, we define a formula that *replaces* pressure with fugacity:

$$d\mu = K T d(\ln f) \quad (31.76)$$

Immediately following definition (31.76), we solve for f to write

$$f = \exp \left(\frac{\mu - \mu_0}{K T} \right).$$

Chapter 32

Variational Mechanics

Chapter 33

Classical Field Theory

Part VI

Modern Physics

Chapter 34

Special Relativity

1 Minkowski Space

1.1 Special Relativity

Prior to 1905, scientists were satisfied to explain the physical universe with the laws of mechanics, thermodynamics, and electrodynamics. No physicist openly suspected, for instance, that space itself - the empty host for matter, would be intricately shaped the matter that occupies it. Nor would anyone have seriously entertained that the flow of time has something to do with the reference frame from which it is measured.

Albert Einstein would be the first to challenge the very bedrock of physics at the level of space and time. He pointed out that the laws of physics run into trouble unless we let go of the separateness and the rigidity of space and time. Instead, these must be regarded as a whole called *spacetime*, and the correction to the field of physics due to spacetime is known as *special relativity*.

In the following we briefly review the case for special relativity and explore a few of its consequences and so-called paradoxes.

Light as a Wave

Understanding the nature of light has been a priority to physicists over the centuries. Early measurements of the speed of light were hitting around $c \approx 3 \times 10^8$ m/s.

Meanwhile, the equations of electrodynamics, as com-

plied by Maxwell, read

$$\begin{aligned}\nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0} \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{B} &= \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t},\end{aligned}$$

where $\mathbf{E}$ and $\mathbf{B}$ are the respective electric and magnetic fields, ρ is charge density, and $\mathbf{J}$ is the displacement current. The constant ϵ_0 is the electric permittivity of free space, and the constant μ_0 is the magnetic permeability of free space.

In a region without charges and currents, Maxwell's equations combine to yield simultaneous differential equations:

$$\begin{aligned}\epsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} &= \nabla^2 \mathbf{E} \\ \epsilon_0 \mu_0 \frac{\partial^2 \mathbf{B}}{\partial t^2} &= \nabla^2 \mathbf{B}\end{aligned}$$

Both the $\mathbf{E}$ - and $\mathbf{B}$ -field seem to adhere to a wave equation, giving rise to an *electromagnetic wave*. Comparing to a mechanical wave equation, the propagation speed of an electromagnetic wave is evidently

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}.$$

Amazingly, the values of ϵ_0 , μ_0 are deeply related to the speed of light. Measurements of these quantities were available to Maxwell, and surely enough, the value of c calculated to 3×10^8 m/s in agreement with the measured

value. Compelled by this connection, Maxwell himself commented:

'We have strong reason to conclude that light itself - including radiant heat and other radiation, if any - is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field according to electromagnetic laws.'

Aether Hypothesis

To understand light in terms of electromagnetic waves, speculation quickly arose over the medium of propagation. Light, having a faster speed than any known wave, should only move through something immensely rigid. Paradoxically though, such a medium needs to be thin enough to not be 'noticed' by ordinary matter. This ghostly inhabitant of all visible space was deemed the *Luminiferous Aether*.

Physicists reasoned the aether must be at rest with respect to some frame of reference, and surely Earth constantly moves with respect to such a frame. If the Aether is really there it should follow that light rays deviate away from straight lines by an 'aether wind' when observed from any Earth-based laboratory. Using measurements of the off-target deflection, perhaps Earth's absolute motion through the aether could be deduced, and the sense of 'absolute rest' in the universe could be established.

Michelson-Morley Experiment

An experiment to detect the presence of aether wind was devised in 1880-1881 by German physicist A. A. Michelson, and refined in partnership with E. W. Morley by 1887 in the United States. They devised an experiment in where a light beam was split into two perpendicular components, with each ray reflected back and recombined into a pair of interfering beams. With the so-called Michelson-Morley apparatus, Earth's motion through the aether would be detected by measuring moving 'fringes' in the recombined beam. To eliminate any bias from the physical alignment of the apparatus, it was designed as free-floating on a bed of mercury, able to rotate around a center axis.

Conducting their experiment, Michelson and Morley were shocked to find that the fringe pattern did not change when varying any experimental parameters. Time of day, duration of the experiment, orientation of the apparatus - none of these changed the result. In other words, no aether wind was measured, and the aether hypothesis was in trouble.

Magnet and Ring

It's unclear whether Einstein knew of the Michelson-Morley experiment; all Einstein needed was electrodynamics. As a direct example of Faraday's law, it's well known that a permanent magnet in motion causes an electric field to arise around the magnet. Such a field can be captured by a ring-shaped conductor, oriented perpendicular to the magnet's motion, resulting an induced current in the ring. By symmetry, the same effect arises when the magnet is instead held at rest while ring moves along the axis of the magnet.

Einstein was the first to ponder the notion of a permanent magnet 'at rest', or a loop at rest for that matter, is not an absolute idea at all. For example, a magnet that lies on the floor of a uniformly-coasting train car does not 'know' it's moving, and does not induce a local electric field. A ring traveling along with the magnet will have no induced current. When viewed from a non-moving frame of reference however, such as a train station, the magnet does appear to induce a field around it. (You could test this by making a large-enough loop for the whole train to pass through.) It should be impossible for the magnet to simultaneously emit an electric field and to not emit an electric field - unless we discard the notion of 'absolute' motion altogether and accept that relative motion is all that matters.

This insight allowed Einstein to discard the aether hypothesis. If the stationary aether exists, then all magnets, even those 'at rest', should have an induced electric field caused by aether wind. This is experimentally not the case - instead only the relative motion between the magnet and the loop is significant.

Emission Theories of Light

Several attempts were made to modify the equations of electrodynamics (by altering the so-called retardation time) in order to correct the problems pointed out by Einstein. In the Maxwell emission theory of light, all electromagnetic radiation propagates at speed c from the fixed point in aether from which it was emitted. In a modified version named after Ritz, all electromagnetic radiation propagates at speed c with respect to a moving point that follows the instantaneous motion of the source at emission time.

After plenty of effort to defend any version of emission theory by Einstein and others, it was decided that such a theory isn't strong enough to fix electrodynamics, and all versions of emission theory were abandoned.

Enter Vacuum

With the aether hypotheses sufficiently deflated, physicists were left with their set of wave equations, but no medium to support the waves. Further studies from optics indicate that materials such as glass or water are in fact a hindrance to light wave propagation. For lack of a better option, the idea that *a medium isn't needed at all* eventually gained traction. That is, it seems that light ‘prefers’ to travel through vacuum.

Catching up to Light

Einstein popularized the use of *gedanken*, the German word for ‘thought experiment’, to reason through problems involving the nature of light. In the following, we’ll examine what it means to ‘catch up to light’.

Begin by considering a light source, such as a laser or light bulb at rest in a laboratory. Rays emitted from the light source, which take off at speed c , are made of oscillating $\mathbf{E}$ - and $\mathbf{B}$ -fields oriented mutually perpendicular to the propagation direction in accordance with electrodynamics.

Next consider a train or other reference frame that can move fast enough to keep pace with the speed of light. With the laboratory light source still on, what happens if we enter the car, accelerate to speed c , and observe the light rays in their ‘own’ reference frame?

Since light rays are made of oscillating $\mathbf{E}$ - and $\mathbf{B}$ -fields we expect to measure the waves to ‘slow down’ as the car approaches speed c , and finally become stationary at exactly c . Einstein realized this ‘frozen light wave’ is clearly nonsense, as Maxwell’s equations do not predict stationary oscillatory waves. Moreover, experiment has never turned up such a result.

Even worse, there is no evidence in theory or experiment of a light ray slowing down at all (in vacuum). The only observation that fits into existing electrodynamics is for the speed of light to never deviate from c (in vacuum), regardless of the frame in which it is measured.

Galilean Transformation Breakdown

Newtonian physics handles relative motion between reference frames via the *Galilean transformation*

$$\begin{aligned} t' &= t \\ x' &= x - vt \end{aligned}$$

implying

$$\frac{dx'}{dt'} = \frac{dx}{dt} - v,$$

where v is the relative speed between the boosted reference frame and the unprimed reference frame.

If the above is true, one can imagine an experiment where a beam of light is observed in the two inertial frames, with dx'/dt' and dx/dt being the observed speed of light in the inertial frames. With no limit on v however, one can imagine tuning the relative frame velocity to $v = dx/dt \rightarrow c$, and thus $dx'/dt' \rightarrow 0$. This corresponds to the ‘frozen light wave’ situation again. Due to this, there’s something inherently incorrect about the Galilean transformation.

Einstein’s Postulates

In effort to reconcile all of the problems brought about by studying space, time, and light, Einstein decided to try a ‘hard reset’ of physics based on two postulates:

1. The laws of physics are equivalent in all inertial reference frames.
2. The speed of light c is a universal constant.

1.2 Spacetime

The 1 + 3-dimensional Minkowski space is mapped by *spacetime coordinates* q^μ such that $q^0 = ct$, $q^1 = x$, $q^2 = y$, $q^3 = z$, or

$$q^\mu = (ct, x, y, z) = (ct, \mathbf{x}).$$

Each variable is typically a function of an affine parameter λ .

Events

A specific location in spacetime is called an *event*. For a pair of events 1 and 2, we catalog these via q_1^μ , q_2^μ with an index subscript. The differential between the pair of events is written

$$\Delta q^\mu = q_2^\mu - q_1^\mu.$$

Inertial Frame

Minkowski space refers to any non-accelerated reference frame, also known as an *inertial frame*. An inertial frame refers to any frame at rest, any frame with a constant position offset, or any frame that moves at constant velocity. These are all related by the term *Lorentz group*.

Minkowski Metric

By convention, the metric $\eta_{\mu\nu}$ for Minkowski Space is given by:

$$\begin{aligned}\eta_{00} &= -1 \\ \eta_{11} &= \eta_{22} = \eta_{33} = 1 \\ \eta_{\mu\nu} &= 0 \text{ if } \mu \neq \nu\end{aligned}$$

Some texts instead work in 3 + 1 dimensions by placing a positive sign on the 00-component with negative signs on all other components. The distinction between the two regimes is called the *signature*. That is, we work with the signature

$$(-, +, +, +) .$$

Spacetime Interval

In terms of spacetime coordinates, the differential interval reads

$$dS^2 = \eta_{\mu\nu} dq^\mu dq^\nu = -c^2 dt^2 + d\mathbf{x} \cdot d\mathbf{x} ,$$

where

$$d\mathbf{x} \cdot d\mathbf{x} = dx^2 + dy^2 + dz^2 ,$$

and we see dS is not equal to arc length due to the presence of a time component. Instead, dS^2 is the *differential spacetime interval*.

The spacetime interval can be written in non-differential form by associating dq^μ with Δq^μ such that

$$\begin{aligned}\Delta t &= t_2 - t_1 \\ \Delta \mathbf{x} &= \mathbf{x}_2 - \mathbf{x}_1 ,\end{aligned}$$

and correspondingly

$$\Delta S^2 = -c^2 \Delta t^2 + \Delta \mathbf{x} \cdot \Delta \mathbf{x} .$$

Spacelike Events

Events with $\Delta S^2 > 0$ are called *spacelike*, because the spatial component outweighs the time component:

$$\Delta S^2 = -c^2 \Delta t^2 + \Delta \mathbf{x} \cdot \Delta \mathbf{x} > 0$$

Spacelike events are not causally connected, which means they are sufficiently separated in space such that neither event 1, 2 has influence on the other.

Timelike Events

Events with $\Delta S^2 < 0$ are called *timelike*, because the time component outweighs the spatial component:

$$\Delta S^2 = -c^2 \Delta t^2 + \Delta \mathbf{x} \cdot \Delta \mathbf{x} < 0$$

Timelike events are causally connected, which means that one event *could* have been informed by the other event.

Lightlike Events

Events with $\Delta S^2 = 0$ are called *lightlike*, as only a photon has $c = \Delta x / \Delta t$.

1.3 Proper Time

When it comes to analyzing the motion of massive bodies, we are interested in pairs of events that are timelike, corresponding to a negative spacetime interval. To quantify this, define a quantity τ called the *proper time* such that

$$dS^2 = -c^2 d\tau^2 .$$

Note the definition of proper time is also true in the general case, and we'll work with a general metric $g_{\mu\nu}$ as a reminder when appropriate to do so.

By writing

$$-c^2 d\tau^2 = -c^2 dt^2 + d\mathbf{x} \cdot d\mathbf{x} ,$$

we see the left side lacks a $\Delta \mathbf{x}'$ -like term, therefore we associate the proper time τ with the rest frame time interval as a matter of definition.

Gamma Factor

Starting from

$$dS^2 = -c^2 dt^2 + d\mathbf{x} \cdot d\mathbf{x} ,$$

replace dS^2 with the definition of proper time and divide by dt^2 to get

$$c^2 \left(\frac{d\tau}{dt} \right)^2 = c^2 - v^2 ,$$

where

$$\mathbf{v} = \frac{d\mathbf{x}}{dt} .$$

The velocity vector $\mathbf{v}$ is the change in position vector with respect to the time parameter t , not necessarily the proper time.

It's interesting to get a formula that relates $d\tau$ to dt . Solving the above for $dt/d\tau$ gives the so-called *gamma factor*:

$$\frac{dt}{d\tau} = \frac{1}{\sqrt{1 - v^2/c^2}} = \gamma_v$$

Problem 1

For low speeds $v \ll c$, show that:

$$\gamma \approx 1 + \frac{v^2}{2c^2} + \frac{3v^4}{8c^4} + \frac{5v^6}{16c^6} + \dots$$

Problem 2

Use the gamma factor to check that $dS^2 = -c^2 d\tau^2$.

Answer:

$$\begin{aligned} dS^2 &= -c^2 dt^2 + dx^2 \\ &= -c^2 dt^2 \left(1 - \frac{v^2}{c^2}\right) \\ &= \frac{-c^2 dt^2}{\gamma^2} = -c^2 d\tau^2 \end{aligned}$$

Time Dilation

Rearranging the above, we can also write

$$dt = \gamma d\tau.$$

Since γ is always greater than one, it follows that the proper time $d\tau$ is always *smaller* than any other time dt . Likewise, an arbitrary time Δt is always larger, or 'dilated' in comparison to $d\tau$. This effect is called time dilation.

Problem 3

Consider a device called a *light clock* that consists of a vertical tube of length L whose inner surface is mirrored. Trapped inside the tube is a vertically-moving pulse of light reflecting rapidly between the mirrored ends, where one 'tick' corresponds to the light traversing the length of the tube, i.e. one tick per L/c seconds, or $L = c\Delta t$.

Next, consider an observer moving horizontally with speed v with respect to the light clock. Show that the motion of the trapped light pulse is characterized by

$$\sqrt{L^2 + (v\Delta t')^2} = c\Delta t',$$

and find the corresponding 'tick' in the moving frame. Answer: $\Delta t' = \gamma\Delta t$

Length Contraction

The mechanism behind time dilation has a consequence for length as well. Consider an inertial frame moving at speed v with respect to a stationary rod of length L_p . The subscript 'p' denotes the *proper length* of the rod.

In the frame of the rod, the time required for an observer in the moving frame to pass the rod is given by $L_p = v\Delta t$. In the moving frame, let us not assume rod has the same length by writing $L' = v\Delta t'$. Taking the ratio of these, we find

$$\frac{L'}{L_p} = \frac{\Delta t'}{\Delta t}.$$

Since we're interested in the moving frame, associate $\Delta t'$ with the proper time $\Delta\tau$, and the above becomes:

$$L' = \frac{L_p}{\gamma}$$

Evidently, the apparent length of the rod is shorter than the proper length. Since there's nothing particular about a stationary rod however, it follows that all objects appear to become shorter, or *contract* in length along the direction of motion.

Atmospheric Muons

Cosmic rays that interact with Earth's atmosphere have a chance of generating muon particles (somewhat like heavy electrons). Being unstable particles, muons decay in about 2.2×10^{-6} sec. Interestingly, muons generated in the high atmosphere, approximately 6000 m, can still be detected at ground level. Using Newtonian mechanics to calculate the downward speed of a detected muon, the result is at least

$$v_{\text{Newton}} = \frac{\Delta x}{\Delta t} = \frac{6.0 \times 10^3 \text{ m}}{2.2 \times 10^{-6} \text{ s}},$$

which exceeds the speed of light.

The correct speed calculation must involve the gamma factor and the notion of proper time. For this, we assign $\Delta\tau = 2.2 \times 10^{-6}$ s, and then we write the *relativistic velocity*

$$u_{\text{rel}} = \frac{\Delta x}{\gamma\Delta\tau} = \frac{1}{\gamma} v_{\text{Newton}}.$$

Since $\gamma > 1$, the relativistic velocity is knocked back to a value that must be less than c .

Correspondingly, from the rest frame of the muon, the distance $\Delta x'$ to the ground is calculated from

$$u_{\text{rel}} = \frac{\Delta x'}{\Delta\tau}.$$

Comparing the two forms for u_{rel} , we find

$$\Delta x' = \frac{1}{\gamma} \Delta x ,$$

which means the ground ‘appears’ (to the muon) closer than it really is.

Four-Velocity

On any manifold, the derivative of the spacetime coordinates q^μ with respect to proper time is called the *proper velocity*:

$$U^\mu = \frac{dq^\mu}{d\tau} = d_\tau q^\mu$$

Scaling the velocity with the gamma factor yields a velocity vector with respect to proper time τ :

$$\gamma \mathbf{v} = \frac{dt}{d\tau} \frac{d\mathbf{x}}{dt} = \frac{d\mathbf{x}}{d\tau}$$

Meanwhile, the proper velocity can be written

$$U^\mu = \frac{dq^\mu}{d\tau} = \left(c \frac{dt}{d\tau}, \frac{d\mathbf{x}}{d\tau} \right) = (\gamma c, \gamma \mathbf{v}) ,$$

also known as the *four-velocity*.

The norm of the proper velocity always comes out to $-c^2$ via

$$\begin{aligned} U_\mu U^\mu &= g_{\mu\nu} U^\mu U^\nu = -\gamma^2 c^2 + \gamma^2 v^2 \\ &= c^2 \left(\frac{v^2 - c^2}{c^2 - v^2} \right) = -c^2 , \end{aligned}$$

which is to say the norm of the four-velocity is a spacetime invariant:

$$U_\mu U^\mu = -c^2$$

Isolating Proper Time

By definition of the proper time, the differential spacetime interval implies

$$g_{\mu\nu} dq^\mu dq^\nu = -c^2 d\tau^2 ,$$

In terms of the spacetime invariant, one may also solve the above for the differential proper time:

$$d\tau = \frac{1}{c} \sqrt{-g_{\mu\nu} dq^\mu dq^\nu}$$

Note the use of the general metric $g_{\mu\nu}$ in place of the Minkowski metric $\eta_{\mu\nu}$. The above is in fact a general formula that isn't strictly limited to non-accelerated reference frames.

Problem 4

Consider a particle moving according to

$$\begin{aligned} ct &= a\lambda \\ x &= \frac{a}{2} \cos(\lambda) \\ y &= \frac{a}{2} \sin(\lambda) \\ z &= 0 , \end{aligned}$$

where a is a real constant and λ is a parameter of motion. Compute the proper time between $\lambda = 0$ and $\lambda = 2\pi$. Answer:

$$\begin{aligned} \tau &= \int_{\lambda=0}^{\lambda=2\pi} \frac{1}{c} \sqrt{-\eta_{\mu\nu} dq^\mu dq^\nu} \\ &= \int_0^{2\pi} \frac{1}{c} \sqrt{a^2 - a^2/4} d\lambda \\ &= \frac{a\sqrt{3}\pi}{c} \end{aligned}$$

Affine Parameter

As briefly mentioned, the spacetime coordinates q^μ are to depend on an affine parameter λ such that $d\lambda \propto dS$. To assert this, consider a parameterization $q^\mu(\lambda)$ and a given value λ_0 .

The differential of q^μ near λ_0 is given by

$$\begin{aligned} q^\mu(\lambda) &= q^\mu(\lambda_0 + d\lambda) - q^\mu(\lambda_0) \\ &= \left(\frac{d}{d\lambda} q^\mu(\lambda) \right) \Big|_{\lambda_0} d\lambda , \end{aligned}$$

where the parenthesized quantity is the instantaneous *tangent vector* at λ_0 .

Using this form in the spacetime interval, i.e.

$$dS^2 = \eta_{\mu\nu} dq^\mu dq^\nu ,$$

we find

$$\begin{aligned} dS^2 &= - \left(\left(\frac{d}{d\lambda} q^0(\lambda) \right) \Big|_{\lambda_0} \right)^2 (d\lambda)^2 \\ &\quad + \left(\left(\frac{d}{d\lambda} \mathbf{q}(\lambda) \right) \Big|_{\lambda_0} \right)^2 (d\lambda)^2 , \end{aligned}$$

which by the chain rule is equivalent to

$$\begin{aligned} dS^2 &= -(d\lambda)^2 \left(\frac{dt}{d\lambda} \right)^2 \left(c^2 - \left(\frac{d\mathbf{q}}{dt} \right)^2 \right) \\ &= -c^2 (d\lambda)^2 \left(\frac{d\tau}{d\lambda} \right)^2 . \end{aligned}$$

From the above we see that if λ , τ relate by two unknown coefficients such that

$$\lambda = A\tau + B ,$$

we have

$$dS^2 = -c^2 \cancel{A^2} d\tau^2 \frac{1}{\cancel{A^2}} = -c^2 d\tau^2 ,$$

recovering the definition of proper time.

Problem 5

The trajectory of a particle moving at speed $c/2$ is parameterized by

$$\begin{aligned} ct &= 2\lambda \\ x &= c\lambda . \end{aligned}$$

Determine the proper time τ in terms of λ . Answer:

$$\begin{aligned} \Delta S^2 &= -c^2 2^2 \Delta\lambda^2 + c^2 \Delta\lambda^2 \\ -c^2 \Delta\tau^2 &= c^2 \lambda^2 (-3) \\ \tau &= \sqrt{3}\lambda \end{aligned}$$

1.4 Four-Vectors

In the same way we define generalized coordinates q^μ or the proper velocity U^μ as a type (1, 0) tensor, we can define other tensors P^μ , F^μ , K^μ , also called *four-vectors*.

Using what we know of tensors, the Minkowski metric $\eta_{\mu\nu}$, and the proper time, we can wring out a handful of useful tools of special relativity.

Four-Momentum

Supposing the coordinates q^μ trace the motion of a particle of mass m , the product

$$P^\mu = mU^\mu$$

is called the *four-momentum*. Alternatively, the four-momentum is written

$$P^\mu = (\gamma mc, \mathbf{p}) ,$$

where *relativistic momentum* vector $\mathbf{p}$ is given by

$$\mathbf{p} = \gamma m \mathbf{v} .$$

The norm of the four-momentum is also an invariant given by

$$P_\mu P^\mu = -m^2 c^2 .$$

Energy-Momentum Relation

Expanding $P_\mu P^\mu = -m^2 c^2$ using the Minkowski metric, one writes

$$-\gamma^2 m^2 c^2 + p^2 = -m^2 c^2 ,$$

or, eliminating negative signs:

$$(\gamma mc^2)^2 = p^2 c^2 + m^2 c^4$$

Both sides of the equation have units of energy squared, thus identify the scalar E with

$$\begin{aligned} E &= \gamma mc^2 \\ E &= \sqrt{p^2 c^2 + m^2 c^4} , \end{aligned}$$

called the *relativistic energy*.

In terms of the total energy, the four-momentum is

$$P^\mu = \left(\frac{E}{c}, \mathbf{p} \right) .$$

Another quantity emerges by contracting P_μ with q_μ , coming out to

$$q^\mu P_\mu = -Et + \mathbf{p} \cdot \mathbf{x} .$$

We'll subsequently show in several ways that the above is another spacetime invariant.

Rest Energy

For the special case $v = 0$, $p = 0$, $\gamma = 1$, the energy equation does not settle to zero, but instead to the *rest energy* of the particle:

$$E_0 = mc^2$$

Astonishingly, this tells that a particle has energy simply for existing, regardless of any motion or external potentials.

Kinetic Energy

Subtracting the rest energy mc^2 from the total energy γmc^2 gives an equation for the relativistic kinetic energy T :

$$T = E - E_0 = mc^2 (\gamma - 1)$$

Proper Force

Starting with the proper velocity U^μ , take a τ -derivative to yield the *proper acceleration* $dU^\mu/d\tau$, and multiply through by mass m to write the *proper force* F^μ :

$$\begin{aligned} F^\mu &= m \frac{d}{d\tau} U^\mu = \frac{d}{d\tau} P^\mu \\ &= \left(\frac{1}{c} \frac{dE}{d\tau}, \frac{d\mathbf{p}}{d\tau} \right) = \gamma \left(\frac{1}{c} \frac{dE}{dt}, \frac{d\mathbf{p}}{dt} \right) \end{aligned}$$

We can do more to simplify the dE/dt term. In Newtonian physics, it's readily shown for constant $\mathbf{F}$ that

$$\frac{dE}{dt} = \frac{d}{dt} \int \mathbf{F} \cdot d\mathbf{x} = \frac{d}{dt} \int (\mathbf{F} \cdot \mathbf{v}) dt = \mathbf{F} \cdot \mathbf{v}$$

As it turns out, the same is true in the relativistic case:

$$\begin{aligned} \frac{d}{dt} (E^2) &= \frac{d}{dt} (p^2 c^2 + m^2 c^4) \\ 2 \left(\gamma m c^2 \right) \frac{dE}{dt} &= c^2 2 \frac{d\mathbf{p}}{dt} \cdot \mathbf{p} \\ \frac{dE}{dt} &= \mathbf{F} \cdot \mathbf{v} \end{aligned}$$

In the above we defined the *relativistic force* vector $\mathbf{F} = d\mathbf{p}/dt$. With this, the proper force is written in terms of instantaneous power and relativistic force:

$$F^\mu = \gamma \left(\frac{\mathbf{F} \cdot \mathbf{v}}{c}, \mathbf{F} \right)$$

A contraction between the proper force and the proper velocity equals zero, which is as invariant as they come:

$$F^\mu U_\mu = -\gamma^2 \mathbf{F} \cdot \mathbf{v} + \gamma^2 \mathbf{F} \cdot \mathbf{v} = 0$$

This can also be shown via the product rule:

$$\begin{aligned} F^\mu U_\mu &= m \left(\frac{dU^\mu}{d\tau} \right) U_\mu \\ &= \frac{m}{2} \frac{d}{d\tau} (U^\mu U_\mu) \\ &= \frac{d}{d\tau} (-c^2) = 0 \end{aligned}$$

1.5 Photons

Starting with

$$E = \gamma m c^2,$$

we can attempt to describe a *photon* with $v \rightarrow c$ and $m \rightarrow 0$. Meanwhile, the energy-momentum relation with $m = 0$ simply yields

$$E = pc.$$

Comparing the two energies, we interpret the photon momentum as

$$p = \lim_{\substack{v \rightarrow c \\ m \rightarrow 0}} \gamma m c.$$

Here, it makes sense to acknowledge the wave nature of light using $\lambda f = c$, where λ is the wavelength and f is the frequency, giving

$$p = \lim_{\substack{v \rightarrow c \\ m \rightarrow 0}} \gamma m \lambda f.$$

The momentum of a photon ought to increase with its frequency, or more strongly, be proportional to its frequency. This dependence is carried exactly by the frequency f on the right, leaving everything else to be repackaged:

$$p = \left(\lim_{\substack{v \rightarrow c \\ m \rightarrow 0}} \gamma m \lambda \right) f$$

Multiply both sides by c to restore the energy of the photon:

$$E = \left(\lim_{\substack{v \rightarrow c \\ m \rightarrow 0}} \gamma m \lambda c \right) f$$

The parenthesized quantity cannot be evaluated on paper. As it turns out, the constant of proportionality between the energy of a photon and the frequency of a photon, named *Planck's constant*, denoted h , is neither zero nor infinite.

For the energy of a photon, we write

$$E = hf,$$

and for the momentum,

$$p = \frac{h}{\lambda}.$$

Alternatively, one can use the 'angular' formulation by introducing the angular frequency ω and wavenumber k such that

$$\begin{aligned} \omega &= 2\pi f \\ k &= \frac{2\pi}{\lambda} \end{aligned}$$

and

$$\omega = ck.$$

Along with these comes Planck's reduced constant $\hbar$:

$$\hbar = \frac{h}{2\pi}$$

Taking the above together, we can write the energy of a photon

$$E_\gamma = \hbar\omega ,$$

and the three-momentum vector of a photon

$$\mathbf{p}_\gamma = \hbar\mathbf{k} ,$$

where $\mathbf{k}$ contains the wavenumber and a unit vector for the direction of travel via $\mathbf{k} = k\hat{\mathbf{k}}$. In terms of k , the energy of a photon is

$$E_\gamma = \hbar ck .$$

Photon Four-Vector

For photons, the four-momentum is renamed to K^μ and is called the *photon four-vector*:

$$K^\mu = \left(\frac{E_\gamma}{c}, \hbar\mathbf{k} \right) = \hbar \left(\frac{\omega}{c}, \mathbf{k} \right)$$

The norm of the photon four-vector is straightforwardly shown to be zero:

$$K_\mu K^\mu = \hbar^2 \left(\frac{-\omega^2}{c^2} + k^2 \right) = 0$$

A contraction with the position vector yields another invariant quantity:

$$q^\mu K_\mu = \hbar(-\omega t + \mathbf{k} \cdot \mathbf{x})$$

The above can be attained quickly by integrating $E_\gamma = \hbar f$ with respect to time. For this we have

$$\hbar\omega \int dt = \hbar k \int c dt ,$$

implying

$$-\omega t + kx = \text{constant} .$$

1.6 Lorentz Transformation

As a type (1, 0) tensor, the spacetime coordinates q^μ are subject to the tensor transformation law when changing to a primed coordinate system:

$$q^{\mu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} q^\mu$$

For Minkowski space, such a change of coordinates is called the *Lorentz transformation*, sometimes called a *Lorentz boost*.

Derivation

Consider an inertial frame O and a second inertial frame O' that moves at constant speed v along the x -axis with respect to O . For simplicity we set no relative motion on the y - and z -axes, and there is no need to concern over how y and z transform, i.e. $y' = y, z' = z$.

To figure out the restriction on t' and x' , the most careful way forward is to assume each is a function of all unprimed variables:

$$ct' = f(t, x)$$

$$x' = g(t, x)$$

Note the factor c is included to assure units of distance. On the grounds that Minkowski space exhibits translational symmetry, consider two events 1 and 2, and conclude

$$ct'_2 - ct'_1 = f(t_2 - t_1, x_2 - x_1)$$

$$x'_2 - x'_1 = g(t_2 - t_1, x_2 - x_1) ,$$

suggesting f, g are linear functions. We therefore introduce four unknown coefficients A, B, C, D , and write:

$$ct' = At + Bx$$

$$x' = Ct + Dx$$

Two of the constants can be eliminated straightforwardly. The origin of the primed frame is located at $x' = 0$ and has position $x = vt$ with respect to the unprimed frame. From the x' -equation we learn $C = -Dv$, leading to

$$x' = D(x - vt) .$$

Likewise, the origin of the unprimed frame is located at $x = 0$, and has position $x' = -vt'$ with respect to the primed frame. With this we find $A = Dc$, leading to

$$ct' = D \left(ct + \frac{B}{D}x \right) .$$

To clean up our work, let the ratio B/D equal a new coefficient F_v , and also rename D as γ_v . With this, we have

$$ct' = \gamma_v (F_v x + ct)$$

$$x' = \gamma_v (x - vt) ,$$

and the two unknowns are γ_v and F_v , each a function of velocity.

To get a handle on F_v , we'll exploit the property of the Lorentz group that qualifies two successive Lorentz transformations as equivalent to one effective Lorentz transformation. In particular, suppose the primed frame O' moves with speed v_1 with respect to the unprimed frame

O . Also, suppose a double-primed frame O'' moves with speed v_2 with respect to O' . The pair of equations above is written twice:

$$\begin{aligned} ct' &= \gamma_{v_1} (F_{v_1} x + ct) \\ x' &= \gamma_{v_1} (x - v_1 t) \\ ct'' &= \gamma_{v_2} (F_{v_2} x' + ct') \\ x'' &= \gamma_{v_2} (x' - v_2 t') \end{aligned}$$

Substitute the first pair of equations into the second and simplify to write:

$$\begin{aligned} ct'' &= \gamma_{v_1} \gamma_{v_2} ((F_{v_1} + F_{v_2}) x + (1 - F_{v_2} v_1/c) ct) \\ x'' &= \gamma_{v_1} \gamma_{v_2} ((1 - v_2 F_{v_1}/c) x - (v_1 + v_2) t) \end{aligned}$$

Comparison to the unprimed system of equations tells us that the t -coefficient in the first equation equals the x -coefficient in the second equation, or

$$1 - F_{v_2} v_1/c = 1 - v_2 F_{v_1}/c ,$$

telling us

$$\frac{F_{v_1}}{v_1} = \frac{F_{v_2}}{v_2} .$$

Since the left side and right side separate all 1- and 2-values, it follows that each side is equal to the same constant $1/w$ such that

$$F_v = \frac{v}{w} ,$$

where w has units of speed. Updating our progress, we now have:

$$\begin{aligned} ct' &= \gamma_v \left(\frac{vx}{w} + ct \right) \\ x' &= \gamma_v (x - vt) \end{aligned}$$

What remains is to determine γ_v . The trick for this is to write the inverse Lorentz transformation from the boosted frame to the unprimed frame. For this we need $v \rightarrow -v$, and:

$$\begin{aligned} ct &= \gamma_{-v} \left(\frac{-vx'}{w} + ct' \right) \\ x &= \gamma_{-v} (x' + vt') \end{aligned}$$

Substitute the first pair of equations into the second and simplify to find

$$\begin{aligned} ct &= \gamma_v \gamma_{-v} \left(1 + \frac{v^2}{wc} \right) ct \\ x &= \gamma_v \gamma_{-v} \left(1 + \frac{v^2}{wc} \right) x , \end{aligned}$$

and inevitably

$$\gamma_v \gamma_{-v} = \frac{1}{1 + v^2/wc} .$$

Finally, we lock the speed constant w , which is still arbitrary, to the negative of the speed of light such that $w = -c$. Also, due to the symmetry of the Lorentz transformation, it follows that γ_v and γ_{-v} are numerically the same and equal to the same factor γ . Finally we (re)discover the gamma factor:

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}}$$

The final product of the derivation is the Lorentz transformation in 1 + 1 dimensions

$$\begin{aligned} ct' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(ct - \frac{vx}{c} \right) \\ x' &= \frac{1}{\sqrt{1 - v^2/c^2}} (x - vt) , \end{aligned}$$

or, in matrix form:

$$\begin{bmatrix} ct' \\ x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma v/c & 0 & 0 \\ -\gamma v/c & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} ct \\ x \\ y \\ z \end{bmatrix}$$

The operation that inverts the above, called the inverse Lorentz transformation, toggles the prime symbols and the sign on v :

$$\begin{aligned} ct &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(ct' + \frac{vx'}{c} \right) \\ x &= \frac{1}{\sqrt{1 - v^2/c^2}} (x' + vt') \end{aligned}$$

Beta Parameter

Defining a dimensionless parameter

$$\beta = \frac{v}{c} ,$$

the Lorentz transformation can be written in especially tight form:

$$\begin{aligned} ct' &= \gamma (ct - \beta x) \\ x' &= \gamma (-\beta ct + x) \end{aligned}$$

Event Pairs

For a pair of events 1, 2, we can define

$$\begin{aligned}\Delta t &= t_2 - t_1 \\ \Delta x &= x_2 - x_1,\end{aligned}$$

and then take two instances of the Lorentz transformation to write:

$$\begin{aligned}c\Delta t' &= \gamma \left(c\Delta t - \frac{v\Delta x}{c} \right) \\ \Delta x' &= \gamma (\Delta x - v\Delta t)\end{aligned}$$

Problem 6

Some texts introduce the Lorentz transformation as a generalized Galilean transformation

$$\begin{aligned}\Delta x' &= \gamma (\Delta x - v\Delta t) \\ \Delta x &= \gamma (\Delta x' + v\Delta t'),\end{aligned}$$

where γ is still unknown. To determine γ , consider a pulse of light that emerges from the origin and obeys $\Delta x = c\Delta t$, $\Delta x' = c\Delta t'$. (i) Substitute these into the Lorentz transformation and solve for the ratio $\Delta t'/\Delta t$. (ii) From the above, derive the pair of equations:

$$\begin{aligned}c\Delta t' &= \gamma \left(c\Delta t - \frac{v\Delta x}{c} \right) \\ c\Delta t &= \gamma \left(c\Delta t' + \frac{v\Delta x'}{c} \right)\end{aligned}$$

Simultaneous Events

If a pair of events 1, 2 in the unprimed reference frame are simultaneous in time, we have $\Delta t = 0$ by definition. This is not the case in a primed frame however, as the above tells us:

$$c\Delta t'_{\text{simul.}} = -\gamma \frac{v\Delta x}{c}$$

To interpret this, one can imagine two pulses of light separated by Δx emitted simultaneously in the unprimed frame. Due to the motion of the primed frame, the pulses are not observed simultaneously.

For another case, we could have a pair of events 1, 2 in the unprimed frame that are simultaneous in space, but separated by a time $\Delta t = \Delta\tau$. For this, we have $\Delta x = 0$, yet the Lorentz transformation implies

$$\Delta x'_{\text{simul.}} = -\gamma v\Delta\tau.$$

Time Dilation

Consider a pair of events 1, 2 that satisfy $\Delta x = 0$, i.e. the events are simultaneous in space but not simultaneous in time $\Delta t = \Delta\tau$. This can be imagined as two successive flashes emitted from a bulb fixed in the unprimed frame. In this case, the Lorentz transformation produces

$$\Delta t' = \gamma\Delta\tau,$$

the familiar time dilation effect.

Length Contraction

Consider a rod of length $\Delta x = L_p$ that is at rest in the unprimed reference frame. The subscript 'p' denotes the proper length of the rod. From a boosted reference frame moving at relative speed v , suppose we take simultaneous measurements of the rod such that $\Delta t' = 0$, and $\Delta x'$ is the observed length.

Using $\Delta t' = 0$, the Lorentz transformation tells us

$$0 = c\Delta t - \frac{vL_p}{c},$$

and eliminating Δt gives

$$\Delta x' = \gamma \left(L_p - \frac{v^2}{c^2} L_p \right) = \frac{L_p}{\gamma},$$

the familiar length contraction effect.

Lorentz Invariant

Recalling that the differential spacetime interval dS^2 is invariant with respect to an arbitrary change of coordinates, it should follow that the spacetime interval ΔS^2 doesn't change when subject to a Lorentz transformation. For this reason we say the spacetime interval is a *Lorentz invariant*.

To check the invariance of ΔS^2 explicitly, use the one-dimensional Lorentz transformation to write

$$\begin{aligned}-c^2\Delta t'^2 + \Delta x'^2 &= \\ \gamma^2 \left(-\left(c\Delta t - \frac{v\Delta x}{c} \right)^2 + (\Delta x - v\Delta t)^2 \right) &= \\ = \gamma^2 \left(1 - \frac{v^2}{c^2} \right) (-c^2\Delta t^2 + \Delta x^2) &= \\ = -c^2\Delta t^2 + \Delta x^2\end{aligned}$$

to conclude $\Delta S'^2 = \Delta S^2$.

Velocity Addition

As per our setup, consider an inertial frame O and a second inertial frame O' that moves with constant speed v with respect to O . Now, introduce a particle that moves at constant speeds $u_x > v$, u'_x in each respective frame.

The speed u_x of the particle in the unprimed frame is given by:

$$\begin{aligned} u_x &= \frac{\Delta x}{\Delta t} \\ &= \frac{\gamma (\Delta x' + v \Delta t')}{\gamma (\Delta t' + v \Delta x' / c^2)} \\ &= \frac{u'_x + v}{1 + u'_x v / c^2} \end{aligned}$$

Evidently, the classical notion of velocity addition needs to respect the Lorentz transformation. The inverse to the above reads

$$u'_x = \frac{u_x - v}{1 - u_x v / c^2} .$$

Note that if the particle has a y -component to its velocity, we still have $\Delta y = \Delta y'$ but the same isn't true for u_y , u'_y . Instead, one calculates

$$u_y = \frac{\Delta y'}{\Delta t} = \frac{u'_y}{\gamma (1 + u'_y v / c^2)}$$

for u_y , and correspondingly for u'_y :

$$u'_y = \frac{u_y}{\gamma (1 - u_y v / c^2)}$$

Rapidity

Knowing the nonlinearity of velocity addition while seeking a quantity that *does* accumulate linearly, define a parameter ϕ called the *rapidity* such that

$$\begin{aligned} v &= c f(\phi) \\ u'_x &= c f(\phi') \\ u_x &= c f(\phi' + \phi) , \end{aligned}$$

where f is a function we must determine. Mapping the above to the formula for velocity addition, one writes

$$f(\phi' + \phi) = \frac{f(\phi') + f(\phi)}{1 + f(\phi') f(\phi)}$$

This form should be familiar from the hyperbolic trig identity

$$\tanh(\phi' + \phi) = \frac{\tanh(\phi') + \tanh(\phi)}{1 + \tanh(\phi') \tanh(\phi)} ,$$

begging the association

$$f(\phi) = \tanh(\phi) .$$

In terms of the rapidity, any speed v is expressed as

$$v(\phi) = c \tanh(\phi) .$$

By construction, the rapidity parameter can be the sum of terms

$$\phi = \phi_1 + \phi_2 + \phi_3 + \dots ,$$

yet the speed v never exceeds c .

The gamma factor can also be written in terms of rapidity, namely

$$\cosh(\phi) = \gamma .$$

To go with this, a simple formula for $\sinh(\phi)$ is straightforwardly derived:

$$\sinh(\phi) = \gamma \frac{v}{c} = \gamma \beta$$

Finally, the Lorentz transformation takes particularly simple form using the rapidity as a parameter:

$$\begin{aligned} ct' &= \cosh(\phi) ct - \sinh(\phi) x \\ x' &= -\sinh(\phi) ct + \cosh(\phi) x \end{aligned}$$

Energy-Momentum Formulation

The Lorentz transformation has been framed in terms of the spacetime coordinates ct , x , etc. An alternative formulation of the Lorentz transformation can be framed in terms of the energy and the momentum.

Staying in one dimension, consider an inertial frame O and a second inertial frame O' that moves with speed v with respect to O . Also let a particle of mass m move across both frames at speed $u > v$, u' with energies E , E' and momentum values p , p' respectively.

Begin by writing the energy and momentum of the particle in the unprimed frame:

$$\begin{aligned} E &= \gamma_u mc^2 = \frac{mc^2}{\sqrt{1 - u^2/c^2}} \\ p &= \gamma_u mu = \frac{mu}{\sqrt{1 - u^2/c^2}} \end{aligned}$$

In the boosted frame, the energy is $E' = \gamma_{u'} mc^2$, simplifying to

$$E' = mc^2 / \sqrt{1 - \frac{1}{c^2} \left(\frac{u - v}{1 - uv/c^2} \right)^2} ,$$

where the formula for velocity addition has been used to replace u' . Simplify inside the square root to find:

$$\begin{aligned} E' &= \frac{mc^2 (1 - uv/c^2)}{\sqrt{(1 - v^2/c^2)(1 - u^2/c^2)}} \\ &= \gamma_v \gamma_u mc^2 (1 - uv/c^2) \\ &= \gamma_v (E - vp) \end{aligned}$$

Repeating the exercise for the momentum, begin with $p' = \gamma_{u'} mu'$ and make similar substitutions to find:

$$\begin{aligned} p' &= \frac{mu'}{\sqrt{1 - u'^2/c^2}} \\ &= \gamma_v \gamma_u m (u - v) \\ &= \gamma_v (p - vE/c^2) \end{aligned}$$

It's cleaner to work with pure energy units, so multiply the momentum equation by a factor of c to produce the energy-momentum formulation of the Lorentz transformation:

$$\begin{aligned} E' &= \gamma (E - vp) \\ cp' &= \gamma (cp - vE/c) \end{aligned}$$

This is easily inverted by toggling the prime symbols and the sign on v .

Previously, the contraction

$$q^\mu P_\mu = -Et + \mathbf{p} \cdot \mathbf{x}$$

was advertised as a Lorentz invariant, which we establish now for 1 + 1 dimensions. Constructing the primed version of the above and simplifying, we have:

$$\begin{aligned} -E't' + p'x' &= -\gamma (E - vp) \gamma (t - vx/c^2) \\ &\quad + \gamma (p - vE/c^2) \gamma (x - vt) \\ &= -Et + px \end{aligned}$$

Three-Dimensional Motion

The Lorentz transformation as derived takes place in effectively 1 + 1 dimensions, with x being the direction of motion. To adapt this for three dimensions, we consider two inertial reference frames in relative motion characterized by the velocity vector $\mathbf{v}$.

In the unprimed system, a position vector

$$\mathbf{r} = x \hat{\mathbf{x}} + y \hat{\mathbf{y}} + z \hat{\mathbf{z}}$$

can be resolved into parallel-to- $\mathbf{v}$ and perpendicular-to- $\mathbf{v}$ components:

$$\mathbf{r} = \mathbf{r}_\parallel + \mathbf{r}_\perp$$

The Lorentz transformation only deals with the parallel component of the above, with the perpendicular component coming for the ride. Adjusting for this, we write:

$$\begin{aligned} ct' &= \gamma \left(t - \frac{\mathbf{r} \cdot \mathbf{v}}{c} \right) \\ (\mathbf{r})' &= \mathbf{r}_\perp + \gamma (\mathbf{r}_\parallel - \mathbf{v}t) \end{aligned}$$

To simplify, make the replacements

$$\begin{aligned} \mathbf{r}_\parallel &= \left(\frac{\mathbf{r} \cdot \mathbf{v}}{v^2} \right) \mathbf{v} \\ \boldsymbol{\beta} &= \mathbf{v}/c \end{aligned}$$

to get

$$\begin{aligned} ct' &= \gamma (ct - \mathbf{r} \cdot \boldsymbol{\beta}) \\ (\mathbf{r})' &= \mathbf{r} + (\gamma - 1) \boldsymbol{\beta} \left(\frac{\mathbf{r} \cdot \boldsymbol{\beta}}{\beta^2} \right) - \gamma \boldsymbol{\beta} ct. \end{aligned}$$

For transparency, the above can be expressed in a condensed matrix-like notation:

$$\begin{bmatrix} ct' \\ \mathbf{r}' \end{bmatrix} = \begin{bmatrix} \gamma & -\gamma \boldsymbol{\beta}^T \\ -\gamma \boldsymbol{\beta} & I + (\gamma - 1) \boldsymbol{\beta} \boldsymbol{\beta}^T / \beta^2 \end{bmatrix} \begin{bmatrix} ct \\ \mathbf{r} \end{bmatrix}$$

1.7 Spacetime Diagrams

A *Minkowski diagram*, also known as a *spacetime diagram*, is a visual representation of spacetime in 1 + 1 dimensions. For this we consider a pair of inertial reference frames, i.e. two coordinate systems mapped by (ct, x) , (ct', x') respectively. Assume each reference frame shares a common origin, and also that the primed reference frame is boosted with speed v along x .

Axes

Per custom, we let the horizontal axis in the spacetime diagram correspond to distance x , and the vertical axis corresponds to ct . In addition, the axes x' and ct' are fitted into the diagram, however these are *not perpendicular*. Instead, write the Lorentz transformation

$$ct' = \gamma \left(ct - \frac{vx}{c} \right)$$

and note the x' -axis corresponds to all $t' = 0$. Thus the x' axis is given by

$$x(t)_{x'\text{-axis}} = \left(\frac{c}{v} \right) ct.$$

By the same reasoning, the ct' -axis corresponds to all $x' = 0$, and is given by

$$x(t)_{ct'-\text{axis}} = \left(\frac{v}{c}\right) ct.$$

The line $x = ct$ is the same as the line $x' = ct'$, which corresponds to a photon moving along in the positive direction in both frames.

As shown in Figure 34.1, the ct' - and x' -axes make the same angle α with respect to the ct - and x -axes, where α is given by

$$\tan(\alpha) = \frac{v}{c}.$$

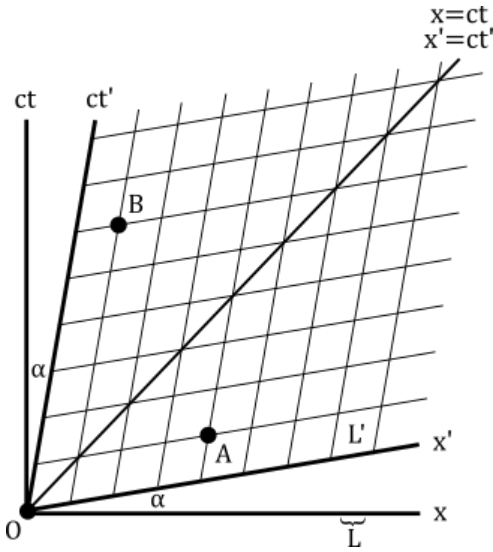


Figure 34.1. Minkowski Diagram.

Events

A single event in spacetime maps to one point on the spacetime diagram. Denoting the origin in the Figure as event O , any event on line $x = ct = ct'$ is lightlike with O . Any event A that is down-and-right of said line is spacelike with O . Any event B that is up-and-left of said line is timelike with O . As depicted, events A and B are timelike with each other.

Scale

In the spacetime diagram, the sense of unit length in the two coordinate systems is not the same. In the unprimed frame, consider a pair of simultaneous ($\Delta t = 0$) events separated

by length $\Delta x = L$. By the Lorentz transformation, the same events in the boosted frame are characterized by

$$\begin{aligned}\Delta x' &= \gamma L \\ c\Delta t' &= \gamma \left(\frac{-vL}{c}\right).\end{aligned}$$

Next impose the Pythagorean relation

$$L'^2 = (c\Delta t')^2 + \Delta x'^2$$

and solve for L' to find:

$$L' = L \sqrt{\frac{1 + v^2/c^2}{1 - v^2/c^2}}$$

The same result can be attained by considering two different events such that $\Delta x = 0$ and $c\Delta t = L$. For this, we would have

$$\begin{aligned}\Delta x' &= \gamma(-v\Delta t) = \gamma \left(\frac{-vL}{c}\right) \\ c\Delta t' &= \gamma c\Delta t = \gamma L,\end{aligned}$$

which essentially repeats the above calculation.

Light Cone

For any event O in spacetime, the set of all events timelike with O is called the *light cone* of the event. Treating the event as the origin, the boundary of the light cone is defined by the lines $x = \pm ct$.

Worldline

A physical object moving in spacetime traces a path called a *worldline*. An object at rest has a straight worldline parallel to the ct -axis. When in motion, the slope of the worldline is given by

$$m = \frac{c\Delta t}{\Delta x} = \frac{c}{v}.$$

In general, the slope never deviates more than 45 degrees from the object's rest path.

Problem 7

Consider a worldline $q^\mu(\lambda)$ that passes through the event

$$E = (0, 0, 1 \text{ m}, 0),$$

where the motion is parameterized by

$$\begin{aligned} ct &= \lambda \\ x &= \lambda/2 \\ y &= (1 \text{ m}) \sqrt{1 - (\lambda/1 \text{ m})^2} \\ z &= 0. \end{aligned}$$

Find (i) the tangent vector at event E , and (ii) the tangent vector in a boosted frame with speed v in the x -direction.

Answer:

$$\begin{aligned} \left. \frac{dx^\mu}{d\lambda} \right|_E &= \left(1, \frac{1}{2}, \frac{-\lambda}{\sqrt{1 - (\lambda/1 \text{ m})^2}}, 0 \right) \Big|_{\lambda=0} \\ \frac{d}{d\lambda} (ct') &= \frac{d}{d\lambda} \left(\gamma \left(c\lambda - \frac{v\lambda}{2c} \right) \right) = \gamma \left(1 - \frac{v}{2c} \right) \\ \frac{d}{d\lambda} (x') &= \frac{d}{d\lambda} \left(\gamma \left(\frac{\lambda}{2} - \frac{v\lambda}{c} \right) \right) = \gamma \left(\frac{1}{2} - \frac{v}{c} \right) \end{aligned}$$

1.8 Paradoxes

Abusers of special relativity theory tend to be trapped by various paradoxes. Here we highlight several tricky situations that are resolved by applying the theory carefully.

Ladder Paradox

Consider a barn or similar building with ‘front’ and back’ doors, separated by length D , that can open or close instantly. Consider also a ladder of length L , oriented horizontally, that is able to move relativistically toward the front door of the barn. Things get interesting when we insist that $L > D$, meaning the ladder would be too large to fit in the barn at rest.

If the ladder has sufficient velocity, its length appears contracted according to observers in the barn, so there does exist (at least) an instant in which the entire ladder fits within the barn. On the other hand, if we ask an observer on the ladder what happens, the length of the barn is contracted, and it would seem that the ladder is far too long to fit inside the barn.

The so-called ‘ladder paradox’ addresses the question of whether or not the ladder truly fits within the barn. To illustrate further, suppose an observer at rest in the barn operates a switch to simultaneously close the front and back doors during a moment when the ladder is entirely enclosed, but then re-opens the rear door to avoid a collision. If we

imagine the frame of the ladder however, it might seem that the front door closes down on the ladder.

Whether or not the ladder contacts a door cannot be a matter of reference frame - certain facts must be invariant. In order to avoid a paradox, we must note that, in the reference frame of the ladder, the doors do not operate simultaneously. Instead, the rear door shuts and opens before the ladder has fully entered the barn. The front door shuts only after the ladder passes the doorway, by which time the front of the ladder has moved well-outside the back door. The ladder in fact touches no door, and there is no paradox.

For another take on the ladder paradox, let us explore what would happen if the back door remains closed, i.e. stopping the ladder while inside the barn. According to an observer in the barn, the ladder becomes fully enclosed so that the front door shuts. On hitting the back wall, the velocity instantaneously drops to zero and all length contraction vanishes. The ladder has no choice but to snap or explode between the two doors.

In the frame of the ladder however, contact with the rear door takes place with part of the ladder still not having reached the front door, so how can the ladder still be trapped within the barn and explode between the doors? The resolution involves noting that the ladder does not instantaneously decelerate as it does in the frame of the barn. The deceleration occurs sequentially along its length such that the front of the ladder stops first, whereas the back of the ladder hasn’t ‘received the news’ to decelerate until after it’s already inside the barn. Clearly an object won’t hold together under this circumstance, so the ladder blows up in both reference frames, as it should.

Twin Paradox

Imagine two identical twins, one of whom enters a relativistically-enabled rocket ship, travels the solar system at high speed, and returns to Earth. The Earth-bound twin, who considers himself at rest, would conclude that time dilation caused the traveling twin to age more slowly - he returns to Earth younger than he ‘should’ be. On the other hand however, the traveling twin could insist that Earth was moving through space with respect to his rocket, and the Earth-bound twin should be the younger of the two. Clearly, both twins can’t be correct, hence the ‘twin paradox’.

The resolution to the twin paradox is simple: the traveling twin must turn around at some point in his journey, so he moves with two inertial frames. The Earth-bound twin remains in the same frame throughout the experiment, thus the situation is not symmetric. The traveling twin unambiguously ages less than the Earth-bound twin.

To illustrate, suppose a twin named Albert always stays on Earth. A second twin, named Barney, travels 5 lightyears straight out into space at speed $c/2$ and returns back to Earth at speed $c/4$. Determine (i) how much Albert has aged during Barney's trip, and (ii) how much Barney has aged during his own trip.

According to Albert on Earth, the duration of the trip is the sum of two time intervals:

$$\Delta t_A = \frac{5 \text{ ly}}{c/2} + \frac{5 \text{ ly}}{c/4} = 30 \text{ yr}$$

According to Barney in the rocket, the two legs of the trip obey the Lorentz transformation

$$\begin{aligned}\Delta t'_{B_1} &= \frac{10 \text{ yr} - 5 \text{ ly}/2c}{\sqrt{1 - 1/4}} \\ \Delta t'_{B_2} &= \frac{20 \text{ yr} - 5 \text{ ly}/4c}{\sqrt{1 - 1/16}},\end{aligned}$$

so the total time spent is

$$\Delta t_B = 5 \left(\sqrt{3} + \sqrt{15} \right) \text{ yr} \approx 28.03 \text{ yr}.$$

Barney returns to Earth younger than Albert.

Problem 8

Use the Lorentz invariant $\Delta S'^2 = \Delta S^2$ for Albert and Barney's reference frames to derive the same answer for Δt_B . Answer:

$$\begin{aligned}-c^2 \Delta t_1'^2 &= -c^2 \Delta t_1^2 + \Delta x^2 \\ -c^2 \Delta t_2'^2 &= -c^2 \Delta t_2^2 + \Delta x^2 \\ c \Delta t' &= \sum_{j=1,2} \sqrt{c^2 \Delta t_j^2 - \Delta x^2}\end{aligned}$$

Problem 9

Suppose Barney gets a new ship that can travel at constant speed. What speed would be required to make the same round trip so that Barney ages 1% as much as Albert? Answer:

$$\begin{aligned}\Delta t' &= 1\% \Delta t \\ v &= c \sqrt{1 - (1/100)^2} \approx c \times 0.99994999875 \dots\end{aligned}$$

1.9 Relativistic Mechanics

Previously, we used the proper time τ to define several of the players for relativistic mechanics, namely the four-velocity U^μ , the four-momentum $P^\mu = mU^\mu$, and the proper force $F^\mu = m dU^\mu/d\tau$. It's instructive to check these items against the Lorentz transformation, so we tackle this now.

Relativistic Force

Consider an inertial frame O , along with a second inertial frame O' that moves with speed v with respect to O . Then, a particle of mass m , whose instantaneous velocity is $u > v$, u' respectively, is accelerated in the same direction as v .

According to the primed frame, the force on the particle is

$$F' = m \frac{du'}{dt'} = m \frac{du'}{dt} \frac{dt}{dt'}.$$

Replace u' using the velocity addition formula, and substitute dt via the Lorentz transformation:

$$F' = m \frac{d}{dt} \left(\frac{u - v}{1 - uv/c^2} \right) \frac{d}{dt'} \left(\gamma \left(t' + \frac{vx'}{c^2} \right) \right)$$

Keeping mind v and γ are constant, the above simplifies to

$$\begin{aligned}F' &= m \frac{du}{dt} \frac{(1 - uv/c^2) + (u - v)(v/c^2)}{(1 - uv/c^2)^2} \\ &\quad \times \gamma \left(1 + \frac{u'v}{c^2} \right).\end{aligned}$$

Next, tune the speed v such that $v = u$ always, thereby maintaining $u' = 0$ and $F' = F$. With this, the force becomes

$$\begin{aligned}F &= m \frac{1}{(1 - v^2/c^2)^{3/2}} \frac{dv}{dt} \\ &= m \frac{d}{dt} \left(\frac{v}{\sqrt{1 - v^2/c^2}} \right) \\ &= m \frac{d}{dt} (\gamma v).\end{aligned}$$

Relativistic Momentum

Knowing that force is classically equal to the time derivative of the momentum, the same should be true in the relativistic case. Of course, we need to include γ in the definition of the relativistic momentum:

$$\mathbf{p} = \gamma m \mathbf{v}$$

Problem 10

For low speeds $v \ll c$, show that, for one spatial dimension:

$$p \approx mv + \frac{mv^3}{2c^2} + \frac{3mv^5}{8c^4} + \frac{5mv^7}{16c^6} + \dots$$

Relativistic Energy

The work done on a particle, i.e. the energy added, is equal to the integral of the relativistic force projected into the displacement from point 1 to point 2:

$$\begin{aligned} E &= \int_{x_1}^{x_2} \mathbf{F} \cdot d\mathbf{x} = m \int_{t_1}^{t_2} \frac{d}{dt} (\gamma \mathbf{v}) \cdot \mathbf{v} dt \\ &= m \int_{t_1}^{t_2} \left(\frac{d\gamma}{dt} \mathbf{v} + \gamma \frac{d\mathbf{v}}{dt} \right) \cdot \mathbf{v} dt \end{aligned}$$

As a side problem, note that $\mathbf{v} \cdot d\mathbf{v}/dt$ is the same as $v dv/dt$:

$$\begin{aligned} v \frac{dv}{dt} &= \sqrt{v_x^2 + v_y^2 + v_z^2} \left(\frac{d}{dt} \sqrt{v_x^2 + v_y^2 + v_z^2} \right) \\ &= v_x \frac{dv_x}{dt} + v_y \frac{dv_y}{dt} + v_z \frac{dv_z}{dt} \\ &= \mathbf{v} \cdot \frac{d\mathbf{v}}{dt} \end{aligned}$$

Then, the energy integral can be written without vectors:

$$E = m \int_{t_1}^{t_2} \left(\frac{d\gamma}{dt} v + \gamma \frac{dv}{dt} \right) v dt = m \int_{t_1}^{t_2} \frac{d}{dt} (\gamma v) v dt$$

Using the template

$$\int U dV = UV| - \int V dU,$$

integrate the above by parts via

$$\begin{aligned} U &= u \\ dV &= \frac{d}{dt} (\gamma v) dt \\ dU &= du \\ V &= \gamma v, \end{aligned}$$

so the energy reads

$$E = \gamma m v^2 \Big|_{v_1}^{v_2} - m \int_{v_1}^{v_2} \frac{v}{\sqrt{1 - v^2/c^2}} dv,$$

and the remaining integral is solved by elementary means.

Without loss of generality, assume the particle starts from rest and has final velocity v . After simplifying, the relativistic energy resolves to:

$$\begin{aligned} E &= \gamma m v^2 + \frac{1}{\gamma} m c^2 \\ &= \gamma m c^2 \end{aligned}$$

Rest and Kinetic Energy

The relativistic energy can be regarded as the sum of the rest energy and the kinetic portion via

$$E = E_0 + T,$$

where

$$E_0 = m c^2$$

and

$$T = m c^2 (\gamma - 1).$$

Problem 11

For slow speeds $v \ll c$, show that:

$$T \approx \frac{m v^2}{2} \left(1 + \frac{3v^2}{4c^2} + \frac{5v^4}{8c^4} + \dots \right)$$

Energy-Momentum Relation

Starting with the relativistic momentum $\mathbf{p} = \gamma m \mathbf{v}$, square both sides and replace v^2 using the gamma factor:

$$\begin{aligned} p^2 &= \gamma^2 m^2 v^2 = \gamma^2 m^2 c^2 \left(1 - \frac{1}{\gamma^2} \right) \\ &= \frac{E^2}{c^2} - m^2 c^2 \end{aligned}$$

Solve for E to recover the energy-momentum relation:

$$E = \sqrt{p^2 c^2 + m^2 c^4}$$

Collision of Identical Masses

Consider an observer at rest who prepares two identical bodies (balls of clay perhaps), each of mass m , to move toward each other in opposing directions. If each body has the same initial kinetic energy T , the total energy of the system is

$$E = 2 (m c^2 + T).$$

When two objects collide and merge together without losing energy to heat, sound, etc., the resulting mass M is at rest, having energy $E = M c^2$. Use energy conservation to eliminate E to find

$$M = 2m + \frac{2T}{c^2}.$$

Evidently, the resulting object has mass M that is greater than the sum $m + m$ by a correction $2T/c^2$. Or, replace T to also find

$$M = 2m + 2m (\gamma - 1) = 2\gamma m.$$

Note that the combination γm is sometimes referred to as the ‘relativistic mass’. This is a confusing practice however, as the mass m is an invariant quantity. For this reason, the term ‘rest mass’ is completely synonymous with plain ‘mass’

Decay of a Motionless Particle

Suppose a particle of mass M at rest spontaneously decays into two particles, where each has mass $m_{1,2}$, momentum $\mathbf{p}_{1,2}$, and energy $E_{1,2}$.

From conservation of momentum we can write

$$\mathbf{p}_1 + \mathbf{p}_2 = 0,$$

and from conservation of energy we also have

$$Mc^2 = \sqrt{p_1^2 c^2 + m_1^2 c^4} + \sqrt{p_2^2 c^2 + m_2^2 c^4}.$$

Square both sides of the energy equations and simplify to find

$$M^2 = m_1^2 + m_2^2 + 2 \left(\frac{E_1 E_2}{c^4} - \frac{\mathbf{p}_1 \cdot \mathbf{p}_2}{c^2} \right).$$

The above is straightforwardly solved for E_1 or E_2 . To find E_1 , replace E_2 via

$$E_2 = Mc^2 - E_1,$$

and replace $\mathbf{p}_1 \cdot \mathbf{p}_2$ with $-p_1^2$. With this, we find

$$E_1 = c^2 \left(\frac{M^2 + m_1^2 - m_2^2}{2M} \right),$$

and correspondingly

$$E_2 = c^2 \left(\frac{M^2 - m_1^2 + m_2^2}{2M} \right).$$

Relativistic Doppler Shift 1+1D

Consider an inertial frame O and a second inertial frame O' that moves with speed v with respect to O . Fixed to the origin of O is a machine that emits light at proper frequency f .

Working in one spatial dimension, write the invariant

$$-E't' + p'x' = -Et + px,$$

and recast all E - and p - variables in terms of frequency:

$$-f't' + f'\frac{x'}{c} = -ft + f\frac{x}{c}$$

Use the Lorentz transformation to replace each instance of x' and t' to find, after simplifying:

$$f' \gamma (-ct + x) \left(1 + \frac{v}{c} \right) = f(-ct + x)$$

Replace γ using its definition to ultimately find

$$f' = f \frac{\sqrt{1 - v/c}}{\sqrt{1 + v/c}},$$

known as the *relativistic doppler shift*.

Since the frame O' is moving away from O , the observed frequency obeys $f' < f$, an effect called *redshift*. On the other hand, we can reverse the sign on v to have the observer in O' approaching the source, where we get $f' > f$, called *blueshift*.

Relativistic Doppler Shift 1+2D

Consider a light ray propagating in space with momentum vector $\mathbf{p} = \hbar \mathbf{k}$ as observed in an inertial reference frame O . In a second reference frame O' with relative velocity $v \hat{\mathbf{x}}$, the momentum there reads $\mathbf{p}' = \hbar \mathbf{k}'$.

As a sinusoidal wave, the electric field components E and E' (not to be confused with energy) of the light ray read

$$E = E_0 \cos(\mathbf{k} \cdot \mathbf{x} - \omega t)$$

$$E' = E'_0 \cos(\mathbf{k}' \cdot \mathbf{x}' - \omega' t')$$

We'll now consider two-dimensional motion of such a light ray. That is, let the relative motion among O , O' be along the x -direction, thus the vectors $\mathbf{k}$, $\mathbf{k}'$ live somewhere in the xy -plane. Each $\mathbf{k}$ -vector makes a respective angle θ , θ' with respect to the x - and x' -axes. With this, the above becomes:

$$E = E_0 \cos \left(2\pi \left(\frac{x \cos(\theta) + y \sin(\theta)}{\lambda} - ft \right) \right)$$

$$E' = E'_0 \cos \left(2\pi \left(\frac{x' \cos(\theta') + y' \sin(\theta')}{\lambda'} - f't' \right) \right)$$

Use the Lorentz transformation

$$ct' = \gamma \left(ct - \frac{vx}{c} \right)$$

$$x' = \gamma (x - vt)$$

$$y' = y$$

while remembering that the argument inside each cosine function is the same invariant. Matching coefficients on x , y , and t gives three simultaneous equations, after simplifying:

$$\begin{aligned}\frac{\gamma}{\lambda'} \left(\cos(\theta') + \frac{v}{c} \right) &= \frac{\cos(\theta)}{\lambda} \\ \frac{\sin(\theta')}{\lambda'} &= \frac{\sin(\theta)}{\lambda} \\ f' \gamma \left(1 + \frac{v}{c} \cos(\theta') \right) &= f\end{aligned}$$

The third equation is the generalization of the relativistic doppler shift in 1 + 1 dimensions. For $\theta' = 0$ we reproduce the result derived for 1 + 1 dimensions. The case $\theta' = \pi$ corresponds to the source and the observer approaching each other.

1.10 Relativistic Lagrangian

A version of Lagrangian mechanics that accomodates special relativity begins with the action integral

$$S = -mc^2 \int_{\tau_1}^{\tau_2} d\tau,$$

where m is the mass of a body and τ is the proper time. Using $\gamma = dt/d\tau$, the same integral is recast in a non-rest frame:

$$S = -mc^2 \int_{t_1}^{t_2} \sqrt{1 - \frac{\dot{q}^2}{c^2}} dt,$$

where $\dot{q}$ is the relative speed that we ordinarily label as v .

Introducing a potential energy term $U(q)$, we pick out the relativistic lagrangian:

$$L = -mc^2 \sqrt{1 - \frac{\dot{q}^2}{c^2}} - U(q)$$

At low speeds $\dot{q} \ll c$, the relativistic Lagrangian reduces to

$$L \approx -mc^2 + \frac{mc^2}{2} \frac{\dot{q}^2}{c^2} + \frac{mc^2}{8} \frac{\dot{q}^4}{c^4} + \dots - U(q),$$

which subtracts off the rest energy and shows that the trailing kinetic terms are negligible:

$$L \approx \frac{1}{2} m \dot{q}^2 - U(q)$$

Equations of Motion

Inserting the relativistic Lagrangian L into the Euler-Lagrange equation

$$\frac{\partial L}{\partial \mathbf{q}} = \frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{q}}},$$

we find

$$-\frac{\partial U}{\partial \mathbf{q}} = m \frac{d}{dt} \left(\frac{\dot{\mathbf{q}}}{\sqrt{1 - \dot{q}^2/c^2}} \right) = \mathbf{F}(\mathbf{q}),$$

giving the formula for the relativistic momentum:

$$\mathbf{p} = \gamma m \dot{\mathbf{q}} = \frac{\partial L}{\partial \dot{\mathbf{q}}}$$

The statement of energy conservation is written

$$\begin{aligned}E &= \dot{\mathbf{q}} \cdot \frac{\partial L}{\partial \dot{\mathbf{q}}} - L = \dot{\mathbf{q}} \cdot \mathbf{p} + \frac{mc^2}{\gamma} + U(\mathbf{q}) \\ &= m \left(\gamma v^2 + \frac{c^2}{\gamma} \right) + U(\mathbf{q}) \\ &= \gamma mc^2 + U(\mathbf{q}),\end{aligned}$$

which is the sum of the rest, kinetic, and potential energies.

1.11 Relativistic Collisions

We've seen in the case of collision of identical masses and the decay of a massive particle that mass is not conserved in the relativistic scheme. Instead, the relativistic energy

$$E = \gamma mc^2 = \sqrt{p^2 c^2 + m^2 c^4}$$

is the conserved quantity, which tangles up mass with energy.

To go with mass-energy conservation, the relativistic momentum $\mathbf{p} = \gamma m \mathbf{v}$ is also conserved. Conveniently, these quantities are the components of the four-momentum

$$P^\mu = \left(\frac{E}{c}, \mathbf{p} \right)$$

which must be conserved across spacetime events, and in particular, collisions. Before and after a relativistic collision of particles, we can work safely with the conservation relation

$$P_{\text{initial}}^\mu = P_{\text{final}}^\mu.$$

Compton Scattering

Consider a particle of mass m at rest in a laboratory reference frame. A prepared photon of wavelength λ_i transfers energy and momentum to the particle if a collision takes place. As a result, a new photon with increased wavelength λ_f emerges at angle θ with respect to the incident axis, and the particle carries the photon's missing momentum away at angle $-\phi$ with respect to the incident axis at speed v .

Setting up four-vector notation for the initial and 'final' states of collision, write:

$$\text{Initial photon state: } \gamma_i = (E_i/c, \hbar \mathbf{k}_i)$$

$$\text{Initial particle state: } M = (mc, 0)$$

$$\text{Final photon state: } \gamma_f = (E_f/c, \hbar \mathbf{k}_f)$$

$$\text{Final particle state: } P = (E_p/c, \mathbf{p})$$

Using the above, the collision is entirely characterized by

$$\gamma_i + M = \gamma_f + P$$

Many relations for Compton scattering can be derived from the above. For practice, multiply each side by the contravariant copy of itself and use the distributive property of contraction to write

$$\gamma_i^\alpha - m^2 c^2 + \gamma_i M = \gamma_f^\alpha - m^2 c^2 + \gamma_f P$$

$$\gamma_i M = \gamma_f P.$$

The norm of a photon four-vector is zero, and the Minkowski metric is responsible for the signage. This process is the Minkowski space equivalent to 'squaring both sides' of an equation.

By the above process, the statement of energy conservation is recovered by projecting M across the conservation statement, giving

$$-E_i m - m^2 c^2 = -E_f m - E_p m,$$

or, as expected:

$$E_i + mc^2 = E_f + E_p$$

To derive the Compton scattering formula, rearrange to write

$$\gamma_i - \gamma_f = P - M$$

and square both sides:

$$\gamma_i^\alpha + \gamma_f^\alpha - 2\gamma_i \gamma_f = -2m^2 c^2 - 2PM$$

$$2 \frac{E_i E_f}{c^2} - 2\hbar^2 \mathbf{k}_i \cdot \mathbf{k}_f = -2m^2 c^2 + 2E_p m$$

$$\frac{E_i E_f}{c^2} (1 - \cos(\theta)) = m(E_i - E_f)$$

Rearrange to attain the Compton scattering formula:

$$\frac{1}{E_f} - \frac{1}{E_i} = \frac{1 - \cos(\theta)}{mc^2}$$

In terms of wavelength, the Compton scattering formula reads

$$\lambda_f - \lambda_i = \frac{h}{mc} (1 - \cos(\theta)),$$

where h/mc is the *Compton wavelength*.

A formula involving the particle recoil angle ϕ is attained starting with

$$\gamma_f - \gamma_i + P = M,$$

where squaring both sides and simplifying leads to

$$-\gamma_i \gamma_f - \gamma_i P + \gamma_f P = 0.$$

Replace $\gamma_f P$ with $\gamma_i M$ and keep simplifying:

$$\frac{E_i E_f}{c^2} (1 - \cos(\theta)) + \frac{E_i E_p}{c^2} - \hbar \mathbf{k}_i \cdot \mathbf{p} - E_i m = 0$$

Next evaluate the remaining dot product and cancel E_i to ultimately get

$$E_f - mc^2 + E_p = E_f \cos(\theta) + pc \cos(\phi),$$

or

$$E_i = E_f \cos(\theta) + pc \cos(\phi).$$

Of course, we can divide this by c to recover the statement of momentum conservation along the incident axis:

$$\hbar k_i = \hbar k_f \cos(\theta) + p \cos(\phi)$$

Correspondingly, we also have

$$0 = \hbar k_f \sin(\theta) - p \sin(\phi).$$

Problem 12

Use the Compton scattering formula to derive:

$$E_f = \frac{mc^2}{1 - \cos(\theta) + mc^2/E_i}$$

Problem 13

Let T denote the kinetic energy carried off by the particle. Derive the formula:

$$T = \frac{E_i^2}{E_i + mc^2/(1 - \cos(\theta))}$$

Problem 14

Derive the formula for the recoil angle:

$$\cos(\phi) = \frac{T(1 + mc^2/E_i)}{\sqrt{(mc^2 + T)^2 - m^2c^4}}$$

Problem 15

Derive the kinetic energy formula:

$$T = \frac{2mc^2 \cos^2(\phi)}{(1 + mc^2/E_i)^2 - \cos^2(\phi)}$$

Problem 16

Derive the ratio summarizing the photon on the left and the target on the right:

$$\frac{E_f}{E_i} = \frac{E_p - pc \cos(\phi)}{mc^2}$$

Problem 17

Derive the formula relating θ to ϕ :

$$\tan(\phi) = \left(\frac{mc^2}{E_i + mc^2} \right) \left(\frac{\sin(\theta)}{1 - \cos(\theta)} \right)$$

Pion Production

Consider a single ‘target’ proton of mass m_p that is stationary with respect to a laboratory reference frame. An incoming photon of energy E_γ collides with the proton and is absorbed. A π^0 particle, or *pion*, is created in the collision and carries away momentum $\mathbf{p}_\pi$ at angle θ with respect to the incident axis. The proton also recoils with momentum $\mathbf{p}_p$ at an angle $-\phi$.

Setting up four-vector notation for the initial and ‘final’ states of collision, write:

$$\text{Initial photon state: } \gamma = (E_\gamma/c, \hbar \mathbf{k}_\gamma)$$

$$\text{Initial proton state: } M = (m_p c, 0)$$

$$\text{Pion state: } \pi = (E_\pi/c, \mathbf{p}_\pi)$$

$$\text{Final proton state: } P = (E_p/c, \mathbf{p}_p)$$

From the above, let us plan to solve for the pion momentum p_π in terms of incoming energy E_γ , particle masses m_p , m_π , and the pion recoil angle θ . The collision is entirely characterized by

$$\gamma + M = \pi + P.$$

It’s a worthwhile exercise to explicitly attain the statements of energy and momentum conservation from the

four-vectors. First note the collision is also characterized by $\gamma - \pi = P - M$, which implies

$$-m_\pi^2 c^2 - 2\gamma\pi = -2m_p^2 c^2 + 2m_p E_p.$$

Meanwhile, the four-vector relation $\gamma + M - \pi = P$ expands to

$$-m_\pi^2 c^2 - 2E_\gamma m_p + 2E_\pi m_p - 2\gamma\pi = 0.$$

Eliminate the common terms between the two results to find the energy conservation statement

$$E_\gamma + m_p c^2 = E_\pi + E_p.$$

Note this result is attained trivially by distributing M across $\gamma + M = \pi + P$.

To gain the statement of momentum conservation, expand $\gamma + M = \pi + P$ to find

$$2\gamma M = -m_\pi^2 c^2 + 2\pi P.$$

Also expand $\gamma - \pi = P - M$ to get

$$-m_\pi^2 c^2 - 2\gamma\pi - 2\gamma P + 2\pi P = 0,$$

and eliminate common terms to learn

$$\gamma M = \gamma P + \gamma\pi,$$

which could have been trivially attained. Simplifying, we have

$$\begin{aligned} -E_\gamma m_p &= -\frac{E_\gamma E_p}{c^2} + \hbar \mathbf{k}_\gamma \cdot \mathbf{p}_p - \frac{E_\gamma E_\pi}{c^2} + \hbar \mathbf{k}_\gamma \cdot \mathbf{p}_\pi \\ -m_p c &= -E_p/c + p_p \cos(\phi) - E_\pi/c + p_\pi \cos(\theta), \end{aligned}$$

or

$$\frac{E_p + E_\pi - m_p c^2}{c} = p_\pi \cos(\theta) + p_p \cos(\phi),$$

readily producing the statement of momentum conservation:

$$\hbar k_\gamma = p_\pi \cos(\theta) + p_p \cos(\phi)$$

Correspondingly, we also have

$$0 = p_\pi \sin(\theta) - p_p \sin(\phi).$$

Returning to the issue of solving for p_π , take the four-vector relation

$$\gamma - \pi = P - M$$

and simplify like mad until the form

$$X + Y p_\pi = Z E_\pi$$

is attained, where:

$$\begin{aligned} X &= \frac{m_\pi^2 c^2}{2} + m_p E_\gamma \\ Y &= \frac{\hbar E_\gamma \cos(\theta)}{c} \\ Z &= \frac{E_\gamma}{c^2} + m_p \end{aligned}$$

Next, square both sides of the above (the regular way) to get

$$X^2 + 2XYp_\pi + Y^2 p_\pi^2 = Z^2 E_\pi^2,$$

and replace the pion energy term with the energy-momentum relation. Simplifying again, one finds

$$Ap_\pi^2 + Bp_\pi + C = 0,$$

where

$$\begin{aligned} A &= Y^2 + Z^2 c^2 \\ B &= 2XY \\ C &= X^2 - Z^2 m_\pi^2 c^4. \end{aligned}$$

Finally, the pion momentum is

$$p_\pi = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}.$$

Antiproton Production

At nonrelativistic speeds, two protons that collide will exchange energy and momentum in a more-or-less classical fashion. At relativistic speeds however, two colliding protons have a change to produce an extra proton along with an antiproton.

Such a reaction can be tracked with momentum four-vectors via

$$P_i + P_t = P_1 + P_2 + P_3 + P_A,$$

where P_i represents an incoming proton that collides with a stationary 'target' proton represented by P_t . The products P_{123} are the three ordinary protons, and P_A represents the antiproton.

The question on hand is, what is the minimum energy E_i that the incoming proton needs in order for the reaction to take place? Proceed by squaring both sides of the above, where on the left we find

$$\begin{aligned} (P_i + P_t)^2 &= P_i^2 + P_t^2 + 2P_i P_t \\ &= -2m_p^2 c^2 - 2E_i m_p \end{aligned}$$

To easily square the right side of the four-vector equation, exploit the invariance of the four-momentum norm to choose the center-of-momentum reference frame of the products. In such a frame, the three protons and the antiproton emerge from the collision *at rest* if we are looking for the minimal E_i . Thus, we can write

$$(P_1 + P_2 + P_3 + P_A)^2 = -4^2 m_p^2 c^2.$$

Putting both results together, we find

$$\begin{aligned} (P_i + P_t)^2 &= (P_1 + P_2 + P_3 + P_A)^2 \\ -2m_p^2 c^2 - 2E_i m_p &= -16m_p^2 c^2, \end{aligned}$$

or

$$E_i = 7m_p c^2.$$

Of course, this result includes the rest mass of the incident proton. The kinetic energy supplied must be at minimum

$$T_i = 6m_p c^2.$$

Suppose a better-equipped laboratory prepares the same reaction using no stationary target proton, but instead using two incoming protons of equal and opposite speeds. In this case, we have

$$(P_1 + P_2)^2 = -2m_p^2 c^2 - \frac{2E_i^2}{c^2} - 2\left(\frac{E_i^2}{c^2} - m_p^2 c^2\right).$$

The center-of-momentum frame for the products is the laboratory frame itself, thus by the invariance of the norm of the four-momentum we write

$$-2m_p^2 c^2 - \frac{2E_i^2}{c^2} - 2\left(\frac{E_i^2}{c^2} - m_p^2 c^2\right) = -16m_p^2 c^2,$$

reducing to

$$E_i = 2m_p c^2.$$

This is the energy of each proton, thus the laboratory must supply a total of twice this amount minus the rest mass of the incident protons:

$$E = 4m_p c^2 - 2m_p c^2 = 2m_p c^2$$

1.12 Tensor Formalism

Returning to the generalized coordinate transformation that motivates the Lorentz transformation, recall

$$q^{\mu'} = \frac{\partial q^{\mu'}}{\partial q^\mu} q^\mu.$$

As it turns out, the most general coordinate transformation contains ten distinct members: (i) four translations in $1 + 3$ spacetime dimensions, (ii) three Lorentz boosts, (iii) three rotations about spatial axes. Together, these are called the *Poincare group*.

For Minkowski space, the derivative term in the coordinate transformation is relabeled to the *Lorentz transformation tensor* Λ such that

$$q^{\mu'} = \Lambda_{\mu}^{\mu'} q^{\mu}.$$

Lorentz Transformation Tensor

We can (re)discover the properties of Λ by considering an inertial frame O and a second inertial frame O' that moves at constant speed v along the x -axis with respect to O . Writing the spacetime interval in both inertial frames tells us

$$\eta_{\alpha\gamma} dq^{\alpha} dq^{\gamma} = \eta_{\mu\nu} \Lambda_{\alpha}^{\mu} \Lambda_{\gamma}^{\nu} dq^{\alpha} dq^{\gamma},$$

or

$$\eta_{\alpha\gamma} = \eta_{\mu\nu} \Lambda_{\alpha}^{\mu} \Lambda_{\gamma}^{\nu}.$$

From the above, the case $\alpha = \beta = 0$ tells us

$$1 = \left(\Lambda_0^0\right)^2 - \left(\Lambda_0^1\right)^2,$$

and meanwhile $\alpha = \beta = 1$ gives

$$1 = \left(\Lambda_1^1\right)^2 - \left(\Lambda_1^0\right)^2.$$

These results encourage the association

$$\begin{aligned}\Lambda_0^0 &= \Lambda_1^1 = \cosh(\phi) = \gamma \\ \Lambda_0^1 &= \Lambda_1^0 = -\sinh(\phi) = -\gamma \frac{v}{c},\end{aligned}$$

where ϕ is the rapidity parameter defined by

$$\tanh(\phi) = \frac{v}{c}.$$

Multiple Boosts

In addition to the inertial frames O , O' , consider a third inertial frame O'' that is double-boosted with respect to O with rapidity parameter ϕ' . For this situation, we write the double coordinate transformation

$$q^{\mu''} = \Lambda(\phi')_{\mu'}^{\mu''} \Lambda(\phi)_{\mu}^{\mu'} q^{\mu}.$$

The case $\mu'' = 0$ gives

$$\begin{aligned}ct'' &= (\cosh(\phi') \cosh(\phi) + \sinh(\phi') \sinh(\phi)) ct \\ &\quad - (\cosh(\phi') \sinh(\phi) + \sinh(\phi') \cosh(\phi)) x,\end{aligned}$$

which simplifies to

$$ct'' = \cosh(\phi' + \phi) ct - \sinh(\phi' + \phi) x.$$

Note that a similar exercise can be done for the case $\mu'' = 1$ to get a similar result.

As expected, the rapidity is cumulative across multiple boosts. Defining the sum

$$\Phi = \phi' + \phi,$$

we can say of the Lorentz transformation tensor

$$\Lambda(\Phi)_{\mu}^{\mu''} = \Lambda(\phi')_{\mu'}^{\mu''} \Lambda(\phi)_{\mu}^{\mu'}.$$

Inverse Lorentz Transformation

In tensor notation, the inverse Lorentz transformation is given by

$$q^{\mu} = \tilde{\Lambda}_{\mu'}^{\mu} q^{\mu'}.$$

Substituting $q^{\mu'}$ gives

$$q^{\mu} = \tilde{\Lambda}_{\mu'}^{\mu} \Lambda_{\nu}^{\mu'} q^{\nu}.$$

The factor multiplying q^{ν} can only be the Kronecker delta δ_{ν}^{μ} , telling us

$$\tilde{\Lambda}_{\mu'}^{\mu} \Lambda_{\nu}^{\mu'} = \delta_{\nu}^{\mu}.$$

2 Rindler Space

2.1 Uniform Acceleration

Minkowski space is the habitat of inertial reference frames in relative motion with constant speed. If the speed is not constant, corresponding to an accelerated observer, we may treat the accelerated frame as an instantaneous boosted frame. With this, we can deploy the full apparatus of Minkowski space at each instant.

Consider an inertial frame O and a second frame O' that is accelerated with respect to O such that

$$\frac{dv}{dt} = \tilde{a} ,$$

where $\tilde{a}$ is constant. The rapidity ϕ is recast as a function of the proper time of the accelerated frame as $\phi(\tau)$. Likewise, the nontrivial components of the Lorentz transformation become

$$\begin{aligned}\Lambda_0^0 &= \Lambda_1^1 = \cosh(\phi(\tau)) = \gamma(\tau) \\ \Lambda_0^1 &= \Lambda_1^0 = -\sinh(\phi(\tau)) = -\gamma(\tau) \frac{v(\tau)}{c} ,\end{aligned}$$

where

$$q^{\mu'}(\tau) = \Lambda(\tau)_{\mu}^{\mu'} q^{\mu}(\tau) .$$

Working in 1 + 1 dimensions, let us ignore any y - and z -components for simplicity, and all four-vectors become two-vectors for this purpose. Writing the two-velocity, we have

$$U^{\mu} = (\gamma c, \gamma v) = (c \cosh(\phi(\tau)), c \sinh(\phi(\tau))) .$$

Two-Acceleration

The proper acceleration is the τ -derivative of U^{μ} . For constant acceleration, this is

$$a^{\mu} = \frac{dU^{\mu}}{d\tau} = c \left(\sinh(\phi) \frac{d\phi}{d\tau}, \cosh(\phi) \frac{d\phi}{d\tau} \right) ,$$

where a^{μ} is also known as the *two-acceleration*.

The x -component of the four-acceleration is in fact constant, which we prove by substitution via

$$\begin{aligned}a^x &= c \cosh(\phi) \frac{d\phi}{d\tau} = \gamma \frac{d}{d\tau} \operatorname{arctanh}\left(\frac{v}{c}\right) \\ &= \frac{\gamma}{1 - v^2/c^2} \frac{dv}{d\tau} = \frac{dv}{dt} = \tilde{a} .\end{aligned}$$

Knowing a^x is constant, use the Lorentz transformation

$$a^{\mu'} = \Lambda_{\mu}^{\mu'} a^{\mu}$$

to find

$$\tilde{a} = a^{x'} = \Lambda_0^1 a^t + \Lambda_1^1 a^x = c \frac{d\phi}{d\tau} .$$

Evidently, the rapidity ϕ is an affine parameter with respect to the proper time:

$$\phi(\tau) = \frac{\tilde{a}\tau}{c} + \phi_0$$

For convenience we let $\phi_0 = 0$.

Also, we see that a contraction with the two-velocity is zero

$$a^{\mu} U_{\mu} = 0 ,$$

and that the norm of a^{μ} is a spacetime invariant

$$a^{\mu} a_{\mu} = c^2 \left(\frac{d\phi}{d\tau} \right)^2 = \tilde{a}^2 .$$

In terms of the proper time, the two-velocity reads

$$U^{\mu} = c (\cosh(\tilde{a}\tau/c), \sinh(\tilde{a}\tau/c)) .$$

Integrate once with respect to proper time to recover a formula for the coordinates $q^{\mu}(\tau)$:

$$\begin{aligned}ct(\tau) &= \frac{c^2}{\tilde{a}} \sinh(\tilde{a}\tau/c) \\ x(\tau) &= \frac{c^2}{\tilde{a}} \cosh(\tilde{a}\tau/c) - \frac{c^2}{\tilde{a}}\end{aligned}$$

Note the integration constants are chosen so the origin occurs at $\tau = 0$.

Problem 18

For low speeds $v \ll c$, show that:

$$x(\tau) = \frac{c^2}{\tilde{a}} \left(1 + \frac{1}{2} \left(\frac{\tilde{a}\tau}{c} \right)^2 + \cdots - 1 \right) \approx \frac{1}{2} \tilde{a} \tau^2$$

Accelerated Particle

Starting at $t = 0$, suppose a particle of mass m is accelerated uniformly in the x -direction. Using relativistic mechanics, we write the force

$$F = m\tilde{a} = m \frac{d}{dt} (\gamma v) ,$$

where F is constant.

Setting up an integral for the velocity as a function of t , we have

$$\tilde{a} = \frac{d}{dt} \left(\frac{v}{\sqrt{1 - v^2/c^2}} \right) ,$$

simplifying to

$$v(t) = \frac{c}{\sqrt{1 + c^2/\tilde{a}^2 t^2}}.$$

Integrating once more with respect to time gives the position:

$$x(t) = ct\sqrt{1 + \frac{c^2}{\tilde{a}^2 t^2}} - \frac{c^2}{\tilde{a}}$$

Switching domain to proper time, the above reads

$$x(t(\tau)) = ct\sqrt{1 + 1/\sinh^2(\tilde{a}\tau/c)} - \frac{c^2}{\tilde{a}},$$

which simplifies to the same form $x(\tau)$ above.

The kinetic energy T gained by the accelerated particle is given by $T = E - E_0$. To gain another formula for T , first write

$$\begin{aligned} q^\mu P_\mu &= -Et + px \\ &= -\gamma mc^2 t + \gamma m v x \\ &= -mc^2 t', \end{aligned}$$

and take a τ -derivative to get

$$\begin{aligned} U^\mu P_\mu + q^\mu F_\mu &= -mc^2 \frac{dt'}{d\tau} \\ -mc^2 - \gamma F v t + \gamma F x &= -\gamma mc^2 \\ -E_0 + F\gamma(x - vt) &= -E \end{aligned}$$

Evidently, the kinetic energy is given by

$$T = -F\gamma(x - vt) = -Fx'.$$

2.2 Rindler Coordinates

Accelerated Rocket

Consider an inertial frame O and a second frame O' , such as a rocket mapped by coordinates $c\tau, x'$, that undergoes uniform acceleration $\tilde{a}$. The floor is always at $x' = 0$. Assume the floor of the rocket is observed at $x = 0$ at time $t = \tau = 0$, and its position is observed at $x(t)$ as time evolves.

Rindler Coordinates

By tradition, we replace x' with ξ (Greek ξ) to denote the height from floor as observed within the rocket. The ordered pair $(c\tau, \xi)$ is the *Rindler coordinate system*. The task now is to come up with the equations $t(\tau, \xi), x(\tau, \xi)$ that allows a stationary observer to analyze motion in the frame of the rocket.

To proceed, we acknowledge there are *two* contributions to the total boost at $x' = \xi$. First, the floor of the rocket obeys:

$$\begin{aligned} ct_{\text{floor}}(\tau) &= \frac{c^2}{\tilde{a}} \sinh(\tilde{a}\tau/c) \\ x_{\text{floor}}(\tau) &= \frac{c^2}{\tilde{a}} \cosh(\tilde{a}\tau/c) - \frac{c^2}{\tilde{a}} \end{aligned}$$

Second, the location $x' = \xi$ has its own Lorentz boost with respect to the floor, thus we invoke the inverse Lorentz transformation to write

$$\begin{aligned} ct(\tau, \xi) &= \gamma(ct_0 + vx_0/c) + ct_{\text{floor}}(\tau) \\ x(\tau, \xi) &= \gamma(x_0 + vt_0) + x_{\text{floor}}(\tau), \end{aligned}$$

where $t_0 = 0$ and $x_0 = \xi$ by construction.

Equations of Motion

Simplifying the above gives a pair of equations of motion to describe all points within the moving rocket:

$$\begin{aligned} ct(\tau, \xi) &= \left(\frac{c^2}{\tilde{a}} + \xi\right) \sinh(\tilde{a}\tau/c) \\ x(\tau, \xi) &= \left(\frac{c^2}{\tilde{a}} + \xi\right) \cosh(\tilde{a}\tau/c) - \frac{c^2}{\tilde{a}} \end{aligned}$$

The equations of motion can be inverted to solve for the Rindler coordinates in terms of the Minkowski coordinates:

$$\begin{aligned} c\tau(t, x) &= \frac{c^2}{\tilde{a}} \operatorname{arctanh}\left(\frac{ct}{x + c^2/\tilde{a}}\right) \\ \xi(t, x) &= -\frac{c^2}{\tilde{a}} + \sqrt{(x + c^2/\tilde{a})^2 - (ct)^2} \end{aligned}$$

The τ -derivative of the equations of motion yields the components of the two-velocity:

$$\begin{aligned} U^0 &= c \left(1 + \frac{\xi\tilde{a}}{c^2}\right) \cosh(\tilde{a}\tau/c) \\ U^1 &= c \left(1 + \frac{\xi\tilde{a}}{c^2}\right) \sinh(\tilde{a}\tau/c) \end{aligned}$$

Rindler Diagram

Much like a Minkowski diagram, a *Rindler diagram* is a visualization of the equations of motion within the frame of the accelerated frame. Starting with the equations for $ct(\tau, \xi), x(\tau, \xi)$, we combine these to write

$$x = ct \coth(\tilde{a}\tau/c) - \frac{c^2}{\tilde{a}}.$$

From the above we discern that contours of constant τ correspond to straight lines in the Rindler diagram. Changing τ changes the slope of each line, but all of them pass through $x = -c^2/\tilde{a}$ at $t = 0$. For a more intuitive sense of the τ -countours, solve for ct to write

$$ct = x \tanh(\tilde{a}\tau/c) + \frac{c^2}{\tilde{a}} \tanh(\tilde{a}\tau/c) .$$

Start again with the equations of motion and eliminate the hyperbolic terms to find

$$\left(\frac{x + c^2/\tilde{a}}{c^2/\tilde{a} + \xi} \right)^2 - \left(\frac{ct}{c^2/\tilde{a} + \xi} \right)^2 = 1 .$$

For countours of constant ξ , the above describes a family of hyperbolas offset distance $-c^2/\tilde{a}$ from the origin.

Figure 34.2 is the so-called Rindler diagram that shows contours of constant τ , ξ in the context of a stationary reference frame.

The equations of motion are non-analytic at $x = -c^2/\tilde{a}$, and events with $x < -c^2/\tilde{a}$ are tossed out of the analysis. Thus the Rindler spacetime does not cover all of Minkowski spacetime. Certain events that take place in the inertial frame can *never* have a causal relation with events in the accelerated frame.

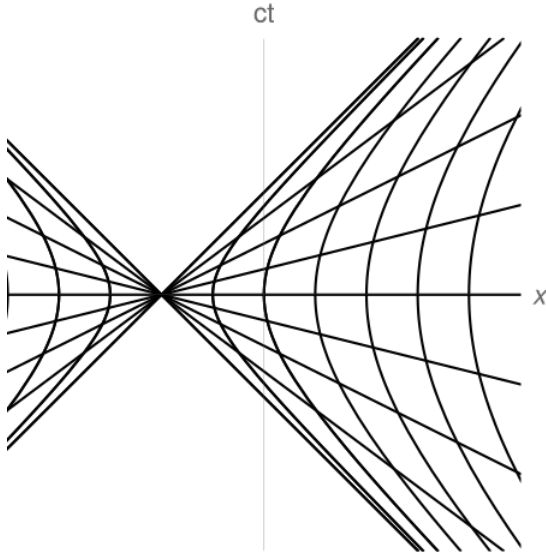


Figure 34.2. Rindler diagram showing contours of constant τ (lines) and constant ξ (hyperbolas). Straight lines intersect at $x = -c^2/\tilde{a}$.

Rindler Metric

In terms of the Rindler coordinates, the differential space-time interval dS^2 is written:

$$\begin{aligned} dS^2 = -c^2 d\tau^2 + dx^2 = & - \left(\left(\frac{c^2}{\tilde{a}} + \xi \right) \frac{\tilde{a}}{c} \cosh\left(\frac{\tilde{a}\tau}{c}\right) d\tau + \sinh\left(\frac{\tilde{a}\tau}{c}\right) d\xi \right)^2 \\ & + \left(\left(\frac{c^2}{\tilde{a}} + \xi \right) \frac{\tilde{a}}{c} \sinh\left(\frac{\tilde{a}\tau}{c}\right) d\tau + \cosh\left(\frac{\tilde{a}\tau}{c}\right) d\xi \right)^2 \end{aligned}$$

After simplifying, we end up with

$$dS^2 = -c^2 d\tau^2 \left(1 + \frac{\xi \tilde{a}}{c^2} \right)^2 + d\xi^2 .$$

Reading from the above, we pick out the components of the Rindler metric in the coordinates (τ, ξ) :

$$\begin{aligned} g_{00} &= - \left(1 + \frac{\xi \tilde{a}}{c^2} \right)^2 \\ g_{11} &= 1 \end{aligned}$$

Importantly, note that the same result can be calculated from the tensor transformation law

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} \eta_{\mu\nu} ,$$

where $\eta_{\mu\nu}$ is the Minkowski metric.

Rindler Inverse Metric

The Rindler inverse metric is calculated from

$$g^{\mu\alpha} g_{\alpha\nu} = \delta^\mu_\nu ,$$

implying

$$-1 \left(1 + \xi \tilde{a}/c^2 \right)^2 g^{00} + g^{11} = 2 ,$$

and thus:

$$\begin{aligned} g^{00} &= -1 / \left(1 + \xi \tilde{a}/c^2 \right)^2 \\ g^{11} &= 1 \end{aligned}$$

2.3 Properties of Rindler Space

Local Clock Speed

Rindler space has the curious property that the sense of time in the accelerated frame depends on the distance ξ from the origin in that frame. At a fixed height ξ above the floor, let

the local time is denoted $\tau_{(\xi)}$, so the differential spacetime interval reads

$$-c^2 d\tau_{(\xi)}^2 = -c^2 d\tau^2 \left(1 + \frac{\xi \tilde{a}}{c^2}\right)^2 + 0,$$

reducing to

$$d\tau_{(\xi)} = d\tau \left(1 + \frac{\xi \tilde{a}}{c^2}\right).$$

Evidently, any interval $d\tau_{\xi}$ is always greater than the corresponding interval $d\tau$ on the floor of the rocket. That is, if you are standing up in the accelerating rocket, your head is aging faster than your feet. This effect is called *gravitational redshift*.

Local Acceleration

The local acceleration two-vector at the point $x' = \xi$ is defined as

$$a_{(\xi)}^{\mu} = \frac{d^2 x^{\mu}}{d\tau_{(\xi)}^2},$$

where only the floor of the rocket accelerates at the constant $\tilde{a}$. This can be calculated from U^{μ} along with the chain rule to give:

$$\begin{aligned} a_{(\xi)}^0 &= \left(\frac{\tilde{a}}{1 + \xi \tilde{a}/c^2} \right) \sinh(\tilde{a}\tau/c) \\ a_{(\xi)}^1 &= \left(\frac{\tilde{a}}{1 + \xi \tilde{a}/c^2} \right) \cosh(\tilde{a}\tau/c) \end{aligned}$$

From the two-acceleration, we see the norm of the two-acceleration is a function of ξ via

$$\|a_{(\xi)}\|^2 = \left(\frac{\tilde{a}}{1 + \xi \tilde{a}/c^2} \right)^2,$$

implying the local acceleration

$$\tilde{a}_{(\xi)} = \frac{\tilde{a}}{1 + \xi \tilde{a}/c^2}.$$

The local acceleration $\tilde{a}_{(\xi)}$ can also be attained by Taylor-expanding the equations of motion $ct(\tau, \xi)$, $x(\tau, \xi)$ when $\tilde{a}\tau/c$ is a small quantity. For this, we find

$$ct \approx \left(\frac{c^2}{\tilde{a}} + \xi \right) \left(\left(\frac{\tilde{a}\tau}{c} \right) + \frac{1}{3!} \left(\frac{\tilde{a}\tau}{c} \right)^3 + \dots \right),$$

which gives, to first order in the small quantity:

$$t \approx \tau \left(1 + \frac{\xi \tilde{a}}{c^2} \right)$$

Repeating for the x -equation, we have

$$x \approx \left(\frac{c^2}{\tilde{a}} + \xi \right) \left(1 + \frac{1}{2!} \left(\frac{\tilde{a}\tau}{c} \right)^2 + \dots \right) - \frac{c^2}{\tilde{a}},$$

where retaining up to the order-two terms gives

$$x \approx \xi + \frac{1}{2} \left(\frac{\tilde{a}}{1 + \xi \tilde{a}/c} \right)^2 t^2.$$

Photon Redshift

In deriving the formula for the relativistic doppler shift, the intermediate form

$$f' \gamma \left(1 + \frac{v}{c} \right) = f$$

was attained, where f is the proper frequency of a pulse of light and f' is the observed frequency. At low speeds $v \ll c$, we can ignore the gamma factor but still have

$$f' \left(1 + \frac{v}{c} \right) \approx f.$$

Thus, there will still be a changing redshift or blueshift effect if the primed frame is accelerating.

Stationary Object

If an object remains fixed at position $x_0 > -c^2/\tilde{a}$ in the inertial frame, then the position of the object in the accelerated frame is given by $\xi(\tau)$. From the equations of motion, we derive

$$\xi(\tau) = \frac{x_0 + c^2/\tilde{a}}{\cosh(\tilde{a}\tau/c)} - \frac{c^2}{\tilde{a}}.$$

In the limit $\tau \rightarrow \infty$, the apparent position ξ becomes asymptotic to $-c^2/\tilde{a}$.

The apparent velocity of the stationary object is given by the two-velocity

$$V^{\mu} = \left(c \frac{d\tau}{dt}, \frac{d\xi}{dt} \right),$$

where the derivative is with respect to t , the proper time of the inertial frame. Calculating $d\xi/dt$ first, we have

$$\begin{aligned} \frac{d\xi}{dt} &= \frac{-c^2 \tilde{a}}{\sqrt{(x_0 + c^2/\tilde{a})^2 - (ct)^2}} \\ &= \frac{-c(c^2/\tilde{a} + \xi) \sinh(\tilde{a}\tau/c)}{(c^2/\tilde{a} + \xi)} \\ &= -c \sinh(\tilde{a}\tau/c). \end{aligned}$$

To find $d\tau/dt$, take the t -derivative of the x -equation:

$$\frac{dx_0}{dt} = \frac{d\xi}{dt} \cosh(\tilde{a}\tau/c) + c \left(1 + \frac{\xi\tilde{a}}{c}\right) \sinh(\tilde{a}\tau/c) \frac{d\tau}{dt}$$

After substituting $d\xi/dt$, the above simplifies to

$$\frac{d\tau}{dt} = \frac{\cosh(\tilde{a}\tau/c)}{1 + \xi\tilde{a}/c}.$$

With these results, the two-velocity is:

$$V^\mu = c \left(\frac{\cosh(\tilde{a}\tau/c)}{1 + \xi\tilde{a}/c}, -\sinh(\tilde{a}\tau/c) \right)$$

Accounting for the Rindler metric $g_{\mu\nu}$, the covariant version of V reads

$$V_\mu = g_{\mu\nu}V^\nu = -c (\cosh(\tilde{a}\tau/c), \sinh(\tilde{a}\tau/c)),$$

and we see the norm of V produces the familiar spacetime invariant:

$$V_\mu V^\mu = -c^2$$

Many Observers

For an interesting problem, consider the set of all observers that experience one and the same acceleration $\tilde{a}$. At time $t = 0$, each observer is assumed to be in a rest position $\tilde{x}$, and each has its own sense of proper time $\theta(ct, x)$.

For a given observer, the equations of motion read

$$\begin{aligned} ct(\theta, \tilde{x}) &= \frac{c^2}{a} \sinh(\tilde{a}\theta/c) \\ x(\theta, \tilde{x}) &= \tilde{x} + \frac{c^2}{\tilde{a}} \cosh(\tilde{a}\theta/c) - \frac{c^2}{\tilde{a}}, \end{aligned}$$

and the differential spacetime interval works out as:

$$\begin{aligned} dS^2 &= -c^2 dt^2 + dx^2 \\ &= -c^2 d\theta^2 + 2c \sinh(\tilde{a}\theta/c) d\tilde{x} d\theta + d\tilde{x}^2 \end{aligned}$$

2.4 Alternate Spaces

Alternative Rindler Space

Consider the equations of motion for Rindler space:

$$\begin{aligned} ct &= \frac{c^2}{\tilde{a}} e^{\xi\tilde{a}/c^2} \sinh(\lambda\tilde{a}/c) \\ x &= \frac{c^2}{\tilde{a}} e^{\xi\tilde{a}/c^2} \cosh(\lambda\tilde{a}/c) \end{aligned}$$

Calculating the differential spacetime interval, one readily finds

$$dS^2 = e^{2\xi\tilde{a}/c^2} (-c^2 d\lambda^2 + d\xi^2).$$

From the above, the alternative Rindler metric is:

$$\begin{aligned} g_{\lambda\lambda} &= -e^{2\xi\tilde{a}/c^2} \\ g_{\lambda\xi} &= g_{\xi\lambda} = 0 \\ g_{\xi\xi} &= e^{2\xi\tilde{a}/c^2} \end{aligned}$$

The same result is attained using the transformation law

$$g_{\mu'\nu'} = \frac{\partial q^\mu}{\partial q^{\mu'}} \frac{\partial q^\nu}{\partial q^{\nu'}} g_{\mu\nu}.$$

Problem 19

A vector V has constant components $V^\lambda = 1$, $V^\xi = 0$. Calculate the components V^t , V^x . Answer:

$$\begin{aligned} V^t &= c \frac{\partial t}{\partial \lambda} V^\lambda + c \frac{\partial t}{\partial \xi} V^\xi = c e^{\xi\tilde{a}/c^2} \cosh(\lambda\tilde{a}/c) \\ V^x &= \frac{\partial x}{\partial \lambda} V^\lambda + \frac{\partial x}{\partial \xi} V^\xi = c e^{\xi\tilde{a}/c^2} \sinh(\lambda\tilde{a}/c) \end{aligned}$$

Milne Space

The *Milne spacetime* is a toy model of an expanding universe in 1 + 1 dimensions using the parameterization

$$\begin{aligned} ct &= T \cosh(kz) \\ x &= T \sinh(kz), \end{aligned}$$

where k is a constant. The corresponding differential spacetime interval is:

$$dS^2 = -c^2 dT^2 + c^2 T^2 k^2 dz^2$$

Friedman-Robertson-Walker Space

The Friedman-Robertson-Walker space, or FRW space, is a homogeneously expanding or contracting Minkowski space. Jumping right to the differential spacetime interval, one has:

$$dS^2 = -c^2 dt^2 + a^2(t) \delta_{jk} dq^j dq^k,$$

where $a(t)$ characterizes the acceleration and the indices j , k cover the three spatial dimensions.

Chapter 35

Quantum Mechanics

Chapter 36

Statistical Mechanics

1 Introduction

Statistical mechanics concerns with the behavior of matter and energy when the number of parts in a system is very large. It begins by admitting that any attempt to keep a ledger of the position and momentum of *every* particle in a system, even on a computer, becomes intractable beyond certain limits. Also, one can imagine a grand mechanical ledger being redundant to a simpler *statistical* analysis.

Entropy

Recall from elementary thermodynamics that Carnot successfully identified an entirely reversible thermodynamic cycle composed of isothermal and adiabatic processes. In one cycle of a so-called Carnot engine, all instances of heat energy Q transferred to the system divided by the temperature T at each instance, sums to zero:

$$\sum_{\text{Carnot cycle}} \frac{Q}{T} = 0$$

Traversing an open path or an irreversible path in PV -space will give a non-zero result on the right, identified as the entropy:

$$\sum \frac{Q}{T} = \Delta S$$

In the differential limit, the sum becomes an integral of over small increments dQ/T :

$$\Delta S = \int \frac{dQ}{T}$$

Another exhibit comes from analyzing free expansion of ideal gas from an initial volume into confined vacuum.

For initial volume V_i of gas that expands freely into a final volume V_f , the entropy change simplifies to:

$$\Delta S = NK \ln \left(\frac{V_f}{V_i} \right) = K \ln \left(\frac{V_f^N}{V_i^N} \right)$$

Note that all information about the system is enclosed in the natural logarithm, which in this case is a dimensionless ratio of volumes - just a number.

Configurations

A system of one particle confined to a volume only large enough to contain that particle has no options for change, and we say there is only one available *configuration* in the system. If the system volume is doubled, there are suddenly *two* available configurations - the particle can occupy either half of the available volume. It clearly follows that the number of configurations available to a single-particle system scales with the volume of the system. Start adding more particles, and the number of configurations grows rapidly. The total number of configurations available to a system is called the combinatoric *multiplicity*, denoted Ω (Greek Omega).

Distinguishability

An important consideration is whether any two members in a system are *distinguishable* or *indistinguishable*. For instance, the number of configurations of two identical letters XX is precisely one - it would be *physically* redundant to swap the order of each X and claim a second configuration. However, if given the distinguishable pair X_1X_2 , then it is *not* futile to swap each member, giving X_2X_1 as the second configuration.

The multiplicity of a system having two particles occupying any of N available volume units is easy to write. The first particle has all N ‘choices’ of empty volume units, so we begin with $\Omega_1 = N$. The second particle can’t choose from all N volume units, but instead $N - 1$, indicating $\Omega_2 = N - 1$.

If the particles are distinguishable, the total multiplicity is simply the product of Ω_1 and Ω_2 :

$$\Omega_{\text{dist}} = \Omega_1 \Omega_2 = N (N - 1)$$

For particles that are identical though, it follows that Ω_{dist} over-counts the multiplicity by a factor of 2. The total multiplicity is then:

$$\Omega_{\text{indist}} = \frac{N (N - 1)}{2}$$

Extensive and Intensive

Recall from elementary thermodynamics that a system exhibits two classes of variables - extensive and intensive. *Extensive* quantities, such as mass, volume, and energy, scale when resizing the system. By contrast, *intensive* quantities, such as density and temperature, don’t scale by resizing the system.

In a class of its own, the multiplicity is *super-extensive*, meaning it grows by multiplication. In order to ‘knock down’ the multiplicity into an everyday extensive quantity, the most natural maneuver is to take the logarithm of Ω , so that multiplication

$$\Omega_T = \Omega_1 \Omega_2$$

becomes addition:

$$\ln (\Omega_T) = \ln (\Omega_1) + \ln (\Omega_2)$$

Rethinking Entropy

Our walk-through of combinatorics suggests an information-based definition of entropy for two reasons. (i) For an ideal gas expanding freely into vacuum, we found that entropy is equal to the Boltzmann constant K times the natural log of some number that has to do with configurations. (ii) The natural log of the multiplicity is an extensive quantity. It follows that entropy shall be *defined* generally as K times the natural log of the multiplicity:

$$S = K \ln (\Omega) \tag{36.1}$$

2 Microcanonical Ensemble

A thermodynamic system that is not allowed to exchange volume V , number of particles N , or energy U with its environment is a *microcanonical ensemble*. One should recall from elementary thermodynamics that the entropy S of such a system can only increase with time, which is indeed the case here. The entropy is crucial to calculate, as the myriad of thermodynamic variables are computed from S or its derivatives.

Microstate

In a microcanonical ensemble, consider an instantaneous ‘snapshot’ of the whole system, and write the position and momentum of *every* particle involved, within the Heisenberg limit $\Delta x \Delta p \geq \hbar$. Such a (potentially enormous) body of information is called a *microstate*.

A fundamental assumption of the microcanonical ensemble is that *all microstates are equally accessible*. This is to say that microstates are uniformly sampled by the system as time evolves.

Macrostate

System-wide, or otherwise ‘global’ quantities of the system are called *macrostates*. One can readily imagine that many, sometimes an *infinite* number of microstates can correspond to one macrostate. For instance, a box of gas in the macrostate with fixed pressure, fixed volume, and fixed temperature, can correspond to very many microstates. A corollary to this is that any observed macrostate, such as the air density in a room, is one corresponding to the greatest number of microstates.

Ergodic Hypothesis

A fundamental assumption of microcanonical ensemble theory is the *ergodic hypothesis*, stating that the time-average $\bar{x}$ of an observable is equivalent to the ensemble average $\langle x \rangle$ of the observable. We demonstrate this by considering a simple harmonic oscillator of mass m having position $q(t)$ and momentum $p(t)$ at fixed energy E obeying the (Hamiltonian) differential equation

$$H(q, p) = \frac{p^2}{2m} + \frac{m\omega^2}{2} q^2,$$

solved by:

$$\begin{aligned} q(t) &= A \cos(\omega t) \\ p(t) &= -m\omega A \sin(\omega t) \\ A &= \sqrt{\frac{2E}{m\omega^2}} \end{aligned}$$

We easily calculate the time-average of the mean-square position $q(t)^2$ using the definition

$$\overline{q^2} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T q(t)^2 dt,$$

which comes out to

$$\overline{q^2} = \lim_{T \rightarrow \infty} \frac{A^2}{T} \int_0^T \frac{\cos(2\omega t) + 1}{2} dt = \frac{A^2}{2} = \frac{E}{m\omega^2}.$$

We next calculate the ensemble average $\langle q^2 \rangle$ using

$$\langle q^2 \rangle = \frac{\int q^2 \rho(q, p) dq dp}{\int \rho(q, p) dq dp},$$

where the density

$$\rho(q, p) \propto \delta(H(q, p) - E)$$

is a Dirac delta function that obeys

$$\int \delta(\phi(y)) dy = \frac{1}{\phi'(y_0)}.$$

Starting with the denominator, find:

$$\begin{aligned} \int \rho(q, p) dq dp &= \int \delta(H(q, p) - E) dq dp \\ &= \int \frac{1}{\left(\frac{\partial p}{\partial q}\right) \Big|_{\frac{p^2}{2m} + \frac{m\omega^2}{2} q^2 - E = 0}} dq \\ &= \int_{\pm} \frac{m dq}{\sqrt{2m(E - m\omega^2 q^2/2)}} \\ &= 2 \int \frac{p dt}{|p|} = \frac{2}{\omega} \int \omega dt = \frac{2}{\omega} \int_0^\pi d\theta \\ &= \frac{2\pi}{\omega} \end{aligned}$$

Moving on to the numerator, the calculation is similar until the last step:

$$\begin{aligned} \int q^2 \rho(q, p) dq dp &= \frac{2}{\omega} \int q^2 \omega dt \\ &= \frac{2A^2}{\omega} \int \cos^2(\omega t) \omega dt \\ &= \frac{2A^2}{\omega} \frac{\pi}{2} = \frac{\pi A^2}{\omega} \end{aligned}$$

Finally, the ensemble average $\langle q^2 \rangle$ comes to

$$\langle q^2 \rangle = \frac{\pi A^2}{\omega} \frac{\omega}{2\pi} = \frac{A^2}{2} = \frac{E}{m\omega^2},$$

matching the time-average.

2.1 Equilibrium

Having fixed volume, number of particles, and energy, the dynamic quantities in the microcanonical ensemble tend to be the temperature T , pressure P , and chemical potential μ .

Thermal Equilibrium

Consider a microcanonical system A having multiplicity Ω_A , internal energy U_A , and entropy $S_A = K \ln(\Omega_A)$, and also a second system B characterized by Ω_B , U_B , and $S_B = K \ln(\Omega_B)$.

When placed in contact, the systems A and B are allowed to exchange internal energy while maintaining a constant total $U_C = U_A + U_B$, with corresponding differential $dU_A = -dU_B$. (The volume and number of particles in each subsystem must remain the same.) The entropy is extensive and obeys $S_C = S_A + S_B$, and meanwhile the multiplicity is super-extensive and follows $\Omega_C = \Omega_A \Omega_B$.

As each system achieves thermal equilibrium, the energies U_A and U_B are not fixed, but instead hover around average values $\bar{U}_A$ and $\bar{U}_B$. The temperature of each system finds a common value $T_C = T_A = T_B$, which is defined from elementary thermodynamics in terms of S and U :

$$T_A = \left(\frac{\partial S_A}{\partial U_A} \right)_{V_A, N_A, \bar{U}_A}^{-1}$$

$$T_B = \left(\frac{\partial S_B}{\partial U_B} \right)_{V_B, N_B, \bar{U}_B}^{-1}$$

Encoded in the above definitions is the notion that the entropy is maximized at thermal equilibrium.

To proceed, let us choose a dependent variable (from system A for simplicity), and frame the total entropy as

$$S_C(U_A) = K \ln(\Omega_C(U_A)).$$

By Taylor-expanding in the neighborhood $U_A \approx \bar{U}_A$, the entropy becomes

$$S_C(U_A) \approx S_C(\bar{U}_A) + \left(\frac{\partial S_C}{\partial U_A} \right)_{\bar{U}_A} (U_A - \bar{U}_A) + \frac{1}{2!} \left(\frac{\partial^2 S_C}{\partial U_A^2} \right)_{\bar{U}_A} (U_A - \bar{U}_A)^2 + \dots,$$

where we stop expanding after at least second order, because the first-order coefficient is identically zero:

$$\left(\frac{\partial S_C}{\partial U_A} \right)_{\bar{U}_A} = \frac{\partial S_A}{\partial U_A} + \frac{\partial S_B}{\partial U_A} = \frac{\partial S_A}{\partial U_A} - \frac{\partial S_B}{\partial U_B}$$

$$= \frac{1}{T_A} - \frac{1}{T_B} = 0$$

The second-order coefficient is recast by brute force, giving

$$\left(\frac{\partial^2 S_C}{\partial U_A^2} \right)_{\bar{U}_A} = \frac{\partial}{\partial U_A} \left(\frac{\partial S_C}{\partial U_A} \right) = \frac{\partial}{\partial U_A} \left(\frac{1}{T_A} - \frac{1}{T_B} \right)$$

$$= -\frac{1}{T_A^2} \frac{\partial T_A}{\partial U_A} - \frac{1}{T_B^2} \frac{\partial T_B}{\partial U_B},$$

which contains the definition of the heat capacity at constant volume for each subsystem near equilibrium. Condensing the second-order coefficient into a variable $-K\sigma_A^{-2}$, we have

$$\left(\frac{\partial^2 S_C}{\partial U_A^2} \right)_{\bar{U}_A} = -\frac{1}{T^2} \left(\frac{1}{(C_V)_A} + \frac{1}{(C_V)_B} \right) = \frac{-K}{\sigma_A^2}.$$

In terms of σ_A^2 , the total entropy near equilibrium reads:

$$S_C(U_A) \approx S_C(\bar{U}_A) - \frac{K}{2\sigma_A^2} (U_A - \bar{U}_A)^2 \quad (36.2)$$

The total multiplicity Ω_C of the combined system is given by inverting the definition $S_C = K(\ln \Omega_C)$, or

$$\Omega_C \approx e^{S_C(\bar{U}_A)/K} e^{-(U_A - \bar{U}_A)^2 / 2\sigma_A^2},$$

which is Gaussian in the U_A -variable. As normalized, the above becomes a probability density

$$p(U_A) = \left(\frac{1}{2\pi\sigma_A^2} \right)^{1/2} e^{-(U_A - \bar{U}_A)^2 / 2\sigma_A^2}.$$

Problem 1

Use the thermodynamic identity $dU = TdS - PdV + \mu dN$ to derive

$$T = \left(\frac{\partial S}{\partial U} \right)_{V, N}^{-1}.$$

Problem 2

Use Gaussian integrals to show:

$$\int_{\text{all } U_A} p(U_A) dU_A = 1$$

$$\langle U_A \rangle = \bar{U}_A$$

$$\langle U_A^2 \rangle = \bar{U}_A^2 + \sigma_A^2$$

Problem 3

From elementary thermodynamics, we know the ideal gas has internal energy $U = 3NKT/2$. For connected systems A and B of total energy U , (i) use the equilibrium condition to write U_A in terms of U , N_A , and N_B . (ii) Solve for σ_A explicitly and show that σ_A/U_A scales with $N^{-1/2}$. Answer:

$$\begin{aligned}
 T_A &= T_B \\
 \frac{2U_A}{3N_A K} &= \frac{2(U - U_A)}{3N_B K} \\
 U_A &= U \frac{N_A}{N} \\
 \frac{1}{\sigma_A^2} &= \frac{3N_A}{2U_A^2} + \frac{3N_B}{2(U - U_A)^2} \\
 &= \frac{3}{2} \frac{N^2}{U^2} \left(\frac{1}{N_A} + \frac{1}{N_B} \right) = \frac{3}{2U^2} \frac{N^3}{N_A N_B} \\
 \frac{\sigma_A}{U_A} &= \sqrt{\frac{3}{2}} \frac{U \sqrt{N_A N_B}}{N \sqrt{N}} \frac{N}{U N_A} \\
 &= \sqrt{\frac{3}{2}} \sqrt{\frac{N_B}{N_A}} \frac{1}{\sqrt{N}} \propto \frac{1}{\sqrt{N}}
 \end{aligned}$$

Mechanical Equilibrium

Extending the previous example, consider a composite microcanonical system C made from two systems A , B that are allowed to exchange energy and volume (but not particles). In particular, we have

$$\begin{aligned}
 0 &= dU_A - dU_B \\
 0 &= dV_A - dV_B \\
 0 &= dN_A = dN_B
 \end{aligned}$$

such that

$$\begin{aligned}
 \Omega_C &= \Omega_A(U_A, V_A, N_A) \\
 &\times \Omega_B(U_C - U_A, V_C - V_A, N_B) .
 \end{aligned}$$

The system C tends toward thermal equilibrium with $T_A = T_B$ and also mechanical equilibrium with $P_A = P_B$, written as:

$$\begin{aligned}
 P_A &= T \left(\frac{\partial S_A}{\partial V_A} \right)_{U_A, N_A, \bar{V}_A} \\
 P_B &= T \left(\frac{\partial S_B}{\partial V_B} \right)_{U_B, N_B, \bar{V}_B}
 \end{aligned}$$

Using $S_C = K \ln(\Omega_C)$ and Taylor-expanding in the neighborhood $V_A \approx \bar{V}_A$, the entropy is approximately

$$\begin{aligned}
 S_C(V_A) &\approx S_C(\bar{V}_A) + \left(\frac{\partial S_C}{\partial V_A} \right)_{\bar{V}_A} (V_A - \bar{V}_A) \\
 &+ \frac{1}{2!} \left(\frac{\partial^2 S_C}{\partial V_A^2} \right)_{\bar{V}_A} (V_A - \bar{V}_A)^2 + \dots ,
 \end{aligned}$$

and we draw attention to the first-order coefficient:

$$\begin{aligned}
 \left(\frac{\partial S_C}{\partial V_A} \right)_{\bar{V}_A} &= \frac{\partial S_A}{\partial V_A} + \frac{\partial S_B}{\partial V_A} = \frac{\partial S_A}{\partial V_A} - \frac{\partial S_B}{\partial V_B} \\
 &= \frac{P_A - P_B}{T} = 0
 \end{aligned}$$

Problem 4

Use the thermodynamic identity $dU = TdS - PdV + \mu dN$ to derive

$$P = T \left(\frac{\partial S}{\partial V} \right)_{U, N} .$$

Diffusive Equilibrium

A system allowed to exchange energy, volume, and particles will tend toward diffusive equilibrium. For this situation, the thermodynamic identity $dU = TdS - PdV + \mu dN$ readily leads to

$$\mu = -T \left(\frac{\partial S}{\partial N} \right)_{U, V} ,$$

which must be the same for each subsystem A , B .

2.2 Microcanonical Paramagnetism**Magnetic Materials**

A simple application of microcanonical ensemble theory is *paramagnetism*. Magnetism is a phenomenon caused by circulating electric currents in matter or vacuum. Most solids generally fall into one of three categories: paramagnetic, diamagnetic, or ferromagnetic.

Ferromagnetic materials have the strongest coupling to magnetic phenomena, and are most familiar to everyday experience. Inside a ferromagnetic material, atoms are arranged into macroscopic *domains*, wherein electrons orbit their respective nuclei such that the angular momentum vector of each electron points the same way. A strong permanent magnet consists of many domains, where any given cluster of domains has a net magnetic moment that is strong in a direction we decide to call ‘north’.

Paramagnetic materials exhibit weak affinity to magnetism, and do not carry permanent magnetic fields. A paramagnetic material responds attractively to an external magnetic field, wherein electrons inside the material align their axis of orbit in order to increase the total magnetic field near the paramagnet. It follows that a paramagnetic material is attracted to an external field and moves toward it.

A diamagnetic material acts in the opposite manner as a paramagnet - it tries to oppose an external magnetic field and is repelled from it. These phenomena only occur when the temperature of the material is low enough, otherwise thermal noise prevents domain alignment.

Paramagnet in External Field

Consider a large- N paramagnetic material of fixed volume made of non-interacting spin $1/2$ particles, each having magnetic moment $\mu = 2\mu_B S$, where μ_B is the Bohr magneton and S is the spin of a given particle. When an external magnetic field $\mathbf{B} = +B\hat{z}$ is applied, a number $N_\uparrow$ particles tend to align their spin axis along the external field's direction, where a minority $N_\downarrow$ of particles go against the field. (At low temperatures or in strong fields, none of the spin vectors support an angle with respect to $\hat{z}$.) The paramagnet in turn acts as a two-state system with total internal energy $U = -N_\uparrow\mu B + N_\downarrow\mu B$ among its $N = N_\uparrow + N_\downarrow$ particles.

As a two-state system, the multiplicity Ω in terms of U and N reads

$$\begin{aligned}\Omega(U, N) &= \frac{N!}{N_\uparrow! N_\downarrow!} \\ &= \frac{N!}{\left(\frac{N}{2} - \frac{U}{2\mu B}\right)! \left(\frac{N}{2} + \frac{U}{2\mu B}\right)!}.\end{aligned}\quad (36.3)$$

Proceeding carefully, let $x = U/2\mu B$ and apply Stirling's approximation

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

to arrive at

$$\Omega(x, N) \approx \frac{N^N \sqrt{N/2\pi}}{\left(\frac{N^2}{4} - x^2\right)^{N/2} \left(\frac{N}{2} - x\right)^{-x} \left(\frac{N}{2} + x\right)^x \sqrt{\frac{N^2}{2} - x^2}},$$

which is more easily handled if we take the natural logarithm:

$$\begin{aligned}\ln(\Omega(U, N)) &\approx N \ln(N) + \frac{1}{2} \ln\left(\frac{N}{2\pi}\right) \\ &\quad - \left(\frac{N+1}{2}\right) \ln\left(\frac{N^2}{4} - x^2\right) \\ &\quad + x \left(\ln\left(\frac{N}{2} - x\right) - \ln\left(\frac{N}{2} + x\right)\right)\end{aligned}\quad (36.4)$$

Proceed by Taylor-expanding inside the $\ln$ -terms, using the approximation that $N \gg x$. Doing so, we find

$$\begin{aligned}\ln\left(\frac{N}{2} \pm x\right) &\approx \ln\left(\frac{N}{2}\right) \pm \frac{2x}{N} \\ \ln\left(\frac{N^2}{4} - x^2\right) &\approx 2 \ln\left(\frac{N}{2}\right) - \left(\frac{2x}{N}\right)^2,\end{aligned}$$

and the multiplicity boils down to (ignoring the last very small term)

$$\ln(\Omega(U, N)) \approx N \ln(2) + \ln\sqrt{\frac{2}{\pi N}} - \frac{2x^2}{N} + \frac{2x^2}{N^2}.$$

While already in the $\ln(\Omega)$ domain, multiply by the Boltzmann constant to write the entropy of the two-state paramagnet

$$S = K \left(N \ln(2) + \frac{1}{2} \ln\left(\frac{2}{\pi N}\right) - \frac{U^2}{2N\mu^2 B^2} \right), \quad (36.5)$$

which decreases as more spins align. The corresponding multiplicity is

$$\Omega(U, N) = 2^N \sqrt{\frac{2}{\pi N}} e^{-2(U/2\mu B)^2/N}. \quad (36.6)$$

Note that by dividing out the total multiplicity (total number of microstates) 2^N , Ω becomes a Gaussian probability density in the dimensionless quantity $U/2\mu B$:

$$p(U/2\mu B, N) = \sqrt{\frac{2}{\pi N}} e^{-2(U/2\mu B)^2/N}$$

Having fixed volume and number of particles, the formula $1/T = (\partial S/\partial U)_{V,N}$ gives the temperature as

$$KT = -\frac{N}{U} \mu^2 B^2,$$

which has a minus sign to balance out the overall negativity of U . This relation is handy near $U = 0$, but completely ignores the boundary cases $U = \pm N\mu B$.

To refine the calculation, we must return to Equation (36.4) without using Taylor expansion. Instead of writing a massively ugly formula for the entropy, go straight to the temperature calculation by taking a derivative with respect to $x = U/2\mu B$, which yields

$$\frac{\partial S}{\partial x} = K \left(\frac{x}{\frac{N^2}{4} - x^2} + \ln \left(\frac{\frac{N}{2} - x}{\frac{N}{2} + x} \right) \right),$$

where the first term has been ignored due to the factor of N^{-2} . This is equivalent to leaving out the $\sqrt{2\pi N}$ -like terms in Stirling's approximation. Adjusting for U instead of x , we find

$$KT = \frac{2\mu B}{\ln \left(\frac{N\mu B - U}{N\mu B + U} \right)}, \quad (36.7)$$

which equals zero at the boundaries $U = \pm N\mu B$.

Negative Temperature

Curiously, a two-state paramagnet *can have negative temperature* if conditions are right. If the system is prepared so that most spins oppose the external magnetic field, the temperature is indeed negative, however not colder than absolute zero. The temperature undergoes a major discontinuity as the energy passes through $U = 0$, at which point the temperature is ambiguously either/and $T = -\infty$ or $T = \infty$. If the energy is a small positive number $U \approx \mu B$, the temperature is around $-\infty$ (spins alternate up and down). If the energy is instead negative with $U \approx -\mu B$, the temperature is approximately ∞ . This is not so far-fetched, because a paramagnet that can still manage $U = 0$ despite the external B -field should be quite hot!

2.3 Entropic Spring

Consider a stretchable material classified as a polymer chain, such as a rubber band, that is maintained under tension F in the x -direction. In particular, we imagine the chain made of $N \gg 1$ links, each of length a and connected end-to-end, oriented as follows: (i) The $N_{\parallel}$ links oriented parallel to x each contribute a to the total length of the chain. (ii) The $N_{\perp}$ links oriented perpendicular to x do not contribute to the length L of the chain. Using the variables on hand, we write

$$L = Nx,$$

where

$$x = a \frac{N_{\parallel}}{N}.$$

Such a setup is the *entropic spring model*. Let us also assume that the system is in an environment of fixed temperature T . Since each link doesn't 'care' about the orientation of other links, the internal energy is taken as $U = 0$. The relevant thermodynamic potential is in fact the Gibbs free energy in the form

$$G = -FL - TS,$$

where $-FL$ is playing the one-dimensional role of PV . The negative sign occurs because pressure is usually a push from outside the system. The force F in this case is a pull.

As a microcanonical system, we can write the multiplicity of links parallel to x :

$$\Omega(N_{\parallel}, N) = \frac{N!}{N_{\parallel}! (N - N_{\parallel})!}$$

The next target is the entropy, which first requires Stirling's approximation to arrive at

$$\Omega(N_{\parallel}, N) \approx N \ln(N) - N_{\parallel} \ln(N_{\parallel}) - (N - N_{\parallel}) \ln(N - N_{\parallel}),$$

to which $S = K \ln(\Omega)$ is readily applied.

The Gibbs free energy is minimized at equilibrium. For this, we calculate

$$\begin{aligned} \frac{dG}{dL} &= \frac{d}{dL} (-FL - TS)_{x=\langle x \rangle} \\ 0 &= -F - KT \left(\frac{d\Omega}{dN_{\parallel}} \frac{dN_{\parallel}}{dL} \right)_{x=\langle x \rangle}, \end{aligned}$$

simplifying to an equation of state

$$F = \frac{-KT}{a} \ln \left(\frac{a}{\langle x \rangle} - 1 \right).$$

We next solve for $\langle x \rangle$ and subsequently write the length of the chain as $\langle L \rangle = N \langle x \rangle$ while under force:

$$\langle L \rangle = \frac{Na}{1 + e^{-Fa/KT}}$$

The case $F = 0$ corresponds to $a = 2 \langle x \rangle$, or $N = 2N_{\parallel}$. This means that without any force, statistically half of the links are in the parallel orientation. Using $L = Nx$, it follows that the rest length L_0 of the chain is

$$L_0 = \frac{Na}{2}.$$

Define

$$\Delta = \langle L \rangle - L_0$$

so the equation of state becomes

$$F = \frac{-KT}{a} \ln \left(\frac{Na - 2\Delta}{Na + 2\Delta} \right).$$

For small Δ , we note that

$$\ln \left(\frac{Na - 2\Delta}{Na + 2\Delta} \right) \approx \frac{-4\Delta}{Na},$$

and the above simplifies to a Hookean regime:

$$F = \frac{4KT}{Na^2} (\langle L \rangle - L_0)$$

Note the effective spring constant is proportional to temperature, as one might expect for a polymer. Metals, on the other hand, lose stiffness at high temperature.

2.4 Einstein Solid

Using quantum mechanics, an accurate and precise description of matter is ‘easy’ enough to attain for simple systems such as isolated atoms, simple harmonic oscillators, and perfect crystals. Choosing how to model a system is a careful game, because we want to capture the physically interesting phenomena, while filtering out details that won’t factor into the investigation.

In a model called the *Einstein solid*, matter is assumed to consist of N non-interacting quantum harmonic oscillators, with each oscillator having energy

$$\epsilon_i(n_i) = \hbar\omega \left(n_i + \frac{1}{2} \right) \quad n_i = 0, 1, 2, \dots,$$

where ω is the angular vibration frequency, and n_i is the quantum excitation level of the i th oscillator, having allowed values 0, 1, 2, etc. The Einstein solid model allows us to ignore nucleons, electrons, spin, momentum, and so on.

Total Energy

The total energy E of an Einstein solid is the sum

$$E = \sum_{i=0}^N \epsilon_i(n_i) = \hbar\omega \sum_{i=0}^N \left(n_i + \frac{1}{2} \right),$$

which is *not* zero when all excitation numbers n_i are zero.

Zero-Point Energy

The ground-state quantity

$$E_0 = \frac{1}{2} N \hbar\omega$$

is called the *zero-point energy*, and is thermodynamically inaccessible. An oscillator cannot lose or modify its zero-point energy without destroying the system, so we ignore it in the same way we ignore the rest mass-energy mc^2 in calculations.

Internal Energy

The internal energy of an Einstein solid subtracts off the zero-point energy, and is given by:

$$U = E - E_0 = \hbar\omega \sum_{i=1}^N n_i \quad (36.8)$$

Multiplicity

Begin with a ground-state Einstein solid with zero internal energy. If we add some integer number u of energy units $\hbar\omega$ to the system, the internal energy of the solid jumps to $U = u\hbar\omega$. The particular distribution of energy units among individual oscillators is not defined, which means we much track *all* possible configurations available to the system, captured by the multiplicity $\Omega(U, N)$.

To proceed, we shall depict the boundary between two oscillators as a vertical slash ($|$), and one energy unit $\hbar\omega$ as a dot ($\bullet$). Taking a system with $N = 4$ as an example, we depict the unique ground state by four empty spaces (three slashes) and no dots:

$$U = 0 \quad ||| \quad \Omega(0, 4) = 1$$

If we add one energy unit to the system, we find exactly four ways to arrange the dot between various slashes

$$\bullet ||| \quad | \bullet || \quad || \bullet | \quad ||| \bullet,$$

so we conclude there are four configurations available, or $\Omega(1, 4) = 4$. That is, any one of the four oscillators has $n_j = 1$ while $n_{i \neq j} = 0$. Adding a second energy unit to the system, the internal energy becomes $U = 2\hbar\omega$, and there are ten ways to arrange two dots among the slashes

$$\begin{aligned} &\bullet \bullet ||| \quad \bullet | \bullet || \quad \bullet || \bullet | \quad \bullet ||| \bullet \quad | \bullet \bullet || \\ &| \bullet | \bullet | \quad | \bullet || \bullet \quad || \bullet \bullet | \quad || \bullet | \bullet \quad ||| \bullet \bullet, \end{aligned}$$

so we write $\Omega(2, 4) = 10$, and we could continue this way for any number of oscillators containing any number of energy units.

In the pictorial representation of the Einstein solid, there are $N - 1$ slashes and $u = U/\hbar\omega$ dots, or $N - 1 + u$ total symbols. Framing the problem this way, we may borrow directly from two-state analysis to write the multiplicity (think of dots and slashes as heads and tails on a coin). Of the $N - 1 + u$ symbols, u of them contribute to the internal energy. The multiplicity is exactly:

$$\Omega(u, N) = \frac{(N - 1 + u)!}{u! (N - 1)!} \quad (36.9)$$

To anticipate calculations that follow, we may as well eliminate the factorial terms using Stirling's approximation

$$\ln(n!) \approx n \ln n - n + \ln \sqrt{2\pi n},$$

so the multiplicity becomes

$$\begin{aligned} \ln(\Omega(u, N)) &= (N - 1) \ln \left(1 + \frac{u}{N - 1}\right) \\ &+ u \ln \left(1 + \frac{N - 1}{u}\right) \\ &+ \frac{1}{2} \ln \left(\frac{N - 1 + u}{2\pi u (N - 1)}\right). \end{aligned}$$

In the large- N limit, the final term will be negligible compared to the first two. Of course, $N - 1$ may also be replaced by N . Assuming further that $u \gg N$, the natural log of the multiplicity is approximately

$$\ln(\Omega(u, N)) \approx N(1 + \ln(u/N)),$$

implying

$$\Omega(u, N) \approx \left(\frac{ue}{N}\right)^N.$$

Thermodynamic Properties

With the natural log of the the multiplicity on hand, multiply by the Boltzmann constant to write the entropy of an N -particle Einstein solid

$$S = K \left(N \ln \left(1 + \frac{u}{N}\right) + u \ln \left(1 + \frac{N}{u}\right) \right), \quad (36.10)$$

where the internal energy is $U = u\hbar\omega$.

Having fixed volume and number of particles, the formula $1/T = (\partial S/\partial U)_{V,N}$ gives the temperature as

$$KT = \frac{\hbar\omega}{\ln(1 + N\hbar\omega/U)}.$$

In the $u \gg N$ limit, this resolves to a formula for the average energy per element in the solid:

$$KT = \frac{U}{N}$$

In light of the equipartition theorem, the left side is the sum of two factors $KT/2$, as the harmonic oscillator has two degrees of freedom. In the limit that u cannot be assumed to be much greater than N , we may still solve for U/N to derive the *Planck* result for the average energy per oscillator

$$\frac{U}{N} = \frac{\hbar\omega}{\exp(\hbar\omega/KT) - 1},$$

reproducing the $u \gg N$ case for high temperatures.

Solving for the internal energy U in terms of T comes out to

$$U = \frac{N\hbar\omega}{e^{\hbar\omega/KT} - 1}.$$

A temperature derivative of U gives the heat capacity at constant volume:

$$C_V = \frac{dU}{dT} = \frac{N(\hbar\omega)^2 e^{\hbar\omega/KT}}{KT^2 (e^{\hbar\omega/KT} - 1)^2}$$

Combined Solids

Consider a combined Einstein solid composed of two equal-volume subsystems A and B such that

$$N/2 = N_A = N_B$$

$$U_A = \frac{U}{2} + Q$$

$$U_B = \frac{U}{2} - Q,$$

where $U = u\hbar\omega$ is the total internal energy and N is the total number of particles. The quantity Q is a small energy imbalance between the two systems, defined as q counts of $\hbar\omega$. Taking the trivial case $q = 0$, the entropy of each $S_j(u_j, N_j, q)$ of subsystem reads

$$S_A\left(\frac{u}{2}, \frac{N}{2}, 0\right) = S_B\left(\frac{u}{2}, \frac{N}{2}, 0\right) = \frac{S(u, N, 0)}{2},$$

which is half of the total entropy, indicating the multiplicity to be

$$\Omega(u, N, 0) = \Omega\left(\frac{u}{2}, \frac{N}{2}, 0\right)^2.$$

The next job is to develop the nontrivial case $\Omega(u, N, q)$ for relatively small q . Begin by splitting the multiplicity

inspired from Equation (36.10) to account for subsystems A and B ,

$$\Omega(u, N, q) = \left(1 + \frac{u/2 + q}{N/2}\right)^{N/2} \left(1 + \frac{u/2 - q}{N/2}\right)^{N/2} \left(1 + \frac{N/2}{u/2 + q}\right)^{u/2+q} \left(1 + \frac{N/2}{u/2 - q}\right)^{u/2-q},$$

simplifying to, in the small- q limit:

$$\Omega(u, N, q) = \Omega(u, N, 0) e^{-2Nq^2/u^2} \quad (36.11)$$

Dividing away the Ω -term and introducing a new normalization constant $A(u, N)$, the multiplicity becomes a Gaussian distribution in the q -variable:

$$p(u, N, q) = A(u, N) e^{-2Nq^2/u^2},$$

where the normalization constant is determined by

$$1 = A(u, N) \int_{-\infty}^{\infty} e^{-2Nq^2/u^2} dq$$

$$A(u, N) = \sqrt{\frac{2N}{\pi u^2}}.$$

This allows calculations of $\langle q \rangle$, $\langle q^2 \rangle$, and the peak width (or standard deviation or variance):

$$\langle q \rangle = A(u, N) \int_{-\infty}^{\infty} q e^{-2Nq^2/u^2} dq = 0$$

$$\langle q^2 \rangle = A(u, N) \int_{-\infty}^{\infty} q^2 e^{-2Nq^2/u^2} dq = \frac{u^2}{4N}$$

$$\sigma_q = \sqrt{\langle q^2 \rangle - \langle q \rangle^2} = \frac{u}{2\sqrt{N}}$$

To express everything in terms of the dimensionless quantities u_A and u_B , start by writing

$$u_A = \frac{u}{2} + q \quad u_B = \frac{u}{2} - q$$

$$u_A^2 = \frac{u^2}{4} - uq + q^2 \quad u_B^2 = \frac{u^2}{4} - uq + q^2,$$

and use the ‘sum of the averages’ rule on each quantity (noting that u is already an average), accounting for the results above:

$$\langle u_k \rangle = \frac{u}{2} \pm \langle q \rangle = \frac{u}{2}$$

$$\langle u_k^2 \rangle = \frac{u^2}{4} \pm u \langle q \rangle + \langle q^2 \rangle = \frac{u^2}{4} + \frac{u^2}{4N}$$

$$\sigma_{u_A} = \sigma_{u_B} = \sigma_q$$

2.5 Quantum Ideal Gas

The ideal gas is defined as a collection of non-interacting particles occupying a container, which happens to be a generalization of a well-studied ‘confined quantum particle’ system (the infinite well). Consider a one-dimensional ‘box’ of length L occupied by a single particle of mass m . The walls at $x = 0$ and $x = L$ are impenetrable. The behavior of the particle is determined by the Schrodinger equation, namely

$$i\hbar \frac{\partial}{\partial t} \psi(x, t) = \left(\frac{-\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x, t) \right) \psi(x, t),$$

where ψ is the wavefunction of the particle, and V is the potential energy of the particle, which is zero inside the box.

The confined-particle solution to the Schrodinger equation is

$$\psi(x, t) = \sqrt{\frac{2}{L}} e^{-i\epsilon t/\hbar} \sin(kx),$$

where the particle’s energy ϵ and wavenumber

$$k = \frac{p}{\hbar}$$

are related by

$$\epsilon = \frac{\hbar^2 k^2}{2m}.$$

Explicitly, these are:

$$\epsilon = \frac{n^2 \hbar^2}{8mL^2}$$

and

$$k = \frac{n\pi}{L}$$

The number n is any positive nonzero integer. (Exclude the negative k -branch as we are dealing with standing-wave solutions.)

Generalizing the container to be three dimensional by $L \rightarrow L^3$, it’s readily shown that the particle’s energy becomes

$$\epsilon = \frac{\hbar^2}{8mL^2} (n_x^2 + n_y^2 + n_z^2).$$

Increasing the number of particles from one to N , the total internal energy U becomes the sum

$$U = \sum_{j=1}^N \epsilon_j,$$

which motivates inquiry into the total excitation number among all particles, which we define as a ‘dimensionless energy’ quantity:

$$(N^*)^2 = \sum_{j=1}^N (n_x^2 + n_y^2 + n_z^2) = \frac{8mL^2U}{h^2}$$

Multiplicity

The number of microstates Ω that correspond to an ideal gas of given energy U with fixed volume $L^3 = V$ and number of particles N must begin as the integral over all possible position and momentum states as:

$$\Omega = \frac{1}{N!} \frac{\int dx_1 \dots dx_{3N} \int dp_1 \dots dp_{3N}}{h^{3N}} \quad (36.12)$$

The factor of $1/N!$ has been inserted to account for the indistinguishability of ideal gas particles, and Planck’s constant $h = 2\pi\hbar$ is present for normalization purposes. This is simply a generalization of the $N = 1$ case, where the number of microstates available to a single particle is the product of ‘space bins’ and ‘momentum bins’, simplified by the Heisenberg uncertainty relation

$$\Omega_1 = \frac{L_x}{\Delta x} \frac{L_p}{\Delta p} \approx \frac{L_x L_p}{h}.$$

The position integral is easy to evaluate, as we simply get N multiples of the volume V of the gas, or

$$\int dx_1 \dots dx_{3N} = V^N.$$

The momentum integral is less obvious. First recast the dp_i factors in terms of the integers n_i , and write

$$\Omega = \frac{1}{N!} \frac{V^N}{h^{3N}} \left(\frac{\hbar\pi}{L} \right)^{3N} \int dn_1 \dots dn_{3N}.$$

The remaining integral is the surface area of a hypersphere having $3N - 1$ dimensions and radius N^* . As it turns out, such an integral resolves to:

$$\int dn_1 \dots dn_{3N} = \frac{2\pi^{3N/2} (N^*)^{3N-1}}{(3N/2 - 1)!}$$

The anticipated multiplicity of the ideal gas of energy U and fixed volume V and particle number $N \gg 1$ reads

$$\Omega(U, V, N) = \frac{1}{N!} \frac{V^N}{h^{3N}} \frac{2\pi^{3N/2}}{(3N/2)!} (2mU)^{3N/2} \quad (36.13)$$

Problem 5

A container of volume V is filled with one mole of ideal gas. Calculate the probability that a region with volume $V/1000$ will be completely void of particles. Answer:

$$P = \frac{\Omega(U, V \frac{999}{1000}, N)}{\Omega(U, V, N)} = \left(\frac{999}{1000} \right)^{6 \times 10^{23}} \approx 0$$

Entropy

With the multiplicity in hand, take the natural logarithm and multiply by the Boltzmann constant K to attain the entropy of the ideal gas. (Of course, use Stirling’s approximation to handle the factorials.) After doing so, we find

$$S = NK \left(\ln \left(\frac{V}{N} \left(\frac{4\pi mU}{3Nh^2} \right)^{3/2} \right) + \frac{5}{2} \right), \quad (36.14)$$

a result known as the *Sackur-Tetrode* equation.

As a historical note, the pioneers of statistical mechanics hadn’t developed the insight to slip the factor of $1/N!$ into the multiplicity. The entropy they were calculating was wrong, namely the factor of $5/2$ incorrectly appeared as $3/2$, a mistake remembered as the *Gibbs paradox*.

Problem 6

Write the multiplicity Ω of the ideal gas in terms of the thermal de Broglie wavelength. Answer:

$$\Omega = \frac{2(3N/2)^{3N/2}}{N! (3N/2)!} \left(\frac{V}{\lambda_0^3} \right)^N$$

State Variables

From elementary thermodynamics, take the so-called first thermodynamic identity

$$dU = TdS - PdV + \mu dN$$

to derive three relations:

$$\begin{aligned} \left(\frac{\partial S}{\partial U} \right)_{V,N} &= \frac{1}{T} \\ \left(\frac{\partial S}{\partial V} \right)_{U,N} &= \frac{P}{T} \\ \left(\frac{\partial S}{\partial N} \right)_{U,V} &= -\frac{\mu}{T} \end{aligned}$$

The first two relations are trivial to evaluate after inserting equation (36.14), recovering well-known results for the ideal gas:

$$\begin{aligned} U &= \frac{3}{2} NKT \\ PV &= NKT \\ H &= U + PV = \frac{5}{2} NKT \end{aligned}$$

From the above we calculate the heat capacity at constant volume and at constant pressure:

$$\begin{aligned} C_V &= \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{3}{2} NK \\ C_P &= \left(\frac{\partial H}{\partial T} \right)_{P,N} = \frac{5}{2} NK \end{aligned}$$

The chemical potential also simplifies:

$$\begin{aligned} \mu &= -KT \ln \left(\frac{S}{N} - \frac{5}{2} \right) \\ &= -KT \ln \left(\frac{V}{N} \left(\frac{4\pi m U}{3N h^2} \frac{3}{2} NKT \right)^{3/2} \right) \\ &= -KT \ln \left(\frac{V}{N \lambda_0^3} \right) \end{aligned}$$

Length λ_0 is known as the *thermal de Broglie wavelength*, namely

$$\lambda_0 = \frac{h}{\sqrt{2\pi mKT}}. \quad (36.15)$$

Accounting for dilute (non-ideal) solutions, we tack on a constant μ_0 to the chemical potential to write

$$\mu = \mu_0 - KT \ln \left(\frac{V}{N \lambda_0^3} \right),$$

or as a function of pressure,

$$\mu = \mu_0 + KT \ln \left(\frac{P}{P_0} \right),$$

where $P_0 = KT/\lambda_0^3$.

Other Dimensions

The ideal gas multiplicity equation (36.13) can be condensed via

$$\Omega(U, V, N) = \left(L^3 \right)^N U^{3N/2} f(N),$$

where generalizing the number of dimensions from 3 to α , we get

$$\Omega(U, V, N) = L^{\alpha N} U^{\alpha N/2} f(N).$$

Proceeding in analog to the three-dimensional case, the state variables for a gas occupying α dimensions are related by:

$$\begin{aligned} U &= \frac{\alpha}{2} NKT \\ PL^\alpha &= NKT \end{aligned}$$

Problem 7

Write the multiplicity of an ideal gas occupying an α -dimensional box with side length L . Answer:

$$\Omega(U) = \frac{L^{\alpha N}}{h^{\alpha N}} \frac{2\pi^{\alpha N/2}}{N! (\alpha N/2)!} (2mU)^{\alpha N/2}$$

Problem 8

Consider a three-dimensional ideal gas consisting of N_3 particles occupying a volume L^3 that can exchange energy (but not matter) with a two-dimensional ideal gas consisting of N_2 particles on a surface of area L^2 . Let the total energy U of the system equal the sum $U_3 + U_2$, the energies of the individual subsystems. Find (i) the total number of microstates of the system as a function of E_2 , and find (ii) the equilibrium value of U_2 and $U_3 = U - U_2$. Answer:

$$\Omega = \left(\sqrt{U - U_2} \right)^{3N_3} \left(\sqrt{U_2} \right)^{2N_2} f(L, N_2, N_3)$$

$$\left(\frac{\partial S}{\partial U_2} \right)_{L^2, N} = 0$$

$$\bar{U}_2 = \frac{N_2 U}{N_2 + 3N_3/2}$$

$$\bar{U}_3 = \frac{3N_3 U/2}{N_2 + 3N_3/2}$$

3 Canonical Ensemble

A physical system with constant volume V and constant number of particles N in an environment of fixed temperature T is a *canonical ensemble*. The internal energy U is not constant.

Helmholtz Free Energy

Consider two canonical systems A and B that combine to a composite system C , where the internal energies sum to a constant $U = U_A + U_B$ such that $dU_B = -dU_A$.

The total multiplicity in terms of U_A reads

$$\Omega_C(U_A) = \Omega_A(U_A) \Omega_B(U - U_A) .$$

Taking the natural logarithm of both sides gives

$$\ln(\Omega_C(U_A)) = \ln(\Omega_A(U_A)) + \ln(\Omega_B(U - U_A)) .$$

Evaluating to first order near $U_A \approx 0$ and $U_B \approx U$, we find

$$\ln(\Omega_C(U_A)) \approx \ln(\Omega_A(U_A)) - U_A \frac{\partial \ln(\Omega_B)}{\partial U_B} .$$

Multiplying through by K converts the above to an entropy equation

$$S_C(U_A) \approx S_A(U_A) - U_A \left(\frac{\partial S_B}{\partial U_B} \right)_{V, N, U_B \approx U} ,$$

where the derivative term is identically equal to $1/T$. So far then, we have

$$-TS_C(U_A) \approx U_A - TS_A(U_A) .$$

From elementary thermodynamics, note the right side is simply the Helmholtz free energy of system A , and happens to be the quantity that the canonical ensemble seeks to minimize. Therefore we deduce

$$A = -KT \ln(\Omega_C(U_A))$$

is a formula for the Helmholtz free energy. Moreover, we see that A is minimized when the combined entropy is maximized.

3.1 Boltzmann Factor

Returning to the first-order cononical ensemble relation

$$\ln \Omega_C(U_A) \approx \ln \Omega_A(U_A) - U_A \frac{\partial \ln \Omega_B}{\partial U_B}$$

for connected systems, use the identity

$$\frac{\partial \ln(\Omega_B)}{\partial U_B} = \frac{1}{KT}$$

and exponentiate both sides:

$$e^{\ln \Omega_C(U_A)} \approx e^{\ln \Omega_A(U_A)} e^{-U_A/KT}$$

$$\Omega_C(U_A) \approx \Omega_A(U_A) e^{-U_A/KT}$$

$$\Omega_C(U_A) \propto e^{-U_A/KT}$$

Note too that $\Omega_A \ll \Omega_C$ and could have been ignored. The final term on the right is known as the *Boltzmann factor*.

For a second motivation, recall the entropy of the universe is the sum of the system entropy plus that of its environment:

$$dS_{\text{uni}} = \left(\frac{1}{T} (dU + PdV - \mu dN) \right)_{\text{sys}} + dS_{\text{env}}$$

We are in the regime of constant N , and any change dV is much less than dU for the system or the environment. Near equilibrium with $dS_{\text{uni}} = 0$, we can write the above equation for two neighboring states 1, 2 as

$$0 = \frac{\Delta U_{\text{sys}}}{T} + \Delta S_{\text{env}} .$$

To continue, recall from elementary statistics that the ratio

$$\left(\frac{\Omega(2)}{\Omega(1)} \right)_{\text{env}} = \left(\frac{\text{Pr}(2)}{\text{Pr}(1)} \right)_{\text{env}}$$

holds, which says the probability $P(j)$ of finding the environment in state j is proportional to the number of corresponding microstates available in the environment.

Using the definition $S = K \ln(\Omega)$, we derive

$$\begin{aligned} \left(\frac{\text{Pr}(2)}{\text{Pr}(1)} \right)_{\text{env}} &= e^{(S(2) - S(1))/K} \\ &= e^{-(U(2) - U(1))/KT} , \end{aligned}$$

or

$$\left(\frac{\text{Pr}(2)}{\text{Pr}(1)} \right)_{\text{env}} = \left(\frac{e^{-(U(2))/KT}}{e^{-(U(1))/KT}} \right)_{\text{sys}} .$$

From this, we deduce that the probability of finding the system in state j is (once again) proportional to the Boltzmann factor:

$$\text{Pr}(j) \propto e^{-U(j)/KT}$$

Two-State Paramagnet

Recall that a two-state paramagnet in an external magnetic field has N total particles where $N_\uparrow$ particles align with the external field, and $N_\downarrow$ oppose the field such that $N = N_\uparrow + N_\downarrow$. The internal energy of the system is given by $U = -N_\uparrow\mu B + N_\downarrow\mu B$.

In the $N_\uparrow, N_\downarrow$ notation, the ratio inside the logarithm in Equation (36.7) is equivalent to the ratio $N_\uparrow/N_\downarrow$. Solving for this, we get

$$\frac{N_\uparrow}{N_\downarrow} = e^{2\mu B/KT},$$

where using $N = N_\uparrow + N_\downarrow$, we find

$$\begin{aligned}\frac{N_\uparrow}{N} &= \frac{e^{\mu B/KT}}{e^{\mu B/KT} + e^{-\mu B/KT}} \\ \frac{N_\downarrow}{N} &= \frac{e^{-\mu B/KT}}{e^{\mu B/KT} + e^{-\mu B/KT}}.\end{aligned}$$

Each ratio $N_{\uparrow,\downarrow}/N$ can be interpreted as the probability $g_{\uparrow,\downarrow}$ of finding a random particle in the up- or down-state, and the energy of any one particle shall be denoted ϵ such that $\epsilon(N_\uparrow) = -\mu B$ and $\epsilon(N_\downarrow) = \mu B$. The above pair of equations reduces to

$$\Pr(\epsilon) = \frac{1}{Z} e^{-\epsilon/KT}$$

such that

$$\epsilon(N_{\uparrow,\downarrow}) = \mp\mu B,$$

where the exponential term is the Boltzmann factor and

$$Z = e^{\mu B/KT} + e^{-\mu B/KT}$$

is the normalization constant.

3.2 Partition Function

Converting from the multiplicity Ω to a probability is done by dividing by a normalization constant Z . In terms of the Boltzmann factor, the probability of the system having energy state U_j is

$$\Pr(j) = \frac{1}{Z} e^{-U_j/KT}, \quad (36.16)$$

where Z is called the canonical *partition function*, given by

$$Z = \sum_j e^{-U_j/KT}, \quad (36.17)$$

such that the probability obeys the normalization condition

$$1 = \sum_j \Pr(j).$$

If more than one microstate corresponds to the same energy U_j , then the sum may be recast over all energies by introducing a degeneracy factor as

$$Z = \sum_j g(j) e^{-U_j/KT},$$

where $g(j)$ is a dimensionless number that counts the number of states for a given energy.

Supposing U_A lives on a continuous energy spectrum instead of a discrete one, the sum becomes an integral according to

$$Z = \int g(U) e^{-U/KT} dU.$$

Recasting over position- and momentum-space, and generalizing to α dimensions, the partition function is

$$Z = \frac{1}{h^\alpha} \int \int e^{-H(x,p)/KT} d^\alpha x d^\alpha p, \quad (36.18)$$

where $H(x, p)$ is the Hamiltonian of the system. Compare to Equation (36.12).

3.3 Non-interacting Particles

Consider a system of two particles A and B having internal energy $U = U_A(s_A) + U_B(s_B)$, where $s_{A,B}$ denotes the state of particle A, B respectively.

Distinguishable Particles

If particle A is somehow different than particle B , they are a *distinguishable* pair. The combined partition function $Z_{\text{indist.}}$ must involve the sum over all states s_A, s_B such that:

$$\begin{aligned}Z_{\text{dist.}} &= \sum_{s_A} \sum_{s_B} e^{-(U_A(s_A) + U_B(s_B))/KT} \\ &= \sum_{s_A} e^{-U_A(s_A)/KT} \sum_{s_B} e^{-U_B(s_B)/KT} \\ &= Z_A Z_B\end{aligned}$$

Evidently the combined partition function is the product $Z_A Z_B$, which easily generalizes to N particles:

$$Z_{\text{indist.}} = \prod_{k=1, \dots, N} Z_k$$

Indistinguishable Particles

Particles that are identical have a curious implication for physical systems: there is no discernible way to know if distinguishable particles trade places. From combinatorial arguments, it follows that the partition function for indistinguishable particles must be

$$Z_{\text{indist.}} = \frac{1}{N!} \prod_{k=1, \dots, N} Z_k .$$

Problem 9

Consider two groups A, B of non-interacting indistinguishable particles. Show that the total partition function is

$$Z = \left(\frac{1}{N_A!} z_{1_A}^{N_A} \right) \left(\frac{1}{N_B!} z_{1_B}^{N_B} \right) ,$$

where $z_{1_A}^{N_A}, z_{1_B}^{N_B}$ are the one-particle partition functions of each group.

Two-State System

A ubiquitous and nontrivial partition function applies to the two-state system. For a special case, suppose a single particle can occupy two energy levels $U = \pm\epsilon$, and these are the only two states (no degeneracy). For this we write the partition function

$$z_1 = e^{-\epsilon/KT} + e^{\epsilon/KT} = 2 \cosh \left(\frac{\epsilon}{KT} \right) .$$

Now, let us introduce a second distinguishable particle with access to the same energy levels $U = \pm\epsilon$, and the two particles do not interact. The combined energy of the system can be any of $-2\epsilon, 0, 2\epsilon$. The ground state however is not unique, as there are two distinct ways for $\pm\epsilon + \pm\epsilon$ to sum to zero. Thus, the two-particle partition function is

$$z_2 = e^{2\epsilon/KT} + 2e^0 + e^{-2\epsilon/KT} ,$$

simplifying to

$$z_2 = 2^2 \cosh^2 \left(\frac{\epsilon}{KT} \right) = z_1^2 ,$$

which we'd expect for distinguishable particles. For a total of N distinguishable particles, the partition function is

$$Z = z_1^N = 2^N \cosh^N \left(\frac{\epsilon}{KT} \right) .$$

3.4 Observables

Borrowing more from probability theory, we have that an *observable* variable X has an ensemble average $\langle X \rangle$ that is calculated via

$$\langle X \rangle = \sum_j \hat{X}_j \Pr(j) ,$$

where $\hat{X}_j$ is the corresponding operator. The brackets on $\langle X \rangle$ are often omitted for brevity.

Internal Energy

The internal energy U of a canonical system is straightforwardly written as an ensemble average:

$$\langle U \rangle = \frac{1}{Z} \sum_j U_j e^{-U_j/KT}$$

This can be simplified considerably. Begin by noticing

$$U = \frac{-1}{Z} \sum_j \left(\frac{\partial}{\partial (KT)^{-1}} e^{-U_j/KT} \right) ,$$

or

$$U = \frac{-1}{Z} \frac{\partial}{\partial (KT)^{-1}} Z .$$

Using the chain rule, we finally have:

$$U = - \frac{\partial}{\partial (KT)^{-1}} \ln(Z) \quad (36.19)$$

Problem 10

Calculate the internal energy of a two-state system consisting of N distinguishable particles. Examine the extreme temperature limits $T \rightarrow \infty, T \rightarrow 0$ and compare the first result to the high-temperature formula for a two-state paramagnet. Answer:

$$U = -N\epsilon \tanh(\epsilon/KT)$$

$$U_{T \rightarrow \infty} = \frac{-N\epsilon^2}{KT}$$

$$U_{T \rightarrow 0} = -N\epsilon$$

Energy Fluctuations

In many scenarios, the internal energy is sharply-peaked around its average value $\langle U \rangle$, where we approximate the profile to be Gaussian, i.e.

$$\Pr(U) = \tilde{A} e^{-(U - \langle U \rangle)^2 / 2\sigma^2} ,$$

where $\tilde{A}$ is a normalization constant.

Meanwhile, the continuous probability in terms of the Boltzmann factor is

$$\Pr(U) = g(U) e^{-U/KT}.$$

Equating these and taking the natural logarithm we find

$$\ln(g(U)) - \frac{U}{KT} = \ln(\tilde{A}) - \frac{(U - \langle U \rangle)^2}{2\sigma^2}.$$

Taking a single partial derivative with respect to U , and then evaluating at $U \rightarrow \langle U \rangle$, the above gives

$$\frac{1}{KT} = \frac{\partial}{\partial U} \ln(g(U)) \Big|_{U \rightarrow \langle U \rangle}.$$

The second-order derivative is more interesting. Carrying this out, we find

$$-\frac{1}{KT^2} \frac{\partial T}{\partial U} = \frac{\partial}{\partial U} \left(\frac{\partial}{\partial U} \ln(g(U)) \Big|_{U \rightarrow \langle U \rangle} \right) - \frac{1}{\sigma^2},$$

where $\partial T / \partial U$ is the reciprocal of the heat capacity at constant volume, letting us conclude

$$\sigma = \sqrt{KT^2 C_V}.$$

For example, an ideal gas at about 300°C has $KT \approx (1/40) \text{ eV}$ (electron-volts). The fluctuation width σ evaluates to about $2 \times 10^{10} \text{ eV}$, making $\sigma/U \approx 10^{-12}$.

Entropy

To work out the entropy of the canonical ensemble, note first that despite the energies U_j not being equal, all states j are equally-accessible. From this, we reason that probability of finding the system in one particular state is the reciprocal of the multiplicity:

$$\Pr(j) = \frac{1}{\Omega}$$

By definition, the entropy is calculated from

$$\langle S \rangle = \sum_j \hat{S} \Pr(j),$$

where the operator $\hat{S}$ is given by

$$\hat{S} = K \ln(\Omega) = -K \ln(\Pr(j)).$$

Proceed by substitution and write

$$S = -K \sum_j \ln\left(\frac{1}{Z} e^{-U_j/KT}\right) \frac{1}{Z} e^{-U_j/KT},$$

and simplify to find

$$S = \frac{K}{Z} \ln(Z) \sum_j e^{-U_j/KT} + \frac{1}{T} \langle U \rangle.$$

Solving for $U - TS$, also known as the Helmholtz free energy, we derive a powerful relationship to the partition function:

$$A = -KT \ln(Z) \quad (36.20)$$

Beta Parameter

The quantity $(KT)^{-1}$ is sufficiently ubiquitous to warrant its own abbreviation, namely

$$\beta = \frac{1}{KT}. \quad (36.21)$$

This makes the internal energy easier to write, as Equation (36.19) reduces to

$$U = -\frac{\partial}{\partial \beta} \ln(Z).$$

Take a T -derivative across Equation (36.21) to write a useful identity:

$$\frac{d\beta}{dT} = \frac{-1}{KT^2} = \frac{-\beta}{T}$$

Problem 11

It's easy to show from the thermodynamic identity

$$dA = -SdT - PdV + \mu dN$$

that various derivatives of the Helmholtz free energy yield variables of state:

$$-S = \left(\frac{\partial A}{\partial T} \right)_{V,N}$$

$$-P = \left(\frac{\partial A}{\partial V} \right)_{T,N}$$

$$\mu = \left(\frac{\partial A}{\partial N} \right)_{T,V}$$

Simplify or otherwise verify each of the above. Answer:

$$\begin{aligned} S &= \frac{\partial}{\partial T} (KT \ln(Z)) \\ &= K \ln(Z) + KT \frac{\partial}{\partial \beta} \ln(Z) \left(\frac{d\beta}{dT} \right) \\ &= \frac{-A + U}{T} \end{aligned}$$

$$P = KT \frac{\partial}{\partial V} \ln(Z)$$

$$= \frac{1}{Z} \sum_j \left(-\frac{\partial U_j}{\partial V} \right) e^{-U_j/KT} = \langle P \rangle$$

$$\mu_{\text{dist.}} = -KT \frac{\partial}{\partial N} \ln(z_1^N)$$

$$= KT \ln\left(\frac{1}{z_1}\right)$$

$$\mu_{\text{indist.}} = -KT \frac{\partial}{\partial N} \ln\left(\frac{z_1^N}{N!}\right)$$

$$= -KT \ln(z_1) + KT \ln(N!)$$

$$\approx -KT \ln(z_1) + KT \frac{\partial}{\partial N} (N \ln(N) - N)$$

$$= KT \ln\left(\frac{N}{z_1}\right)$$

3.5 Applications

Entropic Spring

In the entropic spring model, we consider a polymer chain of $N \gg 1$ links under a tension force F along the x -direction, where any link orients either parallel or perpendicular to x . Links parallel to x contribute length a to the chain, with total length L . Links perpendicular to x do not contribute to length.

We then write two energy values

$$\epsilon_{\parallel} = -Fa$$

$$\epsilon_{\perp} = 0,$$

and thus the partition function of a single link is given by

$$z_1 = 1 + e^{Fa/KT}.$$

Let $\langle x \rangle$ equal the average length per segment. Using the definition, we proceed with:

$$\langle x \rangle = \sum_{k=\parallel, \perp} \hat{x}(k) \Pr(k)$$

$$= \frac{(0)1 + (a)e^{Fa/KT}}{1 + e^{Fa/KT}}$$

$$= \frac{a}{1 + e^{-Fa/KT}}$$

Using $\langle L \rangle = N \langle x \rangle$, we find the average length to be

$$\langle L \rangle = \frac{Na}{1 + e^{-Fa/KT}}.$$

In the high-temperature limit, the above readily reduces to

$$\langle L \rangle = L_0 + \frac{NFa^2}{4KT},$$

where $L_0 = Na/2$ corresponds to zero force applied. Solve for F to recover a Hookean statement:

$$F = \frac{4KT}{Na^2} (\langle L \rangle - L_0)$$

Equipartition Theorem

In elementary thermodynamics, one references the idea of equipartition of energy to explain, for instance, factors like $3/2$ as it appears in the internal energy equation $U = (3/2)NKT$ for the ideal gas. In particular, one has that the internal energy per molecule gains a factor $KT/2$ for each independent quadratic degree of freedom.

The plan is to prove that the equipartition theorem is a special case of a case where molecules are governed by the Hamiltonian

$$H(x, p) = H(p) + bx^\alpha.$$

For a single molecule, write the partition function using Equation (36.18):

$$z_1 = \frac{1}{h} \int e^{-H(p)/KT} dp \int e^{-bx^\alpha/KT} dx$$

Let the momentum integral be represented by a function $f(T, V)$ (absorbing the Planck factor as well). Also let $w = bx^\alpha/KT$ such that $dw = \alpha bx^{\alpha-1} dx/KT$. With this we attain

$$z_1 = f(T, V) \frac{1}{\alpha} \left(\frac{KT}{b}\right)^{1/\alpha} \int_0^\infty w^{1/\alpha-1} e^{-w} dw.$$

The final integral doesn't need to be explicitly evaluated, but nonetheless can be shortened to $\Gamma(1/w)$.

Generalizing to N total particles, the partition function is written

$$Z = F(T, V, N) \left(\frac{KT}{b}\right)^{N/\alpha} \left(\frac{\Gamma(1/w)}{w}\right)^N,$$

where $F(T, V, N)$ is another arbitrary function that contains at least f and an optional factor of $N!$ that depends on the distinguishability of molecules.

With Z on hand, we next compute the internal energy using Equation (36.19). Anticipating the derivative, note that the term involving the gamma function is strictly numerical

and is ignored. The F -term has unknown T -dependence, however we can ignore this as well, as it is purely a kinetic term while we're interested in internal energy only. Thus we need to calculate

$$U = -\frac{\partial}{\partial (KT)^{-1}} \ln \left(\left(\frac{KT}{b} \right)^{N/\alpha} \right),$$

or

$$U = \frac{N}{\alpha} \frac{\partial}{\partial \beta} \ln (\beta b),$$

finally yielding

$$\frac{U}{N} = \frac{KT}{\alpha}.$$

That is, the energy per particle is generally equal to KT/α , where $\alpha = 2$ generates the equipartition theorem.

Maxwell Speed Distribution

In an ideal gas, the probability of observing a particle with speed v in the interval $v_1 < v < v_2$ is given by

$$P(v_1 < v < v_2) = \int_{v_1}^{v_2} D(v) dv,$$

where $D(v) dv$ is the Maxwell speed distribution:

$$D(v) dv = \left(\frac{m}{2\pi KT} \right)^{3/2} 4\pi v^2 e^{-mv^2/KT} dv$$

The derivation of $D(v)$ begins with slipping the kinetic energy of a particle into the Boltzmann factor

$$e^{-U/KT} = e^{-mv^2/2KT},$$

which is the basis for the normalization condition

$$1 = A \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-mv^2/2KT} dv_x dv_y dv_z,$$

with A constant. The integral is over all of velocity space in three-dimensional Cartesian coordinates. Switching to spherical coordinates and noticing the lack of an angular variable, the above reduces to

$$1 = A \int_0^{\infty} e^{-mv^2/2KT} 4\pi v^2 dv.$$

Next let $w^2 = mv^2/2KT$ to write

$$1 = 4\pi A \left(\frac{2KT}{m} \right)^{3/2} \int_0^{\infty} w^2 e^{-w^2} dw,$$

and notice the remaining Gaussian integral is resolved via

$$\int_0^{\infty} w^2 e^{-w^2} dw = \frac{\sqrt{\pi}}{4}.$$

This allows solving for A and the form for $D(v)$ is recovered.

3.6 Canonical Ideal Gas

The ideal gas system can be analyzed as a microcanonical ensemble.

Partition Function

For a single particle in a volume V , the partition function is

$$z_1 = \frac{1}{h^3} \int d^3x \int e^{-p^2/2mKT} d^3p.$$

The spatial integral trivially evaluates to V . The momentum integral can be recast in spherical coordinates:

$$z_1 = \frac{V}{h^3} 4\pi \int_0^{\infty} p^2 e^{-p^2/2mKT} dp$$

Next let $w^2 = p^2/2mKT$ to write

$$z_1 = \frac{V}{h^3} 4\pi (2mKT)^{3/2} \int_0^{\infty} w^2 e^{-w^2} dw,$$

where the remaining Gaussian integral resolves to $\sqrt{\pi}/4$. For z_1 , we finally have

$$z_1 = \frac{V}{h^3} (2\pi mKT)^{3/2}.$$

The single-particle partition function can be recast in terms of the thermal de Broglie wavelength using Equation (36.15). With this, z_1 is simply

$$z_1 = \frac{V}{\lambda_0^3}.$$

For a gas of N indistinguishable particles, the total partition function is

$$Z = \frac{1}{N!} (z_1)^N.$$

State Variables

The internal energy per particle of the ideal gas is calculated from Equation (36.19) via

$$\frac{U}{N} = -\frac{\partial}{\partial (KT)^{-1}} \ln(z_1) ,$$

or

$$U = N \frac{\partial}{\partial \beta} \ln \left(\frac{h^3}{V} \left(\frac{\beta}{2\pi m} \right)^{3/2} \right) = \frac{3}{2} NKT ,$$

as expected.

Calculating the Helmholtz free energy A from the partition function Z requires Stirling's approximation to write

$$\ln(Z) \approx -N \ln(N) + N + N \ln \left(\frac{V}{\lambda_0^3} \right)$$

such that

$$A = -NKT \left(1 + \ln \left(\frac{V}{N\lambda_0^3} \right) \right) , \quad (36.22)$$

or equivalently,

$$A = -NKT \ln \left(\frac{eV}{N\lambda_0^3} \right) .$$

Using $A = U - TS$, the entropy readily simplifies to

$$S = NK \left(\ln \left(\frac{V}{N\lambda_0^3} \right) + \frac{5}{2} \right) ,$$

in agreement with the Sackur-Tetrode Equation (36.14). One may alternatively calculate the entropy from

$$S = - \left(\frac{\partial A}{\partial T} \right)_{V,N} .$$

Explicitly:

$$\begin{aligned} S &= \frac{-A}{T} - 3NKT \frac{d}{dT} \ln(\lambda_0) \\ &= \frac{-A}{T} + \frac{3}{2} NKT \frac{d}{dT} \ln \left(\frac{2\pi mKT}{h^2} \right) \\ &= (U - A) / T \end{aligned}$$

Recalling that the V -derivative of A is $-P$, i.e.,

$$-P = \left(\frac{\partial A}{\partial V} \right)_{T,N} ,$$

insert Equation (36.22) to recover the ideal gas law

$$PV = NKT .$$

The T -derivative of U is the heat capacity:

$$C_V = \left(\frac{\partial U}{\partial T} \right)_{V,N} = \frac{3}{2} NKT$$

Finally, the chemical potential for indistinguishable particles is given by

$$\mu = KT \ln \left(\frac{N}{z_1} \right) = -KT \ln \left(\frac{V}{N\lambda_0^3} \right) ,$$

in agreement with the microcanonical analysis.

Relativistic Gas

When a particle moves relativistically with momentum p , the kinetic energy $T = p^2/2m$ is replaced by $T = pc$, where c is the speed of light. Otherwise following the same derivation that applies to the canonical ideal gas, one finds the partition function of a single particle of relativistic ideal gas to be

$$z_1 = 4\pi V \left(\frac{KT}{hc} \right)^3 \int_0^\infty w^2 e^{-w} dw ,$$

where the remaining integral resolves to 2. If the gas is made of indistinguishable particles, the total partition function is

$$Z = \frac{1}{N!} \left(8\pi V \left(\frac{KT}{hc} \right)^3 \right)^N .$$

Problem 12

Suppose a relativistic ideal gas comprised of $3N$ particles is confined to a one-dimensional channel of length L . Show that the partition function is:

$$Z = \frac{1}{(3N)!} \left(2L \left(\frac{KT}{hc} \right) \right)^{3N}$$

Gas in Gravity

Consider an ideal gas of $N \gg 1$ molecules of mass m at temperature T in a cylinder of base area B and height L .

Suppose too that each molecule is subject to Earth's local gravity and is some height $z < L$ above the ground. The single-particle partition function is written

$$z_1 = \frac{B}{h^3} \int_0^L e^{-mgz/KT} dz \int e^{-p^2/2mKT} d^3p ,$$

which contains a new Boltzmann factor $e^{-mgz/KT}$, where mgz is the gravitational potential energy of the particle.

Handling each integral separately, we have

$$\int_0^L e^{-mgz/KT} dz = \left(\frac{KT}{mg} \right) (1 - e^{-mgL/KT}) ,$$

along with

$$\frac{1}{h^3} \int e^{-p^2/2mKT} d^3p = \left(\frac{2\pi mKT}{h^2} \right)^{3/2} = \frac{1}{\lambda_0^3} ,$$

so the single-particle partition function is

$$z_1 = \frac{V}{\lambda_0^3} \left(\frac{KT}{mgL} \right) (1 - e^{-mgL/KT}) .$$

Per standard ideal gas analysis, we take all particles as indistinguishable and write the total partition function as

$$Z = \frac{1}{N!} (z_1)^N .$$

To anticipate energy calculations, it helps to write the natural logarithm of the partition function as

$$\ln(Z) = Z_0 + N \ln \left(\frac{1 - e^{-mgL\beta}}{mgL\beta} \right) ,$$

where Z_0 applies to the unperturbed ideal gas. Then, the Helmholtz free energy is

$$A = A_0 - NKT \ln \left(\frac{1 - e^{-mgL/KT}}{mgL/KT} \right) ,$$

where A_0 applies to the unperturbed ideal gas.

The internal energy U is calculated straightforwardly from Equation (36.19), leading to

$$U = U_0 + NKT - \frac{NmgL}{1 - e^{-mgL/KT}} ,$$

where U_0 applies to the unperturbed ideal gas. From the above we discern the following limits:

$$U_{L \rightarrow \infty} = \frac{5}{2} NKT$$

$$U_{mgL \ll KT} = U_0$$

The volume derivative of the Helmholtz free energy is related to the pressure, and we use this to calculate the force at the top of the cylinder:

$$\begin{aligned} F = BP &= - \frac{\partial A}{\partial L} \Big|_{T,N} \\ &= NKT \frac{\partial}{\partial L} \left(\ln \left(\frac{BL}{N\lambda_0^3} \right) + \ln \left(\frac{1 - e^{-mgL/KT}}{mgL/KT} \right) \right) \\ &= \frac{NKT}{L} \left(\frac{mgL/KT}{e^{mgL/KT} - 1} \right) \end{aligned}$$

From the above we discern the following limits:

$$F_{L \rightarrow \infty} = 0$$

$$F_{mgL \ll KT} = \frac{NKT}{L}$$

The average height $\langle z \rangle$ of a molecule from the base of the cylinder is calculated via

$$\langle z \rangle = \frac{1}{z_g} \int_0^L z e^{-mgz/KT} dz ,$$

where normalization is achieved by

$$z_g = \int_0^L e^{-mgz/KT} dz .$$

Instead of a brute-force calculation, notice that a tricky derivative of the Boltzmann factor gets the same job done. In particular:

$$\begin{aligned} \langle z \rangle &= \frac{1}{z_g} \left(\frac{-KT}{m} \right) \int_0^L \frac{\partial}{\partial g} (e^{-mgz/KT}) dz \\ &= \left(\frac{-KT}{m} \right) \frac{\partial}{\partial g} \ln(z_g) \end{aligned}$$

Notice too that z_g was evaluated previously, and we could substitute into the above and carry on with precision. In the approximation $L \rightarrow \infty$, as one may use for Earth's atmosphere, then z_g reduces to KT/mg , thus

$$\langle z \rangle \approx \left(\frac{-KT}{m} \right) \frac{\partial}{\partial g} \ln \left(\frac{KT}{mg} \right) = \frac{KT}{mg} .$$

Problem 13

If the average thickness of Earth's atmosphere is 10 km, use $g = 9.8 \text{ m/s}^2$ and $T = 300 \text{ K}$ to estimate the mass of a molecule of air. Answer: $m \approx 4.23 \times 10^{-26} \text{ kg}$

Interacting Particles

Consider a gas of N particles whose interactions are governed by a homogeneous function of the n th degree:

$$W(\lambda \mathbf{r}_1, \dots, \lambda \mathbf{r}_N) = \lambda^n W(\mathbf{r}_1, \dots, \mathbf{r}_N)$$

Due to the scale invariance inherent to V , one may claim the partition function to have the form

$$Z(T, V) = T^a f(V T^b) ,$$

where f is an arbitrary function, and a, b are functions of N and n .

Putting $Z(T, V)$ through the meat grinder, we first find

$$\ln(Z) = a \ln(T) + \ln\left(f(VT^b)\right),$$

betraying the Helmholtz free energy, internal energy, and pressure of such a gas:

$$\begin{aligned} A &= -KT \left(a \ln(T) + \ln\left(f(VT^b)\right) \right) \\ U &= -\frac{\partial}{\partial \beta} \ln(Z) = KT \left(a + \frac{f'}{f} b VT^b \right) \\ P &= -\left(\frac{\partial A}{\partial V} \right)_{T,N} = KT \left(\frac{f'}{f} T^b \right) \end{aligned}$$

Eliminating the f' -dependence from the above yields the equation of state:

$$U = aKT + bPV$$

This is as far as we get without getting dirty. Now we work toward finding a, b by first writing the partition function of a single particle:

$$z_1 = \int_0^V e^{-\lambda^n W/KT} d^3x \int e^{-p^2/2mKT} d^3p$$

Proceed with the substitutions

$$\begin{aligned} x &\rightarrow x' \lambda \\ p &\rightarrow p' \lambda^{n/2} \end{aligned}$$

so z_1 becomes

$$\begin{aligned} z_1 &= \left(\lambda^3 \int_0^{V/\lambda^3} e^{-\lambda^n W/KT} d^3x' \right) \\ &\quad \times \left(\lambda^{3n/2} \int e^{-\lambda^n p'^2/2mKT} d^3p' \right), \end{aligned}$$

or, accounting for all N particles,

$$Z = \lambda^{3N(1+n/2)} Z(T\lambda^{-n}, V\lambda^{-3}).$$

Using the given formula for the partition function, we further have

$$Z = \lambda^{3N(1+n/2)} (T\lambda^{-n})^a f\left((V\lambda^{-3})(T\lambda^{-n})^b\right).$$

In order to retain $Z(T, V) = T^a f(VT^b)$, it can only follow that

$$\begin{aligned} a &= 3N \left(\frac{1}{n} + \frac{1}{2} \right) \\ b &= \frac{-3}{n}. \end{aligned}$$

In terms of n , the equation of state now reads

$$\frac{n}{3}U + PV = NKT \left(1 + \frac{n}{2} \right).$$

Van der Waals Model

The Van der Waals model is a generalized ideal gas model that introduces an effective volume

$$V \rightarrow V - Nb,$$

where b is a constant pertaining to the gas. Also, the Leonard-Jones attraction between particles imparts an average energy

$$\epsilon = \frac{-aN}{V},$$

where a is another constant.

Using the standard apparatus, we find the relevant partition function to be

$$Z = \frac{1}{N!} \left(\frac{V - Nb}{\lambda_0^3} e^{aN/VKT} \right)^N$$

and furthermore

$$\ln(Z) \approx N + N \ln \left(\frac{V - Nb}{N\lambda_0^3} \right) + \frac{aN^2}{VKT}.$$

The internal energy is the negative β -derivative of $\ln(Z)$, which trivially works out to

$$U = U_0 - \frac{aN^2}{V},$$

where U_0 applies to the unperturbed ideal gas. Next, the Helmholtz free energy is

$$A = -NKT \left(1 + \ln \left(\frac{V - Nb}{N\lambda_0^3} \right) \right) - \frac{aN^2}{V},$$

or equivalently

$$A = A_0 - NKT \ln \left(1 - \frac{Nb}{V} \right) - \frac{aN^2}{V},$$

where A_0 applies to the unperturbed ideal gas.

Using $A = U - TS$, the entropy is given by

$$S = \frac{U_0 - A_0}{T} + NK \ln \left(1 - \frac{Nb}{V} \right),$$

or equivalently

$$S = NK \left(\ln \left(\frac{V - Nb}{N\lambda_0^3} \right) + \frac{5}{2} \right),$$

which is not dependent on a .

Let P_0 equal the pressure of the unperturbed ideal gas. Then, the pressure in the Van der Waals model is

$$P = - \left(\frac{\partial A}{\partial V} \right)_{T,N}$$

$$P = P_0 + \frac{NKT}{1 - Nb/V} \left(\frac{Nb}{V^2} \right) - \frac{aN^2}{V^2}.$$

Simplifying delivers the famed Van der Waals equation:

$$\left(P + \frac{aN^2}{V^2} \right) (V - Nb) = NKT$$

In conditions of constant temperature and constant number of particles, the thermodynamic identity $dG = -SdT + VdP + \mu dN$ can be integrated via

$$G = \int VdP.$$

Writing the Van der Waals equation as

$$P = \frac{NKT}{V - Nb} - \frac{aN^2}{V^2},$$

we have

$$dP = \frac{-NKT}{(V - Nb)^2} dV + \frac{2aN^2}{V^3} dV.$$

Carrying out the integral gives the Gibbs free energy up to an integration constant that is not a function of volume or pressure:

$$G = -NKT \ln(V - Nb) + \frac{N^2 b KT}{V - Nb} - \frac{2aN^2}{V} + C(T)$$

Problem 14

Calculate the Gibbs free energy of the Van der Waals gas using $G = U + PV - TS$. Use the result to show that the integration constant in the Gibbs free energy is given by

$$C(T) = NKT \ln(N\lambda_0^3).$$

Internal Freedom

Consider a gas of N non-interacting particles with internal degrees of freedom. Each particle has an internal spectrum of energies ϵ_j with degeneracies g_j , where $j = 0, 1, 2, \dots, n$.

The single-particle partition function is

$$z_1 = \left(\frac{V}{\lambda_0^3} \right) z_{\text{int}},$$

where

$$z_{\text{int}} = \sum_{j=0}^n g_j e^{-\epsilon_j/KT}.$$

As a gas of indistinguishable particles, the total partition function is

$$Z = \frac{1}{N!} \left(\frac{V}{\lambda_0^3} z_{\text{int}} \right)^N.$$

Letting A_0 , U_0 , S_0 equal the unperturbed Helmholtz free energy, internal energy, and entropy, we have:

$$A = A_0 - NKT \ln(z_{\text{int}})$$

$$U = U_0 + NKT^2 \frac{\partial}{\partial T} \ln(z_{\text{int}})$$

$$S = S_0 + NK \ln(z_{\text{int}}) + NKT \frac{\partial}{\partial T} \ln(z_{\text{int}})$$

If z_{int} is not a function of the volume, then the ideal gas law still holds:

$$P = - \left(\frac{\partial A}{\partial V} \right)_{T,N} = \frac{NKT}{V}$$

This makes quick work of the enthalpy

$$H = U + PV = \frac{5}{2} NKT + NKT \frac{\partial}{\partial T} \ln(z_{\text{int}}),$$

and now the heat capacities can be worked out:

$$C_V = \left(\frac{\partial U}{\partial T} \right)_{V,N}$$

$$= \frac{3}{2} NK + 2NKT \frac{\partial}{\partial T} \ln(z_{\text{int}})$$

$$+ NKT^2 \frac{\partial^2}{\partial T^2} \ln(z_{\text{int}})$$

$$C_P = \left(\frac{\partial H}{\partial T} \right)_{P,N} = C_V + NK$$

3.7 Simple Harmonic Oscillator

The simple harmonic oscillator (SHO) is any system characterized by an attractive potential

$$V(x) = \frac{m\omega^2 x^2}{2},$$

where motion is sinusoidally centered around $x = 0$, m is the mass, and ω is the angular frequency of the oscillator.

By approximating solid materials as a collection of harmonic oscillators, it follows that each oscillator is fixed in place, in which case all particles are distinguishable. Moreover, the volume is approximately constant, and the pressure is identically zero.

Classical Oscillator System

A classical system consisting of N identical simple harmonic oscillators is governed by the Hamiltonian

$$H = \sum_{j=1}^N \left(\frac{p_j^2}{2m} + \frac{m\omega^2 x_j^2}{2} \right),$$

where x_j , p_j are the position and momentum of the j th oscillator.

For a single oscillator, the partition function is straightforwardly calculated:

$$\begin{aligned} z_1 &= \frac{1}{h} \int_{-\infty}^{\infty} e^{-m\omega^2 x^2/2KT} dx \int_{-\infty}^{\infty} e^{-p^2/2mKT} dp \\ &= \frac{1}{h} \sqrt{\frac{2KT}{m\omega^2}} \sqrt{2mKT} \left(\int_{-\infty}^{\infty} e^{-w^2} dw \right)^2 \end{aligned}$$

The parenthesized Gaussian integral resolves to $\sqrt{\pi}$. Also remembering $\hbar = h/2\pi$, the single-oscillator partition function reduces to

$$z_1 = \frac{KT}{\hbar\omega} = (\beta\hbar\omega)^{-1}.$$

Generalizing to N distinguishable oscillators, the total partition function is given by

$$Z = z_1^N = \left(\frac{KT}{\hbar\omega} \right)^N = (\beta\hbar\omega)^{-N}.$$

With the partition function on hand, we swiftly calculate the Helmholtz free energy A and the internal energy U :

$$A = -KT \ln(Z) = NKT \ln \left(\frac{\hbar\omega}{KT} \right)$$

$$U = -\frac{\partial}{\partial \beta} \ln(Z) = NKT$$

The entropy of the classical SHO system is given by

$$\begin{aligned} S &= - \left(\frac{\partial A}{\partial T} \right)_{V,N} = \frac{U - A}{T} \\ &= NK \left(1 - \ln \left(\frac{\hbar\omega}{KT} \right) \right) \end{aligned}$$

Having no pressure, the enthalpy is no different from the internal energy:

$$H = U + PV = NKT$$

Using $G = H - TS$, we find the Gibbs free energy to be the same as the Helmholtz free energy:

$$G = NKT \ln \left(\frac{\hbar\omega}{KT} \right)$$

Finally, the chemical potential is given by:

$$\mu = \left(\frac{\partial A}{\partial N} \right)_{T,V} = \frac{G}{N} = KT \ln \left(\frac{\hbar\omega}{KT} \right)$$

Problem 15

Calculate the grand free energy

$$\Phi = U - TS - \mu N$$

for the classical SHO system. Answer: 0

Quantum Oscillator System

A system of simple harmonic oscillators can be analyzed as a quantum system. Using the same potential energy $V(x) = m\omega^2 x^2/2$, all dynamics are given by the Schrodinger equation, with a notable result being the energy ϵ_n of the oscillator as a function of excitation number n :

$$\epsilon_n = \hbar\omega \left(n + \frac{1}{2} \right)$$

Using $\exp(\epsilon_n/KT)$ as the Boltzmann factor, the partition function z_1 of a single oscillator is

$$z_1 = \sum_{n=0}^{\infty} e^{-\epsilon_n/KT} = e^{-\hbar\omega/2KT} \sum_{n=0}^{\infty} \left(e^{-\hbar\omega/KT} \right)^n,$$

which accounts for all energies accessible to the oscillator. The final sum is a geometric series, and z_1 simplifies as

$$z_1 = \frac{e^{-\hbar\omega/2KT}}{1 - e^{-\hbar\omega/KT}} = \frac{1}{2 \sinh(\hbar\omega/2KT)}.$$

Accounting for all N particles, the partition function for a canonical ensemble of quantum oscillators is

$$Z = \frac{e^{-N\hbar\omega/2KT}}{(1 - e^{-\hbar\omega/KT})^N} = \left(2 \sinh\left(\frac{\hbar\omega}{2KT}\right)\right)^{-N}.$$

With the partition function on hand, we calculate the Helmholtz free energy A and the internal energy U :

$$\begin{aligned} A &= N \left(\frac{\hbar\omega}{2} + KT \ln(1 - e^{-\hbar\omega/KT}) \right) \\ &= NKT \ln \left(2 \sinh\left(\frac{\hbar\omega}{2KT}\right) \right) \\ U &= N\hbar\omega \left(\frac{1}{e^{\hbar\omega/KT} - 1} + \frac{1}{2} \right) \\ &= N\hbar\omega \coth\left(\frac{\hbar\omega}{2KT}\right) \end{aligned}$$

The combination $1/(e^{\hbar\omega/KT} - 1)$ is significant enough to warrant its own symbol, for which we reserve the *occupation number* $\langle n \rangle$:

$$\langle n \rangle = \frac{1}{e^{\hbar\omega/KT} - 1}$$

In terms of the occupation number, the internal energy is written

$$U = N\hbar\omega \left(\langle n \rangle + \frac{1}{2} \right),$$

which is in many ways analogous to the energy formula ϵ_n .

Let $\langle n \rangle = 0$ correspond to the zero point energy U_0 , particularly

$$U_0 = \frac{N\hbar\omega}{2}.$$

Then, the energies on hand become

$$\begin{aligned} A &= U_0 + KT \ln(1 - e^{-\hbar\omega/KT}) \\ U &= U_0 + N\hbar\omega \langle n \rangle. \end{aligned}$$

Using $A = U - TS$, the entropy works out to be

$$S = \frac{N\hbar\omega}{T} \langle n \rangle - NK \ln(1 - e^{-\hbar\omega/KT}).$$

To recover the classical results, take the limit $KT \gg \hbar\omega$ so the occupation number becomes

$$\langle n \rangle_{KT \gg \hbar\omega} \approx \frac{KT}{\hbar\omega}.$$

Problem 16

Calculate the enthalpy H , Gibbs free energy G , chemical potential μ , and grand free energy Φ of the quantum oscillator system and show that each reduces to its classical analog at high temperature.

Quantum Oscillator Gas

Consider an ideal gas where each indistinguishable particle is a quantum harmonic oscillator. For such a system we begin with the partition function of the gas

$$Z = \frac{1}{N!} \left(\frac{V}{\lambda_0^3} z_{\text{int}} \right)^N,$$

where

$$z_{\text{int}} = \frac{1}{2 \sinh(\hbar\omega/2KT)}.$$

We showed previously that the internal energy is given by

$$U = U_0 + NKT^2 \frac{\partial}{\partial T} \ln \left(\frac{1}{2 \sinh(\hbar\omega/2KT)} \right),$$

simplifying as:

$$\begin{aligned} U &= U_0 + N \frac{\partial}{\partial \beta} \ln \left(2 \sinh\left(\frac{\beta\hbar\omega}{2}\right) \right) \\ &= \frac{3}{2}NKT + \frac{1}{2}N\hbar\omega \coth\left(\frac{\hbar\omega}{2KT}\right) \end{aligned}$$

The heat capacity at constant volume is the T -derivative of the above, thus we have, after simplifying:

$$C_V = \frac{3}{2}NK + NK \frac{(\hbar\omega/2KT)^2}{\sinh^2(\hbar\omega/2KT)}$$

From the heat capacity we discern the following limits:

$$\begin{aligned} C_V(T \rightarrow 0) &= \frac{3}{2}NK \\ C_V(T \rightarrow \infty) &= \frac{5}{2}NK \end{aligned}$$

3.8 Canonical Paramagnetism

Classical Paramagnet

Consider a ‘solid’ system of distinguishable non-interacting particles in conditions of fixed temperature T and fixed external magnetic field $\mathbf{B} = B\hat{z}$. In the classical model, each particle’s magnetic moment $\boldsymbol{\mu}$ may orient at an arbitrary angle θ with respect to the z -axis, having energy

$$U = -\boldsymbol{\mu} \cdot \mathbf{B} = \mu_z B_z = -\mu B \cos(\theta).$$

With momentum playing no role in magnetization, the partition function for a single particle in the system is written as a spatial integral over the shell of a three-sphere:

$$z_1 = 2\pi \int_0^\pi e^{\mu B \cos(\theta)/KT} \sin(\theta) d\theta$$

Quickly solve this by letting $x = \mu B \cos(\theta)/KT$ to find

$$z_1 = \frac{4\pi KT}{\mu B} \sinh\left(\frac{\mu B}{KT}\right).$$

The average magnetization $\langle \mu_z \rangle$ of a single particle is written:

$$\langle \mu_z \rangle = \frac{2\pi}{z_1} \int_0^\pi \mu \cos(\theta) e^{\mu B \cos(\theta)/KT} \sin(\theta) d\theta$$

Using the same substitution, the integral resolves to

$$\langle \mu_z \rangle = \mu \left(\coth\left(\frac{\mu B}{KT}\right) - \frac{KT}{\mu B} \right),$$

an application of the *Langevin function*.

In the high-temperature limit, the parenthesized quantity reduces to $\mu B/3KT$, known as the *Curie law*, which states that the magnetization of a paramagnetic particle is proportional to B and inversely proportional to T :

$$\langle \mu_z \rangle \propto \frac{\mu^2 B}{K T}$$

The factor before B/T is called the *Curie constant* for the given material.

Quantum Two-State Paramagnet

The quantum two-state paramagnet has two energy levels $U = \pm\epsilon$, where $\epsilon = -\mu B$. As a typical two-state system, we know the partition function of a single particle is given by

$$z_1 = 2 \cosh\left(\frac{\mu B}{KT}\right).$$

Made from N distinguishable particles, the total partition function of the quantum two-state paramagnet is

$$Z = z_1^N = 2^N \cosh^N\left(\frac{\mu B}{KT}\right),$$

and the internal energy is easily shown to be

$$U = -N\mu B \tanh\left(\frac{\mu B}{KT}\right).$$

The average magnetization $\langle \mu \rangle$ is straightforwardly calculated via

$$\langle \mu \rangle = \frac{1}{z_1} \sum_{j=1,2} \mu_j e^{-\mu_j B/KT},$$

where $\mu_j = \pm\mu$. One may evaluate the sum manually, or notice that a proper derivative of z_1 emits the same structure:

$$\begin{aligned} \langle \mu \rangle &= \frac{-1}{z_1} \frac{\partial}{\partial (B/KT)} \sum_{j=1,2} e^{-\mu_j B/KT} \\ &= -\frac{\partial}{\partial (B/KT)} \ln(z_1) \\ &= -\mu \tanh\left(\frac{\mu B}{KT}\right) \end{aligned}$$

Quantum Many-State Paramagnet

In a paramagnet exposed to an external magnetic field $\mathbf{B} = B \hat{z}$, each particle's spin vector has a quantized projection on the z -axis. If a given particle's total spin is J , it follows that the z -projected spin is given by $m_s \hbar$, where m_s is the spin quantum number ranging from $-J$ to J in integer or half-integer steps.

The Hamiltonian of one particle is

$$H = -\mu_z B = -\gamma m_s B,$$

where γ is the *gyromagnetic ratio* of the particle species. For an isolated electron we have $\gamma = g_e \mu_B / \hbar$, where $g_e \approx 2.00$ and μ_B is the *Bohr magneton*, given by $\mu_B = e\hbar/2m_e$, where m_e is the electron mass and e is the electron charge.

The partition function of a single particle is straightforwardly written as the sum of all possible Boltzmann factors, namely

$$z_1 = \sum_{m_s=-J}^{m_s=J} \left(e^{-\gamma B/KT} \right)^{m_s}.$$

Let $q = \gamma B/KT$ and break apart the sum:

$$\begin{aligned} z_1 &= \sum_{m_s=-J}^{m_s=0} (e^{-q})^{m_s} + \sum_{m_s=0}^{m_s=J} (e^{-q})^{m_s} - 1 \\ &= \sum_0^J (e^q)^x + \sum_0^J (e^{-q})^x - 1 \end{aligned}$$

Then, using the geometric series, we have

$$z_1 = \frac{1 - e^{q(J+1)}}{1 - e^q} + \frac{1 - e^{-q(J+1)}}{1 - e^{-q}} - 1,$$

simplifying ultimately to

$$z_1 = \frac{\sinh(\gamma B (J + 1/2) / KT)}{\sinh(\gamma B / 2KT)} .$$

By computing the observable $\langle M \rangle \propto \langle \gamma m_s \rangle$, it can be shown that the magnetization of a paramagnet with N particles per unit volume is

$$\langle M \rangle = N \gamma J B_J(x) ,$$

where $x = \gamma J B / KT$, and the function $B_J(x)$ is called the *Brillouin function*, which is a generalization of the Langevin function as follows:

$$B_J(x) = \frac{2J+1}{2J} \coth\left(\frac{2J+1}{2J}x\right) - \frac{1}{2J} \coth\left(\frac{x}{2J}\right)$$

If the magnetic field strength B is sufficiently low and the temperature T is sufficiently high, it can be shown that the

Curie constant for any J is

$$C = \frac{\partial \langle M \rangle}{\partial B} = \frac{\mu_B^2}{3K} N g_e^2 J (J + 1) .$$

The special case $J = 1/2$ makes the Brillouin function simplify to $\tanh(x)$, which in the small- x limit reduces to x , matching the previously-found result for the two-state paramagnet.

Classical behavior is recovered from the quantum analysis by allowing the spin vector to occupy the whole continuum of solid angle, accomplished by letting $J \rightarrow \infty$. In this case, the Brillouin function reduces, as expected, to the Langevin function.

4 Grand Canonical Ensemble

So far we have studied in detail the microcanonical ensemble (constant V, N, U) and the canonical ensemble (constant V, N, T). Each of these forbid particles from flowing between the system and the environment.

Naturally, there is another creature called the *grand canonical ensemble* that accounts for systems with non-constant N at constant volume.

Grand Free Energy

Consider two grand canonical systems A and B that combine to a composite system C . The internal energy and number of particles are allowed to flow between the systems, meaning

$$\begin{aligned} U &= U_A + U_B \\ N &= N_A + N_B . \end{aligned}$$

All in terms of A -variables, the entropy of system C reads

$$S_C = S_A(N_A, U_A) + S_B(N - N_A, U - U_A) .$$

From the entropy we affirm the condition for thermal equilibrium

$$\begin{aligned} \frac{\partial S_C}{\partial U_A} = 0 &= \left(\frac{\partial S_A}{\partial U_A} \right)_{V_A, N_A} - \left(\frac{\partial S_B}{\partial U_A} \right)_{V_B, N_B} \\ 0 &= \frac{1}{T_A} - \frac{1}{T_B} , \end{aligned}$$

along with diffusive equilibrium:

$$\begin{aligned} \frac{\partial S_C}{\partial N_A} = 0 &= \left(\frac{\partial S_A}{\partial N_A} \right)_{U_A, V_A} - \left(\frac{\partial S_B}{\partial N_A} \right)_{U_B, V_B} \\ 0 &= \frac{-\mu_A + \mu_B}{T} \end{aligned}$$

We seek an overall quantity that is minimized at equilibrium in analogy to the Helmholtz and Gibbs free energies. To proceed, write the composite entropy as a Taylor expansion in the vicinity $U_A \approx 0, N_A \approx 0$:

$$\begin{aligned} S_C &\approx S_A(U_A, N_A) + S_B(U_B, N_B) \\ &\quad - U_A \left(\frac{\partial S_B}{\partial U_A} \right)_{V_B, N_B} - N_A \left(\frac{\partial S_B}{\partial N_A} \right)_{U_B, V_B} \end{aligned}$$

In the above, we may suppose system B is much larger than system A , so that S_B is essentially a very large constant

and is ignored. Replacing derivatives with known quantities, the above tells us

$$S_C(U_A, N_A) \approx S_A - \frac{U_A}{T} + \frac{\mu N_A}{T} ,$$

or

$$TS_C = TS_A - U_A + \mu N_A .$$

Notice that the remaining quantity on the right is the negative grand free energy of system A , thus we find, dropping the subscript:

$$TS_C = -\Phi = -(U - TS + \mu N)$$

Using the definition $S = K \ln(\Omega)$, conclude that

$$\Phi = -KT \ln(\Omega_C) .$$

That is, we find that the system tries to minimize the grand free energy.

4.1 Gibbs Factor

Returning to the first-order grand canonical ensemble relation

$$S_C(U_A, N_A) \approx S_A - \frac{U_A}{T} + \frac{\mu N_A}{T}$$

for connected systems, divide through by K and exponentiate both sides:

$$e^{S_C(U_A)/K} \approx e^{S_A(U_A)/K} e^{(-U_A + \mu N_A)/KT}$$

Note too that $S_A \ll S_C$ and is ignored. The final term on the right is known as the *Gibbs factor*.

For another motivation, one may retrace the development of the canonical ensemble while including the μdN term as it appears in

$$dS_{\text{uni}} = \left(\frac{1}{T} (dU + PdV - \mu dN) \right)_{\text{sys}} + dS_{\text{env}} .$$

This distills to the notion that finding the system in state j is proportional to the Gibbs factor:

$$\text{Pr}(j) \propto e^{(-U(j) + \mu N(j))/KT}$$

Note that $N(j)$ is the number of particles in the system when in state j with energy $U(j)$. If the system has multiple species of particles, make the replacement

$$\mu N \rightarrow \sum_k \mu_k N_k .$$

4.2 Grand Partition Function

Much as the canonical probability relates to the Boltzmann factor, we use the Gibbs factor to write the probability of a particle in state j as

$$\Pr(j) = \frac{1}{Z_G} e^{(-U(j) + \mu N(j))/KT},$$

where z_G is the *grand partition function*:

$$Z_G = \sum_j e^{(-U(j) + \mu N(j))/KT}$$

In terms of the Boltzmann factor, the grand partition function is also written

$$Z_G = \sum_j z_B(j) z_f^{N(j)},$$

where z_B is the Boltzmann factor and z_f is a fugacity term defined as

$$z_f = e^{\mu/KT}.$$

If there are multiple species of particles present, the grand partition function accordingly becomes:

$$\begin{aligned} Z_G &= \sum_j e^{(-U(j) + \mu_1 N_1(j) + \mu_2 N_2(j) + \dots)/KT} \\ &= \sum_j \sum_k e^{(-U(j) + \mu_k N_k(j))/KT} \end{aligned}$$

4.3 Observables

In a statistical system, recall that a variable X has an ensemble average $\langle X \rangle$ that is calculated via

$$\langle X \rangle = \sum_j \hat{X}(j) \Pr(j),$$

where $\hat{X}(j)$ is the corresponding operator.

Internal Energy

The internal energy U of a grand canonical system is straightforwardly written:

$$\langle U \rangle = \frac{1}{Z_G} \sum_j U(j) e^{(-U(j) + \mu N(j))/KT}$$

As with the canonical ensemble analysis, the above can be simplified using a trick of differentiation. Begin by noticing

$$U = \frac{-1}{Z_G} \left(\frac{\partial}{\partial (KT)^{-1}} Z_G \right)_{z_f, V},$$

simplifying (using the chain rule) to:

$$U = - \left(\frac{\partial}{\partial (KT)^{-1}} \ln(Z_G) \right)_{z_f, V} \quad (36.23)$$

Problem 17

Show that the internal energy of the grand canonical ensemble can be written:

$$U = KT^2 \left(\frac{\partial}{\partial T} \ln(Z_G) \right)_{z_f, V}$$

Number of Particles

For a system of N particles, the ensemble average $\langle N \rangle$ at constant V and constant T is calculated via

$$\langle N \rangle = \frac{1}{Z_G} \sum_j N(j) e^{(-U(j) + \mu N(j))/KT}.$$

After simplifying, the above can be written:

$$N = \left(KT \frac{\partial}{\partial \mu} \ln(Z_G) \right)_{V, T} \quad (36.24)$$

Entropy

The statistical entropy is straightforwardly calculated from

$$\begin{aligned} S &= -K \sum_j \ln \left(\frac{1}{Z_G} e^{(-U(j) + \mu N(j))/KT} \right) \\ &\times \frac{1}{Z_G} e^{(-U(j) + \mu N(j))/KT}, \end{aligned}$$

simplifying to

$$S = \frac{U}{T} + K \ln \left(\frac{Z_G}{z_f^N} \right). \quad (36.25)$$

Helmholtz Free Energy

Using the definition $A = U - TS$, one readily reads from Equation (36.25) that the Helmholtz free energy of the grand canonical system is given by:

$$A = -KT \ln \left(\frac{Z_G}{z_f^N} \right) \quad (36.26)$$

Grand Free Energy

Using the definitions $\Phi = U - TS - \mu N$ and $A = U - TS$, we derive a formula for the grand free energy as

$$\Phi = -KT \ln \left(\frac{Z_G}{z_f^N} \right) - NKT \ln \left(e^{\mu/KT} \right),$$

reducing to

$$\Phi = -KT \ln (Z_G). \quad (36.27)$$

Equation of State

Recalling finally that the identity for the grand free energy $\Phi = -PV$ always holds, this combines with Equation (36.27) to make an equation of state:

$$PV = KT \ln (Z_G) \quad (36.28)$$

4.4 Grand Canonical Ideal Gas

For a canonical ideal gas of N particles with volume V at temperature T , recall that the partition function is given by

$$Z_B = \frac{1}{N!} \left(\frac{V}{\lambda_0^3} \right)^N,$$

where λ_0 is the thermal de Broglie wavelength

$$\lambda_0 = \frac{h}{\sqrt{2\pi mKT}}.$$

For the grand canonical ensemble, the corresponding partition function reads:

$$Z_G = \sum_k \frac{1}{N_k!} \left(\frac{V}{\lambda_0^3} \right)^{N_k} z_f^{N_k}$$

The equation for Z_G takes exponential form if we let the upper limit on k go to infinity. To see this, let $y = Vz_f/\lambda_0^3$ so that

$$Z_G = \sum_{k=0}^{\infty} \frac{y^{N_k}}{N_k!} = e^y = e^{Vz_f/\lambda_0^3},$$

or

$$\ln (Z_G) = \frac{Vz_f}{\lambda_0^3}.$$

Immediately from $\ln (Z_G)$, we calculate the internal energy and the pressure

$$U = \frac{3}{2} NK \ln (Z_G)$$

$$P = \frac{KT}{V} \ln (Z_G),$$

and eliminate the $\ln (Z_G)$ -term to find

$$U = \frac{3}{2} PV.$$

We have yet to show that $\ln (Z_G)$ is equivalent to the number of particles N . Using Equation (36.24), we find

$$N = \frac{VKT}{\lambda_0^3} \frac{\partial}{\partial \mu} \left(e^{\mu/KT} \right) = \frac{Vz_f}{\lambda_0^3},$$

and the equivalence is established.

Problem 18

Show that $N = \ln (Z_G)$ by working backward from the chemical potential of the ideal gas:

$$\mu = -KT \ln \left(\frac{V}{N\lambda_0^3} \right)$$

Part VII

Computation

Chapter 37

Iterative Methods

1 Matrix Toolkit

Review

The linear system $A\mathbf{x} = \mathbf{b}$ is ubiquitous in multivariate systems, and rears its face in a majority of numerical analysis applications. We'll restrict analysis to systems that can be exactly solved, which is to say the system is determined by n equations and n unknowns, corresponding to a square matrix A with n rows and n columns.

Often, we're interested in 'solving' for $\mathbf{x}$, which requires having the inverse of A , namely A^{-1} such that

$$(A^{-1}A)\mathbf{x} = \mathbf{x} = A^{-1}\mathbf{b}.$$

Pursuing a popular technique, define the sub-matrix S_{jk} that removes the j th row and the k th column from the original matrix A . (The width and height of S are each one less than those of A .) Define the matrix minor M_{jk} as the determinant of S_{jk} .

Also, let there be another matrix B that relates to A via

$$B_{jk} = (-1)^{j+k} M_{kj}.$$

That is, the jk th component of the matrix B is equal to the determinant of the S_{jk} sub-matrix multiplied by a constant. The reason for defining B this way is that the product AB equals the determinant of A multiplied by the identity matrix:

$$AB = (\det A) I$$

Multiply A^{-1} into each side to find a tight formula for the inverse:

$$A^{-1} = \frac{1}{\det A} B$$

A second way for determining the components x_j involves the augmented matrix

$$A|b = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} & b_1 \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} & b_2 \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} & b_3 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} & b_n \end{bmatrix},$$

which 'tacks on' the vector $\mathbf{b}$ to the right side of A so the final object has n rows and $n + 1$ columns.

The augmented matrix permits three row operations: (i) exchange of two rows, (ii) multiply a row by a scalar, (iii) replace a row by the sum of itself and another row. Using row operations, it's possible to bring the augmented matrix $A|b$ into upper triangular form:

$$A|b = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} & b_1 \\ 0 & A_{22} & A_{23} & \cdots & A_{2n} & b_2 \\ 0 & 0 & A_{33} & \cdots & A_{3n} & b_3 \\ 0 & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & A_{nn} & b_n \end{bmatrix}$$

Keep in mind the components A_{jk} , b_j are not the same between the original and its transformed version.

Starting with the upper triangular form, the matrix A is reduced again by similar steps to bring about the lower triangular form. This reduces A to having only diagonal entries:

$$A|b = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 & b_1 \\ 0 & A_{22} & 0 & 0 & 0 & b_2 \\ 0 & 0 & A_{33} & 0 & 0 & b_3 \\ 0 & 0 & 0 & \cdots & 0 & \cdots \\ 0 & 0 & 0 & 0 & A_{nn} & b_n \end{bmatrix}$$

Finally, attain the row-reduced echelon form by normalizing the j th row by A_{jj} to arrive at the form $I|x$, where I is the identity matrix.

To summarize, the process

$$A|b \rightarrow I|x$$

exposes the components of x as the right-most column of the augmented matrix $I|x$. This, of course, is Gaussian elimination.

Remarkably, row operations can also be used to calculate the inverse of a matrix. For an $n \times n$ matrix A , it turns out that the augmented matrix $A|I$, having n rows and $2n$ columns, may undergo the same treatment as $A|b$. That is, perform row operations to attain the transformation:

$$A|I \rightarrow I|A^{-1}$$

This is Gauss-Jordan elimination.

Pseudocode

The contents of a matrix A can be replaced by the contents of another matrix B by the following pseudocode:

```
# Requires matrix B
# Returns matrix A

function Equate(B)
  for each j in rows of B
    for each k in columns of B
      A(j, k) = B(j, k)
  return A
```

This process requires matrices A and B to be of the same dimension. Note how the ‘inner’ loop goes over the columns of B and the ‘outer’ loop goes by row, which is analogous to reading text on a page. Of course, there are many more ways to replace the j, k th component of matrix B , such as running the loops backwards.

Augmenting

For two matrices A and B having the same number of rows but no restriction of the number of columns, the augmented matrix $A|B$ is attained by the following:

```
# Requires A, B (same no. of rows)
# Returns AB
```

```
function Augment(A, B)
  r = rows in A or B
  c1 = number of columns in A
  c2 = number of columns in B
  for j from 1 to r
    for k from 1 to c1
      AB(j, k) = A(j, k)
    for k from 1 to c2
      AB(j, k + c1) = B(j, k)
  return AB
```

Matrix Multiplication

A matrix A having J rows and N columns multiplies into another matrix B with N rows and K columns to give a new matrix C with J rows and K columns. The j kth component of the resulting matrix is given by:

$$C_{jk} = \sum_{n=1}^N A_{jn} B_{nk}$$

Of course, there are $j \times k$ calculations like this to do, which is a dauntless endeavor for a computer. As pseudocode, the whole matrix C is calculated by:

```
# Requires A size J*N
# Requires B size N*K
# Returns C size JK

function Multiply(A, B)
  for each j in rows of A
    for each k in columns of B
      for each n in columns of A
        C(j, k) = C(j, k) + A(j, n) * B(n, k)
  return C
```

Sub-Matrix

For a matrix A , the sub-matrix S_{jk} that removes the j th row and k th column is established by the following.

```

# Requires matrix A size J*K
# Requires Row j, Column k
# Returns matrix S size (J-1)(K-1)

function SubMatrix(A, j, k)
    x = 0
    for each u in rows of A
        if (u <> j)
            y += 1
            x = 0
            for each v in columns of A
                if (v <> k)
                    x += 1
                    S(x, y) = A(u, v)
    return S

```

Determinant

For the square matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

the determinant of A is

$$\det A = A_{11}A_{22} - A_{12}A_{21}.$$

The most trivial matrix has $A = A_{11}$, in which case $\det A = A_{11}$.

For an $n \times n$ matrix, the determinant can be written from a recursive equation:

$$\det A = \sum_{\substack{j \leq n \\ k=1}}^n (-1)^{j+k} A_{jk} M_{jk}$$

Recall that M_{jk} is the determinant of the matrix minor S_{jk} , hence the recursive nature of the above.

The matrix determinant calculation is represented in the pseudocode below. Importantly, note that certain environments have trouble making perfect sense of $(-1)^{j+k}$. For this, the instruction `if (k MOD 2) = 0` is used instead to achieve the same ends, where MOD is the ‘modulus’ operator.

```

# Requires matrix A size n^2
# Returns Det(A)

function Determinant(A)
    n = side(A)
    if n = 1: det = A(1, 1)
    if n = 2: det = A(1, 1) * A(2, 2) -
                A(1, 2) * A(2, 1)
    if n > 2
        define matrix S size (n-1)x(n-1)
        for k from 1 to n
            S = SubMatrix(A, 1, k)
            m = Det(S)
            if (k MOD 2) = 0
                m = -m
            det = det + A(1, k) * m
    return det

```

Inverse Matrix

For an invertible matrix A , recall that the inverse A^{-1} is calculated by

$$A^{-1} = \frac{1}{\det A} B,$$

where the components of B depend on the sub-matrix minors M_{kj} as detailed above.

The pseudocode below steps through the inverse calculation and stores the result as `Inv`. Take note of the MOD operator performing a similar role as it does in the determinant calculation.

```

# Requires matrix A size n^2
# Returns inverse of A

function Inverse(A)
    define matrix S size (n-1)x(n-1)
    detA = Determinant(A) # nonzero
    for each j in rows of A
        for each k in columns of A
            S = Submatrix(A, j, k)
            m = Determinant(S)
            if ((j + k) MOD 2) = 1
                m = -1 * m
            Inv(j, k) = m / detA
    return Inv

```

Augmented Upper Triangular

The augmented matrix $A|b$ with n rows and $n + 1$ columns can be reduced to the upper triangular form (same dimensions) by the pseudocode below.

```
# Requires Ab size n*m
# Returns upper triangular

function UT(Ab)
  n = number of rows in Ab
  m = number of columns in Ab
  for j from 1 to n
    for k from (j+1) to n
      f = Ab(k, j) / Ab(j, j)
      for p from 1 to m
        Ab(k, p) = Ab(k, p) -
          - f * Ab(j, p)
  return Ab
```

For a bonus, this method also works for the $A|I \rightarrow I \rightarrow A^{-1}$ system with n rows and $2n$ columns. The structures b and I are interchangeable.

Augmented Lower Triangular

Modifying the previous example with surgical edits to the loop limits and step direction give a way to compute the lower triangular form of $A|b$ or $A|I$:

```
# Requires Ab size n*m
# Returns lower triangular

function LT(Ab)
  n = number of rows in Ab
  m = number of columns in Ab
  for j from n to 2 step -1
    for k from (j-1) to 1 step -1
      f = Ab(k, j) / Ab(j, j)
      for p from 1 to m
        Ab(k, p) = Ab(k, p) -
          - f * Ab(j, p)
  return Ab
```

Note that in the pursuit of $A|I \rightarrow I|A^{-1}$, the upper triangular and lower triangular calculations can be done in any order.

RREF Matrix

The augmented system $A|B$ can be wrestled into RREF format by the pseudocode that follows:

```
# Requires Ab size n*m
# Returns R(REF)

function RREF(AB)
  n = number of rows in AB
  m = number of columns in AB
  rref = LT(UT(AB))
  # Normalize:
  for j from 1 to n
    for k from n + 1 to m
      R(j, k) = R(j, k) / R(j, j)
    R(j, j) = R(j, j) / R(j, j) # = 1
  return rref
```


2 Curve Fitting

2.1 Regression Analysis

Regression analysis is an attempt to derive meaningful patterns from numerical data. To introduce the subject, we'll explore the scenario of fitting a curve $y = f(x)$ to a set of given data points $\{(x_j, y_j)\}$ in various ways.

Linear Fit

Suppose we're provided with the following set of ordered pairs:

x_j	0.6	1.8	2.8	3.6	4.2	5.6
y_j	1.6	1.6	2.6	2.0	4.0	3.6

Take each pair (x_j, y_j) with $j = 1, 2, \dots, n$ as a point in the Cartesian plane. Further, suppose there was reason to believe that the pattern in the provided points is described by a straight line in the plane

$$y = mx + b,$$

where the slope m and y-intercept b are unknown, and to be found using the data provided.

To advance on the problem, move all variables to one side, and consider n instances of the equation

$$h_j(m, b) = mx_j + b - y_j.$$

That is, h_j measures the vertical distance between the point $mx_j + b$ and y_j . If the approximate line passes directly through (x_j, y_j) , then $h_j(m, b) = 0$ for that point.

As defined, $h_j(m, b)$ could be a positive or a negative value, which would mean negative errors cancel out positive ones. To avoid this, let us work with the square of the vertical distance represented by h_j and call this a new function $F_j(m, b)$:

$$F_j(m, b) = (mx_j + b - y_j)^2$$

The total vertical distance from each point (x_j, y_j) to the line $y = mx + b$ is the sum of all F_j :

$$F(m, b) = \sum_{j=1}^n F_j(m, b) = \sum_{j=1}^n (mx_j + b - y_j)^2$$

Now comes the new idea. The ideal m and b for the data provided should correspond to a minimum in $F(m, b)$. That is, set the partial derivatives of F with respect to these

variables to zero, and the correct m, b are implicated. We then have

$$\begin{aligned} \frac{\partial F}{\partial m} &= 0 = \sum_{j=1}^n 2x_j (mx_j + b - y_j) \\ \frac{\partial F}{\partial b} &= 0 = \sum_{j=1}^n 2 (mx_j + b - y_j) . \end{aligned}$$

To keep the algebra contained, define the quantity

$$X^\alpha Y^\beta = \sum_{j=1}^n x_j^\alpha y_j^\beta,$$

which isn't treated as regular algebraic variable, for instance $(X)(X) \neq X^2$, and $(X)(Y) \neq XY$.

In terms of the sums X, Y , etc., the minimization of $F(m, b)$ gives a system of two equations with two unknowns:

$$bn + mX = Y$$

$$bX + mX^2 = YX$$

The solution is straightforward using matrix methods or traditional:

$$\begin{aligned} m &= \frac{(n)(XY) - (X)(Y)}{(n)(X^2) - (X)(X)} \\ b &= \frac{(Y)(X^2) - (X)(XY)}{(n)(X^2) - (X)(X)} \end{aligned}$$

To finish the example on hand, use a calculator to find $X = 18.6, Y = 15.4, XY = 55.28, X^2 = 73.4$. The final answer is:

$$m \approx 0.479$$

$$b \approx 1.082$$

Exponential Fit

Consider another set of points $\{(x_j, y_j)\}$ in the Cartesian plane that would be best approximated by an exponential fit

$$y(x) = A e^{mx},$$

where the scaling constant A and exponential parameter m are to be determined from the data given.

For this problem, let $A = e^b$, where b is another constant, and the equation becomes $y = e^{mx+b}$. Take the natural log of both sides to write

$$\ln(y) = mx + b.$$

From here, the problem is completely analogous to the straight-line fit, except all y_j are substituted for $\ln(y_j)$. One b is known, reverse the exponential to solve for A .

Power Law Fit

The same trick works for a power law fit, i.e.

$$y(x) = Ax^B,$$

in where

$$\ln(y) = \ln(A) + B \ln(x).$$

This is directly analogous to the linear case provided we let:

$$X^\alpha Y^\beta = \sum_{j=1}^n \ln(x_j)^\alpha \ln(y_j)^\beta$$

Polynomial Fit

For any set of n total points (x_j, y_j) in the Cartesian plane, we can try a polynomial of order $m < n$ to fit the data:

$$y(x) = A_0 + A_1x + A_2x^2 + \cdots + A_mx^m$$

Similar to the straight-line fit, define a vertical distance $h_j(\{A_m\})$ such that

$$h_j = A_0 + A_1x_j + A_2x_j^2 + \cdots + A_mx_j^m - y_j.$$

Square this distance and sum over all n data points:

$$\begin{aligned} F(\{A_m\}) &= \sum_{j=1}^n h_j^2 \\ &= \sum_{j=1}^n (A_0 + A_1x_j + \cdots + A_mx_j^m - y_j)^2 \end{aligned}$$

The best-fitting polynomial is the one that minimizes F with respect to all A_j simultaneously. Writing these out, one finds

$$\begin{aligned} \frac{\partial F}{\partial A_0} &= 2 \sum_{j=1}^n (A_0 + A_1x_j + \cdots + A_mx_j^m - y_j) \\ \frac{\partial F}{\partial A_1} &= 2 \sum_{j=1}^n x_j (A_0 + A_1x_j + \cdots + A_mx_j^m - y_j) \\ \frac{\partial F}{\partial A_2} &= 2 \sum_{j=1}^n x_j^2 (A_0 + A_1x_j + \cdots + A_mx_j^m - y_j), \end{aligned}$$

availing the pattern

$$\frac{\partial F}{\partial A_k} = 2 \sum_{j=1}^n x_j^k (A_0 + A_1x_j + \cdots + A_mx_j^m - y_j),$$

for any $k \leq m$.

Each derivative is zero on the left, and the universal factor of 2 drops out. Not forgetting to distribute the x_j^k term into each sum, all of the above information is best written in matrix notation. In particular, we have

$$M = \begin{bmatrix} n & X & X^2 & \cdots & X^m \\ X & X^2 & X^3 & \cdots & X^{m+1} \\ X^2 & X^3 & X^4 & \cdots & X^{m+2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X^m & X^{m+1} & X^{m+2} & \cdots & X^{2m} \end{bmatrix},$$

such that

$$M \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \cdots \\ A_m \end{bmatrix} = \begin{bmatrix} Y \\ YX \\ YX^2 \\ \cdots \\ YX^m \end{bmatrix}.$$

There is a lot of information to juggle with if you're insane enough to do this by hand. Regardless, system can be solved by finding the row-reduced echelon form of:

$$\begin{bmatrix} n & X & X^2 & \cdots & X^m & Y \\ X & X^2 & X^3 & \cdots & X^{m+1} & YX \\ X^2 & X^3 & X^4 & \cdots & X^{m+2} & YX^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X^m & X^{m+1} & X^{m+2} & \cdots & X^{2m} & YX^m \end{bmatrix}$$

For one example, consider the following list of $n = 11$ ordered pairs:

x_j	0	1	2	3	4	5
y_j	1	6	17	34	57	86

x_j	6	7	8	9	10
y_j	121	162	209	262	321

Supposing there were reason to choose $m = 3$ for a quadratic approximation, we would then write

$$\begin{bmatrix} n & X & X^2 \\ X & X^2 & X^3 \\ X^2 & X^3 & X^4 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} Y \\ YX \\ YX^2 \end{bmatrix},$$

which, using the numbers on hand, evaluates to

$$\begin{bmatrix} 11 & 55 & 385 \\ 55 & 385 & 3025 \\ 385 & 3025 & 25333 \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 1276 \\ 9900 \\ 82434 \end{bmatrix}.$$

By standard means, we ultimately find $A_0 = 1$, $A_1 = 2$, $A_2 = 3$, or:

$$y(x) = 1 + 2x + 3x^2$$

2.2 Interpolation

In regression analysis, a set of n total data points $\{(x_j, y_j)\}$, leads to a best-fit polynomial (or other curve) that passes near, but not necessarily through each data point. By a more powerful process called *interpolation*, it's possible to find a curve that passes through each data point. With some effort, such a curve can be made continuous and smooth in its domain.

Rectangle Approximation

The crudest interpolation of the provided data points is the rectangular approximation, which draws a horizontal line for all $n - 1$ points spanning from x_j to x_{j+1} such that

$$f_0(x_j \leq x < x_{j+1}) = y_j$$

Of course, we can also draw lines to the left instead of the right by a shift of index:

$$f_0(x_j \leq x < x_{j+1}) = y_{j+1}$$

Connect the Dots

A slightly more informative approximation to the provided data points is the linear interpolation, which is a fancy name for connect the dots. By standard straight line methods, successive data points are connect by lines given by

$$f_1(x) = y_j + (x - x_j) \left(\frac{y_{j+1} - y_j}{x_{j+1} - x_j} \right).$$

This has the appearance of a first terms of a Taylor approximation and also the form $y = mx + b$. All of these are equivalent to linear order.

There is another way to express the line $f_1(x)$ that appears mighty peculiar at first:

$$f_1(x) = y_j \left(\frac{x_{j+1} - x}{x_{j+1} - x_j} \right) + y_{j+1} \left(\frac{x - x_j}{x_{j+1} - x_j} \right)$$

Make sure the two expressions for $f_1(x)$ are equivalent.

Quadratic Interpolation

The linear interpolation can be improved by including an x^2 -like term in $f(x)$, giving a quadratic interpolation in terms of undermined coefficients:

$$f_2(x) = A_0 + A_1x + A_2x^2$$

One way to find the unknown coefficients is to pick three consecutive points, such as (x_j, y_j) with $j = 0, 1, 2$. This

generates three equations and three unknowns, which can be solved by standard means.

For a different approach to the problem, let us recite a trick named after Lagrange, which extends the 'peculiar' straight line method written above. The quadratic approximation is called the *Lagrange integrating polynomial*, and is given by

$$f_2(x) = y_0L_0(x) + y_1L_1(x) + y_2L_2(x),$$

where the $L_j(x)$ are called the *Lagrange interpolating basis functions*:

$$L_0 = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

$$L_1 = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$

$$L_2 = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

This can be straightforwardly reconciled with a traditional attack on the problem. In the original function $f_2(x)$, observe that A_2 is exactly one half of the second derivative of the function.

From ordinary calculus, we know the second derivative can be expressed via

$$f''(x) = \lim_{h \rightarrow 0} \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}.$$

This formula is a bit oversimplifying in the sense that h doesn't necessarily equal any given pair $x_j - x_k$, never mind the limit.

Proceed by differentiating $f_2(x)$ twice (only the x^2 term survives). To keep the algebra sane, define

$$\Delta x_{jk} = x_j - x_k,$$

and we have

$$\frac{1}{2}f_2''(x) = \frac{y_0}{\Delta x_{01}\Delta x_{02}} + \frac{y_1}{\Delta x_{10}\Delta x_{12}} + \frac{y_2}{\Delta x_{20}\Delta x_{21}}.$$

Establish a common denominator and rewrite:

$$\frac{1}{2}f_2''(x) = \frac{y_0\Delta x_{12} - y_1\Delta x_{02} + y_2\Delta x_{01}}{\Delta x_{01}\Delta x_{02}\Delta x_{12}}$$

This is technically as far as the comparison can go. To an approximation though, we can have

$$h = x_1 - x_0 = x_2 - x_1 = \frac{x_2 - x_0}{2}$$

and the two second derivative formulas agree.

Higher Orders

The Lagrange interpolating basis functions readily generalize to higher orders. With the products

$$Q_n(x) = \prod_{j \neq n} (x - x_j)$$

$$R_n(x) = \prod_{j \neq n} (x_n - x_j) ,$$

the n th function is

$$L_n(x) = \frac{Q_n(x)}{R_n(x)} .$$

In terms of the Lagrange interpolating basis functions, the corresponding $y_n(x)$ is

$$y_n(x) = \sum_{j=0}^n y_j L_j(x) .$$

2.3 Spline

There is a handy method that combines polynomial interpolation with derivative matching. That is, if we have a set of n points $\{(x_j, y_j)\}$, where $j = 1, 2, 3, \dots, n$, it's possible to come up with a curve that connects all of the points continuously and smoothly.

Cubic Approximation

In the domain $x_j < x < x_{j+1}$, a cubic interpolation going through the points x_j, x_{j+1} is assumed, taking generic form

$$f_j(x) = \sum_{j=0}^3 A_j x^j .$$

Continuity Conditions

For continuity across the entire function f , we must have

$$f_{j-1}(x_j) = f_j(x_j)$$

$$f_j(x_{j+1}) = f_{j+1}(x_{j+1}) ,$$

which says the same thing twice. (Substitute $j \rightarrow j + 1$ in the first equation to recover the second.)

For continuity in derivative of f , differentiate each of the above once with respect to x

$$f'_{j-1}(x_j) = f'_j(x_j)$$

$$f'_j(x_{j+1}) = f'_{j+1}(x_{j+1}) ,$$

and then once more:

$$f''_{j-1}(x_j) = f''_j(x_j)$$

$$f''_j(x_{j+1}) = f''_{j+1}(x_{j+1}) ,$$

If there are n total points provided, there must be $n - 1$ polynomials in the total interpolation, with a total of $4(n - 1)$ unknowns. However, only $4(n - 2)$ equations come from continuity arguments, and we need four more.

At the first point $j = 1$ and the last point $j = n$, we set

$$f_1(x_1) = y_1$$

$$f_{n-1}(x_n) = y_n .$$

With the leftover freedom to impose two more equations, choose the second derivative to be zero at each endpoint:

$$f''_1(x_1) = 0$$

$$f''_{n-1}(x_n) = 0$$

This setup is called the *natural spline*.

Second Derivative Continuity

This is enough to assemble the whole curve. Begin with the abbreviation

$$k_j = f''_{j-1}(x_j) = f''_j(x_j) .$$

Between neighboring k_j, k_{j+1} , the second derivative of $f''_j(x)$ is a straight line connecting the two. Express this line as a two-point Lagrange system:

$$f''_j(x) = k_j \left(\frac{x - x_{j+1}}{x_j - x_{j+1}} \right) + k_{j+1} \left(\frac{x - x_j}{x_{j+1} - x_j} \right)$$

Integrate the above in the x -variable once to get an equation for $f'_j(x)$

$$f'_j(x) = \int f''_j(x) dx + C ,$$

where C is an integration constant. Substituting the above and carrying out the integral, one finds

$$f'_j(x) = \frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^2}{2}$$

$$+ \frac{k_{j+1}}{(x_{j+1} - x_j)} \frac{(x - x_j)^2}{2} + C .$$

Integrate again to get an equation for $f_j(x)$

$$f_j(x) = \frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^3}{3 \cdot 2} + \frac{k_{j+1}}{(x_{j+1} - x_j)} \frac{(x - x_j)^3}{3 \cdot 2} + Cx + D,$$

where D is another integration constant.

The substitutions

$$C = A - B$$

$$D = -Ax_{j+1} + Bx_j$$

makes the above be slightly easier to work with:

$$f_j(x) = \frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^3}{3 \cdot 2} - \frac{k_{j+1}}{(x_j - x_{j+1})} \frac{(x - x_j)^3}{3 \cdot 2} + A(x - x_{j+1}) - B(x - x_j)$$

Integration Constants

The integration constants need to be determined before moving on. Evaluate $f_j(x_j)$ to find

$$y_j = \frac{k_j}{(x_j - x_{j+1})} \frac{(x_j - x_{j+1})^3}{3 \cdot 2} + A(x_j - x_{j+1}),$$

or

$$A = \frac{y_j}{(x_j - x_{j+1})} - k_j \frac{(x_j - x_{j+1})}{3 \cdot 2}.$$

Evaluate $f_j(x_{j+1})$ to find

$$y_{j+1} = \frac{k_{j+1}}{(x_j - x_{j+1})} \frac{(x_j - x_{j+1})^3}{3 \cdot 2} + B(x_j - x_{j+1}),$$

or

$$B = \frac{y_{j+1}}{(x_j - x_{j+1})} - k_{j+1} \frac{(x_j - x_{j+1})}{3 \cdot 2}.$$

Putting the whole solution together:

$$f_j(x) = \frac{k_j}{6} \left(\frac{(x - x_{j+1})^3}{(x_j - x_{j+1})} - (x - x_{j+1})(x_j - x_{j+1}) \right) - \frac{k_{j+1}}{6} \left(\frac{(x - x_j)^3}{(x_j - x_{j+1})} - (x - x_{j+1})(x_j - x_{j+1}) \right) + \frac{y_j(x - x_{j+1}) - y_{j+1}(x - x_j)}{x_j - x_{j+1}}$$

First Derivative Continuity

What remains is to determine the terms k_j in terms of $\{(x_j, y_j)\}$. Write out the derivatives $f'_j(x_j)$ and $f'_{j-1}(x_j)$ and set them equal (as agreed earlier). For an updated $f'_j(x)$, we have

$$f'_j(x) = \frac{k_j}{6} \left(\frac{3(x - x_{j+1})^2}{(x_j - x_{j+1})} - (x_j - x_{j+1}) \right) - \frac{k_{j+1}}{6} \left(\frac{3(x - x_j)^2}{(x_j - x_{j+1})} - (x_j - x_{j+1}) \right) + \frac{y_j - y_{j+1}}{x_j - x_{j+1}}$$

and also, shifting index,

$$f'_{j-1}(x) = \frac{k_{j-1}}{6} \left(\frac{3(x - x_j)^2}{(x_{j-1} - x_j)} - (x_{j-1} - x_j) \right) - \frac{k_j}{6} \left(\frac{3(x - x_{j-1})^2}{(x_{j-1} - x_j)} - (x_{j-1} - x_j) \right) + \frac{y_{j-1} - y_j}{x_{j-1} - x_j}.$$

For shorthand, define

$$\Delta x_{j+} = x_j - x_{j+1}$$

$$\Delta x_{j-} = x_j - x_{j-1}$$

and similar for the y-variables. Then evaluate each derivative equation at x_j and equate the results to get the continuity equation

$$k_{j+1} \Delta x_{j+} + 2k_j (\Delta x_{j+} - \Delta x_{j-}) - k_{j-1} \Delta x_{j-} = 6 \left(\frac{\Delta y_{j-}}{\Delta x_{j-}} - \frac{\Delta y_{j+}}{\Delta x_{j+}} \right).$$

Creating a System

For yet another shorthand, define the coefficients

$$\begin{aligned}\alpha_j &= -\Delta x_{j-} \\ \beta_j &= 2(\Delta x_{j+} - \Delta x_{j-}) \\ \gamma_j &= \Delta x_{j+} \\ \delta_j &= 6 \left(\frac{\Delta y_{j-}}{\Delta x_{j-}} - \frac{\Delta y_{j+}}{\Delta x_{j+}} \right)\end{aligned}$$

which are all known in terms of the provided points. Then, the above can be written

$$\alpha_j k_{j-1} + \beta_j k_j + \gamma_j k_{j+1} = \delta_j,$$

which has three unknowns in general, and we already decided $k_1 = k_n = 0$. Thus only the cases $j = 2, 3, 4, \dots, n-1$ need be written out.

Example n=5

Choosing a modest example with $n = 5$ points provided, the above becomes:

$$\begin{aligned}\alpha_2 0 + \beta_2 k_2 + \gamma_2 k_3 &= \delta_2 \\ \alpha_3 k_2 + \beta_3 k_3 + \gamma_3 k_4 &= \delta_3 \\ \alpha_4 k_3 + \beta_4 k_4 + \gamma_4 0 &= \delta_4\end{aligned}$$

This is a system of three equations and three unknowns k_2, k_3, k_4 , and the problem has been reduced to a regular $A\mathbf{x} = \mathbf{b}$ -like system.

In general, the spline calculation leads to a hefty augmented matrix with $n - 2$ rows and $n - 1$ columns.

2.4 Bézier Curves

Here we develop a way of planning and drawing precise parametric curves in the plane. History has largely settled on the name *Bézier curves* to describe what follows.

Quadratic Case

In the Cartesian plane, consider three given points $\mathbf{p}_j = (x_j, y_j)$, where $j = 0, 1, 2$. Such given points are called *control points*.

Next, suppose there is a quadratic curve $\mathbf{r}(t)$ that passes through the first point, and passes through the last point, skipping the middle $j = 1$ -point.

Letting t be a dimensionless parameter

$$0 \leq t \leq 1,$$

we know

$$\begin{aligned}\mathbf{r}(0) &= \mathbf{p}_0 \\ \mathbf{r}(1) &= \mathbf{p}_2,\end{aligned}$$

and that the slope $\mathbf{r}(t)$ is

$$\frac{d}{dt} \mathbf{r}(t) = \mathbf{v}(t).$$

This describes an infinite family of curves so far, but now impose the restriction that $\mathbf{v}(0)$ is parallel to the difference $\mathbf{p}_1 - \mathbf{p}_0$. Also, let $\mathbf{v}(1)$ be parallel to the difference $\mathbf{p}_2 - \mathbf{p}_1$. For shorthand, let us write this as:

$$\begin{aligned}\mathbf{v}(0) &\propto \mathbf{p}_1 - \mathbf{p}_0 = \mathbf{p}_{01} \\ \mathbf{v}(1) &\propto \mathbf{p}_2 - \mathbf{p}_1 = \mathbf{p}_{12}\end{aligned}$$

If we wish for $\mathbf{r}(t)$ to be quadratic in form, then the velocity vector has at most linear dependence on the parameter t . While you can perhaps guess the form for $\mathbf{v}(t)$ already, let us start a step back and define

$$\mathbf{a} = \frac{d}{dt} \mathbf{v}(t),$$

which is constant for a quadratic curve.

In terms of two unknown parameters α, β , the constant vector $\mathbf{a}$ that characterizes the quadratic curve can be written

$$\mathbf{a} \propto \alpha \mathbf{p}_{01} + \beta \mathbf{p}_{12}.$$

Integrate $\mathbf{a}$ in the t -variable to get the velocity

$$\mathbf{v}(t) \propto \alpha t \mathbf{p}_{01} + \beta t \mathbf{p}_{12} + \gamma,$$

where γ is an integration constant.

Imposing the $t = 0$ and $t = 1$ conditions on the velocity, we swiftly figure out the roles of α, β, γ , particularly

$$\begin{aligned}\alpha &= -A \\ \beta &= A \\ \gamma &= A \mathbf{p}_{01},\end{aligned}$$

where A is an overall proportionality constant. The velocity is given by:

$$\mathbf{v}(t) = A((1-t) \mathbf{p}_{01} + t \mathbf{p}_{12})$$

Now we find the curve $\mathbf{r}(t)$ by integrating the velocity to write

$$\mathbf{r}(t) = A \left(\frac{-(1-t)^2}{2} \mathbf{p}_{01} + \frac{t^2}{2} \mathbf{p}_{12} \right) + \delta,$$

where δ is a constant.

The constants A, δ are determined by the conditions at $t = 0, t = 1$. Writing each of these, we find

$$\begin{aligned} \mathbf{p}_0 &= \frac{A}{2} (-\mathbf{p}_1 + \mathbf{p}_0) + \delta \\ \mathbf{p}_2 &= \frac{A}{2} (\mathbf{p}_2 - \mathbf{p}_1) + \delta, \end{aligned}$$

implying

$$\begin{aligned} A &= 2 \\ \delta &= \mathbf{p}_1. \end{aligned}$$

Knowing the integration constants, we can go back and rewrite the velocity and its derivative:

$$\begin{aligned} \mathbf{v}(t) &= 2(1-t)\mathbf{p}_{01} + 2t\mathbf{p}_{12} \\ \mathbf{a} &= -2\mathbf{p}_{01} + 2\mathbf{p}_{12} \end{aligned}$$

In terms of the given $\mathbf{p}_j$, the position vector simplifies to

$$\mathbf{r}(t) = (1-t)^2 \mathbf{p}_0 + 2t(1-t) \mathbf{p}_1 + t^2 \mathbf{p}_2,$$

and correspondingly:

$$\begin{aligned} \mathbf{v}(t) &= -2(1-t)\mathbf{p}_0 + 2(1-2t)\mathbf{p}_1 + 2t\mathbf{p}_2 \\ \mathbf{a} &= 2(\mathbf{p}_0 - 2\mathbf{p}_1 + \mathbf{p}_2) \end{aligned}$$

Cubic Case

Extending the quadratic case by one, consider four given points $\mathbf{p}_j = (x_j, y_j)$, where $j = 0, 1, 2, 3$. A curve $\mathbf{r}(t)$ passes through $\mathbf{p}_0$ at $t = 0$, and through $\mathbf{p}_3$ at $t = 1$. The velocity obeys

$$\begin{aligned} \mathbf{v}(0) &\propto \mathbf{p}_1 - \mathbf{p}_0 = \mathbf{p}_{01} \\ \mathbf{v}(1) &\propto \mathbf{p}_3 - \mathbf{p}_2 = \mathbf{p}_{23}. \end{aligned}$$

Unlike the quadratic case in where the derivative of $\mathbf{v}(t)$ is a constant $\mathbf{a}$, we now need an $\mathbf{a}(t)$ that is (at most) linearly dependent on t . Work toward this by defining two constants

$$\begin{aligned} \mathbf{p}_{012} &= 2(\mathbf{p}_0 - 2\mathbf{p}_1 + \mathbf{p}_2) \\ \mathbf{p}_{123} &= 2(\mathbf{p}_1 - 2\mathbf{p}_2 + \mathbf{p}_3). \end{aligned}$$

Notice that $\mathbf{p}_{012}$ is same as $\mathbf{a}$ from the quadratic case.

Next, propose a vector $\mathbf{j}$ as the derivative of $\mathbf{a}(t)$

$$\frac{d}{dt} \mathbf{a}(t) = \mathbf{j},$$

which is a constant. Similar to the quadratic case, write $\mathbf{j}$ as a linear combination of vectors

$$\mathbf{j} \propto \alpha \mathbf{p}_{012} + \beta \mathbf{p}_{123},$$

for two (new) unknowns α, β .

It's important to have introduced two and only two unknowns at this stage, and of course all of the provided data points $\{(x_j, y_j)\}$ must occur in $\mathbf{j}$. While $\mathbf{j}$ could have been built in a variety of ways, the above is arguably the most natural.

Now the problem is analogous to the quadratic case with $\mathbf{j}$ playing $\mathbf{a}$'s role, etc. Transcribing the solution for α, β that works for the quadratic case, one writes

$$\begin{aligned} \mathbf{a}(t) &= B(2(1-t)\mathbf{p}_{012} + 2t\mathbf{p}_{123}) \\ \mathbf{j} &= B(-2\mathbf{p}_{012} + 2\mathbf{p}_{123}), \end{aligned}$$

where B is a new proportionality constant.

Integrate $\mathbf{a}(t)$ to get the velocity

$$\mathbf{v}(t) = B\left(-(1-t)^2 \mathbf{p}_{012} + t^2 \mathbf{p}_{123}\right) + \gamma,$$

where γ is a constant.

Integrate once more to get the position up to another constant δ :

$$\mathbf{r}(t) = B\left(\frac{-(1-t)^3}{3} \mathbf{p}_{012} + \frac{t^3}{3} \mathbf{p}_{123}\right) + t\gamma + \delta$$

To determine B and γ , use $\mathbf{r}(0) = \mathbf{p}_0$ and $\mathbf{r}(1) = \mathbf{p}_3$, and then subtract one equation from the other to eliminate δ . This should result in

$$\mathbf{p}_3 - \mathbf{p}_0 = \gamma + \frac{B}{3}(3\mathbf{p}_1 - 3\mathbf{p}_2 + \mathbf{p}_3 - \mathbf{p}_0),$$

which begs the solution

$$\begin{aligned} B &= 3 \\ \gamma &= -3\mathbf{p}_1 + 3\mathbf{p}_2. \end{aligned}$$

Staying with the velocity equation, substitute γ and simplify like mad to get

$$\mathbf{v}(t) = 3(1-t)^2 \mathbf{p}_{01} + 6t(1-t) \mathbf{p}_{12} + 3t^2 \mathbf{p}_{23}.$$

This form is much nicer than the $\mathbf{r}(t)$ derived a few lines above, so let's integrate the velocity again to get the position

$$\mathbf{r}(t) = \int \mathbf{v}(t) dt + \delta,$$

where δ is the same integration constant. Eliminate δ using what's known about $t = 0, t = 1$, and get the final position:

$$\begin{aligned} \mathbf{r}(t) &= (1-t)^3 \mathbf{p}_0 \\ &\quad + 3t(1-t)^2 \mathbf{p}_1 + 3t^2(1-t) \mathbf{p}_2 + t^3 \mathbf{p}_3 \end{aligned}$$

General Case

We've done enough work to hopefully spot a pattern in the quadratic and cubic cases to extend the equations to fourth-order cases and beyond. To set another shorthand notation, define

$$1 - t = w .$$

For the quadratic case:

$$\begin{aligned} \mathbf{r}_2(t) &= 2w \mathbf{p}_{01} + 2t \mathbf{p}_{12} \\ \mathbf{r}_2(t) &= w^2 \mathbf{p}_0 + 2tw \mathbf{p}_1 + t^2 \mathbf{p}_2 \end{aligned}$$

For the cubic case:

$$\begin{aligned} \mathbf{r}_3(t) &= 3w^2 \mathbf{p}_{01} + 6tw \mathbf{p}_{12} + 3t^2 \mathbf{p}_{23} \\ \mathbf{r}_3(t) &= w^3 \mathbf{p}_0 + 3tw^2 \mathbf{p}_1 + 3t^2 w \mathbf{p}_2 + t^3 \mathbf{p}_3 \end{aligned}$$

Looking at the pattern in the numeric coefficients in each position vector $\mathbf{r}(t)$, namely

$$\begin{aligned} \text{quadratic: } \{1, 2, 1\} \\ \text{cubic: } \{1, 3, 3, 1\} , \end{aligned}$$

these are none other than the binomial coefficients.

The n , k th binomial coefficient is given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} .$$

Quickly demonstrating for the cubic case, the above produces:

$$\begin{aligned} \binom{3}{0} &= \frac{3!}{0!(3-0)!} = 1 \\ \binom{3}{1} &= \frac{3!}{1!(3-1)!} = 3 \\ \binom{3}{2} &= \frac{3!}{2!(3-2)!} = 3 \\ \binom{3}{3} &= \frac{3!}{3!(3-3)!} = 1 \end{aligned}$$

Using the so-called 'choose' notation, the respective

position vectors are:

$$\begin{aligned} \mathbf{r}_2(t) &= \binom{2}{0} w^2 \mathbf{p}_0 + \binom{2}{1} tw \mathbf{p}_1 + \binom{2}{2} t^2 \mathbf{p}_2 \\ \mathbf{r}_3(t) &= \binom{3}{0} w^3 \mathbf{p}_0 + \binom{3}{1} tw^2 \mathbf{p}_1 \\ &\quad + \binom{3}{2} t^2 w \mathbf{p}_2 + \binom{3}{3} t^3 \mathbf{p}_3 \end{aligned}$$

General Equations

Each of the above lends to summation notation. Converting to this, we find, after restoring $w = 1 - t$:

$$\mathbf{r}_n(t) = \sum_{j=0}^n \binom{n}{j} (1-t)^{n-j} t^j \mathbf{p}_j$$

Since both the quadratic and the cubic cases are represented by the above, it takes little to imagine that higher n are also accommodated.

In the quadratic and cubic cases above, recall that the center of each derivation is a constant that involves all of the provided control points. For the quadratic case, we discovered

$$\mathbf{a} = \frac{d^2}{dt^2} \mathbf{r}_2(t) = 2(-\mathbf{p}_{01} + \mathbf{p}_{12}) ,$$

and for the cubic case,

$$\mathbf{j} = \frac{d^3}{dt^3} \mathbf{r}_3(t) = 6(-\mathbf{p}_{012} + \mathbf{p}_{123}) .$$

Going from the pattern, one should suspect that the fourth-order case obeys

$$\mathbf{k} = \frac{d^4}{dt^4} \mathbf{r}_4(t) = 24(-\mathbf{p}_{0123} + \mathbf{p}_{1234}) ,$$

and so on for higher orders.

Problem 1

Write down the fourth-order curve $\mathbf{r}_4(t)$ and determine exactly what is meant by $\mathbf{p}_{0123}$, $\mathbf{p}_{1234}$.

3 Numerical Integration

3.1 Riemann Sums

The predecessor to the notion of the integral and the fundamental theorem of calculus is the Riemann sum, which relates the endpoint values of a function $f(x)$ to a sum over the function's slope by

$$f(x_n) - f(x_0) \approx S = \sum_{j=0}^{n-1} f'(x_j^*) \Delta x,$$

where

$$\Delta x = \frac{x_n - x_0}{n},$$

and x_j^* is any x -value within $[x_j, x_{j+1}]$, and

$$x_j = x_0 + j\Delta x.$$

The dimensionless integer j is the index, and n is the total number of bins in the sum.

Since there is freedom in how x_j^* is chosen, there are three standard methods called the left sum, right sum, and midpoint sum:

$$x_j^* = \begin{cases} x_j & \text{Left sum} \\ x_{j+1} & \text{Right sum} \\ (x_j + x_{j+1})/2 & \text{Midpoint sum} \end{cases}$$

Denoting S_{Left} , S_{Right} as the left and right sums respectively, the average of these yields the trapezoid rule:

$$S_{\text{Trap}} = \frac{1}{2} (S_{\text{Left}} + S_{\text{Right}})$$

Of course, all Riemann sums are the same in the continuous limit, which is why the integral need not concern over left, right, mid, etc.

3.2 Simpson's Rule

For approximating the area under a function $f(x)$, an improvement over straight-line methods uses a quadratic function to estimate $f(x)$ at each step, known as *Simpson's rule*. To get started, propose a quadratic form

$$g(x) = Ax^2 + Bx + C,$$

where the coefficients A , B , C depend on $f(x)$ in the neighborhood of x .

Now consider a point x_j somewhere in the region and write the definite integral

$$\int_{x_j-h}^{x_j+h} f(x) dx \approx \int_{x_j-h}^{x_j+h} g(x) dx = I(h).$$

Without filling in the details yet, the result of such an integral is written $I(h)$, where $2h$ is the width of the integration domain. Substituting $g(x)$ into the above and turning the crank gives the form

$$I(h) = \frac{2h}{3} \left(A(3x_j^2 + h^2) + 3Bx_j + 3C \right).$$

Meanwhile, examine a new quantity

$$J(h) = g(x_j - h) + 4g(x_j) + g(x_j + h),$$

which, after substituting $g(x)$, becomes

$$J(h) = \frac{3}{h} I(h).$$

Evidently, the integral $I(h)$ is the same as the sum $J(h)$ up to a factor $3/h$:

$$\begin{aligned} \int_{x_j-h}^{x_j+h} f(x) dx &\approx \int_{x_j-h}^{x_j+h} g(x) dx \\ &= \frac{h}{3} (g(x_j - h) + 4g(x_j) + g(x_j + h)) \end{aligned}$$

Of course, this result only works in the neighborhood on a given x_j .

To apply this over a macroscopic interval, sum over all x_j in steps $2h$, and let $f(x)$ replace the function being evaluated. The integration region is given by

$$\frac{b-a}{n} = 2h,$$

where a , b are the lower and upper limits, and n is the number of bins. The effective bin width is $2h$. In order to have $x_0 - h = a$ and $x_{n-1} + h = b$, the x_j are located via

$$\begin{aligned} x_j &= (a + h) + \frac{j}{n} (b - a) \\ x_j &= (a + h) + j(2h). \end{aligned}$$

Assimilating these changes, the approximation becomes

$$\int_a^b f(x) dx \approx \sum_{j=0}^{n-1} \frac{h}{3} (f(x_j - h) + 4f(x_j) + f(x_j + h)), \quad (37.1)$$

which we may take as a final answer.

Pseudocode

One type of notation we'll employ here is *pseudocode*, which attempts to bridge a computer algorithm to human readability.

Pseudocode is read sequentially (top to bottom), paying particular mind to variables, conditions, and so on, in order to understand what a computer would do if the code were implemented in a proper language. Instructions that are indented are contained in a function, loop, or conditional. Instructions that end with an underscore (`_`) are continued to the next line.

As pseudocode, the Equation (37.1) can be implemented shown in the box that follows. The area being approximated is under the function $f(x) = 4x - x^2$ in the region $0 \leq x \leq 4$ using 15 bins. Lines of code that are indented by two spaces are 'looped over'.

```
f(x) = 4 * x - x * x

a = 0    # lower limit
b = 4    # upper limit
n = 15   # bins
h = (b - a) / (2 * n)
# Init. other variables to zero.

for j from 0 to n - 1
  xj = (a + h) + j * (2 * h)
  f1 = f(xj - h)
  f2 = f(xj)
  f3 = f(xj + h)
  simp += (h/3) * (f1 + 4*f2 + f3)
```

The approximation to the integral is held in the `simp` variable. If the above pseudocode were implemented in a suitable computation environment, one would find:

```
simp = 10.666666666666666
```

This result is indistinguishable from the exact answer to standard computation precision:

$$\int_0^4 (4x - x^2) dx = \frac{32}{3} = 10.666\bar{6}$$

Weighted Average Identity

Starting with Equation (37.1) for Simpson's rule, identify $2h = \Delta x$ and simplify:

$$\begin{aligned} \int_a^b f(x) dx &\approx \\ &\frac{1}{6} \sum_{j=0}^{n-1} \left(f\left(x_j - \frac{\Delta x}{2}\right) + f\left(x_j + \frac{\Delta x}{2}\right) \right) \Delta x \\ &+ \frac{1}{6} \sum_{j=0}^{n-1} 4f(x_j) \Delta x. \end{aligned}$$

The sum has been broken in two parts. The first consists of the left sum S_L and right sum S_R rules added together, which is twice the trapezoid rule S_T . The final sum involving $f(x_j)$ alone is identical to the midpoint sum S_M . Evidently, Simpson's rule is the weighted average of more elementary methods after all:

$$\int_a^b f(x) dx \approx \frac{1}{3} (S_T + 2S_M) \quad (37.2)$$

Setting up another program to approximate the left, right, and midpoint sums, we can also get the trapezoid sum and verify that Simpson's rule obeys the above identity. For an example problem, let us approximate the area under a curve we can verify by hand:

$$\int_{-2}^3 (5x^2 - x) dx = \frac{335}{6} = 55.833\bar{3}$$

Then, making appropriate changes to the above program, we have:

```

f(x) = 5 * x * x - x

a = -2 # lower limit
b = 3  # upper limit
n = 15 # bins
dx = (b - a) / n
# Initialize other variables to zero.

for j from 0 to n - 1
  xj = a + j * dx
  x1 = xj + dx
  xm = (xj + x1) / 2
  left += dx * f(xj)
  right += dx * f(x1)
  mid += dx * f(xm)

# Calculate trap and simp after loop
trap = (left + right) / 2
simp = (1 / 3) * (trap + 2 * mid)

```

With the number of bins set to $n = 15$, the results of such a program turn out as:

```

left  = 52.96296296296295
right = 59.62962962962900
mid   = 55.60185185185184
trap  = 56.29629629629628
simp  = 55.83333333333332

```

Comparing each of these to the exact answer, we see the left sum underestimating, the right sum overestimating,

and so on. Of course, the trailing digits in each may vary slightly, depending on the environment used. Most notably, Simpson's rule seems to get the answer (to this integral) to near-perfect precision.

Problem 2

Let $f(x) = (x^2 + 4x)^{1/3}$, and let $g(x)$ be an antiderivative of $f(x)$. If $g(5) = 7$, find $g(1) = h$.

Answer: Use the fundamental theorem to write

$$g(x) = \int_a^x (t^2 + 4t)^{1/3} dt$$

such that

$$g(5) = \int_a^5 (t^2 + 4t)^{1/3} dt = 7$$

$$g(1) = \int_a^1 (t^2 + 4t)^{1/3} dt = h.$$

Using both equations, we further write

$$7 = h + \int_1^5 (t^2 + 4t)^{1/3} dt,$$

where the remaining integral can be solved by an iterative method. Particularly, we find

$$h = 7 - 10.88222 = -3.88222.$$

4 Newton's Method

Transcendental Equations

Fairly often, solving a problem by analytical means is not a straightforward task, and a whole class of creatures called *transcendental equations* have no analytical solution at all. For these, the best thing we can do is approximate the answer.

For instance, try solving for x in the equation

$$x = \cos(x) ,$$

but don't try too long. While it is possible to manipulate a transcendental equation, there is no satisfactory way to isolate x . One may transform the above into either of

$$\arccos(x) = x$$

$$x = \cos(\cos(\cos(\cdots x \cdots))) ,$$

but each of these are also transcendental. To actually solve the problem on hand, you're better off plotting $y = x$ with $y = \cos(x)$ and hunting for the intersection of the two.

4.1 Newton's Method

A fascinating trick called *Newton's Method* can be used for solving certain problems, including transcendental equations, by numerical estimation.

Borrowing from the example above, consider a function

$$g(x) = x - \cos(x) ,$$

which has solutions $g(x_*) = 0$ for some (or several or many) special x_* . This setup, of course, works for any scenario where $g(x)$ is a differentiable function, and we'll proceed as if working in general.

Expand $g(x)$ to a first-order approximation

$$g_1(x) = g(x_0) + g'(x_0)(x - x_0) ,$$

where x_0 is some value in the domain of $g(x)$, called an *initial guess* that is presumably not equal to x_* . The variable x represents any point near x_0 .

Now, we already know $g(x) = 0$ is hard to deal with, but $g_1(x)$ is *easy* to deal with. Imposing the condition $g_1(x) = 0$ causes x to take on a new value x_1 away from x_0 and presumably closer to x_* as:

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$$

The reason x_1 is an 'improvement' over the initial guess x_0 , i.e. closer to x_* , can be seen geometrically in Figure

37.1. In the Cartesian plane, $g_1(x)$ is the tangent line to the function at x_0 . If x_0 is reasonably close to x_* to begin with, we're dealing with a 'zoomed in' picture of $g(x)$ where things behave linearly *anyway*, supposing the function is well-behaved.

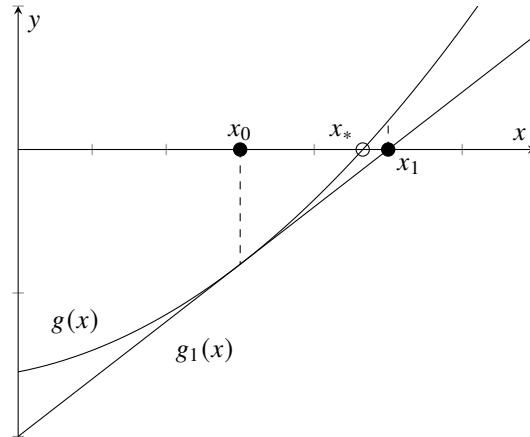


Figure 37.1. Newton's method.

With the improved guess x_1 attained, the process can be repeated to generate x_2 , which forms the initial guess for x_3 , and so on until you get tired. The process is captured in a single *recursive* formula

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} , \quad (37.3)$$

which, just to remind, attempts to solve $g(x_*) = 0$.

X Equals Cos(X)

Finishing the example that got us here, namely

$$g(x) = x - \cos(x)$$

implies

$$g'(x) = 1 + \sin(x) ,$$

thus we write

$$x_{n+1} = x_n - \frac{x_n - \cos(x_n)}{1 + \sin(x_n)} .$$

Choosing a reasonable initial guess such as $x_0 = 0.25$, the evolution of x_n proceeds as:

n	x_n
0	0.25
1	0.8263268718020449
2	0.7406169010184902
3	0.7390856504615118
4	0.7390851332152197
5	0.7390851332151607
6	0.7390851332151607

By the sixth iteration, the approximation for x_* seems to have converged to a number whose precision outruns that of the numerical system used. In conclusion, we find $x_* = \cos(x_*)$ is solved by

$$x_* \approx 0.7390851332151607 \dots$$

Problem 3

Derive the formula for x_{n+1} and find a solution to:

$$\cos(x) - x^2 = \frac{1}{2}$$

Answer:

$$x_{n+1} = x_n + \frac{\cos(x) - x^2 - 1/2}{\sin(x) + 2x} \rightarrow \pm 0.580037$$

Cube Roots

Newton's method need not work only on transcendental equations, as things like cube roots are just as straightforward to churn out as well. The nice part is, you only need a standard four-function calculator to do so. For example, take

$$g(x) = x^3 - 29,$$

solved by the cube root of 29. Setting up the proper recursive formula, we have

$$x_{n+1} = x_n - \frac{x_n^3 - 29}{3x_n^2}$$

With an initial guess of $x_0 = 3$, the evolution of x_n proceeds as:

n	x_n
0	3
1	3.0740740740740740
2	3.0723178299991580
3	3.0723168256861757
4	3.0723168256858470
5	3.0723168256858475
6	3.072316825685847...

Stopping at six iterations, the result seems to be converging near $x_* \approx 3.072\dots$, or

$$(29)^{1/3} \approx 3.072316825685847.$$

Problem 4

Derive the formula for x_{n+1} and find a solution to:

$$x^2 = 8$$

Answer:

$$x_{n+1} = x_n - \frac{x_n^2 - 8}{2x_n} \rightarrow 2.828427$$

Digits of Pi

On a scientific calculator set to radians, type

$$3.14 - \tan(3.14)$$

to get an approximate output

$$\pi \approx 3.14159265\dots$$

The reason this works is left as an exercise for the reader.

Problem 5

A straight one-mile rail is lengthened by one foot and bent into a circular arc so that the endpoints remain separated by one mile. Calculate the maximum height h that the circular rail deviates from the straight line.

Answer: Let $M = 1$ mile, $F = 1$ ft, and use trigonometry to write

$$M + F = R\theta$$

$$R \cos(\theta/2) = R - h$$

$$R \sin(\theta/2) = M/2.$$

Use the third equation to write

$$\sin\left(\frac{5281}{2R}\right) = \frac{5280}{2R}.$$

Use Newton's method to find $R \approx 78335.1$ ft. Use this and the first equation to find θ , which comes out to $\theta \approx 0.0687048$. Finally, use the second equation to attain:

$$h = 78331.5 \text{ ft} \left(1 - \cos\left(\frac{0.0687048}{2}\right)\right) \approx 46.2166 \text{ ft}$$

Problem 6

Suppose a belt is pulled snugly along the equator of Earth, having radius 6.317×10^6 m. If the belt is lengthened by $\Delta = 1$ m and pulled upward at one point, calculate the maximal height h from the ground at the vertex of the belt.

Answer: Observe from trigonometry that

$$(R + h) \cos(\theta) = R$$

$$(R + h) \sin(\theta) = \sqrt{(R + h)^2 - R^2}$$

$$\Delta = 2\sqrt{(R + h)^2 - R^2} - 2R\theta,$$

where the first two equations combine to give

$$\tan(\theta) = \frac{\sqrt{(R + h)^2 - R^2}}{R},$$

and furthermore

$$\Delta = 2R(\tan(\theta) - \theta).$$

Using $\Delta = 1$ m, one can apply Newton's method to the above to approximate the value of θ , which comes out as $\theta \approx 0.0061748$. Then, using the cosine-equation, solve for h and substitute θ to find $h \approx 121.46$ m.

Perhaps not surprisingly, we can anticipate the 'smallness' of θ to skirt around needing Newton's method. By expanding the tangent function near $\theta = 0$, we can write

$$\Delta = 2R\left(\theta + \frac{1}{3}\theta^3 + \dots - \theta\right),$$

implying

$$\theta \approx \left(\frac{3\Delta}{2R}\right)^{1/3} \approx 0.0061749,$$

which is very close to the value previously found. Moreover, by assuming $h \ll R$, the sine-equation reduces to

$$H \approx \frac{R\theta^2}{2} \approx 121.46 \text{ m},$$

in agreement with the rigorously-attained result.

4.2 Second-Order Extension

It's possible to improve the convergence time of Newton's method by including the second order term via

$$g_2(x_0 + h) = g(x_0) + g'(x_0)h + \frac{g''(x_0)}{2!}h^2,$$

where $h = x - x_0$.

Playing a similar game as the first-order case, the original curve is approximated using a parabola instead of a line. Solutions are attained by setting $g_2(x_0 + h) = 0$ and isolating h with the quadratic formula:

$$h = \frac{-g'(x_0)}{g''(x_0)} \pm \frac{g'(x_0)}{g''(x_0)} \sqrt{1 - \frac{2g(x_0)g''(x_0)}{(g'(x_0))^2}}$$

Assume $g(x_0)g''(x_0)$ is small and use Taylor expansion to eliminate the square root. Churning through the algebra gives, in abbreviated notation,

$$h = \frac{-g'}{g''} \pm \left(\frac{g'}{g''} - \frac{g}{g'} - \frac{g^2 g''}{2(g')^3} \right).$$

The second-order result needs to reduce to the first-order result in the small g'' -limit, thus we choose the positive root in the solution for h . In final form, h reads

$$h = \frac{-g}{g'} \left(1 + \frac{gg''}{2(g')^2} \right).$$

Restoring the iterative notation and writing the above as a recursive formula yields a useful improvement to Newton's method:

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} \left(1 + \frac{g(x_n)g''(x_n)}{2(g'(x_n))^2} \right)$$

4.3 Babylonian Method

A procedure less powerful but slightly more straightforward than Newton's method is something that works on roots alone, credited to the ancient Babylonians.

Square Root

Suppose we need to estimate the square root Q of some number Q^2 . Proceed by assuming Q to equal some number $N > Q$ minus a smaller contribution $x \ll Q$ such that

$$N = Q + x,$$

or

$$N^2 = Q^2 + 2Qx + x^2.$$

If x is 'small enough', then the term x^2 is negligible, allowing the first-order equation in x to be written:

$$x \approx \frac{N^2 - Q^2}{2Q}$$

The formula $N = Q + x$ becomes

$$N \approx Q + \frac{N^2 - Q^2}{2Q}.$$

Now, if the left side always evaluates to approximately N , it does so especially well for $Q \approx N$, and it should be true that whatever number we get on the left can become the next Q . In other words, we have a recursive formula

$$Q_{n+1} = Q_n + \frac{N^2 - Q_n^2}{2Q_n},$$

or more simply,

$$Q_{n+1} = \frac{Q_n}{2} + \frac{N^2}{2Q_n}$$

Cube Root

The Babylonian method for cube roots starts the same as the square root case. This time though, we write the third-power expansion of $Q + x$:

$$Q^3 + 3Q^2x + 3x^2Q + x^3 = N^3,$$

and then take the x^2 - and x^3 terms to be negligible. This means x is approximately

$$x \approx \frac{N^3 - Q^3}{3Q^2},$$

and the recursive formula settles to

$$Q_{n+1} = \frac{2}{3}Q_n + \frac{N^3}{3Q_n^2}.$$

Kth Root

One may pursue the generalized Babylonian method for the k th root of the number N . Leaving the details as an exercise, the recursive formula is

$$Q_{n+1} = \left(1 - \frac{1}{k}\right)Q_n + \frac{N^k}{kQ_n^{k-1}}.$$

Perhaps not surprisingly, this result is recovering what Newton's method would have said about the same problem. The above can also be written

$$Q_{n+1} = Q_n - \frac{Q_n^k - N^k}{kQ_n^{k-1}},$$

which is indeed Newton's method applied to

$$g(x) = x^k - N^k,$$

solved by $x_* = N$.

4.4 Multivariate Extension

Newton's method readily generalizes to multiple dimensions, which is to say, to systems of n equations with n unknowns. We assume the system is prepared such that the j th equation obeys

$$g_j(x^1, x^2, \dots, x^n) = 0,$$

where the index j runs from 1 to n . For convenience, define a vector

$$\mathbf{x} = (x^1, x^2, \dots, x^n)$$

so the above is concisely written $g_j(\mathbf{x}) = 0$.

Choosing an arbitrary $\mathbf{x}^k$ and expanding g_j to first order via Taylor expansion leads to

$$g_j(\mathbf{x}^k) \approx g_j(\mathbf{x}_0^k) + \sum_{k=1}^n \frac{\partial g_j(\mathbf{x})}{\partial x^k} \bigg|_{\mathbf{x}_0^k} (\mathbf{x}^k - \mathbf{x}_0^k).$$

Proceed by defining a vector

$$\mathbf{g} = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_n(\mathbf{x})),$$

along with a matrix J such that

$$J_{jk} = \frac{\partial g_j(\mathbf{x})}{\partial x^k} \bigg|_{\mathbf{x}_0},$$

so the above becomes

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x}_0) + J(\mathbf{x} - \mathbf{x}_0).$$

Analogous to the one-dimensional case, we take $\mathbf{x}_0 \rightarrow \mathbf{x}_n$ as an initial guess and set the left side to zero. Then, we can 'solve' for $\mathbf{x}_{n+1}$ in terms of the inverse of J :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - J^{-1} \mathbf{g}(\mathbf{x}_n)$$

There is another way to proceed without calculating the inverse J^{-1} . Instead of solving for $\mathbf{x}_{n+1}$ directly, write

$$J\mathbf{x}_{n+1} = J\mathbf{x}_n - \mathbf{g}(\mathbf{x}_n),$$

and observe the right side is completely determined in terms of $\mathbf{x}_n$. Storing this result in a vector $\mathbf{b}_n$, we have a system

$$J\mathbf{x}_{n+1} = \mathbf{b}_n.$$

As an $A\mathbf{x} = \mathbf{b}$ -like linear system, the above can be solved by finding the row-reduced echelon form (RREF) of the augmented matrix $J|\mathbf{b}_n$.

Nonlinear System

Consider the nonlinear system of equations

$$\begin{aligned} 2x^2 + 3y &= 1 \\ 4x - 7y &= 3, \end{aligned}$$

which, for the sake of this example, has two solutions attainable by standard means:

$$(x, y) = \begin{cases} (-1.58032, -1.33161) \\ (0.72318, -0.0153259) \end{cases}$$

Using Newton's method formalism, the above system reads

$$\begin{aligned} g_1(x) &= 2x^2 + 3y - 1 \\ g_2(x) &= 4x - 7y - 3. \end{aligned}$$

Computing partial derivatives of $g_{1,2}(x)$, we attain the matrix components of matrix J , required for both the inverse method involving J^{-1} and the RREF method involving $J|\mathbf{b}_n$ as described. For this we have $J_{11} = 4x$, $J_{12} = 3$, $J_{21} = 4$, $J_{22} = -7$, or

$$J = \begin{bmatrix} 4x & 3 \\ 4 & -7 \end{bmatrix}.$$

It's also necessary to provide an initial guess $\mathbf{x}_0$ for each solution we're shooting for. In this case, the system represented is the intersection of a line and a parabola, thus there are at most two solutions to attain, and we'll work with the initial values:

$$(x_0, y_0) = \begin{cases} (1, 1) \\ (-1, -1) \end{cases}$$

Proceeding with the inverse method, we first establish

$$J^{-1} = \frac{1}{-28x - 12} \begin{bmatrix} -7 & -3 \\ -4 & 4x \end{bmatrix}$$

by standard means. Next write

$$\begin{aligned} \mathbf{x}_{n+1} &= \mathbf{x}_n \\ &- \frac{1}{-28x_n - 12} \begin{bmatrix} -7 & -3 \\ -4 & 4x_n \end{bmatrix} \begin{bmatrix} 2x_n^2 + 3y_n - 1 \\ 4x_n - 7y_n - 3 \end{bmatrix} \end{aligned}$$

for both initial values $\mathbf{x}_0$.

Starting with $\mathbf{x}_0 = (-1, -1)$, the above calculation recursively proceeds as

$$\begin{bmatrix} -1 \\ -1 \end{bmatrix} \rightarrow \begin{bmatrix} -1.875 \\ -1.5 \end{bmatrix} \rightarrow \begin{bmatrix} -1.610 \\ -1.349 \end{bmatrix} \rightarrow \begin{bmatrix} -1.581 \\ -1.332 \end{bmatrix},$$

which is surely converging to the reported solution $(-1.58, -1.33)$. Repeating for $\mathbf{x}_0 = (1, 1)$, we correspondingly find

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} -0.75 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0.723 \\ -0.0152 \end{bmatrix} \rightarrow \begin{bmatrix} 0.723 \\ -0.0153 \end{bmatrix},$$

which converges to the other stated solution $(0.723, -0.0153)$.

For another approach, we calculate the $\mathbf{x}_{n+1}$ th iteration from $J|\mathbf{b}_n$ using the augmented matrix

$$\begin{bmatrix} 4x_n & 3 & J_{11}x_n + J_{12}y_n - g_1(x_n, y_n) \\ 4 & -7 & J_{21}x_n + J_{22}y_n - g_2(x_n, y_n) \end{bmatrix},$$

simplifying to a structure that can be used recursively:

$$\begin{bmatrix} 4x_n & 3 & 2x_n^2 + 1 \\ 4 & -7 & 3 \end{bmatrix}$$

Inserting $\mathbf{x}_0 = (-1, -1)$, use row operations to attain the RREF matrix

$$\begin{bmatrix} 1 & 0 & -1.875 \\ 0 & 1 & -1.5 \end{bmatrix},$$

and the right-most column reveals $x_1 = -1.875$, $y_1 = -1.5$. Notice that this is the same $\mathbf{x}_1$ yielded by the previous approach to the solution. In fact, all subsequent $\mathbf{x}_n$ for the two methods come out the same in this example.

Finally trying $\mathbf{x}_0 = (1, 1)$ and turning the crank, the solution $\mathbf{x}_1$ is read from

$$\begin{bmatrix} 1 & 0 & 0.75 \\ 0 & 1 & 0 \end{bmatrix}.$$

Again, we attain the same $\mathbf{x}_1$ produced by the previous method, and all subsequent $\mathbf{x}_n$ are also the same.

5 Euler's Method

Numerical methods need not be limited to estimating individual numbers, as estimating entire curves is also fair game. For mechanical or time-dependent systems, a technique called *Euler's method* allows for approximating the path of evolution directly from the differential equation.

To get started, consider a function $f(x)$, and Taylor-expand to first order around a point x_0 :

$$f(x) = f(x_0) + (x - x_0) f'(x_0)$$

Give all x -variables a subscript such that

$$\begin{aligned} x &\rightarrow x_{j+1} \\ x_0 &\rightarrow x_j \\ \Delta x_j &= x_{j+1} - x_j, \end{aligned}$$

and the above becomes

$$f(x_{j+1}) = f(x_j) + \Delta x_j f'(x_j).$$

If some initial value $f_0 = f(x_0)$ is given, we can calculate all subsequent $f(x)$ by linear approximation, somewhat like Newton's method, by taking small intervals Δx_j . For computational convenience, all Δx_j are usually taken as one and the same value Δx .

Exponential Growth and Decay

One of the simplest time-evolving systems is characterized by

$$\frac{d}{dt} f(t) = k f(t),$$

where k carries units of time. As a separable equation, one attains an exact solution via direct integration to find

$$f(t) = f_0 e^{kt}.$$

With or without the exact solution, the problem is framed via Euler's method as

$$\begin{aligned} f(x_{j+1}) &= f(x_j) + k \Delta x_j f(x_j) \\ f(x_{j+1}) &= f(x_j) (1 + k \Delta x_j). \end{aligned}$$

Mechanical Systems

For systems of multiple variables, as is usually the case in mechanics problems, each variable can be iterated in parallel. For example, if a force vector $\mathbf{F}$ causes the velocity

$\mathbf{v}$ and position $\mathbf{x}$ of a particle of mass m to change with time, we have:

$$\begin{aligned} \mathbf{v}_{j+1} &= \mathbf{v}_j + \frac{\mathbf{F}_j}{m} \Delta t \\ \mathbf{x}_{j+1} &= \mathbf{x}_j + \mathbf{v}_j \Delta t \end{aligned}$$

5.1 Forward Euler's Method

Consider the frictionless constant-acceleration scenario characterized by

$$\frac{dv}{dt} = -g,$$

or equivalently

$$\begin{aligned} dv/dt &= -g \\ dx/dt &= v(t). \end{aligned}$$

Expanding each of these to first order, we have

$$\begin{aligned} v(t) &\approx v(t_0) - g \Delta t \\ x(t) &\approx x(t_0) + v(t_0) \Delta t. \end{aligned}$$

To turn these into something useful, we first understand that the quantity Δt is meant to be a 'small enough' number such that $g \Delta t$ and $v(t_0) \Delta t$ are small compared to $v(t_0)$, $x(t_0)$. (This is the essence of the first-order approximation.)

Also, note that the right side of each equation contains all 'known' information, and the left side contains the 'updated' versions of v , x . Much like Newton's method, this setup invites a recursive representation given by

$$v_{j+1} = v_j - g \Delta t \quad (37.4)$$

$$x_{j+1} = x_j + v_j \Delta t \quad (37.5)$$

This setup in particular is called the *forward* Euler's method. Supplying v_0, x_0 as initial values, the above can be used to estimate the subsequent motion as a set of points. The number of points generated has to do with the size of Δt and the total time interval being considered.

5.2 Backward Euler's Method

Recalling the definition of the derivative, namely

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h},$$

note that the definition remains intact by reversing the sign on h :

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x) - f(x-h)}{h}.$$

Moreover, no error is made if we simply shift variables $x \rightarrow x + h$, so we can also write

$$\lim_{h \rightarrow 0} f'(x + h) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h},$$

which essentially recovers the definition.

Recasting the above as a first-order approximation, we get, after rearranging,

$$f(x + h) \approx f(x) + hf'(x + h).$$

Interestingly, the f' -term uses the updated version of x , namely $x + h$ as its argument. This configuration leads to the *backward* Euler's method, and is an implicit formula in the sense that some extra work needs to be done to isolate $f(x + h)$ in terms of initial quantities.

For the problem of freefall kinematics, the backward Euler's method is represented recursively via

$$v_{j+1} = v_j - g\Delta t \quad (37.6)$$

$$x_{j+1} = x_j + v_{j+1}\Delta t, \quad (37.7)$$

which is subtly different from Equations (37.4), (37.5). Note that in a more general case, the g -term would instead be a_{j+1} . We can also simplify so that all $j + 1$ indices are on the left:

$$\begin{aligned} v_{j+1} &= v_j - g\Delta t \\ x_{j+1} &= x_j + v_j\Delta t - g\Delta t^2 \end{aligned}$$

5.3 Energy Considerations

In the absence of friction, freefall kinematics qualifies as an energy-conserving system. At any point during motion, the kinetic energy is given by

$$T(v) = \frac{1}{2}mv^2,$$

where m is the mass of the object that is falling. Meanwhile, freefall near sea level implies the potential energy is

$$U(x) = mgx,$$

where g is the familiar gravity constant. For this situation, conservation of energy means

$$E = T(v) + U(x)$$

is constant.

If conservation of energy is to hold, then the recursive formulas for the forward and backward Euler's method ought to reflect this. At a given step j , the energy is

$$E_j = \frac{1}{2}mv_j^2 + mgx_j.$$

At the next step $j + 1$ the same energy ought to read

$$E_{j+1} = \frac{1}{2}mv_{j+1}^2 + mgx_{j+1}.$$

All is well until we want to substitute v_{j+1} and x_{j+1} into E_{j+1} . That is, nothing says to only use the forward method represented by Equations (37.4), (37.5), or for that matter, nothing forbids the pair of Equations (37.6), (37.7). Which pair is correct? At this point we're obligated to try both, and doing each case carefully, we find:

$$\text{Forward: } E_{j+1} = E_j + \frac{1}{2}mg^2\Delta t^2$$

$$\text{Backward: } E_{j+1} = E_j - \frac{1}{2}mg^2\Delta t^2$$

Evidently, the energy is conserved to zeroth order and first order by each method. There is, however, a second-order error lingering in each result.

It should be noted that error can be minimized when Δt is very small, but this still leaves the question of doing better.

5.4 Mixed Euler's Method

Given how the forward and backward Euler's method produce equal and opposite errors in the total energy, one has to wonder if some mixture of the methods will be better than either alone. Trying the average of the two, we write the *mixed* Euler's method for this problem:

$$v_{j+1} = v_j - g\Delta t \quad (37.8)$$

$$x_{j+1} = x_j + \frac{1}{2}(v_j + v_{j+1})\Delta t \quad (37.9)$$

Equivalently, simplify so that all $j + 1$ indices are on the left:

$$\begin{aligned} v_{j+1} &= v_j - g\Delta t \\ x_{j+1} &= x_j + v_j\Delta t - \frac{1}{2}g\Delta t^2 \end{aligned}$$

Now we check whether energy is conserved from the j th step to the next step $j + 1$. The kinetic term reads

$$\frac{1}{2}mv_{j+1}^2 = \frac{1}{2}mv_j^2 - mgv_j\Delta t + \frac{1}{2}mg^2\Delta t^2,$$

and for the potential term:

$$mgx_{j+1} = mgx_j + mgv_j\Delta t - \frac{1}{2}mg^2\Delta t^2$$

Sum the two equations to indeed find $E_{j+1} = E_j$.

To see each of Euler's methods in action against a con-

crete problem, consider the one-dimensional motion of a body that begins at $x_0 = 10 \text{ m}$ at $t = 0 \text{ s}$ with initial upward speed of $v_0 = 10 \text{ m/s}$. For a time step we'll use $\Delta t = 0.1 \text{ s}$ over a total of 20 iterations. Writing the appropriate *C* program and producing graphs with *gnuplot*, we generate the outputs shown in Figure 37.2.

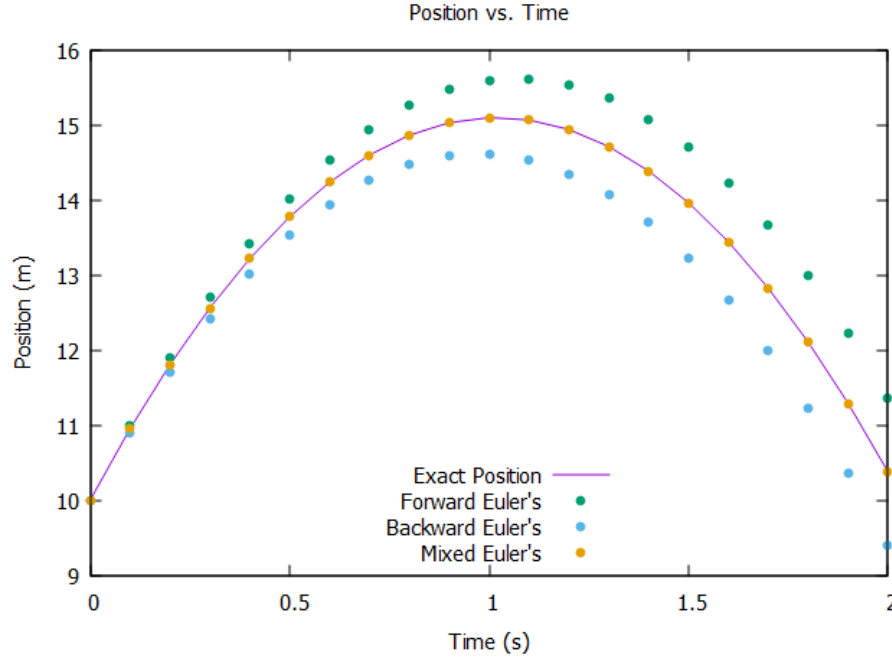


Figure 37.2. Various Euler's method approximations of ideal freefall motion compared to exact solution.

As shown in the Figure, the exact solution is traced by a solid line with the three approximations, namely forward, backward, mixed, appearing as unconnected colored dots. The forward method sails consistently over the exact solution, while the backward method sails under. Perhaps not surprisingly, the mixed method approximation stays perfectly with the exact solution.

Freefall with Air Damping

For the a modified scenario with air damping, the situation is governed by the differential equation

$$\frac{dv}{dt} = -g - bv,$$

where $v(t)$ is the velocity of a falling body, g is the local gravity constant, and b is the linear damping coefficient. Solutions to the system are straightforwardly calculated,

resulting in

$$v(t) = \frac{-g}{b} + \left(v_0 + \frac{g}{b}\right)e^{-bt}$$

$$x(t) = x_0 - \frac{gt}{b} + \frac{1}{b}\left(v_0 + \frac{g}{b}\right)\left(1 - e^{-bt}\right),$$

where after supplying the initial values v_0, x_0 , the motion is completely determined.

We can immediately dispense with any hope of conserving energy, as the effect of friction eats away at the kinetic component without replenishing the potential. Nonetheless, we may still approximate solutions with variations of Euler's method.

For the air damping problem, a set of forward equations are (as always) easy to write explicitly:

$$v_{j+1} = v_j - g\Delta t - bv_j\Delta t$$

$$x_{j+1} = x_j + v_j\Delta t$$

That is, the above is analogous to Equations (37.4), (37.5) and only differ by the presence of the b -term.

As for a backward set of equations, replace downstream j on the right with $j + 1$ to get

$$\begin{aligned} v_{j+1} &= v_j - g\Delta t - bv_{j+1}\Delta t \\ x_{j+1} &= x_j + v_{j+1}\Delta t, \end{aligned}$$

analogous to Equations (37.6), (37.7). Solving for v_{j+1} and x_{j+1} explicitly, we find these to mean

$$\begin{aligned} v_{j+1} &= \frac{v_j - g\Delta t}{1 + b\Delta t} \\ x_{j+1} &= x_j + \left(\frac{v_j - g\Delta t}{1 + b\Delta t} \right) \Delta t. \end{aligned}$$

Finally, we pursue a set of mixed equations by imposing the average on the downstream terms as

$$\begin{aligned} v_{j+1} &= v_j - g\Delta t - \frac{b}{2} (v_j + v_{j+1}) \Delta t \\ x_{j+1} &= x_j + \frac{1}{2} (v_j + v_{j+1}) \Delta t, \end{aligned}$$

which are analogous to Equations (37.8), (37.9).

What we didn't have perviously was the v_{j+1} -equation referencing any variable other than v_j . In the present case, we place the average of v_j and v_{j+1} with the b -term, but one could speculate this method is *too* mixed. Unfortunately for this problem, there is not a notion of energy conservation to verify the construction.

After some effort, the above can be expressed with $j + 1$ on the left and j on the right:

$$\begin{aligned} v_{j+1} &= \frac{v_j (1 - b\Delta t/2) - g\Delta t}{1 + b\Delta t/2} \\ x_{j+1} &= x_j + \left(\frac{v_j - g\Delta t/2}{1 + b\Delta t/2} \right) \Delta t \end{aligned}$$

We're now in position to plot each of the three approximations along with the exact solution for $x(t)$ for the air damping problem. Using the same initial conditions as previous while introducing a damping coefficient of $b = 3 \text{ s}^{-1}$, we generate the output shown in Figure 37.3.

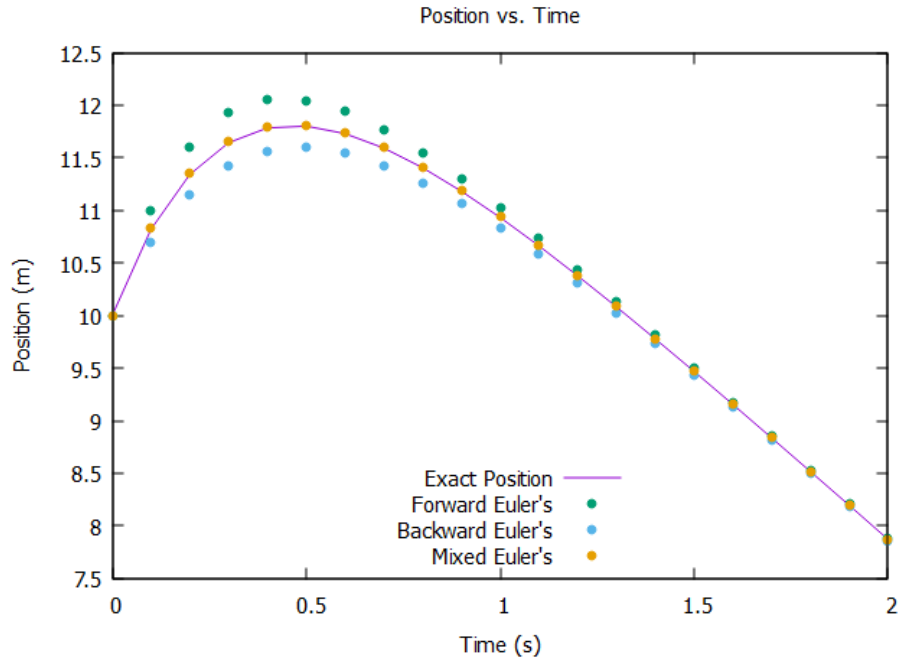


Figure 37.3. Various Euler's method approximations of damped freefall motion compared to exact solution.

In the Figure, note that all approximations agree with the exact solution in the large- t limit, corresponding to the falling body reaching terminal velocity. The slope of the asymptotic line implied ought to be $-g/b$, or roughly -3.3 m/s in the plot.

Comparing the overall performance of each approximation, as seen in the ideal case, the forward method is a little too generous in its output, and the backward method is a little too thrifty. Astonishingly though, the mixed method is spot on with the exact solution.

Other Applications

Arguably most problems in classical mechanics have no analytic solution, particularly for systems involving damping or nonlinear effects.

For freefall in gravity, we've already witnessed

$$\frac{\mathbf{F}}{m} = -g \hat{\mathbf{z}} - b \mathbf{v} ,$$

however a more detailed model of air damping yields the *drag force*:

$$\mathbf{F}_{\text{drag}} = \frac{-C_d}{2} \rho A v^2 \hat{\mathbf{v}}$$

In this, ρ is the air density and A is the forward-facing area of the object in motion. C_d is a constant related to the surface friction.

An object with sufficient spin velocity is subject to interesting forces while moving in air. It can be shown that the force on a near-spherical object that is spinning about an axis $\hat{\boldsymbol{\omega}}$ perpendicular to both (i) its direction of motion and (ii) the gravitational force vector, is the *lift force*:

$$\mathbf{F}_{\text{lift}} = \frac{C_l}{2} \rho A v^2 \frac{\boldsymbol{\omega} \times \mathbf{v}}{|\boldsymbol{\omega} \times \mathbf{v}|}$$

In the above, $\boldsymbol{\omega}$ is the angular velocity. It actually turns out that C_l is not always constant, as the relationship

$$C_l \propto \frac{r\omega}{v}$$

has also been established.

A generalization of the lift force is the so-called *Magnus force*, which permits $\hat{\boldsymbol{\omega}}$ to point anywhere:

$$\mathbf{F}_{\text{Magnus}} = S_0 (\boldsymbol{\omega} \times \mathbf{v})$$

Of course, the trouble is in accurately modeling S_0 , which has to be tackled separately for each problem.

5.5 Matrix Formulation

Simple Harmonic Oscillator

In one dimension, consider the motion of a mass m attached to a weightless Hookean spring with one end fixed at $x = 0$. The total energy E of the system is conserved and is the sum of a kinetic term and a potential term

$$E = \frac{1}{2} m v^2 + \frac{1}{2} k x^2 ,$$

where k is the spring constant. We also write the second-order differential equation

$$m \frac{d^2 x}{dt^2} = -kx .$$

As a second-order system, the above can be written as a pair of first-order equations:

$$\begin{aligned} \frac{dx}{dt} &= v \\ \frac{dv}{dt} &= -\frac{k}{m} x \end{aligned}$$

Of course, solutions to this system are well-known as harmonic oscillations.

Forward and Backward Methods

Now we convert the above into a set of iterative equations. For the forward Euler's method, we write

$$\begin{aligned} x_{j+1} &= x_j + v_j \Delta t \\ v_{j+1} &= v_j - \frac{k}{m} x_j \Delta t , \end{aligned}$$

whereas for the backward Euler's method, there are subtle change in the indices:

$$\begin{aligned} x_{j+1} &= x_j + v_{j+1} \Delta t \\ v_{j+1} &= v_j - \frac{k}{m} x_{j+1} \Delta t , \end{aligned}$$

Solve for x_{j+1} and v_{j+1} to find, for the backward method:

$$\begin{aligned} x_{j+1} &= \frac{x_j + v_j \Delta t}{1 + (k/m) \Delta t^2} \\ v_{j+1} &= \frac{v_j - (k/m) x_j \Delta t}{1 + (k/m) \Delta t^2} \end{aligned}$$

Also, let

$$\lambda = 1 + (k/m) \Delta t^2$$

to ease subsequent calculations.

Momentum Domain

The presence of k/m tends to make the above equations appear somewhat lopsided or asymmetric. Rather than deal with the velocity and the mass as fundamentally separate, use the momentum

$$p_j = mv_j$$

to replace all instances of v_j accordingly.

Mixed Methods

For a mixed method approach, let us propose

$$\begin{aligned} x_{j+1} &= x_j + \frac{1}{2m} (p_j + p_{j+1}) \Delta t \\ p_{j+1} &= p_j - \frac{k}{2} (x_{j+1} + x_j) \Delta t \end{aligned}$$

and work to isolate all $j + 1$ -variables on the left. Define

$$\alpha = \frac{k \Delta t^2}{4m}$$

and the above simplifies to

$$\begin{aligned} x_{j+1} &= \frac{x_j (1 - \alpha) + (1/m) p_j \Delta t}{1 + \alpha} \\ p_{j+1} &= \frac{p_j (1 - \alpha) - k x_j \Delta t}{1 + \alpha} . \end{aligned}$$

In this form, it's straightforward to check that the energy

$$E_{j+1} = \frac{1}{2m} p_{j+1}^2 + \frac{1}{2} k x_{j+1}^2$$

is the same as E_j .

A second mixed approach begins with less ambition. Start with

$$\begin{aligned} x_{j+1} &= x_j + \frac{1}{m} p_j \Delta t \\ p_{j+1} &= p_j - k x_{j+1} \Delta t , \end{aligned}$$

and the second equation reworks to

$$p_{j+1} = p_j \beta - k x_j \Delta t ,$$

where for brevity we have let

$$\beta = 1 - (k/m) \Delta t^2 .$$

Now, this construction is *not* strictly energy preserving, as E_{j+1} does not exactly recover E_j . Rather, it's possible to show that the energy oscillates around a central value without drifting.

Matrix Formulation

Having sets of linear equations with $j + 1$ -indices on the left and j -indices on the right, we can cast each in matrix notation. For each method, we found:

$$\text{Forward: } \begin{bmatrix} x_{j+1} \\ p_{j+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t/m \\ -k \Delta t & 1 \end{bmatrix} \begin{bmatrix} x_j \\ p_j \end{bmatrix}$$

$$\text{Backward: } \begin{bmatrix} x_{j+1} \\ p_{j+1} \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} 1 & \Delta t/m \\ -k \Delta t & 1 \end{bmatrix} \begin{bmatrix} x_j \\ p_j \end{bmatrix}$$

$$\text{Mixed 1: } \begin{bmatrix} x_{j+1} \\ p_{j+1} \end{bmatrix} = \frac{1}{1 + \alpha} \begin{bmatrix} 1 - \alpha & \Delta t/m \\ -k \Delta t & 1 - \alpha \end{bmatrix} \begin{bmatrix} x_j \\ p_j \end{bmatrix}$$

$$\text{Mixed 2: } \begin{bmatrix} x_{j+1} \\ p_{j+1} \end{bmatrix} = \begin{bmatrix} 1 & \Delta t/m \\ -k \Delta t & \beta \end{bmatrix} \begin{bmatrix} x_j \\ p_j \end{bmatrix}$$

6 Runge-Kutta Techniques

Consider a function $x(t)$ whose derivative is a function of both x and t :

$$\frac{d}{dt}x(t) = f(x(t), t)$$

In terms of f , the value of $x(t)$ evaluated at t plus a small time interval Δt is written by first-order Taylor expansion:

$$x(t + \Delta t) \approx x(t) + f(x(t), t) \Delta t$$

Using the same arguments that lead to the backward Euler's method, note the Taylor approximation can also have f evaluated at the 'right' endpoint of the interval Δt :

$$x(t + \Delta t) \approx x(t) + f(x(t + \Delta t), t + \Delta t) \Delta t$$

6.1 Midpoint Method

The forward and backward Euler correspond to using the slope f at the beginning or end of the interval. A natural idea is to involve the slope at the midpoint, a method known as the *second-order Runge-Kutta*, or *RK2* for short.

Using the slope at the midpoint we write:

$$x(t + \Delta t) \approx x(t) + f\left(x\left(t + \frac{\Delta t}{2}\right), t + \frac{\Delta t}{2}\right) \Delta t$$

Also, write the same expansion at $\Delta t/2$ to get

$$x\left(t + \frac{\Delta t}{2}\right) \approx x(t) + \frac{1}{2}f(x(t), t) \Delta t,$$

and insert this into the midpoint equation above:

$$\begin{aligned} x(t + \Delta t) &\approx x(t) \\ &+ f\left(x(t) + \frac{1}{2}f(x(t), t) \Delta t, t + \frac{\Delta t}{2}\right) \Delta t \end{aligned}$$

To clean this up, define the shorthand for the midpoint such that

$$\begin{aligned} t_{\text{mid}} &= t + \frac{\Delta t}{2} \\ x_{\text{mid}} &= x(t) + \frac{1}{2}f(x(t), t) \Delta t, \end{aligned}$$

and we write the result

$$x(t + \Delta t) \approx x(t) + f(x_{\text{mid}}, t_{\text{mid}}) \Delta t.$$

As an iterable equation, the above reads

$$x_{j+1} = x_j + f(x_{\text{mid}}, t_{\text{mid}}) \Delta t.$$

Freefall Acceleration

Consider the case of freefall acceleration characterized by

$$\begin{aligned} \frac{d}{dt}x(t) &= v(t) \\ \frac{d}{dt}v(t) &= -g. \end{aligned}$$

For this we need two deployments of the midpoint method, namely

$$\begin{aligned} x_{j+1} &= x_j + v_{\text{mid}} \Delta t \\ v_{j+1} &= v_j - g \Delta t, \end{aligned}$$

where

$$v_{\text{mid}} = v_j - \frac{1}{2}g \Delta t.$$

Substituting v_{mid} into x_{j+1} , we arrive at:

$$\begin{aligned} v_{j+1} &= v_j - g \Delta t \\ x_{j+1} &= x_j + v_j \Delta t - \frac{1}{2}g \Delta t^2 \end{aligned}$$

Notice these are identical to the mixed Euler's method contained in Equations (37.8)-(37.9). As you may recall, this construction preserves the energy

$$E_j = \frac{1}{2}mv_j^2 + mgx_j$$

across iterations.

Simple Harmonic Oscillator

For the simple harmonic oscillator of mass m in one dimension, we have a conserved energy

$$E = \frac{p^2}{2m} + \frac{1}{2}kx^2,$$

where k is the spring constant. The corresponding equations of motion are

$$\begin{aligned} \frac{d}{dt}x(t) &= \frac{1}{m}p(t) \\ \frac{d}{dt}p(t) &= -kx. \end{aligned}$$

Setting up the appropriate *RK2* system, we write

$$\begin{aligned} x_{j+1} &= x_j + \frac{1}{m}p_{\text{mid}} \Delta t \\ p_{j+1} &= p_j - kx_{\text{mid}} \Delta t, \end{aligned}$$

where

$$p_{\text{mid}} = p_j - \frac{k}{2}x_j\Delta t$$

$$x_{\text{mid}} = x_j + \frac{1}{2m}p_j\Delta t.$$

Define

$$\delta = \frac{k}{2m}\Delta t^2,$$

and solve for x_{j+1} , p_{j+1} entirely in terms of x_j , p_j to get, using matrix notation:

$$\begin{bmatrix} x_{j+1} \\ p_{j+1} \end{bmatrix} = \begin{bmatrix} 1 - \delta & \Delta t/m \\ -k\Delta t & 1 - \delta \end{bmatrix} \begin{bmatrix} x_j \\ p_j \end{bmatrix}$$

The final job is to check whether the above conserves energy across time steps. Running the familiar exercise of calculating E_{j+1} in terms of E_j , we find, after canceling:

$$E_{j+1} = E_j \left(1 + \frac{k^2}{4m^2}\Delta t^4 \right)$$

Evidently, the *RK2* method does not preserve energy in the simple harmonic oscillator system.

6.2 Fourth-Order Method

The *RK2* method extends readily to higher orders of approximation. The derivation is a nightmare but the end product is worth the effort.

Consider the fourth-order expansion of a function $y(t)$ at the point t_0 :

$$y(t_0 + h) = y_0 + hy' + \frac{h^2}{2}y'' + \frac{h^3}{6}y''' + \frac{h^4}{24}y^{(4)} + O(h^5)$$

Writing the derivative of $y(t)$ as

$$\frac{d}{dt}y(t) = f(t, y),$$

the corresponding expansion can be written in integral form:

$$y(t_0 + h) = y_0 + \int_{t_0}^{t_0+h} f(t', y(t')) dt'$$

If we're to use a fourth-order approximation, it's appropriate to assume the integral is approximated by the sum of four evaluations of f , written as:

$$\int_{t_0}^{t_0+h} f(t', y(t')) dt' \approx h(a_1k_1 + a_2k_2 + a_3k_3 + a_4k_4)$$

The constants a_j are to be determined by comparison to the fourth-order Taylor expansion. The values k_j are most generally written

$$k_1 = f(t, y)$$

$$k_2 = f(t + c_2h, y + hb_{21}k_1)$$

$$k_3 = f(t + c_3h, y + h(b_{31}k_1 + b_{32}k_2))$$

$$k_4 = f(t + c_4h, y + h(b_{41}k_1 + b_{42}k_2 + b_{43}k_3)),$$

which introduces new constants c_j and b_{jk} as implied.

To proceed, we'll need to know y' , y'' , etc., out to four orders. These are computed by repeated application of the chain rule, yielding

$$y' = f$$

$$y'' = f_t + f_y f$$

$$y''' = f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_y f_t + f_y^2 f$$

$$y^{(4)} = f_{ttt} + f_{tty}f + 2(f_{tyt}f + f_{tyy}f^2 + f_{ty}f_t + f_{ty}f_y f) + f_{yyt}f^2 + f_{yyy}f^3 + 2f_{yy}f f_t + 2f_{yy}f f_y f + f_{yt}f_t + f_{yy}f f_t + f_y f_{tt} + f_y f_{ty}f + 2f_y f_{yt}f + 2f_y f_{yy}f^2 + f_y^2 f_t + f_y^3 f.$$

Problem 7

Derive the equation for y''' explicitly.

Continue by expanding the weighted sum of stage evaluations. Consider first the term involving k_2 , defined as

$$k_2 = f(t + c_2h, y + hb_{21}k_1).$$

Since $k_1 = f(t, y)$, we substitute $k_1 = f$ and perform a Taylor expansion about (t, y) , keeping terms up to $O(h^3)$:

$$k_2 = f + f_t(c_2h) + f_y(hb_{21}f) + \frac{1}{2}f_{tt}(c_2h)^2 + f_{ty}(c_2h)(hb_{21}f) + \frac{1}{2}f_{yy}(hb_{21}f)^2 + O(h^3)$$

$$= f + h(c_2f_t + b_{21}f_y f) + h^2\left(\frac{c_2^2}{2}f_{tt} + c_2b_{21}f_{ty}f + \frac{b_{21}^2}{2}f_{yy}f^2\right) + O(h^3)$$

The contribution of k_2 to the RK4 ansatz is attained by multiplying by a_2 :

$$\begin{aligned} a_2 k_2 &= a_2 f + h a_2 (c_2 f_t + b_{21} f_y f) \\ &\quad + h^2 a_2 \left(\frac{c_2^2}{2} f_{tt} + c_2 b_{21} f_{ty} f + \frac{b_{21}^2}{2} f_{yy} f^2 \right) \\ &\quad + O(h^3) \end{aligned}$$

Now consider the term involving k_3 , defined as

$$k_3 = f(t + c_3 h, y + h(b_{31} k_1 + b_{32} k_2)) .$$

Substitute $k_1 = f$ and $k_2 = f + O(h)$, which suffices to expand k_3 to second order. Begin by writing

$$\begin{aligned} k_3 &= f + f_t(c_3 h) + f_y \cdot h(b_{31} f + b_{32} f) \\ &\quad + \frac{1}{2} f_{tt}(c_3 h)^2 + f_{ty}(c_3 h) \cdot h(b_{31} f + b_{32} f) \\ &\quad + \frac{1}{2} f_{yy} (h(b_{31} f + b_{32} f))^2 + O(h^3) \end{aligned}$$

Note that $b_{31} f + b_{32} f = (b_{31} + b_{32}) f$, so we simplify:

$$\begin{aligned} k_3 &= f + h(c_3 f_t + (b_{31} + b_{32}) f_y f) \\ &\quad + h^2 \frac{c_3^2}{2} f_{tt} + h^2 c_3 (b_{31} + b_{32}) f_{ty} f \\ &\quad + h^2 \frac{(b_{31} + b_{32})^2}{2} f_{yy} f^2 + O(h^3) \end{aligned}$$

The contribution of k_3 to the RK4 ansatz is then:

$$\begin{aligned} a_3 k_3 &= a_3 f + h a_3 (c_3 f_t + (b_{31} + b_{32}) f_y f) \\ &\quad + h^2 a_3 \frac{c_3^2}{2} f_{tt} + h^2 a_3 c_3 (b_{31} + b_{32}) f_{ty} f \\ &\quad + h^2 a_3 \frac{(b_{31} + b_{32})^2}{2} f_{yy} f^2 + O(h^3) \end{aligned}$$

Finally, consider the term involving k_4 , defined as

$$k_4 = f(t + c_4 h, y + h(b_{41} k_1 + b_{42} k_2 + b_{43} k_3)) .$$

We substitute $k_1 = f$, $k_2 = f + O(h)$, and $k_3 = f + O(h)$, which suffices to expand k_4 to second order. We first write

$$\begin{aligned} k_4 &= f + f_t(c_4 h) + f_y \cdot h(b_{41} f + b_{42} f + b_{43} f) \\ &\quad + \frac{1}{2} f_{tt}(c_4 h)^2 + f_{ty}(c_4 h) \cdot h(b_{41} f + b_{42} f + b_{43} f) \\ &\quad + \frac{1}{2} f_{yy} (h(b_{41} f + b_{42} f + b_{43} f))^2 + O(h^3) \end{aligned}$$

Combining like terms, the sum of the b_{4j} coefficients is

$$b_{41} + b_{42} + b_{43} = B_4 ,$$

so that

$$\begin{aligned} k_4 &= f + h(c_4 f_t + B_4 f_y f) + h^2 \frac{c_4^2}{2} f_{tt} \\ &\quad + h^2 c_4 B_4 f_{ty} f + \frac{B_4^2}{2} f_{yy} f^2 + O(h^3) \end{aligned}$$

The contribution of k_4 to the RK4 ansatz is then

$$\begin{aligned} a_4 k_4 &= a_4 f + h a_4 (c_4 f_t + B_4 f_y f) + h^2 a_4 \frac{c_4^2}{2} f_{tt} \\ &\quad + h^2 a_4 c_4 B_4 f_{ty} f + h^2 a_4 \frac{B_4^2}{2} f_{yy} f^2 + O(h^3) . \end{aligned}$$

Summing the contributions of all four stages, the Runge-Kutta ansatz becomes

$$\begin{aligned} h \sum_{j=1}^4 a_j k_j &= h \left((a_1 + a_2 + a_3 + a_4) f \right. \\ &\quad + h(a_2 c_2 + a_3 c_3 + a_4 c_4) f_t \\ &\quad + h(a_2 b_{21} + a_3(b_{31} + b_{32}) + a_4 B_4) f_y f \\ &\quad + h^2 \frac{1}{2} (a_2 c_2^2 + a_3 c_3^2 + a_4 c_4^2) f_{tt} \\ &\quad + h^2 (a_2 c_2 b_{21} + a_3 c_3 (b_{31} + b_{32}) + a_4 c_4 B_4) f_{ty} f \\ &\quad + h^2 \frac{1}{2} (a_2 b_{21}^2 + a_3 (b_{31} + b_{32})^2 + a_4 B_4^2) f_{yy} f^2 \left. \right) \\ &\quad + O(h^4) . \end{aligned}$$

We now need the fourth-order Taylor series

$$\begin{aligned} y(t + h) &= y_0 + h f + \frac{h^2}{2} (f_t + f_y f) \\ &\quad + \frac{h^3}{6} (f_{tt} + 2 f_{ty} f + f_{yy} f^2) \\ &\quad + \frac{h^4}{24} y^{(4)} + O(h^5) , \end{aligned}$$

where $y^{(4)}$ is expanded above for reference.

By matching coefficients, we come up with a fistfull of restrictions on the unknown coefficients:

$$\begin{aligned} a_1 + a_2 + a_3 + a_4 &= 1 \\ a_2 c_2 + a_3 c_3 + a_4 c_4 &= 1/2 \\ a_2 b_{21} + a_3(b_{31} + b_{32}) + a_4 B_4 &= 1/2 \\ a_2 c_2^2 + a_3 c_3^2 + a_4 c_4^2 &= 1/3 \\ a_2 c_2 b_{21} + a_3 c_3 (b_{31} + b_{32}) + a_4 c_4 B_4 &= 1/3 \\ a_2 b_{21}^2 + a_3 (b_{31} + b_{32})^2 + a_4 B_4^2 &= 1/3 \end{aligned}$$

Solving these algebraic conditions gives, for the a - set of differential equations coefficients

$$\begin{aligned} a_1 &= a_4 = 1/6 \\ a_2 &= a_3 = 1/3, \end{aligned}$$

and furthermore, for the c - and b -coefficients:

$$\begin{aligned} c_2 &= 1/2 & c_3 &= 1/2 & c_4 &= 1 \\ b_{21} &= 1/2 & b_{31} &= 0 & b_{32} &= 1/2 \\ b_{41} &= 0 & b_{42} &= 0 & b_{43} &= 1 \end{aligned}$$

RK4 System

Plugging all of this into the Runge-Kutta ansatz and simplifying, we derive the structure of the fourth-order Runge-Kutta method, also known as *RK4*:

$$\begin{aligned} \frac{d}{dt}y(t) &= f(t, y(t)) = \text{given} \\ y(t+h) &= \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t, y) \\ k_2 &= f\left(t + \frac{h}{2}, y + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t + \frac{h}{2}, y + \frac{h}{2}k_2\right) \\ k_4 &= f(t+h, y+hk_3) \end{aligned}$$

Physical Pendulum

A physical pendulum modeled by a mass m at the end of a massless rod suspended by the endpoint is governed by the

$$\begin{aligned} \frac{d}{dt}\theta(t) &= \omega(t) \\ \frac{d}{dt}\omega(t) &= \frac{-g}{L} \sin(\theta), \end{aligned}$$

where L is the length of the rod and g is the gravity constant. The angle θ is the deflection from the vertical.

To set up an iterative system, we need two deployments of the *RK4* method. That is, we calculate

$$\begin{aligned} \theta_{j+1} &= \theta_j + \frac{1}{6} (k_1^\theta + 2k_2^\theta + 2k_3^\theta + k_4^\theta) h \\ \omega_{j+1} &= \omega_j + \frac{1}{6} (k_1^\omega + 2k_2^\omega + 2k_3^\omega + k_4^\omega) h \end{aligned}$$

from the chain of events:

$$\begin{aligned} k_1^\theta &= \omega_j & k_1^\omega &= \frac{-g}{L} \sin(\theta_j) \\ \theta_2 &= \theta_j + \frac{h}{2} k_1^\theta & \omega_2 &= \omega_j + \frac{h}{2} k_1^\omega \\ k_2^\theta &= \omega_2 & k_2^\omega &= \frac{-g}{L} \sin(\theta_2) \\ \theta_3 &= \theta_j + \frac{h}{2} k_2^\theta & \omega_3 &= \omega_j + \frac{h}{2} k_2^\omega \\ k_3^\theta &= \omega_3 & k_3^\omega &= \frac{-g}{L} \sin(\theta_3) \\ \theta_4 &= \theta_j + h k_3^\theta & \omega_4 &= \omega_j + h k_3^\omega \\ k_4^\theta &= \omega_4 & k_4^\omega &= \frac{-g}{L} \sin(\theta_4) \end{aligned}$$

7 Symplectic Method

In classical mechanics, it is often convenient to reformulate Newton's equations using Hamiltonian mechanics. Traditionally, the state of a particle is described by its position x and velocity v . In the Hamiltonian framework, we instead use generalized coordinates q and canonical momenta p , related through

$$\begin{aligned} q &= x \\ p &= mv, \end{aligned}$$

where m is the particle mass. This change of notation allows us to write the dynamics in terms of the Hamiltonian function

$$H(q, p) = T(p) + U(q),$$

where $T(p)$ is the kinetic energy, typically $p^2/2m$, and $U(q)$ is the potential energy. Hamilton's equations of motion then take the elegant form:

$$\frac{\partial q}{\partial t} = \frac{\partial H}{\partial p} \quad \frac{\partial p}{\partial t} = -\frac{\partial H}{\partial q}$$

Proceeding in an Euler's method-like fashion, we can guess that a set of iterative equations of motion takes the rough form

$$\begin{aligned} q_{j+1} &= q_j + \left(\frac{\partial T}{\partial p} \right)_j \Delta t \\ p_{j+1} &= p_j - \left(\frac{\partial U}{\partial q} \right)_j \Delta t, \end{aligned}$$

but we know to be cautious about the indices on the right.

A naive discretization generally fails to preserve all the important properties of the Hamiltonian across iterations. While exact energy conservation would be ideal, many numerical methods allow the energy to drift over time. Instead, for long-term stability of the dynamics, we often prioritize the preservation of *symplectic structure* - that is, the exact conservation of phase-space volume, as guaranteed by Liouville's theorem. Symplectic integrators are designed precisely to respect this geometric structure, which makes them particularly well-suited for simulating physical systems over long times.

7.1 Symplectic Notation

A system with N degrees of freedom requires N pairs of coordinates q_j, p_j , which means a full description requires

$2N$ items. Let us pack all of this information into a single column vector η such that $\eta^j = q^j$ and $\eta^{j+N} = p^j$, or:

$$\eta = \begin{bmatrix} q \\ p \end{bmatrix} = \begin{bmatrix} q^1 \\ q^2 \\ \vdots \\ p^1 \\ p^2 \\ \vdots \end{bmatrix}$$

In this notation, Hamilton's equations of motion take the form:

$$\frac{\partial \eta^j}{\partial t} = \frac{\partial H}{\partial \eta^{j+N}} \quad \frac{\partial \eta^{j+N}}{\partial t} = -\frac{\partial H}{\partial \eta^j}$$

The J Matrix

Using the symplectic notation, Hamilton's equations of motion can be written as a single entity

$$J \frac{\partial H}{\partial \eta} = \frac{\partial}{\partial t} \eta,$$

where the components of the matrix J are either ± 1 or 0 , with no dynamical information. It's straightforward to show that such a matrix obeys

$$\begin{aligned} JJ &= -I \\ J^T &= J^{-1} = -J \\ |J| &= 1, \end{aligned}$$

where I is the identity matrix.

For an example with $N = 2$ degrees of freedom, the above means:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -\partial_t p^1 \\ -\partial_t p^2 \\ \partial_t q^1 \\ \partial_t q^2 \end{bmatrix} = \begin{bmatrix} \partial_t q^1 \\ \partial_t q^2 \\ -\partial_t p^1 \\ -\partial_t p^2 \end{bmatrix}$$

The M Matrix

For many iterative systems, the evolution to the next step η_{j+1} can be written as a linear combination of the components of η_j , which implies a matrix M at each time step:

$$\begin{bmatrix} q_{j+1} \\ p_{j+1} \end{bmatrix} = M \begin{bmatrix} q_j \\ p_j \end{bmatrix}$$

For instance, the simple harmonic oscillator, when analyzed using a variety of Euler's methods, yields the matrices

$$\begin{aligned}\text{Forward: } M_F &= \begin{bmatrix} 1 & \Delta t/m \\ -k\Delta t & 1 \end{bmatrix} \\ \text{Backward: } M_B &= \frac{1}{\lambda} \begin{bmatrix} 1 & \Delta t/m \\ -k\Delta t & 1 \end{bmatrix} \\ \text{Mixed 1: } M_\alpha &= \frac{1}{1+\alpha} \begin{bmatrix} 1-\alpha & \Delta t/m \\ -k\Delta t & 1-\alpha \end{bmatrix} \\ \text{Mixed 2: } M_\beta &= \begin{bmatrix} 1 & \Delta t/m \\ -k\Delta t & \beta \end{bmatrix},\end{aligned}$$

where:

$$\begin{aligned}\lambda &= 1 + (k/m) \Delta t^2 \\ \alpha &= k\Delta t^2/4m \\ \beta &= 1 - (k/m) \Delta t^2\end{aligned}$$

7.2 Symplectic Condition

Let us set a restriction on the evolution of the system by demanding the structure

$$\begin{aligned}J \frac{\partial}{\partial \eta_j} H(\eta_j) &= \frac{\partial}{\partial t} \eta_j \\ J \frac{\partial}{\partial \eta_{j+1}} H(\eta_{j+1}) &= \frac{\partial}{\partial t} \eta_{j+1}\end{aligned}$$

remains intact.

In the second equation, substitute $\eta_{j+1} = M\eta_j$ to write

$$J \frac{\partial}{\partial M\eta_j} H(M\eta_j) = \frac{\partial}{\partial t} (M\eta_j),$$

and simplify:

$$MJM^T \frac{\partial}{\partial \eta_j} H(\eta_j) = \frac{\partial}{\partial t} (\eta_j)$$

Comparing this to the first equation, we find that the *symplectic condition*

$$MJM^T = J$$

must be satisfied.

It's hard to resist testing the above-listed M -matrices for the symplectic condition. For this we'll need a J matrix of size 2×2 :

$$J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Substituting each matrix above, we find, after some work:

$$\begin{aligned}M_F J M_F^T &= \lambda J \\ M_B J M_B^T &= (1/\lambda) J \\ M_\alpha J M_\alpha^T &= J \\ M_\beta J M_\beta^T &= J\end{aligned}$$

Without surprise, the forward and backward Euler's methods fail to satisfy the symplectic condition. The mixed versions, however, each satisfy the condition. It's straightforward to show that only one of these, namely M_α , conserves energy exactly.

Problem 8

For a 2×2 matrix

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

show that the symplectic condition is equivalent to

$$\det M = ad - bc = 1.$$

Hint:

$$\det(\gamma M) = \gamma^2 \det M$$

7.3 Equations of Motion

Returning to the question of a general set of equations of motion for a general Hamiltonian, we need to decide what to do with

$$\begin{aligned}q_{j+1} &= q_j + \left(\frac{\partial T}{\partial p} \right)_? \Delta t \\ p_{j+1} &= p_j - \left(\frac{\partial U}{\partial q} \right)_? \Delta t.\end{aligned}$$

Using the simple harmonic oscillator for comparison, the equations of motion that brought us the matrices M_α , M_β , respectively, are the mixed cases

$$\begin{aligned}q_{j+1} &= q_j + \frac{1}{2m} (p_j + p_{j+1}) \Delta t \\ p_{j+1} &= p_j - \frac{k}{2} (q_{j+1} + q_j) \Delta t\end{aligned}$$

and

$$\begin{aligned}q_{j+1} &= q_j + \frac{1}{m} p_j \Delta t \\ p_{j+1} &= p_j - k q_{j+1} \Delta t.\end{aligned}$$

Explicit vs. Implicit

In each set of equations, we are lucky enough to be able to solve for q_{j+1}, p_{j+1} *explicitly* in terms of q_j, p_j . This, however, is not always the case. It's more common to have an *implicit* embedding of q_{j+1}, p_{j+1} on the right.

For example, if we write

$$\begin{aligned} q_{j+1} &= q_j + \left(\frac{\partial T}{\partial p} \Big|_{j+1} \right) \Delta t \\ p_{j+1} &= p_j - \left(\frac{\partial U}{\partial q} \Big|_{j+1} \right) \Delta t, \end{aligned}$$

then only the most trivial systems would permit an explicit separation of variables. Otherwise we need a separate tool to solve the chicken-and-egg problem of each variable needing each other variable's updated version.

7.4 Symplectic Integrator

For an option analagous to the M_β case, we propose

$$\begin{aligned} q_{j+1} &= q_j + \left(\frac{\partial T}{\partial p} \Big|_j \right) \Delta t \\ p_{j+1} &= p_j - \left(\frac{\partial U}{\partial q} \Big|_{j+1} \right) \Delta t. \end{aligned}$$

Recall that this construction, when applied to the simple harmonic oscillator, *does* satisfy the symplectic condition, but is *not* guaranteed to keep the energy constant.

A key reason for choosing the above formulation is the implicit nature of the calculation can be skipped. By calculating q_{j+1} *before* p_{j+1} , we use the updated q_{j+1} in the $\partial U / \partial q$ term. With this understanding, we take the above equations as the final answer for the symplectic integrator.

8 Driven Pendulum

One of the most studied systems in classical physics is the physical pendulum: a mass m attached to the end of a rigid rod of length L , swinging freely under the influence of gravity. Except for the case of small oscillations, this system does not admit a simple analytic solution.

From the perspective of iterative numerical methods, it is easy to show - without even running a simulation - that the second-order Runge–Kutta method fails to preserve the system's total energy. In practice, switching to a fourth-order Runge–Kutta method drastically reduces the error and has become the *de facto* standard for simulating pendulum motion.

The problems we tackle here pertain to the *driven* physical pendulum in two dimensions, which means assigning the 'anchor point' a given path of motion $x_0(t)$, $y_0(t)$, and letting the pendulum behave accordingly.

To complicate things more, we skip analysis of the driven 'ordinary' pendulum and jump to the driven double pendulum. To keep things as general as possible, assume a mass m_2 is connected by a rod of length l_2 to a mass m_1 connected to the driving source by a rod of length l_1 .

Considering the driven double pendulum as a training run, we simultaneously address the general case of the driven N pendulum, which has masses m_j connected by rods l_j in series back to the driven point $x_0(t)$, $y_0(t)$.

Coordinates

Setting the origin arbitrarily in the plane, we let the N pendulum be driven from the point $x_0(t)$, $y_0(t)$. The rod l_1 makes an angle θ_1 with respect to the vertical. The end of the rod l_1 locates the mass m_1 at the position $x_1(t)$, $y_1(t)$.

This pattern continues for each mass in the pendulum. That is, the rod l_2 connects to l_1 while making angle θ_2 with the vertical. At the end of rod l_2 is mass m_2 located at $x_2(t)$, $y_2(t)$, and so on. Figure 37.4 shows a configuration for $N = 4$ masses connected by four rods.

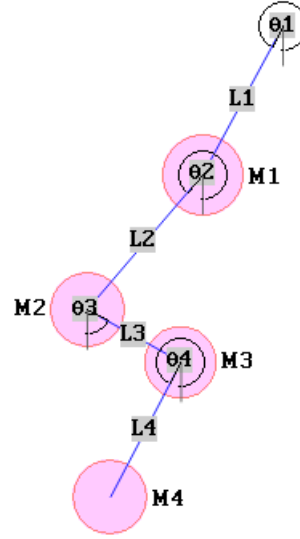


Figure 37.4. Four-body pendulum.

The x - and y - positions of each mass are easy to write. For $N = 2$, we have:

$$\begin{aligned} x_1(t) &= x_0(t) + l_1 \sin(\theta_1) \\ y_1(t) &= y_0(t) - l_1 \cos(\theta_1) \\ x_2(t) &= x_0(t) + l_1 \sin(\theta_1) + l_2 \sin(\theta_2) \\ y_2(t) &= y_0(t) - l_1 \cos(\theta_1) - l_2 \cos(\theta_2) \end{aligned}$$

Generalizing the pattern to N total masses, the j 'th mass in the chain is tracked by

$$\begin{aligned} x_j(t) &= x_0(t) + \sum_{k=1}^j l_k \sin(\theta_k) \\ y_j(t) &= y_0(t) - \sum_{k=1}^j l_k \cos(\theta_k) , \end{aligned}$$

where $j \leq N$.

Derivatives

For a notational convenience, use the 'dot' notation for a single time derivative and the double dot for two time derivatives:

$$\begin{aligned} \dot{x}(t) &= \frac{d}{dt}x(t) \\ \ddot{x}(t) &= \frac{d^2}{dt^2}x(t) \end{aligned}$$

We also handle derivatives of θ with the same notation:

$$\begin{aligned}\dot{\theta}(t) &= \omega(t) = \frac{d}{dt}\theta(t) \\ \ddot{\theta}(t) &= \dot{\omega}(t) = \frac{d}{dt}\omega(t) = \frac{d^2}{dt^2}\theta(t)\end{aligned}$$

8.1 Lagrangian Dynamics

Time evolution of the driven pendulum is calculated via Lagrangian formalism. This entails solving for the total kinetic energy $T(\{\dot{x}_j, \dot{y}_j\})$ and the total potential energy $U(\{x_j, y_j\})$, and assembling the Lagrangian:

$$L = T(\{\dot{x}_j, \dot{y}_j\}) - U(\{x_j, y_j\})$$

Given the coordinate system on hand, it's convenient to use the polar form in place of the Cartesian, so we'll make use of a Lagrangian

$$L = T(\{\omega_j\}) - U(\{\theta_j\}) .$$

In terms of the Lagrangian, equations of motion are given by the Euler-Lagrange equations:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_j} \right) - \frac{\partial L}{\partial \theta_j} &= 0 \\ j &= 1, 2, 3, \dots, N\end{aligned}$$

Kinetic Energy

The hard work begins as we calculate the velocity of each mass in the pendulum. For the case of $N = 2$, we begin with

$$\begin{aligned}\dot{x}_1(t) &= \dot{x}_0 + l_1 \omega_1 \cos(\theta_1) \\ \dot{y}_1(t) &= \dot{y}_0 + l_1 \omega_1 \sin(\theta_1) \\ \dot{x}_2(t) &= \dot{x}_0 + l_1 \omega_1 \cos(\theta_1) + l_2 \omega_2 \cos(\theta_2) \\ \dot{y}_2(t) &= \dot{y}_0 + l_1 \omega_1 \sin(\theta_1) + l_2 \omega_2 \sin(\theta_2) ,\end{aligned}$$

which readily generalizes for any N :

$$\begin{aligned}\dot{x}_j(t) &= \dot{x}_0 + \sum_{k=1}^j l_k \omega_k \cos(\theta_k) \\ \dot{y}_j(t) &= \dot{y}_0 + \sum_{k=1}^j l_k \omega_k \sin(\theta_k)\end{aligned}$$

The kinetic energy needs the square of the velocity, so we're interested in calculating

$$v_j^2(t) = \dot{x}_j^2(t) + \dot{y}_j^2(t)$$

for each mass. For the $N = 2$ case, we find, after some simplifying:

$$\begin{aligned}v_1^2 &= \dot{x}_0^2 + \dot{y}_0^2 \\ &\quad + 2l_1\omega_1(\dot{x}_0 \cos(\theta_1) + \dot{y}_0 \sin(\theta_1)) + l_1^2\omega_1^2 \\ v_2^2 &= \dot{x}_0^2 + \dot{y}_0^2 \\ &\quad + 2l_1\omega_1(\dot{x}_0 \cos(\theta_1) + \dot{y}_0 \sin(\theta_1)) \\ &\quad + 2l_2\omega_2(\dot{x}_0 \cos(\theta_2) + \dot{y}_0 \sin(\theta_2)) \\ &\quad + l_1^2\omega_1^2 + l_2^2\omega_2^2 \\ &\quad + 2l_1l_2\omega_1\omega_2 \cos(\theta_2 - \theta_1)\end{aligned}$$

Assembling

$$T = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 ,$$

we find, letting $M = m_1 + m_2$:

$$\begin{aligned}T &= \frac{1}{2}M(\dot{x}_0^2 + \dot{y}_0^2) \\ &\quad + \dot{x}_0(Ml_1\omega_1 \cos(\theta_1) + m_2l_2\omega_2 \cos(\theta_2)) \\ &\quad + \dot{y}_0(Ml_1\omega_1 \sin(\theta_1) + m_2l_2\omega_2 \sin(\theta_2)) \\ &\quad + \frac{m_1}{2}l_1^2\omega_1^2 \\ &\quad + \frac{m_2}{2}(l_1^2\omega_1^2 + l_2^2\omega_2^2 + 2l_1l_2\omega_1\omega_2 \cos(\theta_2 - \theta_1))\end{aligned}$$

Going beyond $N = 2$, squaring each of $\dot{x}_j$ and $\dot{y}_j$ is a bit of a chore. Starting with the x -equation, we write

$$\dot{x}_j^2 = \dot{x}_0^2 + 2\dot{x}_0 \sum_{k=1}^j l_k \omega_k \cos(\theta_k) + X^2 ,$$

where

$$X = \sum_{k=1}^j l_k \omega_k \cos(\theta_k) .$$

The tricky part is the square of the sum. Working this out carefully, we find

$$\begin{aligned}X^2 &= \sum_{k=1}^j l_k^2 \omega_k^2 \cos^2(\theta_k) \\ &\quad + 2 \sum_{1 \leq k < k' \leq j} l_k l_{k'} \omega_k \omega_{k'} \cos(\theta_k) \cos(\theta_{k'}) ,\end{aligned}$$

which can be used generically:

$$\left(\sum_{k=1}^j a_k \right)^2 = \sum_{k=1}^j a_k^2 + 2 \sum_{1 \leq k < k' \leq j} a_k a_{k'}$$

Repeating the same calculation for the y -equation leads to

$$\begin{aligned}\dot{y}_j^2 &= \dot{y}_0^2 + 2\dot{y}_0 \sum_{k=1}^j l_k \omega_k \sin(\theta_k) \\ &+ \sum_{k=1}^j l_k^2 \omega_k^2 \sin^2(\theta_k) \\ &+ 2 \sum_{1 \leq k < k' \leq j} l_k l_{k'} \omega_k \omega_{k'} \sin(\theta_k) \sin(\theta_{k'}) .\end{aligned}$$

Starting with the definition of the kinetic energy and plugging in the details, we write

$$T = \sum_{j=1}^N \frac{1}{2} m_j v_j^2(t) = \sum_{j=1}^N \frac{1}{2} m_j (\dot{x}_j^2 + \dot{y}_j^2) ,$$

which, after substituting the above, develops into:

$$\begin{aligned}T &= \frac{1}{2} \mathcal{M} (\dot{x}_0^2 + \dot{y}_0^2) \\ &+ \sum_{j=1}^N \left(m_j \dot{x}_0 \sum_{k=1}^j l_k \omega_k \cos(\theta_k) \right) \\ &+ \sum_{j=1}^N \left(m_j \dot{y}_0 \sum_{k=1}^j l_k \omega_k \sin(\theta_k) \right) \\ &+ \frac{1}{2} \sum_{j=1}^N \left(m_j \sum_{k=1}^j l_k^2 \omega_k^2 \right) \\ &+ \sum_{j=1}^N \left(m_j \sum_{1 \leq k < k' \leq j} l_k l_{k'} \omega_k \omega_{k'} \cos(\theta_k - \theta_{k'}) \right)\end{aligned}$$

In the first term, we defined

$$\mathcal{M} = \sum_{j=1}^N m_j .$$

Potential Energy

The potential energy is a bit easier to write than the kinetic energy. For the $N = 2$ case, we have

$$\begin{aligned}U &= m_1 g y_1 + m_2 g y_2 \\ &= M g y_0 - M g l_1 \cos(\theta_1) - m_2 g l_2 \cos(\theta_2) ,\end{aligned}$$

where $M = m_1 + m_2$.

Doing a similar exercise for the driven N -pendulum, we write

$$\begin{aligned}U &= \sum_{j=1}^N m_j g y_j \\ &= \sum_{j=1}^N m_j g \left(y_0 - \sum_{k=1}^j l_k \cos(\theta_k) \right) ,\end{aligned}$$

ending up at

$$U = g \mathcal{M} y_0 - g \sum_{j=1}^N \left(m_j \sum_{k=1}^j l_k \cos(\theta_k) \right) .$$

8.2 Equations of Motion

With the kinetic energy T and potential energy U in hand, equations of motion follow from various derivatives of the Lagrangian $L = T - U$. In particular, we calculate

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_j} \right) - \frac{\partial L}{\partial \theta_j} = 0$$

for each mass involved.

Equations of Motion (N=2)

For the $N = 2$ case, we must evaluate two equations:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_1} \right) - \frac{\partial L}{\partial \theta_1} &= 0 \\ \frac{d}{dt} \left(\frac{\partial L}{\partial \omega_2} \right) - \frac{\partial L}{\partial \theta_2} &= 0\end{aligned}$$

For the first equation of motion, begin by calculating the inner ω_1 partial derivative:

$$\begin{aligned}\frac{\partial L}{\partial \omega_1} &= \frac{\partial T}{\partial \omega_1} - \cancel{\frac{\partial U}{\partial \omega_1}} = \frac{\partial T}{\partial \omega_1} \\ \frac{\partial T}{\partial \omega_1} &= M l_1 (\dot{x}_0 \cos(\theta_1) + \dot{y}_0 \sin(\theta_1)) \\ &+ M l_1^2 \omega_1 + m_2 l_1 l_2 \omega_2 \cos(\theta_2 - \theta_1)\end{aligned}$$

Now we're ready for the full t -derivative:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_1} \right) &= M l_1 (\ddot{x}_0 \cos(\theta_1) - \dot{x}_0 \omega_1 \sin(\theta_1)) \\ &+ M l_1 (\ddot{y}_0 \sin(\theta_1) + \dot{y}_0 \omega_1 \cos(\theta_1)) \\ &+ M l_1^2 \dot{\omega}_1 + m_2 l_1 l_2 \dot{\omega}_2 \cos(\theta_2 - \theta_1) \\ &- m_2 l_1 l_2 \omega_2 (\omega_2 - \omega_1) \sin(\theta_2 - \theta_1)\end{aligned}$$

The θ_1 derivative proceeds as:

$$\begin{aligned}\frac{\partial L}{\partial \theta_1} &= \frac{\partial T}{\partial \theta_1} - \frac{\partial U}{\partial \theta_1} \\ \frac{\partial L}{\partial \theta_1} &= M l_1 \omega_1 (\dot{x}_0 \sin(\theta_1) - \dot{y}_0 \cos(\theta_1)) \\ &\quad - m_2 l_1 l_2 \omega_1 \omega_2 \sin(\theta_2 - \theta_1) + M g l_1 \sin(\theta_1)\end{aligned}$$

Finally, the first equation of motion is the difference:

$$\begin{aligned}0 &= \frac{d}{dt} \left(\frac{\partial L}{\partial \omega_1} \right) - \frac{\partial L}{\partial \theta_1} \\ 0 &= \dot{\omega}_1 M l_1 + \dot{\omega}_2 m_2 l_2 \cos(\theta_2 - \theta_1) \\ &\quad - m_2 l_2 \omega_2^2 \sin(\theta_2 - \theta_1) - M g \sin(\theta_1) \\ &\quad + M (\ddot{x}_0 \cos(\theta_1) + \ddot{y}_0 \sin(\theta_1))\end{aligned}$$

For the second equation of motion, begin by calculating the inner ω_2 partial derivative

$$\begin{aligned}\frac{\partial L}{\partial \omega_2} &= \frac{\partial T}{\partial \omega_2} - \cancel{\frac{\partial U}{\partial \omega_2}} = \frac{\partial T}{\partial \omega_2} \\ \frac{\partial T}{\partial \omega_2} &= m_2 l_2 (\dot{x}_0 \cos(\theta_2) + \dot{y}_0 \sin(\theta_2)) \\ &\quad + m_2 l_2^2 \omega_2 + m_2 l_1 l_2 \omega_1 \cos(\theta_2 - \theta_1),\end{aligned}$$

followed by the full t -derivative:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_2} \right) &= m_2 l_2 (\ddot{x}_0 \cos(\theta_2) - \dot{x}_0 \omega_2 \sin(\theta_2)) \\ &\quad + m_2 l_2 (\ddot{y}_0 \sin(\theta_2) + \dot{y}_0 \omega_2 \cos(\theta_2)) \\ &\quad + m_2 l_2^2 \dot{\omega}_2 + m_2 l_1 l_2 \dot{\omega}_1 \cos(\theta_2 - \theta_1) \\ &\quad - m_2 l_1 l_2 \omega_1 (\omega_2 - \omega_1) \sin(\theta_2 - \theta_1)\end{aligned}$$

The θ_2 derivative proceeds as:

$$\begin{aligned}\frac{\partial L}{\partial \theta_2} &= \frac{\partial T}{\partial \theta_2} - \frac{\partial U}{\partial \theta_2} \\ \frac{\partial L}{\partial \theta_2} &= m_2 l_2 \omega_2 (\dot{x}_0 \sin(\theta_1) - \dot{y}_0 \cos(\theta_1)) \\ &\quad - m_2 l_1 l_2 \omega_1 \omega_2 \sin(\theta_2 - \theta_1) + m_2 g l_2 \sin(\theta_1)\end{aligned}$$

Finally, the second equation of motion is the difference:

$$\begin{aligned}0 &= \frac{d}{dt} \left(\frac{\partial L}{\partial \omega_2} \right) - \frac{\partial L}{\partial \theta_2} \\ 0 &= \dot{\omega}_2 l_2 + \dot{\omega}_1 l_1 \cos(\theta_2 - \theta_1) \\ &\quad + l_1 \omega_1^2 \sin(\theta_2 - \theta_1) - g \sin(\theta_2) \\ &\quad + \ddot{x}_0 \cos(\theta_2) + \ddot{y}_0 \sin(\theta_2)\end{aligned}$$

Equations of Motion (General N)

Calculating the equations of motion for $N > 2$ is made easier with good bookkeeping tools. Jumping right in, the quantity

$$\delta_{nk} = \begin{cases} 1 & n = k \\ 0 & n \neq k \end{cases}$$

is the *Kronecker delta*. Similar to δ , we can also define its counterpart $\tilde{\delta}$:

$$\tilde{\delta}_{nk} = \begin{cases} 0 & n = k \\ 1 & n \neq k \end{cases}$$

Problem 9

Consider the partial contribution to the kinetic energy

$$K = \frac{1}{2} \sum_{j=1}^N \left(m_j \sum_{k=1}^j l_k^2 \omega_k^2 \right).$$

For a given angular frequency ω_n , calculate $\partial K / \partial \omega_n$.

Answer:

$$\begin{aligned}\frac{\partial K}{\partial \omega_n} &= \frac{1}{2} \sum_{j=1}^N \left(m_j \sum_{k=1}^j l_k^2 \frac{\partial (\omega_k^2)}{\partial \omega_k} \delta_{nk} \right) \\ &= \sum_{j=1}^N \left(m_j \sum_{k=1}^j l_k^2 \omega_k \delta_{nk} \right) \\ \frac{\partial K}{\partial \omega_n} &= m_1 \sum_{k=1}^1 l_k^2 \omega_k \delta_{nk} + m_2 \sum_{k=1}^2 l_k^2 \omega_k \delta_{nk} \\ &\quad + m_3 \sum_{k=1}^3 l_k^2 \omega_k \delta_{nk} + \cdots + m_n \sum_{k=1}^n l_k^2 \omega_k \delta_{nk} \\ &\quad + \cdots + m_N \sum_{k=1}^N l_k^2 \omega_k \delta_{nk} \\ \frac{\partial K}{\partial \omega_n} &= \left(\sum_{j=n}^N m_j \right) l_n^2 \omega_n\end{aligned}$$

As expected, the dependence on ω is reduced from power 2 to power 1. The nontrivial part of this result is what happened to the sums: the k -sum vanished, and the j -sum has a new lower limit.

Now we calculate

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_n} \right) - \frac{\partial L}{\partial \theta_j} = 0$$

for each mass involved. For the n th mass in the pendulum, calculate the inner partial derivative first to find:

$$\begin{aligned}\frac{\partial L}{\partial \omega_n} &= \frac{\partial T}{\partial \omega_n} - \frac{\partial U}{\partial \omega_n} \\ &= \sum_{j=n}^N (m_j \dot{x}_0 l_n \cos(\theta_n) + m_j \dot{y}_0 l_n \sin(\theta_n)) \\ &\quad + \sum_{j=n}^N (m_j l_n^2 \omega_n) \\ &\quad + \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k \omega_k \cos(\theta_n - \theta_k) \right)\end{aligned}$$

Then, the surrounding t -derivative works out as:

$$\begin{aligned}\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_n} \right) &= \sum_{j=n}^N (m_j \ddot{x}_0 l_n \cos(\theta_n)) \\ &\quad - \sum_{j=n}^N (m_j \dot{x}_0 l_n \omega_n \sin(\theta_n)) \\ &\quad + \sum_{j=n}^N (m_j \ddot{y}_0 l_n \sin(\theta_n)) \\ &\quad + \sum_{j=n}^N (m_j \dot{y}_0 l_n \omega_n \cos(\theta_n)) \\ &\quad + \sum_{j=n}^N (m_j l_n^2 \dot{\omega}_n) \\ &\quad + \sum_{j=n}^N \left(m_j \sum_{k=1}^j F \right) - \sum_{j=n}^N \left(m_j \sum_{k=1}^j G \right),\end{aligned}$$

where

$$\begin{aligned}F &= \tilde{\delta}_{nk} l_n l_k \dot{\omega}_k \cos(\theta_n - \theta_k) \\ G &= l_n l_k \tilde{\delta}_{nk} \omega_k (\omega_n - \omega_k) \sin(\theta_n - \theta_k) .\end{aligned}$$

Next we need the θ_n -derivative of L . For this we have:

$$\begin{aligned}\frac{\partial L}{\partial \theta_n} &= \frac{\partial T}{\partial \theta_n} - \frac{\partial U}{\partial \theta_n} \\ &= - \sum_{j=n}^N (m_j \dot{x}_0 l_n \omega_n \sin(\theta_n)) \\ &\quad + \sum_{j=n}^N (m_j \dot{y}_0 l_n \omega_n \cos(\theta_n)) \\ &\quad - \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k \omega_n \omega_k \sin(\theta_n - \theta_k) \right) \\ &\quad - g \sum_{j=n}^N m_j l_n \sin(\theta_n)\end{aligned}$$

To assemble the n th equation of motion, write

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \omega_n} \right) = \frac{\partial L}{\partial \theta_n}$$

and simplify like mad. All terms involving first-order derivatives $\dot{x}_0$, $\dot{y}_0$ cancel, as does the mixed $\omega_n \omega_k$ term embedded in G . For a final result, we find the equation of motion for the driven N pendulum:

$$\begin{aligned}0 &= \left(\sum_{j=n}^N m_j \right) (\ddot{x}_0 l_n \cos(\theta_n)) \\ &\quad + \left(\sum_{j=n}^N m_j \right) (\ddot{y}_0 l_n \sin(\theta_n)) \\ &\quad + \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k (\dot{\omega}_k \cos(\theta_n - \theta_k)) \right) \\ &\quad + \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k (\omega_k^2 \sin(\theta_n - \theta_k)) \right) \\ &\quad + \left(\sum_{j=n}^N m_j \right) l_n^2 \dot{\omega}_n - g \left(\sum_{j=n}^N m_j \right) l_n \sin(\theta_n)\end{aligned}$$

8.3 Matrix Formulation

As complicated as the equations of motion are, both for the $N = 2$ case and above, the dependence on any given $\dot{\omega}$ is linear, which means we can proceed with matrix formalism.

Matrix Formulation (N=2)

From the equations of motion for the $N = 2$ case, let

$$\begin{aligned} A_1 &= \frac{m_2}{M} \frac{l_2}{l_1} \cos(\theta_2 - \theta_1) \\ B_1 &= \frac{m_2}{M} \frac{l_2}{l_1} \omega_2^2 \sin(\theta_2 - \theta_1) + \frac{g}{l_1} \sin(\theta_1) \\ &\quad - \frac{1}{l_1} (\ddot{x}_0 \cos(\theta_1) + \ddot{y}_0 \sin(\theta_1)) \\ A_2 &= \frac{l_1}{l_2} \cos(\theta_2 - \theta_1) \\ B_2 &= -\frac{l_1}{l_2} \omega_1^2 \sin(\theta_2 - \theta_1) + \frac{g}{l_2} \sin(\theta_2) \\ &\quad - \frac{1}{l_2} (\ddot{x}_0 \cos(\theta_2) + \ddot{y}_0 \sin(\theta_2)) , \end{aligned}$$

so the equations of motion are contained in a two-dimensional system:

$$\begin{bmatrix} 1 & A_1 \\ A_2 & 1 \end{bmatrix} \begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

Having format $A\dot{\omega} = \mathbf{B}$, the vector $\dot{\omega}$ is isolated by the matrix multiplication of A^{-1} across the equation:

$$\dot{\omega} = A^{-1} \mathbf{B}$$

The inverse of the A -matrix is easy to write for $N = 2$:

$$A^{-1} = \begin{bmatrix} 1 & A_1 \\ A_2 & 1 \end{bmatrix}^{-1} = \frac{1}{1 - A_1 A_2} \begin{bmatrix} 1 & -A_1 \\ -A_2 & 1 \end{bmatrix}$$

Using this, we find

$$\begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \end{bmatrix} = \frac{1}{1 - A_1 A_2} \begin{bmatrix} B_1 - A_1 B_2 \\ -B_1 A_2 + B_2 \end{bmatrix} .$$

Matrix Formulation (General N)

The equations of motion for the N -pendulum are a tangle of θ_n 's, ω_n 's, along with all constraints (masses, lengths), and external forces. Much of this information can be contained in a size $N \times N$ matrix

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1k} & \cdots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \cdots & \cdots & \cdots & \cdots \\ A_{31} & A_{32} & A_{33} & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{n1} & \cdots & \cdots & \cdots & A_{nn} & \cdots & A_{nN} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{N1} & \cdots & \cdots & \cdots & A_{Nk} & \cdots & A_{NN} \end{bmatrix} ,$$

so the equations of motion are written

$$A \begin{bmatrix} \dot{\omega}_1 \\ \dot{\omega}_2 \\ \dot{\omega}_3 \\ \cdots \\ \dot{\omega}_n \\ \cdots \\ \dot{\omega}_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ b_n \\ \cdots \\ b_N \end{bmatrix} ,$$

or in more condensed notation,

$$A\dot{\omega} = \mathbf{b} .$$

Like the $N = 2$ case, the vector $\dot{\omega}$ is isolated by the matrix multiplication of A^{-1} across the equation:

$$\dot{\omega} = A^{-1} \mathbf{b}$$

The inverse of the A -matrix is not so easy to write for arbitrary N , but A^{-1} is readily attained using sub-matrices and determinants.

The job now is to determine all components A_{nk} of the above matrix, along with those of the vector $\mathbf{b}$. This is made easier by switching to index notation, and to begin we'll express the above matrix expression in a way that separates off-diagonal components from diagonal components:

$$\sum_{k=1}^N A_{nk} \dot{\omega}_k \tilde{\delta}_{nk} + \sum_{k=1}^N A_{nk} \delta_{nk} \dot{\omega}_k = b_n$$

Comparing this to the equation of motion for the driven N -pendulum, match the $\dot{\omega}_k$ - and $\dot{\omega}_n$ -terms to write

$$\begin{aligned} \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k (\dot{\omega}_k \cos(\theta_n - \theta_k)) \right) \\ + \left(\sum_{j=n}^N m_j \right) l_n^2 \dot{\omega}_n = b_n . \end{aligned}$$

Meanwhile, the component b_n consists of every term in the equation of motion that does not involve derivatives of

ω . Thus:

$$\begin{aligned} b_n = & - \left(\sum_{j=n}^N m_j \right) (\ddot{x}_0(t) l_n \cos(\theta_n)) \\ & - \left(\sum_{j=n}^N m_j \right) (\ddot{y}_0(t) l_n \sin(\theta_n)) \\ & + g \left(\sum_{j=n}^N m_j \right) l_n \sin(\theta_n) \\ & - \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k (\omega_k^2 \sin(\theta_n - \theta_k)) \right) \end{aligned}$$

For bookkeeping, introduce the δ -like function:

$$\sigma_{kj} = \begin{cases} 0 & k > j \\ 1 & k \leq j \end{cases}$$

Then, we determine off-diagonal components A_{nk} by writing

$$\begin{aligned} \sum_{k=1}^N A_{nk} \dot{\omega}_k \tilde{\delta}_{nk} &= \sum_{j=n}^N \left(m_j \sum_{k=1}^j \tilde{\delta}_{nk} l_n l_k (\dot{\omega}_k \cos(\theta_n - \theta_k)) \right) \\ &= \sum_{j=n}^N \sum_{k=1}^N \sigma_{kj} m_j \tilde{\delta}_{nk} l_n l_k (\dot{\omega}_k \cos(\theta_n - \theta_k)) , \end{aligned}$$

implying, for $n \neq k$:

$$A_{nk} = \sum_{j=n}^N m_j \sigma_{kj} l_n l_k (\cos(\theta_n - \theta_k))$$

Finally we solve for the diagonal components of matrix A. Proceed with

$$\begin{aligned} \sum_{k=1}^N A_{nk} \delta_{nk} \dot{\omega}_k &= \left(\sum_{j=n}^N m_j \right) l_n^2 \dot{\omega}_n \\ &= \sum_{k=1}^N \left(\sum_{j=n}^N m_j l_k^2 \right) \delta_{nk} \dot{\omega}_k , \end{aligned}$$

implying, for $n = k$:

$$A_{nn} = \left(\sum_{j=n}^N m_j \right) l_n^2$$

8.4 Iterative Setup

The physical pendulum, regardless of how many masses are attached, should be handled by a symplectic or higher-order approximation scheme. For this reason, we choose the fourth-order Runge-Kutta method.

Generalized Coordinates

A full description of the evolution of the double pendulum can be fused into $\dot{\omega} = A^{-1} \mathbf{B}$ by interleaving all θ -variables into the time-evolving vector:

$$\frac{d}{dt} \begin{bmatrix} \theta_1 \\ \omega_1 \\ \theta_2 \\ \omega_2 \end{bmatrix} = \begin{bmatrix} \omega_1 \\ (B_1 - A_1 B_2) / (1 - A_1 A_2) \\ \omega_2 \\ (B_2 - A_2 B_1) / (1 - A_1 A_2) \end{bmatrix}$$

Invoking the notion of *generalized coordinates*, pack all θ - and ω -variables into a single vector $\mathbf{q}$, and then the right side is written as a vector $\mathbf{p}$ such that

$$\frac{d}{dt} \mathbf{q}(t) = \mathbf{p}(t) .$$

Note that $\mathbf{p}$ contains ω in the odd entries and $A^{-1} \mathbf{B}$ in the even entries.

As a practical matter, let us tack the time variable t on the end of the vector $\mathbf{q}$ such that:

$$\mathbf{q} = \begin{bmatrix} \theta_1 \\ \omega_1 \\ \theta_2 \\ \omega_2 \\ t \end{bmatrix}$$

Correspondingly, the final entry in $\mathbf{p}$ is

$$p_{4+1} = 1 ,$$

and the relationship $\dot{\mathbf{q}} = \mathbf{p}$ still holds.

For the general- N case, the augmented system is represented by

$$\frac{d}{dt} \mathbf{q}(t) = \frac{d}{dt} \begin{bmatrix} \theta_1 \\ \omega_1 \\ \dots \\ \theta_n \\ \omega_n \\ \dots \\ \theta_N \\ \omega_N \\ t \end{bmatrix} = \mathbf{p}(t) .$$

The generalized coordinate vector $\mathbf{q}$ has $2N + 1$ components, as does vector $\mathbf{p}$. To be clear, the quantities θ, ω relate to q_n via an index k such that:

$$q_n = \begin{cases} \theta_k & k = 2n - 1 \\ \omega_k & k = 2n \\ t & k = 2N + 1 \end{cases}$$

Like the $N = 2$ case, vector $\mathbf{p}$ contains ω in the odd entries and $\dot{\omega} = A^{-1}\mathbf{b}$ in the even entries. The last entry in $\mathbf{p}$ is $p_{2N+1} = 1$:

$$p_n = \begin{cases} \omega_k & k = 2n - 1 \\ \dot{\omega}_k & k = 2n \\ 1 & k = 2N + 1 \end{cases}$$

Driving Force

Examining the equations of motion, one notices the adherence to Galilean relativity. That is, the point $x_0(t), y_0(t)$ sets position of the system without affecting dynamics. The first derivatives $\dot{x}(t), \dot{y}(t)$ are altogether absent. Only the acceleration $\ddot{x}(t), \ddot{y}(t)$ of the anchor point matters physically.

With the freedom to choose any evolution scheme we wish, let's proceed with a toy driving force that traces generalized Lissajous figures characterized by frequencies w_x, w_y (to clearly distinguish from any ω -variables) and scaled by L_x, L_y , respectively:

$$\begin{aligned} x_0(t) &= L_x \cos(w_x t) \\ y_0(t) &= L_y \sin(w_y t) \end{aligned}$$

Take two time derivatives to have the acceleration of the anchor point:

$$\begin{aligned} \ddot{x}_0(t) &= -w_x^2 L_x \cos(w_x t) \\ \ddot{y}_0(t) &= -w_y^2 L_y \sin(w_y t) \end{aligned}$$

RK4 Implementation

In terms of a single pair of generalized coordinates, the fourth-order Runge-Kutta method is written:

$$\begin{aligned} \frac{d}{dt}q(t) &= f(t, q(t)) = p(t) \\ q(t+h) &= \frac{h}{6} (k_1 + 2k_2 + 2k_3 + k_4) \\ k_1 &= f(t, q) \\ k_2 &= f\left(t + \frac{h}{2}, q + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t + \frac{h}{2}, q + \frac{h}{2}k_2\right) \\ k_4 &= f(t+h, q + hk_3) \end{aligned}$$

Of course, we must apply the above structure to each mass involved in the pendulum plus the time component tacked to the end, so we're doing the calculation $2N + 1$ times per step.

Rather than adorn the *RK4* apparatus with yet more subscripts, it's more efficient to turn to a BASIC-like pseudocode notation as shown.

```
FOR j = 1 TO (2*N+1) { qtmp(j) = q(j) }
Calculate p given qtmp
FOR j = 1 TO (2*N+1) { f(1, j) = p(j) }

FOR j = 1 TO (2*N+1) { qtmp(j) = q(j) + (dt/2) * f(1, j) }
Calculate p given qtmp
FOR j = 1 TO (2*N+1) { f(2, j) = p(j) }

FOR j = 1 TO (2*N+1) { qtmp(j) = q(j) + (dt/2) * f(2, j) }
Calculate p given qtmp
FOR j = 1 TO (2*N+1) { f(3, j) = p(j) }

FOR j = 1 TO (2*N+1) { qtmp(j) = q(j) + dt * f(3, j) }
Calculate p given qtmp
FOR j = 1 TO (2*N+1) { f(4, j) = p(j) }

FOR j = 1 TO (2*N+1)
    q(j) = q(j) + (dt/6) * (f(1, j) + 2*f(2, j) + 2*f(3, j) + f(4, j))
```

```
x(0) = Lx * COS(wx * q((2*N+1))) // Update all positions
y(0) = Ly * SIN(wy * q((2*N+1)))

FOR j = 1 TO N
    x(j) = x(j-1) + l(j) * SIN(q(2*j-1))
    y(j) = y(j-1) - l(j) * COS(q(2*j-1))
```

```
xdd = -Lx * (wx^2) * COS(wx * q(2*N+1)) // Anchor point acceleration
ydd = -Ly * (wy^2) * SIN(wy * q(2*N+1))
```

Calculating each successive $f(1, j)$, $f(2, j)$, etc. requires a full update of the vector $\mathbf{p}$. The routine represented by Calculate p given qtmp is shown in the code box that follows.

```

Routine: Calculate p given qtmp
      qtmp -> q

FOR n = 1 TO N // Diagonal components of A
  A(n, n) = 0
  FOR j = n TO N { A(n, n) = A(n, n) + m(j) }
  A(n, n) = A(n, n) * l(n)^2

FOR n = 1 TO N // Off-diagonal components of A
  FOR k = 1 TO N
    IF (n <> k) THEN
      A(n, k) = 0
      FOR j = n TO N
        A(n, k) = A(n, k) + m(j) * sigma(k, j)
          * l(n) * l(k) * COS(q(2*n-1) - q(2*k-1))

FOR n = 1 TO N // Vector b
  b(n) = 0
  FOR j = n TO N
    b(n) = b(n) - m(j) * l(n) * (xdd * COS(q(2*n-1)) + ydd * SIN(q(2*n-1)))
      + g * m(j) * l(n) * SIN(q(2*n-1))
  FOR k = 1 TO j
    IF (n <> k) THEN
      b(n) = b(n) - m(j) * l(n) * l(k)
        * q(2*k)^2 * SIN(q(2*n-1) - q(2*k-1))

AInv = Inverse of A

FOR j = 1 TO N // Vector p
  p(2*j-1) = q(2*j) // Omega in odd entries
  p(2*j) = 0 // Initialize to nonzero for damping model
  FOR k = 1 TO N
    p(2*j) = p(2*j) + AInv(j, k) * b(k) // AInv*b in even entries

p(2 * N + 1) = 1 // Time component

```

Chapter 38

Functional Constructs

1 Turing Machine

... .. Turing (1936) Hilbert's Entscheidungsproblem

2 The Lambda Calculus

The λ -calculus, introduced by Alonzo Church in the 1930s, is a formal system for defining, applying, and reasoning about functions. It is in many ways equivalent to a Turing machine, and achieves Turing completeness with only three syntactic forms. Lambda calculus forms the theoretical foundation of many functional programming languages (i.e., Haskell, Scheme, ML).

2.1 Syntax

In the *untyped* λ -calculus, every term is built from three constructs: variables, abstractions, and applications.

Variables

Variables are symbols representing parameters or placeholders for expressions. For instance:

x, y, f, z_1, foo

Abstraction

A function that takes input x and returns M is written

$\lambda x. M$

where x is the *bound variable* (the function's parameter), and M is the *body* of the function.

Abstraction is the only way to create functions in λ -calculus. Comparing to other regimes, $\lambda x. M$ is like the 'maps to' notation $x \mapsto M$ in mathematics, or $x \Rightarrow M$ in JavaScript arrow functions.

Application

Function application is written by simply placing two terms side by side:

$M N$

This means 'apply function M to argument N .'

Associativity

Application is *left-associative*, thus

$M N P = (M N) P$

means 'apply M to N first, then apply the result to P .' The application

$(\lambda x. M) N$

means: 'call the function $\lambda x. M$ with N as the argument.'

Scope

- A variable x in $\lambda x. M$ is *bound* within M .
- A variable that is not bound is *free*.

For example, if we have $\lambda x. (x y)$, then x is bound and y is free. Whether a variable is bound or free is a matter of its *scope*.

2.2 Special Functions

Using variables, abstraction, and application, we can construct a handful of tools that will be used ubiquitously throughout λ -calculus and its derivatives. It's worth memorizing the following special functions:

Identity

The application

$I = \lambda x. x$

is called the *identity function*, and simply returns its argument unchanged:

$I a = (\lambda x. x) a \rightarrow a$

Constant

The application

$K = \lambda x. \lambda y. x$

is called the *constant function*, which returns the first of its arguments while ignoring the second:

$K a b \rightarrow a$

Substitution

The application

$S = \lambda x. \lambda y. \lambda z. x z (y z)$

is called *substitution*, which takes three arguments $a b c$, and produces the combination $a c (b c)$. That is:

$S a b c \rightarrow a c (b c)$

2.3 Special Operations

Beta Reduction

Beta-reduction captures the idea of applying a function to an argument:

$$(\lambda x.M) N \rightarrow_{\beta} M[x = N]$$

The term $M[x = N]$ denotes substituting N for every occurrence of x in M . In practice, β -reduction always gets rid of the left-most λ .

Alpha Conversion

Two λ -expressions are equivalent up to renaming of bound variables:

$$\lambda x. x =_{\alpha} \lambda y. y$$

This avoids variable name collisions. In mathematics, this is equivalent to a change of dummy variable or integration variable.

Eta Conversion

Eta-conversion expresses functional equivalence

$$\lambda x. (f x) =_{\eta} f$$

when x does not occur as free in f .

Church Encoding

Data structures and numbers can be represented purely as functions. For example:

$$\begin{aligned} 0 &= \lambda f. \lambda x. x \\ 1 &= \lambda f. \lambda x. f x \\ 2 &= \lambda f. \lambda x. f(f x) \end{aligned}$$

These *Church numerals* allow arithmetic to be expressed without primitive numbers.

2.4 Minimal Set

As rock bottom as the S, K functions may appear, these do not form a minimal set. Instead, we'll show that there exists a single entity ι (Greek *iota*) that brings just S and K simultaneously.

To begin, consider two (as yet unnamed) functions A and B . Take ι as a primitive with the sole rewrite

$$\iota(x) \rightarrow ((x)(A))(B),$$

and remember application is left-associative; we will nevertheless keep all parentheses explicit.

Identity

Next we propose the identity term

$$I = (\iota(\iota)) ,$$

and test I on an arbitrary argument (X) using the rewrite rule alone:

$$\begin{aligned} (I(X)) &= ((\iota(\iota))(X)) \\ &\rightarrow (((\iota)(A))(B))(X) \\ &\rightarrow (((((A)(A))(B))(B))(X)) \end{aligned}$$

Looking at the inner part, it suffices to check whether

$$J = (((A(A))(B))(B))$$

does the job of the identity function such that:

$$(J(X)) = (I(X))$$

As formally *opaque* objects, we're stuck without knowing anything about A, B . To proceed, we embed the rewrite system in the λ -calculus notation such that:

$$\iota = \lambda f. f A B$$

This lets us take advantage of β -reduction instead of staying purely at the rewrite-rule level. In this frame, we propose the forms:

$$\begin{aligned} A &= \lambda x. \lambda y. \lambda z. (x(z))(y(z)) \\ B &= \lambda x. \lambda y. x \end{aligned}$$

If A, B look suspiciously like S, K , you wouldn't be wrong. The motive here is to prove that the pair A, B behaves exactly like S, K without using S, K as combinators.

To make sure the rewrite rule stays intact in the λ -notation, we calculate

$$\iota(x) = (\lambda f. f A B) x \rightarrow_{\beta} x A B ,$$

or, using left-associativity,

$$\iota(x) = ((x)(A))(B) ,$$

which agrees with the first thing we wrote about ι .

Now return to the matter of the identity function

$$J = (((A(A))(B))(B)) .$$

Computing this from the inside-out, we have no choice but to invoke the structure of A explicitly:

$$\begin{aligned}
 (A (A)) &= (A A) \\
 &= (\lambda x. \lambda y. \lambda z. (x (z)) (y (z))) A \\
 &\rightarrow_{\beta} \lambda y. \lambda z. A z (y z) \\
 ((A (A)) (B)) &= ((A A) B) \\
 &= (\lambda y. \lambda z. A z (y z)) B \\
 &\rightarrow_{\beta} \lambda z. A z (B z) \\
 (((A (A)) (B)) (B)) &= (((A A) B) B) \\
 &= (\lambda z. A z (B z)) B \\
 J &\rightarrow_{\beta} A B (B B)
 \end{aligned}$$

Comparing $A B (B B)$ to the expression that got us here, one has to wonder where the extra B came from and why there is only one A on the left. This is not the result of an error - we must acknowledge that the perservation of symbol count is *not* maintained in beta reduction.

To handle $(A B)$, we note that

$$\begin{aligned}
 A B &= (\lambda a. \lambda b. \lambda c. (a (c)) (b (c))) B \\
 &\rightarrow_{\beta} (\lambda b. \lambda c. (B (c)) (b (c)))
 \end{aligned}$$

Calculating $(B B)$, we further find:

$$\begin{aligned}
 (B B) &= (\lambda x. \lambda y. x) B \\
 &\rightarrow_{\beta} (\lambda y. B)
 \end{aligned}$$

We're now in position to establish J as equivalent to the identity I . Testing on an arbitrary variable x , we start with

$$\begin{aligned}
 J(x) &= (A B (B B)) (x) \\
 &= ((A B) (B B)) (x),
 \end{aligned}$$

and substitute $A B$ to have

$$\begin{aligned}
 J(x) &= ((\lambda b. \lambda c. (B (c)) (b (c))) (B B)) (x) \\
 &\rightarrow_{\beta} (\lambda c. (B (c)) ((B B) (c))) (x) \\
 &\rightarrow_{\beta} (B (x)) ((B B) (x)) \\
 &= (B (x)) ((\lambda y. B) (x)) \\
 &\rightarrow_{\beta} B (x) B
 \end{aligned}$$

Utilizing B 's exact structure for the first time, continue with

$$J(x) = ((\lambda u. \lambda v. u) (x)) (B)$$

(where we have α -renamed the bound variables of B to u, v to avoid clash with the meta-variable x). Now finish by two β -reductions:

$$((\lambda u. \lambda v. u) (x)) (B) \rightarrow_{\beta} (\lambda v. x) (B) \rightarrow_{\beta} x.$$

Thus $J(x) = x$ for all x , i.e. J is extensionally the identity I .

Deriving S, K

To continue, define three more ι -derived terms (indirectly) involving S, K :

$$\begin{aligned}
 C &= (\iota (I)) \\
 K &= (\iota (C)) \\
 S &= (\iota (K))
 \end{aligned}$$

Use $\iota = \lambda f. f A B$ to expand C :

$$C = (\lambda f. f A B) I \rightarrow_{\beta} I A B = A B$$

Then, by the rewrite rule $\iota(x) \rightarrow ((x)(A)) (B)$ we have:

$$\begin{aligned}
 K &\rightarrow ((C) (A)) (B) \\
 &= C A B = (A B) A B
 \end{aligned}$$

The quantity on the right needs to be worked out manually. We already worked out $(A B)$, namely:

$$(A B) \rightarrow_{\beta} \lambda b. \lambda c. B c (b c)$$

Append A to the right and keep simplifying:

$$\begin{aligned}
 (A B) A &= (\lambda b. \lambda c. B c (b c)) A \\
 &\rightarrow_{\beta} \lambda c. B c (A c)
 \end{aligned}$$

Append B to the right and turn the crank again to find:

$$\begin{aligned}
 (A B) A B &= (\lambda c. B c (A c)) B \\
 &\rightarrow_{\beta} B B (A B) \\
 K &= B B (A B)
 \end{aligned}$$

The $B B$ -term was also previously worked out, so we continue with

$$K = (\lambda y. B) (A B) \rightarrow_{\beta} B,$$

leaving just B . Replacing B with its equivalent λ -expression, we conclude:

$$K = B = \lambda x. \lambda y. x$$

Finally we work out how S fits in. Start with

$$\begin{aligned}
 S &\rightarrow ((K) (A)) (B) \\
 &= K A B = (\lambda x. \lambda y. x) A B \\
 &\rightarrow_{\beta} A,
 \end{aligned}$$

and immedietly conclude S is equivalent to A :

$$S = A = \lambda x. \lambda y. \lambda z. x z (y z)$$

In summary, we find that ι contains S, K such that:

$$\iota = \lambda f. f S K$$

3 The SK Calculus

Central to the λ -calculus are two functions S and K . These, along with other players such as ι and the identity (I), are formally called *combinators*. In particular:

Combinator	Symbol	Expression
Identity	I	$\lambda x. x$
Substitution	S	$\lambda x. \lambda y. \lambda z. x z (y z)$
Constant	K	$\lambda x. \lambda y. x$
iota	ι	$\lambda f. f S K$

As it turns out, we can use the pair S, K as a basis to do any possible computation *without* explicit use of variable storage. These are equivalent in expressive power to the untyped λ -calculus and serve as a foundation for certain minimalist functional programming languages.

Notation

Using a slightly more transparent notation, the S, K combinators read:

$$S[x][y][z] = x[z][y[z]]$$

$$K[x][y] = x$$

The square brackets can help the eye and the computer's input parser sift through statements. Like training wheels though, we can get rid of most of the brackets without ambiguity:

$$S x y z = x z (y z)$$

$$K x y = x$$

Identity

Some introductions to SK -calculus actually call it SKI calculus, where I stands for 'identity', which is it itself a combinator that can be expressed in terms of S, K . This is established by writing:

$$I = S[K][K]$$

For a test, the quantity $I[a]$ must reduce to a , which is indeed the case:

$$\begin{aligned} I[a] &= S[K][K][a] \\ &= K[a][K[a]] \\ &= a \end{aligned}$$

Or, in the streamlined notation:

$$Ia = SKKa = Ka(Ka) = a$$

As an aside, the identity can be more generally written $I = S[K][Z]$, where Z stands for any symbol at all.

What are Combinators?

The S and K combinators, along with any other combinators we encounter, are named for what they do operationally. A combinator simply takes inputs and recombines them as outputs. In the context of λ -calculus, however, combinators aren't simply rearranging symbols; rather these result in new variables, abstractions, and applications. The claim is we can calculate anything this way.

For a modest example, suppose we are handed the data $[a][b][c]$, and are assigned to produce the target expression $a[b[a][c]]$ using a so-far unknown operator O :

$$O(\{S, K\})[a][b][c] = a[b[a][c]]$$

To go about finding O in typical 'left to right' fashion amounts to guessing what to put before $[a][b][c]$, which is unfeasible by hand, perhaps not for a computer.

Rather, it's more straightforward to begin with the target expression and work backwards through λ -calculus. For the example on hand, one might capture the output by writing:

$$F = \lambda a. \lambda b. \lambda c. a (b (a c))$$

The goal then is to systematically convert the λ -notation back to the SK -notation, and the operator O should be evident.

3.1 Elimination Rules

Consider two arbitrary terms M, N standing for any λ -expressions. The following *elimination rules* describe how λ -abstractions can be systematically expressed using only the S and K combinators:

If the body is exactly the variable itself, this is the identity:

$$\lambda x. x = I = S K K$$

If x is not free in M , i.e. the body doesn't mention the bound variable, then $\lambda x. M$ is the constant function $K M$:

$$\lambda x. M = K M$$

To decompose an abstraction of an application into two smaller applications, handled recursively, one writes:

$$\lambda x. (M N) = S (\lambda x. M) (\lambda x. N)$$

This rule is a subtle one. To see this explicitly, we note

$$S f g x = f x (g x) ,$$

and then let

$$\begin{aligned} f &= (\lambda x. M) \\ g &= (\lambda x. N) \end{aligned}$$

so that:

$$S f g = \lambda x. f x (g x) = \lambda x. (M N)$$

Together, these rules allow any λ -expression to be *eliminated* in favor of S and K alone.

Applied Elimination

Returning to the example problem

$$F = \lambda a. \lambda b. \lambda c. a (b (a c)) ,$$

attack the innermost term $\lambda c. a (b (a c))$ as follows, treating a, b as constants with respect to c :

$$\begin{aligned} \lambda c. a (b (a c)) &= S (\underbrace{\lambda c. a}_{K a}) (\underbrace{\lambda c. b (a c)}_{\text{expand}}) \\ (\lambda c. b (a c)) &= S (\underbrace{\lambda c. b}_{K b}) (\underbrace{\lambda c. a c}_{S(K a)I}) \\ &= S (K b) (S (K a) I) \end{aligned}$$

Thus, the result of eliminating λc is

$$G = S (K a) (S (K b) (S (K a) I)) .$$

Now we get rid of the λb term by treating a as constant with respect to b to write:

$$\lambda b. G = \lambda b. \left(\underbrace{S (K a)}_{\text{no } b} \right) \left(\underbrace{S (K b) (S (K a) I)}_{\text{depends on } b} \right)$$

Going term by term, we have

$$\lambda b. S (K a) = K (S (K a))$$

and

$$\begin{aligned} \lambda b. S (K b) (S (K a) I) \\ = S \left(\underbrace{\lambda b. S (K b)}_{\text{expand}} \right) \left(\underbrace{\lambda b. S (K a) I}_{K(S(K a)I)} \right) . \end{aligned}$$

Expanding the first term gives

$$\lambda b. S (K b) = S \left(\underbrace{\lambda b. S}_{K S} \right) \left(\underbrace{\lambda b. K b}_{\text{application}} \right) .$$

and the final application expands as:

$$\lambda b. K b = S \left(\underbrace{\lambda b. K}_{K K} \right) \left(\underbrace{\lambda b. b}_I \right) = S (K K) I$$

Thus, the result of eliminating λb yields:

$$\begin{aligned} H &= S (K (S (K a))) \\ &= S (S (K S) (S (K K) I)) (K (S (K a) I)) \end{aligned}$$

By now the pattern is evident as we finally eliminate the λa term from F . This amounts to applying the same rules, and only terms directly involving a require expansion. That is, whenever a appears inside $K(\dots)$, then a is not free. Occurrences nested as $(K a)$ inside applications trigger the S rule locally. The final result can depend on how this is done, but all should be equivalent. Carrying these steps forward, one finds:

$$\begin{aligned} F &= S (S (K S) (S (K K) I)) (S (K S) K) \\ &= (S (K (S (S (K K) I))) K) \end{aligned}$$

Recall that F is the operator that is supposed to perform as

$$O(\{S, K\})[a][b][c] = a[b[a][c]] .$$

In the square bracket notation, we have:

$$\begin{aligned} O &= s[s[k[s]][s[k[k]][s[k[s]][k]]]] \\ &= [s[k[s[s[k][k]]]][k]] \end{aligned}$$

Running the calculation forward, either by hand or machine, yields the following progression:

```

s[s[k[s]]][s[k[k]]][s[k[s]][k]]][s[k[s[s[k]]][k]]][a][b][c]
s[k[s]]][s[k[k]]][s[k[s]][k]]][a][s[k[s[s[k]]][k]]][k][a][b][c]
k[s][a][s[k[k]]][s[k[s]][k]]][a][s[k[s[s[k]]][k]]][k][a][b][c]
s[s[k[k]]][s[k[s]][k]]][a][s[k[s[s[k]]][k]]][k][a][b][c]
s[k[k]]][s[k[s]][k]]][a][b][s[k[s[s[k]]][k]]][k][a][b][c]
k[k][a][s[k[s]][k]][a][b][s[k[s[s[k]]][k]]][k][a][b][c]
k[s[k[s]][k]][a][b][s[k[s[s[k]]][k]]][k][a][b][c]
s[k[s]][k][a][s[k[s[s[k]]][k]]][k][a][b][c]
k[s][a][k[a]][s[k[s[s[k]]][k]]][k][a][b][c]
s[k[a]][s[k[s[s[k]]][k]]][k][a][b][c]
k[a][c][s[k[s[s[k]]][k]]][k][a][b][c]
a[s[k[s[s[k]]][k]]][k][a][b][c]
a[k[s[s[k]]][k]][a][k[a][b][c]]
a[s[s[k]]][k][a][b][c]
a[s[k][k][b][k[a][b]][c]]
a[k[b][k[b]][k[a][b]][c]]
a[b[k[a][b]][c]]
a[b[a][c]]

```

The final line, namely $a[b[a][c]]$ is the target expression.

$$n = P^n S[K][S][K]$$

3.2 Numbers and Arithmetic

Whole Numbers

Much like the Church encoding for numbers, the S , K combinators can be layered to capture the whole numbers. One way to do so is by the following pattern:

$$\begin{aligned}
0 &= K \\
1 &= S[K] \\
2 &= S[S[K]] \\
3 &= S[S[S[K]]]
\end{aligned}$$

Building the case intuitively, consider the progression:

$$S[K][S][K] \rightarrow K[K][S[K]] \rightarrow K$$

Since we're associating K with the integer 0, we can rewrite 0 as a prefix $S[K]$ acting on a suffix $[S][K]$:

$$0 = S[K][S][K]$$

Now we need some operation to increment $0 \rightarrow 1$, and use the same procedure on 1 to yield 2, and so on. Do this by proposing an operator P that effectively adds one to its argument

$$\begin{aligned}
1 &= P[S[K]][S][K] \\
2 &= P[P[S[K]]][S][K] \\
3 &= P[P[P[S[K]]]][S][K],
\end{aligned}$$

Note that each of these is written as a unique prefix $P[P[\dots[S[K]]]]$ and a universal suffix $[S][K]$. With the universal suffix $(S)(K)$, the whole numbers are the prefixes Q such that

$$Q(S)(K) \rightarrow S^n K.$$

Here, $S^n K$ means n left-nested copies of S applied to K .

To go from n to $n + 1$, the operator P must satisfy:

$$(P Q)(S)(K) = S(Q(S)(K))$$

A λ -term that does exactly this is:

$$P = \lambda q. \lambda r. \lambda y. r((q r) y)$$

This threads the fresh argument r to both positions, then combines with y . Translating with the SK elimination rules gives the point-free SK form

$$P = S(S(K S) K),$$

or in the square bracket notation:

$$P = S[S[K][S]][K]$$

To verify that P acts as the increment operator, compute for arbitrary q, r, y :

$$\begin{aligned}
(P q) r y &= (S(S(K S) K) q) r y \\
&= ((S(K S) K r) (q r)) y \\
&= (K r y) ((q r) y) \\
&= r((q r) y)
\end{aligned}$$

as required. For another sanity check, apply P to $0 = S[K][S][K]$ to get

$$P[S[K]][S][K] \rightarrow S[K],$$

which we identify as 1. Applying P again,

$$P[P[S[K]]][S][K] \rightarrow S[S[K]],$$

which is 2. Continuing in this way establishes the pattern we originally posed:

$$n = S^n[K]$$

Addition Operator

Recall the increment operator

$$\begin{aligned} P &= \lambda q. \lambda r. \lambda y. r((q r) y) \\ &= S[S[K][S]][K] \end{aligned}$$

that satisfies

$$(P q) r y = r((q r) y) .$$

Define the m -fold iterate of P by

$$\begin{aligned} P^0 n &= n \\ P^{m+1} n &= P(P^m n) \end{aligned}$$

We claim that for all prefixes m, n and all r, y :

$$(P^m n) r y = m r(n r y) .$$

That is, P^m applies the step $q \rightarrow r((q r) y)$ exactly m times.

We prove this by induction on m . For $m = 0$, we have

$$(P^0 n) r y = n r y$$

while

$$0 r(n r y) = n r y ,$$

so the base case holds. For all other m , we write:

$$\begin{aligned} (P^{m+1} n) r y &= (P(P^m n)) r y \\ &= ((\lambda q. \lambda r. \lambda y. r((q r) y)) (P^m n)) r y \\ &\rightarrow_{\beta} (\lambda r. \lambda y. r(((P^m n) r) y)) r y \\ &\rightarrow_{\beta} (\lambda y. r(((P^m n) r) y)) y \\ &\rightarrow_{\beta} r(((P^m n) r) y) . \end{aligned}$$

Now apply the induction hypothesis

$$(P^m n) r y = m r(n r y) ,$$

$$(P^{m+1} n) r y = r(m r(n r y)) = (m+1) r(n r y)$$

This matches the numeral law

$$(m+1) r z = r(m r z)$$

and the induction is complete.

With the above, we define the addition operator:

$$\text{add} = \lambda m. \lambda n. \lambda r. \lambda y. m r(n r y)$$

Then, for the universal suffix $r = S, y = K$, we obtain

$$\begin{aligned} (\text{add } m n) S K &= m S(n S K) \\ &= S^m(S^n K) \\ &= S^{m+n} K , \end{aligned}$$

and so:

$$\text{add } m n = P^m n$$

Finally, we need to express the addition operator in pure SK -notation. The details for this get quite wooly, so we'll breeze over the derivation while showing key steps. Begin by treating m, r, n as constants with respect to y , and use the application rule

$$\begin{aligned} \lambda y. m r(n r y) &= S(\lambda y. m r)(\lambda y. n r y) \\ &= S(K(m r))(S(K(n r)) I) . \end{aligned}$$

Next apply the rule again to the whole result. After routine expansions using $I = SKK$ and $S(K K) I = I$, along with the identity

$$S(KS) X = \lambda z. S z(X z) ,$$

one obtains the simplified form:

$$\begin{aligned} \lambda r. (S(K S)(K m)) (S(K S)(K n)) &= \\ S(S(K S)(K m)) (S(K S)(K n)) \end{aligned}$$

Abstracting n and then m from the expression above (omitting only routine S/K bookkeeping and the use of $S(KK)I = I$) yields the point-free SK term:

$$\text{add} = S[K[S]] [S[K[S[K[S]]]] [S[K[K]]]]$$

Given two whole numbers a, b , represented by their SK prefixes, the addition operator combines them as

$$a + b = (\text{add})[a \text{ prefix}][b \text{ prefix}][S][K] .$$

Here the universal suffix $[S][K]$ ensures reduction to the canonical numeral form:

$$\begin{aligned} (\text{add } a \ b) \ S \ K &= a \ S \ (b \ S \ K) \\ &= S^a (S^b K) \\ &= S^{a+b} K . \end{aligned}$$

Thus the addition operator really implements $a + b$ within the SK -calculus.

this in the SK -calculus, we list:

$$\begin{aligned} \text{add} &= S[K[S]] [S[K[S[K[S]]]]][S[K[K]]] \\ 1 \text{ prefix} &= [S[S[K[S]]][K]][S[K]] \\ 2 \text{ prefix} &= [S[S[K[S]][K]] \\ &\quad [S[S[K[S]][K]][S[K]]]] \\ \text{suffix} &= [S][K] \end{aligned}$$

Then, we append the operator, both prefixes, and the universal suffix into a single entity and simplify like crazy. To keep the notation a little simpler, let

$$\begin{aligned} f &= K[S] \\ g &= S[K[S[K[S]]]][S[K[K]]] \\ h &= S[S[K[S]][K]][S[K]] , \end{aligned}$$

For an example, consider the sum $1 + 2 = 3$. To express and the calculation proceeds over thirty steps:

```
s[f][g][h][s[s[f][k]][h]][s][k]
f[h][g[h]][s[s[f][k]][h]][s][k]
s[g[h]][s[s[f][k]][h]][s][k]
g[h][s][s[s[f][k]][h][s]][k]
k[s[f]][h][s[k[k]][h]][s][s[s[f][k]][h][s]][k]
s[f][s[k[k]][h]][s][s[s[f][k]][h][s]][k]
f[s][s[k[k]][h][s]][s[s[f][k]][h][s]][k]
s[s[k[k]][h][s]][s[s[f][k]][h][s]][k]
s[k[k]][h][s][k][s[s[f][k]][h][s][k]]
k[k][s][h[s]][k][s[s[f][k]][h][s][k]]
k[h[s]][k][s[s[f][k]][h][s][k]]
h[s][s[s[f][k]][h][s][k]]
s[f][k][s][s[k][s]][s[s[f][k]][h][s][k]]
f[s][f][s[k][s]][s[s[f][k]][h][s][k]]
s[f][s[k][s]][s[s[f][k]][h][s][k]]
f[s[s[f][k]][h][s][k]][s[k][s][s[s[f][k]][h][s][k]]]
s[s[k][s][s[s[f][k]][h][s][k]]]
s[k[s[s[f][k]][h][s][k]][s[s[s[f][k]][h][s][k]]]
s[s[s[f][k]][h][s][k]]
s[s[f][k][s][h[s]][k]]
s[f[s][f][h[s]][k]]
s[s[f][h[s]][k]]
s[f[k][h[s][k]]]
s[s[h[s][k]]]
s[s[s[f][k][s][s[k][s]][k]]]
s[s[f[s][f][s[k][s]][k]]]
s[s[s[f][s[k][s]][k]]]
s[s[f[k][s[k][s][k]]]
s[s[s[s[k][s][k]]]
s[s[s[k[k][s[k]]]]]
s[s[s[k]]]
```


Perhaps not surprisingly, the final result is the representation of 3 in the SK numerals we developed. I promise to not show so much scratchwork again. If you're crunching through SK -calculus by hand or writing a program to do so, the above could be a useful milestone. Playing with more addition, it turns out that evaluating $3 + 4 = 7$ takes 50 steps, and $30 + 40 = 70$ requires 365 steps.

Multiplication Operator

We may take

$$\text{mul} = S[K[S]][K],$$

also known as the composition combinator B , which satisfies:

$$B x y z \rightarrow x (y z)$$

For a quick check on B , we write

$$\begin{aligned} S (K S) K x y z &= (((K S) x) (K x)) y z \\ &= S (K x) y z \\ &= (K x z) (y z) \\ &= x (y z). \end{aligned}$$

Applied to numeral prefixes m, n and the universal suffix $[S][K]$ we get

$$\begin{aligned} (\text{mul } m n) S K &= (B m n) S K \\ &\rightarrow m (n S) K \\ &= (n S)^m K = S^{mn} K, \end{aligned}$$

so mul indeed computes ordinary multiplication in this encoding.

Multiplication of two numbers is structurally similar to addition, with the only change being a swap of operators. The structure with prefixes and the suffix remains intact:

$$a \times b = (\text{mul})[a \text{ prefix}][b \text{ prefix}][S][K].$$

Despite the relative simplicity of the multiplication operator, getting final answers takes plenty of work. For instance, computing $2 \times 3 = 6$ requires 50 steps. To frame this, one starts with

$$\begin{aligned} P &= S[S[K[S]][K]] \\ 2 &= P[P[S[K]]] \\ 3 &= P[P[P[S[K]]]] \end{aligned}$$

so then:

$$(\text{mul})[P[P[S[K]]]][P[P[P[S[K]]]]][S][K] \rightarrow S^6 K$$

Exponent Operator

We may take

$$\text{exp} = S[K[S[S[K]][K]]][K],$$

which acts as an operator for exponentiation of our SK -encoded numerals.

To check the behavior, note that

$$\begin{aligned} \text{exp } m n S K &= (S[K[S[S[K]][K]]] m n) S K \\ &\rightarrow m ((n S) K) \\ &= m ((S^n K)) K \\ &= (S^n K)^m, \end{aligned}$$

so exp indeed computes exponentiation in this encoding: m^n .

Just as with addition and multiplication, exponentiation fits the general pattern

$$a^b = (\text{exp})[a \text{ prefix}][b \text{ prefix}][S][K].$$

For example, beginning with

$$\begin{aligned} P &= S[S[K[S]][K]] \\ 2 &= P[P[S[K]]] \\ 3 &= P[P[P[S[K]]]] \end{aligned}$$

we may compute

$$(\text{exp})[P[P[S[K]]]][P[P[P[S[K]]]]][S][K] \rightarrow S^8 K,$$

which corresponds to $2^3 = 8$. Or in 116 steps, one can verify $3^2 = 9$.

3.3 Combinatory Operations

For a quick review of what we know of SK -calculus, we take the S, K combinators

$$\begin{aligned} S x y z &= x z (y z) \\ K x y &= x \end{aligned}$$

as a basis for all computation. Not even the identity is fundamental, as $S K K x$ always returns x :

$$I = S K K$$

Review of Arithmetic

For arithmetic operations, we use the numerals $n = S^n K$, or in the prefix-suffix notation

$$n = P^n S K S K ,$$

where P is the increment operator

$$\text{increment} = P = S (S (K S) K) .$$

Then, numerals are written as

$$n \text{ prefix} = P^n S K$$

with:

$$\text{suffix} = S K$$

We take $\tilde{O}$ as the addition, multiplication, or exponent operator for whole numbers, represented by 0. Arithmetic operations are constructed via

$$a \tilde{O} b = (0)[a \text{ prefix}][b \text{ prefix}][S][K] .$$

For a list of operators, we found:

$$\text{add} = S (KS) (S (K (S K S)) I)$$

$$\text{mul} = S (K S) K$$

$$\text{exp} = S (K (S I)) K$$

Self-Apply

One of the most fundamental non-arithmetic operations in SK -calculus is forcing an object to act on itself. Given some x , how do we calculate $x x$? (Recall that in our streamlined notation, $x x$ also means $x[x]$.)

The answer is the **self** operator:

$$\text{self} = S I I ,$$

where $I = S K K$ is the identity. For a quick check, one writes

$$\begin{aligned} S I I x &= I x (I x) \\ &= x (I x) \\ &= x x \end{aligned}$$

Thus $S I I$ is the unique combinator that enforces self-application. Self-application is the simplest non-trivial operation in combinatory logic. While arithmetical operators like **add** or **mul** encode finite algebra, the **self** operator opens the door to infinite structures and recursion.

Restoring the full bracketed notation, the evaluation begins and ends as:

$$S[S[K][K]][S[K][K]][x] \rightarrow x[x] .$$

Cyclic Expression

As with any programming tool, it is possible to create an infinite loop in SK -calculus. In this setting, an ‘infinite loop’ corresponds to a combinatory expression that eventually reproduces itself after some number of reductions. Once this happens, the computation no longer progresses but cycles forever.

One of the simplest such expressions is built from the self-apply operator:

$$\text{loop} = (S I I) (S I I) .$$

Let us unroll this step by step. Beginning in streamlined notation:

$$\begin{aligned} (S I I) (S I I) &= I (S I I) (I (S I I)) \\ &= (S I I) (I (S I I)) \\ &= (S I I) (S I I) . \end{aligned}$$

By the third step the expression has returned to its original form. The evaluation never halts: each attempt to simplify merely reproduces the starting point. This is a genuine cyclic expression. Restoring full bracketed notation, the loop is written:

$$\text{loop} = S[S[K][K]][S[K][K]]$$

such that

$$\text{loop}[\text{loop}] \rightarrow \text{loop}[\text{loop}] .$$

Swap

Another natural question in combinatory logic is whether we can reverse the order of two objects. Given a compound expression $x y$, can we produce $y x$? The combinator that accomplishes this is the **swap** operator:

$$\text{swap} = S (K(S I)) (S (K K) I) ,$$

where $I = S K K$ is the identity.

Testing the definition shows that it indeed flips the order:

$$\begin{aligned} S (K(S I)) (S (K K) I) x y &= (K(S I) x) (S (K K) I x) y \\ &= (S I) (S (K K) I x) y \\ &= I y (x y) \\ &= y x \end{aligned}$$

The final result is clear: `swap` faithfully exchanges the two arguments. This confirms that *SK*-calculus is not limited to building and combining terms, but also permits us to rearrange their order - a subtle but important structural operation that embodies symmetry within combinatory logic. In the full bracketed notation, the transformation begins and ends as:

$$S[K[S[S[K][K]]]][S[K[K]][S[K][K]]][x][y] \rightarrow y[x]$$

Self-Reference

We now arrive at one of the most profound constructions in *SK*-calculus: self-reference. Unlike `self`, which forces an object to act upon itself, self-reference ties an object's behavior to its own description, opening the door to recursion and conditional recursion.

Consider the curious combination

$$R = S(Kx)(SII) ,$$

where x is some as-yet unspecified object. Let us allow R to act on an argument y :

$$\begin{aligned} S(Kx)(SII)y &= (Kxy)(SIIy) \\ &= x(SIIy) \\ &= x(Iy(Iy)) \\ &= x(yy) . \end{aligned}$$

In other words, R transforms y into two consecutive copies of itself before passing them into x . In full bracketed notation, the process reads

$$S[K[x]][S[S[K][K]][S[K][K]]][y] \rightarrow x[y[y]] .$$

The real depth of this construction emerges when we let R act on itself:

$$\begin{aligned} R(R) &= S(Kx)(SII)(S(Kx)(SII)) \\ &= x((SII)(S(Kx)(SII))) \\ &= x(SII(S(Kx)(SII))) \\ &= x(S(Kx)(SII)(S(Kx)(SII))) \\ &= x(R(R)) \end{aligned}$$

The final step shows that we have returned to the original form, but now wrapped once more in x . This establishes a *self-reproducing loop*: R applied to itself yields x applied to $R(R)$, which in turn expands in the same way, indefinitely.

By constructing R , we see that *SK*-calculus is capable not only of finite computation but also of infinite recursion and self-sustaining definitions. The pattern foreshadows the famous fixed-point combinators (such as Curry's Y combinator) that allow functions to reference themselves without explicit naming.

4 Functions

4.1 Functions in Mathematics

In mathematics, a *function* is a mapping from a set of *inputs* (the *domain*) to a set of *outputs* (the *range*), with the rule that each input is associated with exactly one output.

To define a function, one formally one writes:

$$\begin{aligned} f : X &\rightarrow Y \\ x &\mapsto f(x) \end{aligned}$$

Such a definition contains two pieces of information. By writing

$$f : X \rightarrow Y,$$

one speaks of ‘a function from X to Y ’, where X is the set of allowed inputs (domain), and Y is the set of allowed outputs (range). The statement

$$x \mapsto f(x)$$

means ‘from X , take an element x and send it to f to return $f(x)$ ’.

To summarize the definition of a mathematical function, we could say: Define f as a rule that assigns to each element x in X exactly one element $f(x)$ in Y .

A function is often defined by an explicit formula, a table, or an algorithm. Examples include:

$$\begin{aligned} f(x) &= x^2 \\ g(x) &= \sin(x) \end{aligned}$$

The mathematical function is *deterministic* - the same input always produces the same output - and has no notion of ‘memory’ or *state*. This purity is what later inspires *pure functions* in programming.

4.2 Functions in Programming

In programming, a function is a named (or anonymous) block of code that takes arguments and produces a result. The general form is:

$$f(\text{arguments}) \rightarrow \text{result}$$

Note the above is much less strict than the mathematical definition of a function. For this reason, programming functions may:

- Calculate a value from their inputs (*pure functions*)

- Perform actions such as printing to the screen or modifying data in memory (*impure functions*)
- Capture variables from surrounding scopes (*closures*)

Unlike mathematical functions, programming functions can have *side effects* and may depend on or change state.

Purity and State

A function is *pure* if:

1. Its output depends solely on its input arguments.
2. It has no observable side effects.

Pure functions are *referentially transparent*, meaning they can be replaced by their output without changing program behavior.

An *impure* function interacts with state - variables, files, networks, user interfaces - or produces effects beyond returning a value.

Side Effects

Side effects include:

- Modifying global or static variables
- Writing to files or databases
- Printing or logging
- Changing object properties

While often necessary, side effects reduce predictability and testability.

4.3 Functional Programming

Higher-Order Functions

A *higher-order function* is any function that takes another function as an argument, returns a function as its result, or both. They are the building blocks of functional programming, allowing control structures, data transformations, and abstractions to be expressed as reusable operations.

Examples in λ -calculus:

$$\begin{aligned} \text{map} &= \lambda f. \lambda x. s \dots \\ \text{compose} &= \lambda f. \lambda g. \lambda x. f(g\ x) \end{aligned}$$

Function Composition

Function composition combines two or more functions to produce a new one:

$$(h \circ g)(x) = h(g(x))$$

That is, if

$$g : X \rightarrow Y$$

$$h : Y \rightarrow Z$$

then

$$h \circ g : X \rightarrow Z.$$

Composition allows the creation of pipelines where data flows through transformations step by step.

Composition is *associative*, i.e.

$$f \circ (g \circ h) = (f \circ g) \circ h,$$

so grouping does not matter, only the order of application.

Currying

Currying transforms a multi-argument function into a chain of single-argument functions. In λ -calculus:

$$\lambda(x, y). f(x, y) \rightarrow \lambda x. \lambda y. f(x, y)$$

Named after logician Haskell Curry, currying makes partial application straightforward and supports *point-free* style programming.

For an example accessible to many programming languages, consider:

$$\text{add} = \lambda x. (\lambda y. x + y)$$

A curried version of this is built by the concatenation of `add` and one input argument, say 3, to make `add3`. The purpose of `add3` is to add 3 to the single argument sent to the function. We build such a function as follows:

$$\begin{aligned} \text{add3} &= (\lambda x. \lambda y. x + y) 3 \\ &\rightarrow_{\beta} \lambda y. 3 + y \end{aligned}$$

Then, the expression `add3(5)` returns 8.

Partial Application

Currying immediately implies the notion of *partial application*, which is the process of fixing some arguments of a function to create a new function with fewer arguments. We just witnessed the partial application

$$\text{add3} = \text{add}(3)$$

to produce a function that increments its argument by 3.

Point-Free Style

Point-free (or *tacit*) style expresses a function without explicitly mentioning its arguments.

Example:

$$\text{squaresum} = \text{sum} \circ \text{map}(\text{square})$$

instead of:

$$\text{squaresum } x \ s = \text{sum}(\text{map}(\text{square}, x \ s))$$

This style works naturally in a curried and compositional setting, but can reduce readability if overused.

Recursion

Functional programming languages without built-in named recursion can express it using *fixed-point combinators*. The most famous, the *Y* combinator, allows a function to call itself indirectly:

$$Y f = f(Y f)$$

In lazy languages like Haskell, *Y* can be used directly; in strict languages, a variant such as the *Z* combinator avoids immediate infinite application.

Laziness and Strictness

Lazy evaluation delays computation until a result is needed. *Strict evaluation* computes arguments before applying a function.

Many combinator patterns, especially those involving infinite data structures or fixed-point combinators, work naturally in lazy evaluation but require care in strict settings.

4.4 Combinators

In the broadest sense, a *combinator* is a higher-order function with no free variables - it is defined entirely in terms of its arguments and function application. In λ -calculus notation, every variable in a combinator's body is bound by a preceding λ .

Combinators are the building blocks of *combinatory logic*, a variable-free alternative to the λ -calculus.

General vs. Specialised Combinators

We can classify combinators into several overlapping categories:

1. **Basic combinators:** The simplest combinators that perform minimal transformations:

$$I = \lambda x. x \quad (\text{Identity})$$

$$K = \lambda x. \lambda y. x \quad (\text{Constant})$$

$$S = \lambda x. \lambda y. \lambda z. x z (y z) \quad (\text{Substitution})$$

Many combinatory systems take these as primitives.

2. **Derived combinators:** Any combinator that can be expressed in terms of the basic set. For example, the identity I can be expressed in SK calculus as:

$$I = S K K$$

Even complex recursion-supporting combinators can be reduced to S and K.

3. **Fixed-point combinators:** A combinator F is a *fixed-point combinator* if:

$$F f = f (F f)$$

This means F produces a fixed point of its argument function f . The most famous example is the **Y combinator**:

$$Y = \lambda f. (\lambda x. f (x x)) (\lambda x. f (x x))$$

Every fixed-point combinator is a combinator in the general sense, but not every combinator is a fixed-point combinator.

4. **Special-purpose combinators:** These are designed for specific behaviors, such as:

$$\Omega = (\lambda x. x x)(\lambda x. x x)$$

which loops indefinitely (non-terminating self-application). Ω is not a fixed-point combinator - it never produces a usable fixed point - but it is still a combinator.

5 Automata

5.1 Wolfram

5.2 Conway

Chapter 39

Logic Circuits

1 Introduction

It's impossible to exist today without having heard some person assert 'computers are all ones and zeros'. While cryptic and perhaps a bit reductionist, it *is* true, at a certain level, that all of digitized electronics can be reduced down to binary 'on' or 'off' conditions represented by ones and zeros.

Of course, opening an electronic device to reveal its inner entrails will hardly betray its operation, even while the device is running. The intention here is to examine exactly *what* is turning on or off, and how this could possibly be used to manage information and build complex systems.

1.1 Basic Circuit

For a starting point, let us contrive an unspecified but still 'simple' circuit using analog components as sketched in

Figure 39.1. The region labeled *Load* is where one would place any number of resistors, inductors, capacitors, lights, fans, heaters, and so on.

Solid black lines are considered pure conductors, or *wires*. Motivating the circuit is a direct current voltage source, such as a battery, labeled *DC Source*. Connected to the positive terminal on the voltage source is a toggle switch that can *open* (turn off) or *close* (turn on) the circuit. By convention, the *current* flow through the circuit originates from the 'high' (+) side of the voltage source, and flows back in through the 'low' (-) side. (The actual electrons go in reverse of the current.)

The total current entering the load is equal to the total current leaving the load, and this can be measured by placing an *Ammeter* in series with the load. The *Voltmeter* is a silent observer of the circuit with infinite resistance (ideally, at least), which is why it is always situated in parallel with the component(s) being measured.

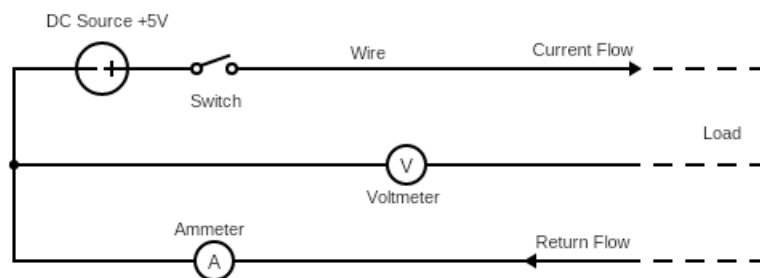


Figure 39.1. Basic circuit.

1.2 Digital Components

The load of an electric circuit may consist of analog components, digital components, or both. Digital components

called *chips* are made from clever configurations of silicon, carbon, and other common or uncommon materials. The end product tends to appear as small ‘black box’ depicted in Figure 39.2, which is a *74LS08 Quad 2-input AND Gate*.

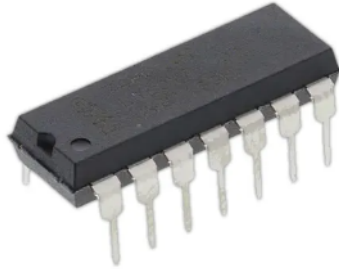


Figure 39.2. 74LS08 Quad 2-input AND Gate.

Making sense of the so-called 74LS08, or any chip for that matter, requires a map called a *pinout diagram* as shown in Figure 39.3.

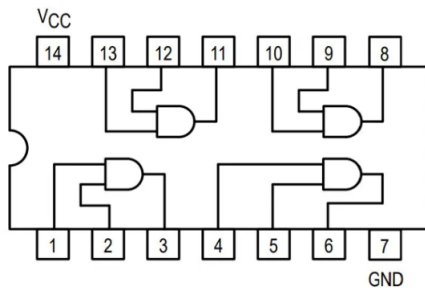


Figure 39.3. Pinout diagram of 74LS08 Quad 2-input AND Gate.

The pinout diagram states the purpose of each ‘leg’, or *pin* of the chip. For the example on hand, we see that pins 14 and 7 correspond to the high- and low voltage sources required for the chip’s operation.

As a matter of program note, we will work in the regime where positive five volts (+5 V) corresponds to an ‘on’ state, using the shorthand symbol 1. Conversely, the ‘off’ state symbolized by 0 corresponds to zero volts (0 V).

1.3 Digital Logic

AND Gate

Studying the pinout diagram of the 74LS08, we can see an embedding of four identical circuits, each having a wire leading from one peg to one location on a D-shaped glyph. This is the symbol for the AND gate, and there are four such gates on the chip shown. Taking the AND gate connected

to pins [1], [2], and [3], it follows that pins [1] and [2] correspond to the *inputs* of the gate. The third pin [3] corresponds to the *output* of the gate. A similar comment applies to pins [4]-[6], and so on for all four gates.

The behavior of the AND gate is specified as follows:

- When both inputs are 1, the output is 1.
- When both inputs are 0, the output is 0.
- When either input is 0, the output is 0.

The same information is represented efficiently in truth table form. The left columns are inputs, the right column is output:

AND		
0	0	0
0	1	0
1	0	0
1	1	1

OR Gate

Similar to the AND gate and implemented on a similar chip is the OR gate. Jumping right to the truth table, the OR gate behaves as follows:

OR		
0	0	0
0	1	1
1	0	1
1	1	1

That is, the OR gate returns 1 when either input is 1, and only returns 0 when both inputs are 0. Notice that the truth table for the OR gate is similar to that of the AND gate, as only the output column varies. In fact, the input pairs (0, 0), (0, 1), (1, 0), (1, 1) are the *only* combinations to consider when it comes to input.

Fourteen More Gates

Writing the output of the AND gate as a minimal list, and doing the same for the OR gate, we have

$$\begin{aligned}\text{AND} &= \{0, 0, 0, 1\} \\ \text{OR} &= \{0, 1, 1, 1\} .\end{aligned}$$

By inspecting the output side of each truth table, we see that the lists $\{0, 0, 0, 1\}$, $\{0, 1, 1, 1\}$ are just two members of *sixteen* possible lists, implying there are fourteen siblings

to the gates already listed. That is, there are sixteen possible logic gates that could represent the unknown Figure 39.4. Throughputs are labeled by letter rather instead of by number.

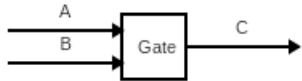


Figure 39.4. Generic logic gate.

All sixteen gates are represented in truth tables below. Certain logic gates are obligatory for completeness of the list but are otherwise trivial, for instance the NULL gate returns zero regardless of its inputs, whereas the A gate simply transmits whatever the first input is doing, and so on. The IMP gate (or any mentioning it) is a bit esoteric and isn't needed for now. There are four gates though, namely AND, OR, NOR, and XOR, that we'll want to keep handy.

NULL		
0	0	0
0	1	0
1	0	0
1	1	0

AND		
0	0	0
0	1	0
1	0	0
1	1	1

NIMP		
0	0	0
0	1	0
1	0	1
1	1	0

A		
0	0	0
0	1	0
1	0	1
1	1	1

| | | | | | | | | | | | |

-(B IMP A)		
0	0	0
0	1	1
1	0	0
1	1	0

B		
0	0	0
0	1	1
1	0	0
1	1	1

XOR		
0	0	0
0	1	1
1	0	1
1	1	0

OR		
0	0	0
0	1	1
1	0	1
1	1	1

| | | | | | | | | | | | |

NOR		
0	0	1
0	1	0
1	0	0
1	1	0

XNOR		
0	0	1
0	1	0
1	0	0
1	1	1

-B		
0	0	1
0	1	0
1	0	1
1	1	0

B IMP A		
0	0	1
0	1	0
1	0	1
1	1	1

| | | | | | | | | | | | |

-A		
0	0	1
0	1	1
1	0	0
1	1	0

IMP		
0	0	1
0	1	1
1	0	0
1	1	1

NAND		
0	0	1
0	1	1
1	0	1
1	1	0

ON		
0	0	1
0	1	1
1	0	1
1	1	1

The astute reader may notice that the output columns in the sixteen tables above express the numbers 0 to 15 in binary.

1.4 Inversion and Splitting

A different kind of component that partners with logic gates is the INV gate, equivalently called a NOT gate, also known as the *inverter*. The INV gate takes a single input and returns a single output. If the input is 1, return 0. If the input is 0, return 1. The inverter has a simple truth table:

INV	
0	1
1	0

Finally, there are instances when it's necessary to 'peel apart' the voltage or current from one wire by transforming into two wires, and this is called *splitting* or *dividing*. Generically, a so-called split can be thought of as a generic logic gate in reverse as shown in Figure 39.5.

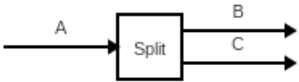


Figure 39.5. Generic divider.

For our purposes, the generic divider need only ‘clone’ the voltage from input *A* and assign the same value to the outputs *B* and *C*. That is, the inside of the ‘Split’ box is simply a junction of three wires.

1.5 Symbols

Logic and Flow

As seen with the AND gate, it turns out that all circuit components already have designated symbols. Several of

these were shown back in Figure 39.1, namely the symbols for a voltage source and various meters.

An alphabet unto their own, logic symbols and related items can be interlaced using wire to spell out intricate circuits, the complexity of which would be extremely difficult to grasp without the aid of symbolic representation. We’ll be making heavy use of six components depicted in Figure 39.6.

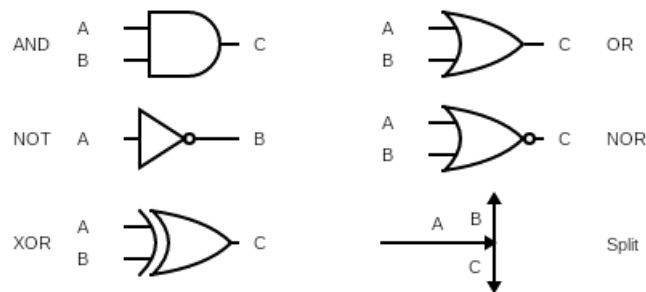


Figure 39.6. Symbols: AND, OR, INV, NOR, XOR, Split

Buttons and Lights

The inevitable concern of inputting information to a circuit and reading information from a circuit is answered by buttons and lights, respectively. Shown in Figure 39.7 are the symbols for the push switch, toggle switch, and light-emitting diode (LED). These are minimalist representations of, for instance, a keyboard and a screen.

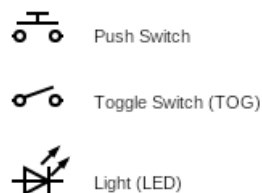


Figure 39.7. Symbols: Push, Toggle, LED

1.6 Logical Equivalencies

One reason for not needing *all* logic gates simultaneously is that some gates can be built from others.

Ruling out XNOR

For a swift example, the output of the XNOR gate is characterized by

$$\text{XNOR} = \{1, 0, 0, 1\},$$

which as the name XNOR may suggest, is the inversion of the XOR gate:

$$\text{XOR} = \{0, 1, 1, 0\}$$

To simulate the output XNOR, one simply need install the XOR gate and send the output through a INV gate. For this reason, it suffices to exclude a XNOR chip from the toolkit altogether, so long as there is sufficient stock of INV and XOR chips.

Another way to replicate a XNOR gate, that does not rely on XOR, is the arrangement of INV, OR, and AND gates depicted in Figure 39.8

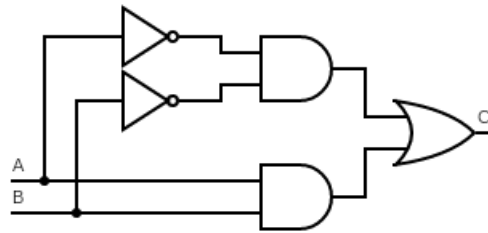


Figure 39.8. Equivalent XNOR gate.

OR Substitute

A case that actually saves skin in the laboratory is one that builds an OR gate from a single AND gate and three INV gates as depicted in Figure 39.9.

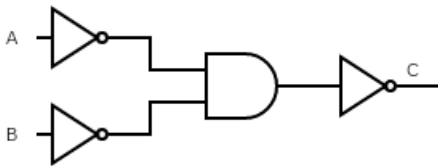


Figure 39.9. Poor man's OR gate.

AND Substitute

It turns out that an AND gate can be made from three NOR gates as depicted in Figure 39.10. Note too that the NOR gate is the linear composition of INV and OR.

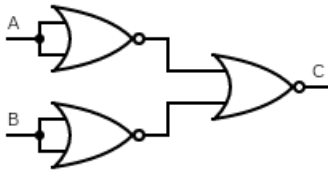


Figure 39.10. Pedant's AND gate.

1.7 Complete and Minimal Sets

The phenomenon of logic gate equivalencies implies a deeper pattern among logic gates than meets the eye. One question that arises, which happens to have multiple answers, concerns finding subsets of all logic gates that can reconstruct the entire table. Even more interestingly, is there a *single* logic that can do the job?

Skipping the stressful derivations, the answers to these questions go as follows:

- AND, INV form a complete set.

- OR, INV form a complete set.

- NAND forms a minimal set.

- NOR forms a minimal set.

Surprisingly, there are in fact two gates, NAND and NOR, that can alone be used (with copies of themselves) to reconstruct all other gates, even the INV gate. Figure 39.11 demonstrates this while introducing the NAND symbol.

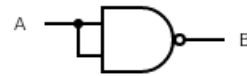


Figure 39.11. Using NAND to make INV.

The full derivation of all logic gates from NAND alone or NOR alone is left as a research project for the reader.

1.8 Feedback

While plenty can be accomplished using logic gates in simple series and parallel configurations, a whole new world opens when we make use of *feedback*, which means using the output of a logic gate to influence its own input.

SR Latch

An important and nontrivial example of a structure that utilizes feedback is the set-reset latch, or SR latch. The latch is composed of two NOR gates mutually arranged as depicted in Figure 39.12. One input leading to each NOR gate comes in as 'normal' input, whereas the second input on each gate is tied to the output of its sibling.

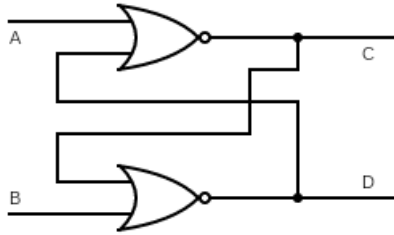


Figure 39.12. Set-Reset latch utilizing feedback.

As it turns out, the SR latch highlights a profound ability of logic circuits that one wouldn't think was possible. Though, working out the truth table for feedback-driven circuits is sadly quite a chore and takes some getting used to.

1.9 Bit Storage

Starting with the SR latch depicted in Figure 39.12, label the inputs using A , B , and also the two outputs C , D . To be definite, associate A , C with the top half, and B , D with the bottom half.

Going with a brute force analysis, we first consider all possible states, even the wrong ones, that could possibly take place in a circuit with two inputs and two outputs. There are $4^2 = 16$ of these in total, and each can be stored in a state vector

$$S_j = \{A_j, B_j, C_j, D_j\},$$

where j ranges from 0 to 15. Now the game becomes eliminating states that could not work for the SR latch on hand.

To dig in, consider the state $S_0 = \{0, 0, 0, 0\}$, which has all inputs and outputs set to 'off'. Consulting the NOR truth table, we find this state to be contradictory, as the output of each NOR gate should be 'on' when both inputs are 'off'. Thus we immediately dismiss the state S_0 .

We can similarly analyze $S_8 = \{1, 0, 0, 0\}$, which has one input set to 'on' with everything else 'off'. This is also a contradictory situation, as the nor gate being fed two 'off' signals results with 'on'. We can by symmetry dismiss $S_4 = \{0, 1, 0, 0\}$. The states $S_5 = \{0, 1, 0, 1\}$ and $S_{10} = \{1, 0, 1, 0\}$ are also dispensable for containing internal contradictions.

A family of states can be knocked away by asking how it's possible for the two outputs be 'on'. This only happens when all inputs are 'off', but outputs are touching inputs,

and another contradiction is reached. It follows that any state of the form $\{A, B, 1, 1\}$ can be excluded, ruling out S_3 , S_7 , S_{11} , S_{15} as viable states.

Setting both inputs to 'on' leads to undefined behavior. The states $S_{13} = \{1, 1, 0, 1\}$ and $S_{14} = \{1, 1, 1, 0\}$ are easily shown to be unattainable. The state $S_{12} = \{1, 1, 0, 0\}$ is not forbidden, but like a coin landing on its side, is an undefined state.

After all that, there are four surviving states allowed for the SR latch:

$$S_1 = \{0, 0, 0, 1\} = \text{reset}$$

$$S_2 = \{0, 0, 1, 0\} = \text{set}$$

$$S_6 = \{0, 1, 1, 0\} = \text{set}$$

$$S_9 = \{1, 0, 0, 1\} = \text{reset}$$

SR Latch Analysis

Listing all possible SR latch states S_1 , S_2 , S_6 , S_9 together, it stands out that the 'rest' state of the circuit, i.e. that which is no inputs coming in, corresponds to *two* different outputs. Left untouched, the circuit could be in either state S_1 or S_2 . For this reason we see how the name 'latch' applies. The purpose of the SR latch is to 'remember' one bit of data.

The state of the SR latch can be toggled by setting one of the inputs to 1 (and optionally back to 0). Looking at states S_6 and S_9 , we see the so-called 'set-reset' operation will toggle the output of the gate 'diagonal' from the input being toggled. This means the output state cannot be toggled by rapidly changing one input. Both inputs must be used to fully toggle the circuit.

By convention, the symbols A , B , C , D are replaced by R , S , Q , and Q' , respectively. The 'set' state applies to $Q = 1$, $Q' = 0$, and the 'reset' state is the opposite of this. In summary, the entirety of the SR latch is represented in the following table:

SR Latch				
R	S	Q	Q'	
0	0	0	1	reset
0	0	1	0	set
0	1	1	0	set
1	0	0	1	reset
1	1	0	0	undefined

2 Simulating Logic Circuits

Electronic devices are widespread and inexpensive thanks to the silicon/digital revolution, and this has led to the gradual buildup of magnificent systems like modern computers.

In this study, we shall walk the path backwards by starting with a computer and writing a program to simulate what happens in simple logic circuits. The end product is to be a kind of ‘virtual breadboard’ designed to capture all of the phenomena described previously, including feedback and data storage. (What we are *not* shooting for is a full-blown simulator with alternating currents, inductance, and all of the detritus of electrodynamics.)

In real life, electric circuits are a rather bare-bones implementation of the laws of electromagnetism and quantum mechanics, and exactly ‘where’ and ‘when’ things occur in the circuit is a matter of the whims of nature. By contrast, our tool of choice is a *procedural programming language*, which is executed linearly and thereby unable to capture the immense simultaneity of a real electric circuit. The best we can do is model the circuit using some kind of data structure and run algorithms on the data to simulate the passage of time. By choosing the right structures and rules, we can capture the circuit’s behavior.

With these considerations, the program we design must be able to:

- (i) load the specification of a circuit made from components and wires.

- (ii) receive real-time user input by simulating switches and buttons on the keyboard.
- (iii) employ an evaluation algorithm that properly updates the circuit state.
- (iv) display all outputs (including a summary of inputs) and the state of the circuit to the screen.

2.1 Labeling Convention

To make some headway on the problem, it’s required to have a sufficient system for labeling the components and wires in a circuit, along with all inputs and outputs. To this end we shall use numbers as labels, thus any number used to identify a component or a wire is used solely for labeling and has no numerical purpose. Formally, such a number is called the *identity*.

For our purposes, it suffices to set up a system with the following qualities:

- A number (in square brackets) [1] to [999] identifies a single component.
- A number (not in square brackets) above 1000 identifies a wire.
- Inputs are labeled in1, in2, in3, etc.
- Outputs are labeled out1, out2, out3, etc.

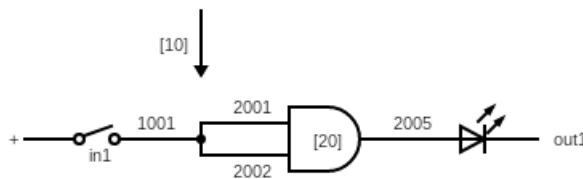


Figure 39.13. Example on/off light switch using labeling convention.

Figure 39.13 demonstrates the labeling convention to depict a simple yet over-engineered on/off switch for a single light. Importantly, notice that the junction of three wires (i.e. Split) is considered a component and has been assigned the label [10]. (The vertical arrow accompanying the [10] is to guide the eye, there is not a stray current in the figure.)

2.2 Data Structures

Begin by defining a generic data type called `Wire` with the following structure:

Type `Wire`:

```
Identity // 1000 to X
Pointer  // [1] to [999]
Value    // 0 or 1
```

The `Identity` is the wire’s own identity (1000-X). The `Pointer` field contains the identity of the downstream component touching the wire. The `value` field records the instantaneous voltage in the wire, just 1 or 0 for our purposes. Note there is not a field to store the wire’s upstream connection.

Next, consider another generic type called `Element` that

contains (i) the specification of a single component, and (ii) all wires touching the component.

```
Type Element:
  Species
  Identity // [1] to [999]
  A As Wire
  B As Wire
  C As Wire
```

The **Species** field stores a text abbreviation of the component being represented. Previously we decided there should be six of these:

```
Species = {AND, INV, NOR, OR, SPL, XOR}
```

For most logic gates, the wires *A*, *B* are inputs and *C* is the output. The INV gate treats *A* as the only input and ignores wire *C*. The Split junction treats *A* as the input and *B*, *C* as output.

Note that the **Element** type does not have an extra field to store the state of the component being represented. The output wire (*C* and/or *B*) does this job.

2.3 Buttons and Lights

More structure is required to handle interaction with the circuit via buttons and lights. For this we need just one generic data type called **Interact** as follows:

```
Type Interact:
  Species
  Identity // 1000 to X
  Pointer // [1] to [999]
  Value // 0 or 1
```

The **Species** field defines which device is being represented. There are three kinds of buttons:

1. PSH: stays on when constantly depressed, off when released
2. TOG: traditional toggle switch
3. PLS: auto-pulse

Figure 39.7 depicts the symbolic difference between the PSH button and the TOG button. Meanwhile, there is only one kind of light, designated LED for *light-emitting diode*. All together, there are four possibilities for the **Species** field:

```
Species = {PSH, TOG, PLS, LED}
```

The **Identity** field applies to the wire used to connect the device to its associated component, and the **Pointer** field is that component's address. The **Value** tracks the state of the interacting device. For buttons, 1 means 'pressed'. For lights, 1 means 'on'.

2.4 Circuit as Data

Using the data types detailed above, three global arrays are used to prepare space for circuit having up to 999 components, 9 buttons, and 9 lights. The total number of wires, i.e. the upper limit in 1000 to *X*, is defined to (far) exceed three times the number of components.

```
Shared Component(1 To 999) As Element
Shared Button(1 To 9) As Interact
Shared Light(1 To 9) As Interact
```

There must be a way to define and label components, buttons and lights, along with all wires connecting these. To do this, circuits are built from records containing six pieces of information:

```
NewComponent ->
  Species, Identity,
  Wire A ID, Wire B ID, Wire C ID,
  Comp B Ptr, Comp C Ptr
```

Consistent with the **Element** type, a new component is assigned a species and an identity, along with the identity of any wire touching its terminals.

The final two 'pointer' fields contain the address of any downstream component that is expected to receive information from the component being defined. This redundancy is worth the investment in the sense that the **Wire** type does not need to specify an upstream component address. When there is no downstream component, that pointer is set to -1.

Buttons and lights require less detail in their respective definitions. Following the **Interact** type, we have:

```
NewButton ->
  Species, Identity, Pointer
```

```
NewLight ->
  Species, Identity, Pointer
```

On/Off Light Switch

Using the structures above, an entire circuit and all aspects of its diagram can be encoded as data. It takes little to imagine this same scheme can support any combination of components.

Returning to the over-engineered on/off light switch first depicted in Figure 39.13, it turns out that the same information, including all labels, is contained in four total records:

Components:

```
("SPL", 10, 1001, 2001, 2002, 20, 20)
```

```
("AND", 20, 2001, 2002, 2005, -1, -1)
```

Interacts:

```
(1, "TOG", 1001, 10)
```

```
(1, "LED", 2005, 20)
```

SR Latch

A fully-labeled SR latch circuit, including two inputs and two outputs, is detailed in Figure 39.14.

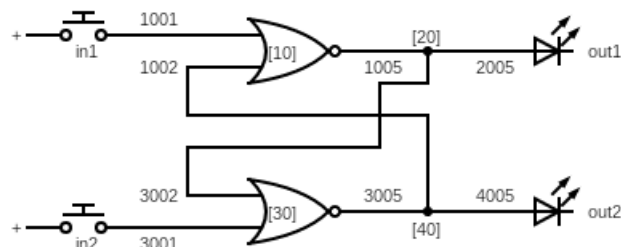


Figure 39.14. Example SR latch circuit diagram using labeling convention.

Interpreting the SR latch circuit diagram as records of data, we may jot down the following (abbreviated) form having four components and four interfaces:

Components:

```
("NOR", 10, 1001, 1002, 1005, -1, 20)
```

```
("NOR", 30, 3001, 3002, 3005, -1, 40)
```

```
("SPL", 20, 1005, 3002, 2005, 30, -1)
```

```
("SPL", 40, 3005, 1002, 4005, 10, -1)
```

Interacts:

```
(1, "PSH", 1001, 10)
```

```
(2, "PSH", 3001, 30)
```

```
(1, "LED", 2005, 20)
```

```
(2, "LED", 4005, 40)
```

2.5 Circuit State

The *state* of the circuit is defined as the ensemble of all **Value** fields among all wires. Note that components themselves not carry state per se, as the **Element** type intentionally lacks a **Value** field. Said another way, the state concerns only the voltages among wires in a circuit, and whatever is happening *inside* a component is irrelevant.

Ground State and Stability

The *ground* state of a circuit occurs when all voltages are off, i.e. the **Value** field for each wire stores 0. As we'll see,

some can circuits remain in the ground state when 'powered on'. Other circuits cannot tolerate a static ground state, and change to a stable state as quickly as possible.

Thinking back to the analysis of the SR latch, we now say 'instability of the ground state' is responsible for the state $\{0, 0, 0, 0\}$ being forbidden.

State Record

While the computer has its own way of remembering the state of a circuit, it is still instructive to have a text readout of the circuit's evolution. For this, define a 'state record' of the format

$$S = \{ \dots [Y] ABC \dots \} ,$$

where $[Y]$ is the identity of a given component, and then the triplet ABC lists the state of that component's A-wire, B-wire, etc.

The complete state record contains a $[Y] ABC$ -like entry for each component. There will be inevitable redundancies in the record, as any one wire often touches two components. The *C*-variable is vestigial and thus set to 0 in the case of INV.

Intermediate State

Continuing the example of the over-engineered on/off light switch and using the labeling scheme from Figure 39.13,

the ground state of the circuit can be represented as:

$$S_{off} = \{[10] 000 [20] 000\}$$

The first component [10] represents the junction of three wires joining the switch in1 to both inputs of the downstream AND gate, component [20].

Closing the switch to turn the light on, it takes little to imagine the final state replaces all 0-bits with 1-bits, as in

$$S_{on} = \{[10] 111 [20] 111\} ,$$

which is surely true, but it's worth going through how that happens despite not yet having detailed the state-updating algorithm.

Immediately after closing the switch but before stability is reached, the circuit (as defined) jumps to the intermediate state

$$S_{med} = \{[10] 100 [20] 000\} ,$$

which means the input to the junction [10] has acknowledged a 'high' signal. Recursing on the intermediate state leads to the on-state after one step.

As it turns out, the intermediate state actually depends on the order in which components are defined in memory, thus the intermediate state is more of a ghost than anything else.

Indeed, if we take the same over-engineered on/off light switch and swap address [10] with address [20], the physically-equivalent circuit is defined:

Components:

("SPL", 20, 1001, 2001, 2002, 10, 10)

("AND", 10, 2001, 2002, 2005, -1, -1)

Interacts:

(1, "TOG", 1001, 20)

(1, "LED", 2005, 10)

Closing the switch on this version, the intermediate state

$$S_{med} = \{[10] 000 [20] 100\}$$

arises as one may expect. Unlike the previous case though, the 'on' state does not come next. Instead, a secondary intermediate state gets in the way

$$S_{sec} = \{[10] 110 [20] 111\} ,$$

from which 'on' follows. Clearly then, the number and order of intermediate states depends on how the circuit is registered in memory.

Variability of the intermediate state is directly analogous to effects in real-life circuits that arise from mixing wire

lengths, component brands, and so on. Unless the technician makes a significant mistake, intermediate states are often imperceptible and require tools more specialized than the ammeter and voltmeter to detect.

Equilibrium

One way to summarize intermediate state analysis is to say that *equilibrium* corresponds to the final state being reached after all intermediate states have been churned through.

2.6 Main Loop

When it comes to actually executing the simulated circuit, it's necessary to have a real-time algorithm, sometimes called a main loop or *game loop*, that will:

- (i) scan for user input (including auto-pulse).
- (ii) update the internal state.
- (iii) display output and summary.

For the simulated to 'feel real' and achieve equilibrium quickly in response to input, the entirety of the game loop should refresh minimally at 30Hz, i.e. 30 frames per second.

2.7 Managing Input

During the game loop, an input-polling routine must scan for keystrokes and inform the circuit whether a button has been pressed.

PSH

For buttons of the PSH species, the following pseudo-code does the job:

```

Let i = index of i'th button (1 to 9)
If key(i) pressed {q = 1}
    Else {q = 0}
r = Button(i).Identity // 1000 to X
j = Button(i).Pointer // [1] to [999]
If (r = Component(j).A.Identity) {
    Component(j).A.Value = q}
ElseIf (r = Component(j).B.Identity) {
    Component(j).B.Value = q}
ElseIf (r = Component(j).C.Identity) {
    Component(j).C.Value = q}

```

Note that the input-polling routine changes the circuit state by editing the voltage value in the wire identified by *r* that connects the button to the component at address *j*. This behavior is applied to each button type.

TOG

The TOG case is slightly more complicated:

```

Let i = index of i'th button (1 to 9)
If key(i) pressed:
  r = Button(i).Identity // 1000 to X
  j = Button(i).Pointer // [1] to [999]
  If (r = Component(j).A.Identity) {
    q = Component(j).A.Value
    If (q = 0) {Component(j).A.Value = 1}
    Else {Component(j).A.Value = 0}
  }
  ElseIf (r = Component(j).B.Identity) {
    q = Component(j).B.Value
    If (q = 0) {Component(j).B.Value = 1}
    Else {Component(j).B.Value = 0}
  }
  ElseIf (r = Component(j).C.Identity) {
    q = Component(j).C.Value
    If (q = 0) {
      Component(j).C.Value = 1}
    Else {Component(j).C.Value = 0}
  }

```

Like its sibling, the TOG routine changes the circuit state.

PLS

The PLS button behaves automatically without requiring user interaction. In real time, the component connected to PLS will receive 2/3 seconds of 'high', and then 1/3 seconds of 'low', repeating forever. The overall frequency is 1 Hz. As pseudo-code, this looks like:

```

Let i = index of i'th button (1 to 9)
If key(i) pressed:
  t = Timer (to ~ms precision)
  t = t - Int(t):
  If (t >= 2/3) {q = 1}
  Else {q = 0}
  r = Button(i).Identity // 1000 to X
  j = Button(i).Pointer // [1] to [999]
  If (r = Component(j).A.Identity) {
    Component(j).A.Value = q}
  ElseIf (r = Component(j).B.Identity) {
    Component(j).B.Value = q}
  ElseIf (r = Component(j).C.Identity) {
    Component(j).C.Value = q}

```

2.8 Updating Internal State

As any input of the circuit is manipulated, the voltage value is immediately changed in the wire leading to the affected

component. Propagation stops there momentarily, leaving a certain 'tension' on the components touching active inputs. Control then passes to the subroutine that updates internal state - the true heart of the simulation that we detail now.

Binary Cases

For each binary logic gate k in the circuit, internal state is updated in two steps:

- (i) Leading into k , the values from input wires A and B are compared using the proper truth table for that component's species. The resulting 1 or 0 is written to k 's output wire C .
- (ii) Downstream from k is a different component j whose input A' or B' must be overwritten to reflect the change in C .

Step (i) can be implemented in a fashion captured in the following pseudo-code:

```

a = Component(k).A.Value
b = Component(k).B.Value
Switch Component(k).Species:
Case "AND"
  Component(k).C.Value = BinaryAND(a, b)
Case "NOR"
  Component(k).C.Value = BinaryNOR(a, b)
Case "OR"
  Component(k).C.Value = BinaryOR(a, b)
Case "XOR"
  Component(k).C.Value = BinaryXOR(a, b)

```

Step (ii) proceeds as follows:

```

j = Component(k).C.Pointer
If (j <> -1) {
  g = Component(k).C.Identity
  c = Component(k).C.Value
  If (Component(j).A.Identity = g) {
    Component(j).A.Value = c}
  ElseIf (Component(j).B.Identity = g) {
    Component(j).B.Value = c}
}

```

Special Cases

There are two exceptions to the above, namely the INV gate and the Split (junction). Each obeys the same two-step process, however special treatment is needed for A , B , and C if used. Jumping right to pseud-code, we have, for the INV-case:

```

a = Component(k).A.Value
// step (i)
Component(k).B.Value = BinaryINV(a)
// step (ii)
j = Component(k).B.Pointer
If (j <> -1) {
    g = Component(k).B.Identity
    b = Component(k).B.Value
    If (Component(j).A.Identity = g) {
        Component(j).A.Value = b}
    ElseIf (Component(j).B.Identity = g) {
        Component(j).B.Value = b}}

```

Finally, the Split-case needs to update two downstream wires and needs its own:

```

a = Component(k).A.Value
// step (i)
Component(k).B.Value = a
Component(k).C.Value = a
// step (ii)
j = Component(k).B.Pointer
If (j <> -1) {

```

```

    g = Component(k).B.Identity
    b = Component(k).B.Value
    If (Component(j).A.Identity = g) {
        Component(j).A.Value = b}
    ElseIf (Component(j).B.Identity = g) {
        Component(j).B.Value = b}}
j = Component(k).C.Pointer
If (j <> -1) {
    g = Component(k).C.Identity
    c = Component(k).C.Value
    If (Component(j).A.Identity = g) {
        Component(j).A.Value = c}
    ElseIf (Component(j).B.Identity = g) {
        Component(j).B.Value = c}}

```

2.9 Output and Summary

With all steps thoroughly detailed, it's straightforward to write source code in a procedural programming language that simulates logic circuit behavior. Using *QB64* as a weapon of choice, the efforts of this study culminate to a single program that executes in a single window as demonstrated in Figure 39.15.

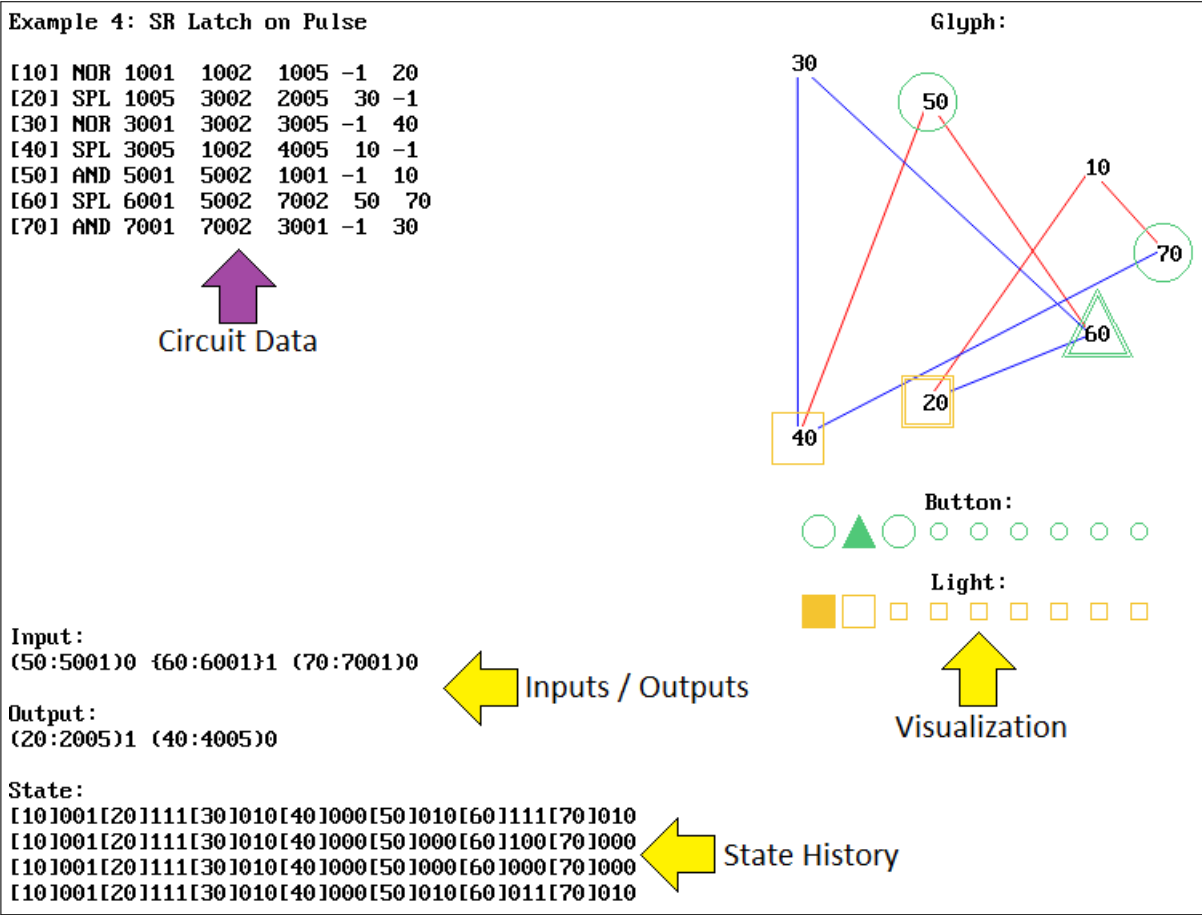


Figure 39.15. Sample output window showing a simulated logic circuit.

The output window has four distinct regions, one that never changes (purple arrow), and three that receive updates often (yellow arrows).

Circuit Data

In the top-left, the precise layout of the circuit is summarized in a way resembling how the circuit is defined by the programmer. All relevant relationships between components and wires can be read from the summary.

Text Readout

The bottom-left region contains a text readout of all active inputs (buttons) and outputs (lights), followed by the state of the circuit and its history.

Glyph

The right side of the Figure is for real-time visualization. In the top-right we place the so-called *glyph*, which is in essence a *bad* circuit diagram not so much unlike the real wires and components in a real circuit. Nonetheless, the glyph shows the specific relationships among components, and also indicates the voltage throughout.

Buttons and Lights

Under the glyph is a row of button and light indicators, and this is the chief output section of the circuit. On a breadboard, the lights would be a row of LEDs. All colors are decidedly yellow, with a filled square () denoting the 'on' state.

The keyboard numerals 1 to 9 are mapped to corresponding buttons in1 to in9. For the example shown, only keys 1 and 3 are of the PSH species as indicated by a circle

(). Button 2, with its triangular indicator (), is of the PLS species and self-operates on a timer. Buttons 4 to 9 are unused in the example, and are displayed as diminished placeholders. Not shown (yet) are buttons of the TOG species, which appear as squares () when used.

3 Switching and Latching

Much ado has been made of the SR latch, which is arguably the simplest nontrivial circuit that can store either 1 or 0. From there, it takes little to imagine that the latching phenomenon can be stacked to encode complex data structures, record storage, lists of instructions, etc. - in other words, modern computing. Starting simple though, we begin here

with a few modifications to the original SR latch.

3.1 SR Latch with Enable

Starting with the original SR latch circuit, suppose the input is piped through a pair of AND gates as shown in Figure 39.16. In addition, the unused inputs of each AND gate are joined and controlled by an additional PSH button.

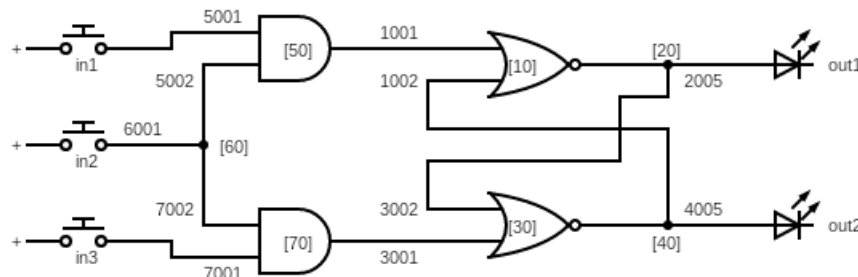


Figure 39.16. Enabled SR latch circuit.

The purpose of the 'enable' feature is to verify the inputs entering the circuit. That is, nothing happens unless the PSH button in2 is depressed, thereby enabling the latch to change state.

As for circuit data, we append the specification of the SR latch with three additional components (two ANDs plus a junction), along with the 'enable' button as shown:

Components:

```
("NOR", 10, 1001, 1002, 1005, -1, 20)
("SPL", 20, 1005, 3002, 2005, 30, -1)
("NOR", 30, 3001, 3002, 3005, -1, 40)
("SPL", 40, 3005, 1002, 4005, 10, -1)
("AND", 50, 5001, 5002, 1001, -1, 10)
("SPL", 60, 6001, 5002, 7002, 50, 70)
("AND", 70, 7001, 7002, 3001, -1, 30)
```

Interacts:

```
(1, "PSH", 5001, 50)
(2, "PSH", 6001, 60)
(3, "PSH", 7001, 70)
(1, "LED", 2005, 20)
(2, "LED", 4005, 40)
```

3.2 SR Latch on Pulse

The circuit represented in Figure 39.15 is another modified SR latch, called the 'pulsed' SR latch. This is identical to the enabled SR latch, except the manual PSH enable button

is replaced by an automatic PLS species. This amounts to the following edit to the *Interacts* section:

Interacts:

```
(2, "PLS", 6001, 60)
```

3.3 D Latch

An elegant modification to the enabled SR latch is the so-called D latch, which conjoins the two circuit inputs while inverting one of them as shown in Figure 39.17, or equivalently as data:

Components:

```
("NOR", 1, 1001, 1002, 1005, -1, 2)
("SPL", 2, 1005, 3002, 2005, 3, -1)
("NOR", 3, 3001, 3002, 3005, -1, 4)
("SPL", 4, 3005, 1002, 4005, 1, -1)
("AND", 5, 5001, 5002, 1001, -1, 1)
("SPL", 6, 6001, 5002, 7002, 5, 7)
("AND", 7, 7001, 7002, 3001, -1, 3)
("SPL", 8, 8001, 7001, 9001, 7, 9)
("INV", 9, 9001, 5001, 0, 5, -1)
```

Interacts:

```
(1, "PSH", 8001, 8)
(2, "PSH", 6001, 6)
(1, "LED", 2005, 2)
(2, "LED", 4005, 4)
```

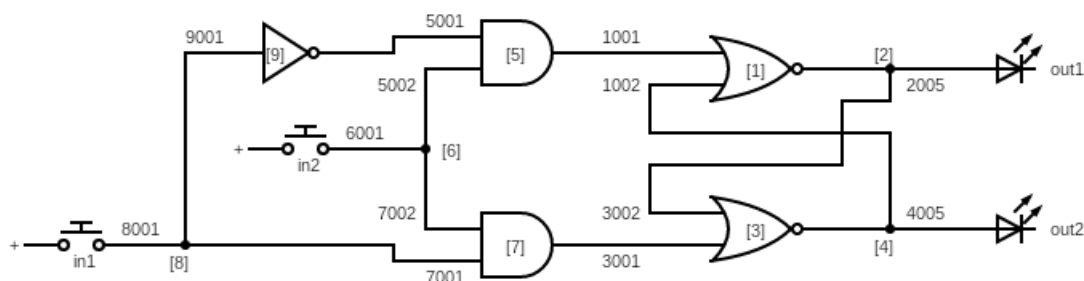


Figure 39.17. D latch circuit.

By convention, the outputs of the circuit labeled out1, out2 are known as Q and Q' , respectively. Because the output values are always in disagreement, the 'main' output is delegated to Q , whereas Q' is typically downplayed as Q 's opposite. The same convention says the input in1 is designated D (think D for 'data'), and the input in2 is designated E for 'enable'.

D Latch Analysis

Analysis of the D latch is slightly more complicated than that of the plain SR latch, but we can play the same state-listing game to understand its behavior. Begin by writing a generic state vector for the D latch, namely

$$S_j = \{D_j, E_j, Q_j, Q'_j\}.$$

Now, one could use a pen-and-paper method to deduce the allowed states of the D-latch. On the other hand, since we're in this to build a simulation, it should be just as good to construct the circuit and perform testing in software space. Regardless of how it is done, one should find, for allowed states:

$$S_1 = \{0, 0, 0, 1\}$$

$$S_2 = \{0, 0, 1, 0\}$$

$$S_5 = \{0, 1, 0, 1\}$$

$$S_9 = \{1, 0, 0, 1\}$$

$$S_{10} = \{1, 0, 1, 0\}$$

$$S_{14} = \{1, 1, 1, 0\}$$

To make sense of these, note that the pair S_1, S_2 begin the same but have the output states swapped. The same can be said for the pair S_9, S_{10} . All four of these states have the E -input set to 0, meaning the circuit is not enabled, thus the state of Q, Q' remain fixed regardless of D 's value.

The states S_5, S_{14} correspond to $E = 1$, thus the circuit is free to exhibit SR-like behavior. When enabled, setting $D = 1$ causes $Q = 1$, whereas $D = 0$ causes $Q = 0$. All of this is summarized in the following D latch truth table:

D Latch			
D	E	Q	Q'
0	0	Q	Q'
1	0	Q	Q'
0	1	0	1
1	1	1	0

Note finally that the state $S_0 = \{0, 0, 0, 0\}$ is not supported in any SR latch or D latch. When a latch-containing circuit is initially turned on from the ground state, latches quickly find their way out of S_0 .

4 Binary Addition

The task of adding two integers can be automated using logic circuits. Given the binary nature of logic gates, numbers involved in calculations addition must be translated to a base-two format before touching a physical circuit.

Starting with the simplest possible case, consider the addition problem

$$A + B = C,$$

where A and B can each be either of 0 or 1. This is a simple-enough problem to list all possible outcomes, which are:

$$\begin{aligned} 0 + 0 &= 00 \\ 0 + 1 &= 01 \\ 1 + 0 &= 01 \\ 1 + 1 &= 10 \end{aligned}$$

The output C is written in two-digit format, and the digits are named, for perhaps obvious reasons, *carry* and *sum*, respectively:

$$C = (\text{Carry}, \text{Sum})$$

4.1 Half Adder

The scenario indulged so far is called the *half adder*, and the task now is to capture its behavior in a circuit. Without

knowing the components just yet, we can nonetheless see that there should be two inputs and two outputs. Further, we can write a truth table for how the circuit ought to behave:

Half Adder			
A	B	Carry	Sum
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

From this, the right-hand columns in isolation read

$$\begin{aligned} \text{Carry} &= \{0, 0, 0, 1\} \\ \text{Sum} &= \{0, 1, 1, 0\}, \end{aligned}$$

which are identical to the output columns of the AND and XOR gates, respectively:

$$\begin{aligned} \text{AND} &= \{0, 0, 0, 1\} \\ \text{XOR} &= \{0, 1, 1, 0\}. \end{aligned}$$

Evidently the truth table of the circuit maps cleanly to our arsenal of components. Thus, let us propose the half adder circuit take the form depicted in Figure 39.18.

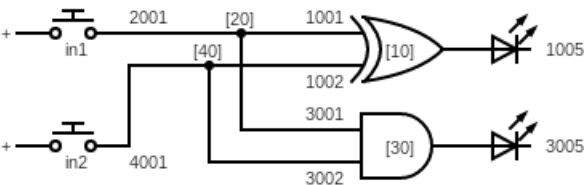


Figure 39.18. Half adder circuit.

As circuit data, the half adder is captured by:

Components:

```
("XOR", 10, 1001, 1002, 1005, -1, -1)
("SPL", 20, 2001, 1001, 3001, 10, 30)
("AND", 30, 3001, 3002, 3005, -1, -1)
("SPL", 40, 4001, 1002, 3002, 10, 30)
```

Interacts:

```
(1, "PSH", 2001, 20)
(2, "PSH", 4001, 40)
(1, "LED", 1005, 10)
(2, "LED", 3005, 30)
```

4.2 Full Adder

While the half adder is perfectly suited for adding a pair of one-bit numbers, it would be nice to perform addition on two-, three-, four-bit numbers and so on. From the outset, we want to do this *without* having to reason out each circuit from first principles.

To break down the task we speak of the *full adder*, which is a streamlining of the half adder suited for integration into larger circuits. In particular, we will see that addition of larger-bit numbers is accomplished by daisy-chaining identical copies of the full adder circuit in a particular arrangement.

To begin, note that the half adder never attains the output state $C = (1, 1)$, which is to say the ‘carry’ and ‘sum’ values are never simultaneously high. For such a state to be attainable, a third input, thus a third button is needed, and here is the beautiful part: the third input shall be the ‘carry’ value from a so-far unmentioned adder-like circuit that is upstream of the one being discussed. It’s purpose is to add 1 to the sum being calculated.

Without knowing which components to use or how they’re arranged, we can still work out the truth table for the full adder:

Full Adder				
in1	in2	Carry-In	Carry-Out	Sum
0	0	0	0	0
0	1	0	0	1
1	0	0	0	1
1	1	0	1	0
0	0	1	0	1
0	1	1	1	0
1	0	1	1	0
1	1	1	1	1

When the Carry-In (synonymous with in3) value is set ‘low’, the full adder does exactly what the half adder does.

With Carry-In set to ‘high’, we consult the bottom half of the truth table to jot down

$$\text{Carry-Out} = \{0, 1, 1, 1\}$$

$$\text{Sum} = \{1, 0, 0, 1\}.$$

Comparing these to the lexicon of logic gates, we find

$$\text{OR} = \{0, 1, 1, 1\}$$

$$\text{XNOR} = \{1, 0, 0, 1\},$$

which is a little troublesome due to the presence of XNOR.

To skirt the problem we exploit the XNOR-equivalent arrangement depicted in Figure 39.8. Ignoring the pair of INV gates momentarily, we see that the bulk of the XNOR-equivalent circuit consists of a pair of AND gates leading to an OR gate, and that is the motif we shall borrow for the full adder.

The exact marriage of the half adder and the Franken-XNOR is a bit tricky to arrange, but nonetheless the full adder circuit is finished in Figure 39.19. An extra NOR gate has been slipped into the circuit as a necessity, and in hindsight of this we see that the full adder really is a configuration of two half adders.

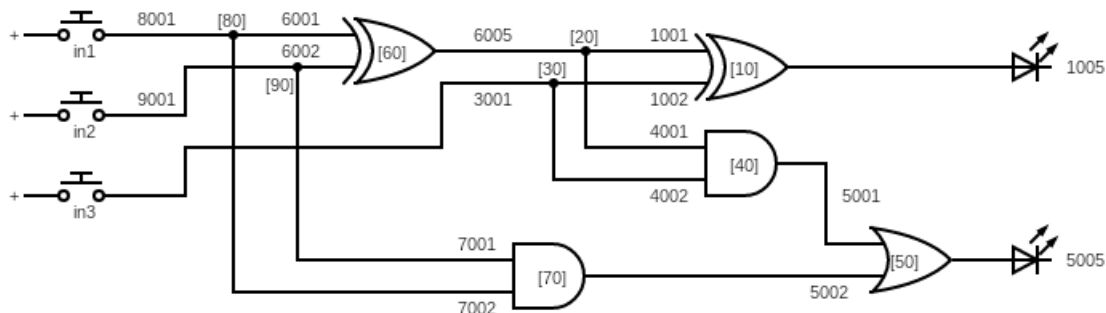


Figure 39.19. Full adder circuit.

Equivalently, the full adder circuit is specified by the following data:

Components:

```
("XOR", 10, 1001, 1002, 1005, -1, -1)
("SPL", 20, 6005, 1001, 4002, 10, 40)
("SPL", 30, 3001, 1002, 4001, 10, 40)
("AND", 40, 4001, 4002, 5001, -1, 50)
("OR", 50, 5001, 5002, 5005, 10, 30)
("XOR", 60, 6001, 6002, 6005, -1, 20)
("AND", 70, 7001, 7002, 5002, -1, 50)
("SPL", 80, 8001, 6001, 7002, 60, 70)
```

```
("SPL", 90, 9001, 6002, 7001, 60, 70)
```

Interacts:

```
(1, "PSH", 8001, 80) // in1
(2, "PSH", 9001, 90) // in2
(3, "PSH", 3001, 30) // carry-in
(1, "LED", 5005, 50) // carry-out
(2, "LED", 1005, 10) // sum
```

4.3 Two Bit Adder

Having so carefully prepared the full adder circuit to play nice among similar structures, the addition of higher-bit

numbers becomes a matter of linking a number of identical adder circuits.

The simplest application is the *two bit adder*, which as the name suggests, adds a pair of two-bit binary numbers. Such a circuit contains five total inputs: a pair for each

input, plus a carry-in bit. There are three outputs: one pair for a two-bit result, plus a carry-out bit. Letting the circuit diagram to do the rest of the talking, the two bit adder is depicted in Figure 39.20.

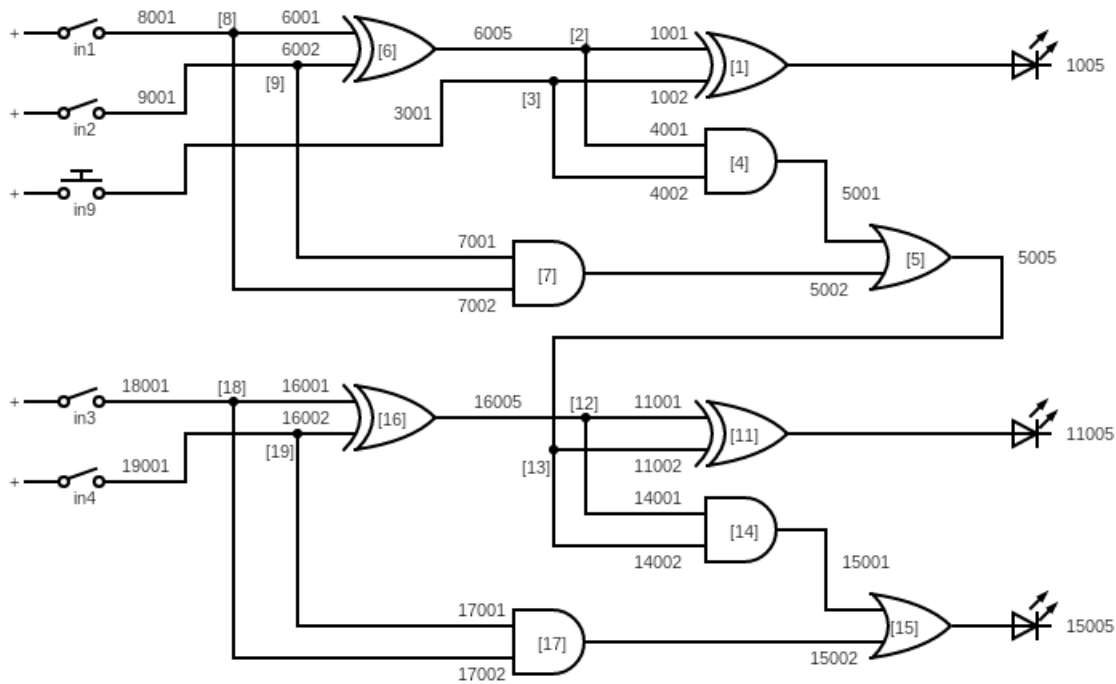


Figure 39.20. Two bit adder circuit.

Note that the four ‘regular’ inputs to the circuit are buttons of the TOG species, where as per usual, the carry-in bit is represented by a PSH button. The whole diagram and all of its labeling is equivalent to the circuit data:

Components:

```
("XOR", 1, 1001, 1002, 1005, -1, -1)
("SPL", 2, 6005, 1001, 4002, 1, 4)
("SPL", 3, 3001, 1002, 4001, 1, 4)
("AND", 4, 4001, 4002, 5001, -1, 5)
("OR", 5, 5001, 5002, 5005, -1, 13)
("XOR", 6, 6001, 6002, 6005, -1, 2)
("AND", 7, 7001, 7002, 5002, -1, 5)
("SPL", 8, 8001, 6001, 7002, 6, 7)
("SPL", 9, 9001, 6002, 7001, 6, 7)
("XOR", 11, 11001, 11002, 11005, -1, -1)
("SPL", 12, 16005, 11001, 14002, 11, 14)
("SPL", 13, 5005, 11002, 14001, 11, 14)
("AND", 14, 14001, 14002, 15001, -1, 15)
```

```
("OR", 15, 15001, 15002, 15005, -1, -1)
("XOR", 16, 16001, 16002, 16005, -1, 12)
("AND", 17, 17001, 17002, 15002, -1, 15)
("SPL", 18, 18001, 16001, 17002, 16, 17)
("SPL", 19, 19001, 16002, 17001, 16, 17)
```

Interacts:

```
(9, "PSH", 3001, 3) // carry in
(1, "TOG", 8001, 8) // a0
(2, "TOG", 9001, 9) // b0
(3, "TOG", 18001, 18) // a1
(4, "TOG", 19001, 19) // b1
(1, "LED", 15005, 15) // carry out
(2, "LED", 11005, 11) // sum1
(3, "LED", 1005, 1) // sum0
```

Two Bit Adder Analysis

Having five total inputs, there are $2^5 = 32$ distinct states of the two bit adder. The outputs are limited to $2^3 = 8$ possibil-

ities though, thus there is significant non-uniqueness in the result as expected for addition. For this reason, coming up with a complete truth table for such a circuit is a redundant thankless chore to be avoided.

Instead, it's quite easy to predict on paper how the circuit ought to appear. The sum handled by the two bit adder can be represented by

$$N_1 + N_2 + C_{in} = C_{out} + S,$$

where the left side is

$$N_1 = \{00, 01, 10, 11\}$$

$$N_2 = \{00, 01, 10, 11\}$$

$$C_{in} = \{0, 1\},$$

and the right is

$$C_{out} = \{0, 1\}$$

$$S = \{00, 01, 10, 11\}.$$

For an example, consider the case

$$N_1 = 10$$

$$N_2 = 01$$

$$C_{in} = 0,$$

which represents the addition problem

$$2 + 1 + 0 = C_{out} + S.$$

The right side is obviously 3, which comes in the form (0, 1, 1). That is, the carry-out bit is 0, and the sum contains 3 in binary, namely 11.

The same problem is represented in Figure 39.21 using buttons and lights. In particular, buttons 1, 3 correspond to N_1 , and buttons 2, 4 correspond to N_2 . The carry-in digit is delegated to the right as button 9.



Figure 39.21. The sum $1 + 2 + 0 = 3$.

Modifying the example, we can set the carry-in bit to 1 instead of 0, which kicks the output state to 4, represented by 100 in binary. This means the S -field contains 00, and the carry-out bit is set to 1 as depicted in Figure 39.22.



Figure 39.22. The sum $1 + 2 + 1 = 4$.

4.4 More Bit Adders

As demonstrated while building the two bit adder, adding yet more bits is a matter of linking yet more full adder circuits. As the adder circuits grow, precisely *how* the inputs correspond to the outputs needs further elaboration that is reserved for the four bit case below.

Three Bits

Without laboring any details and skipping to a result, Figure 39.23 displays a portion of the QB64 output window while running the three-bit adder.

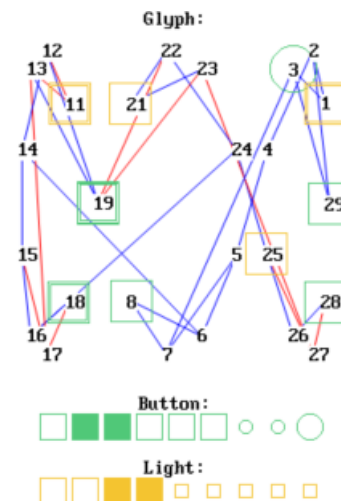


Figure 39.23. Partial output window showing a three bit adder circuit.

Four Bits

Figure 39.24 depicts the four bit adder circuit.

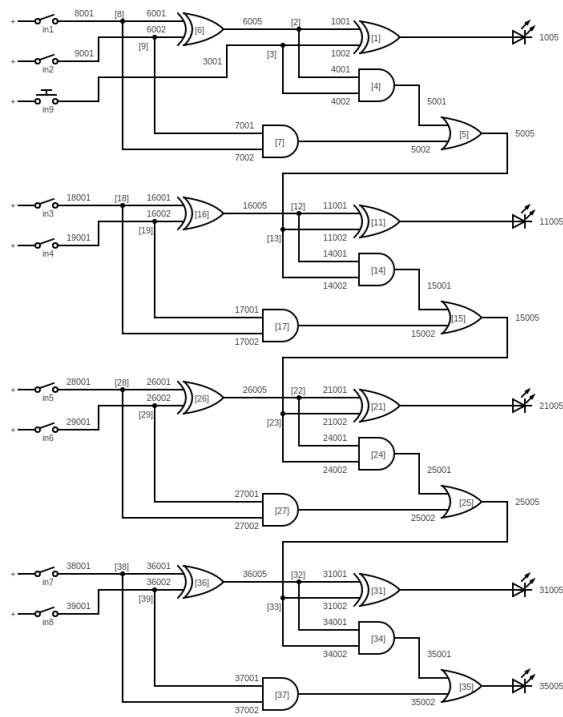


Figure 39.24. Four bit adder circuit.

Having nine total inputs, the four bit adder circuit uses the full range of available buttons. Like the two bit adder, we can begin analysis by writing

$$N_1 + N_2 + C_{in} = C_{out} + S,$$

where, now, the left side is

$$N_1 = \{0000, 0001, \dots, 1111\}$$

$$N_2 = \{0000, 0001, \dots, 1111\}$$

$$C_{in} = \{0, 1\},$$

and the right is

$$C_{out} = \{0, 1\}$$

$$S = \{0000, 0001, \dots, 1111\}.$$

Particularly, the numbers N_1 , N_2 , and S are each one of sixteen possibilities, i.e. the integers 0 to 15. The greatest result the four bit adder can produce is 31, which in binary is 11111.

With the carry digit reserved to button 9 as usual, the circuit receives input as follows: The digits of N_1 are inputted using the keys 1, 3, 5, 7, and the digits of N_2 are inputted using the keys 2, 4, 6, 8. Keeping in mind the context is base two, this further means:

- Buttons 1, 2 are the **ones** column.
- Buttons 3, 4 are the **tens** column.
- Buttons 5, 6 are the **hundreds** column.
- Buttons 7, 8 are the **thousands** column.

For example, suppose we want to compute the sum

$$9 + 13 + 0 = 22.$$

Breaking this down, we have

$$N_1 = 1001$$

$$N_2 = 1101$$

$$C_{in} = 0$$

Since we know the answer is 22, we also have that

$$C_{out} + S = 10110,$$

where the right side is 22 in binary.

As far as circuit input goes, we consult the expressions of N_1 , N_2 , and, reading right to left, discern that we press buttons 1, 2, 6, 7, 8 while leaving the rest alone. By doing this, the lights that go 'high' should be 1, 3, 4.

Figure 39.25. Partial output window demonstrating the sum $9 + 13 = 22$.

Containing the same information depicted in Figure 39.24, the four bit adder circuit data is listed below for completeness:

Components:

```
( "XOR", 1, 1001, 1002, 1005, -1, -1)
( "SPL", 2, 6005, 1001, 4002, 1, 4)
( "SPL", 3, 3001, 1002, 4001, 1, 4)
( "AND", 4, 4001, 4002, 5001, -1, 5)
( "OR", 5, 5001, 5002, 5005, -1, 13)
( "XOR", 6, 6001, 6002, 6005, -1, 2)
( "AND", 7, 7001, 7002, 5002, -1, 5)
( "SPL", 8, 8001, 6001, 7002, 6, 7)
( "SPL", 9, 9001, 6002, 7001, 6, 7)
( "XOR", 11, 11001, 11002, 11005, -1, -1)
( "SPL", 12, 16005, 11001, 14002, 11, 14)
( "SPL", 13, 5005, 11002, 14001, 11, 14)
```

```

("AND", 14, 14001, 14002, 15001, -1, 15)
("OR", 15, 15001, 15002, 15005, -1, 23)
("XOR", 16, 16001, 16002, 16005, -1, 12)
("AND", 17, 17001, 17002, 15002, -1, 15)
("SPL", 18, 18001, 16001, 17002, 16, 17)
("SPL", 19, 19001, 16002, 17001, 16, 17)
("XOR", 21, 21001, 21002, 21005, -1, -1)
("SPL", 22, 26005, 21001, 24002, 21, 24)
("SPL", 23, 15005, 21002, 24001, 21, 24)
("AND", 24, 24001, 24002, 25001, -1, 25)
("OR", 25, 25001, 25002, 25005, -1, 33)
("XOR", 26, 26001, 26002, 26005, -1, 22)
("AND", 27, 27001, 27002, 25002, -1, 25)
("SPL", 28, 28001, 26001, 27002, 26, 27)
("SPL", 29, 29001, 26002, 27001, 26, 27)
("XOR", 31, 31001, 31002, 31005, -1, -1)
("SPL", 32, 36005, 31001, 34002, 31, 34)
("SPL", 33, 25005, 31002, 34001, 31, 34)
("AND", 34, 34001, 34002, 35001, -1, 35)
("OR", 35, 35001, 35002, 35005, -1, -1)

("XOR", 36, 36001, 36002, 36005, -1, 32)
("AND", 37, 37001, 37002, 35002, -1, 35)
("SPL", 38, 38001, 36001, 37002, 36, 37)
("SPL", 39, 39001, 36002, 37001, 36, 37)
Interacts:
(9, "PSH", 3001, 3)    // carry in
(1, "TOG", 8001, 8)    // a0
(2, "TOG", 9001, 9)    // b0
(3, "TOG", 18001, 18)  // a1
(4, "TOG", 19001, 19)  // b1
(5, "TOG", 28001, 28)  // a2
(6, "TOG", 29001, 29)  // b2
(7, "TOG", 38001, 38)  // a3
(8, "TOG", 39001, 39)  // b3
(1, "LED", 35005, 35)  // carry out
(2, "LED", 31005, 31)  // sum3
(3, "LED", 21005, 21)  // sum2
(4, "LED", 11005, 11)  // sum1
(5, "LED", 1005, 1)    // sum0

```

5 Edge Detection

Here we address an important and subtle phenomenon that occurs in real-world circuits called *edge detection*. Consider a circuit where, somewhere in a given wire (either caused by a button or a component), the value (perhaps rapidly) turns on and off as depicted in Figure 39.26.

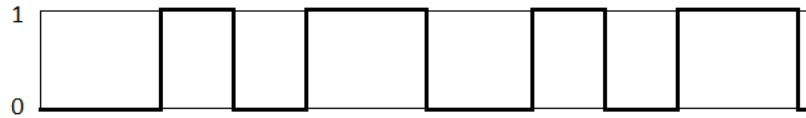


Figure 39.26. Changing value in a wire.

Next, suppose we are interested in capturing *only* the moments when the value changes from 0 to 1. To illustrate, Figure 39.27 overlays a colored edge onto the so-called ‘rising edges’ of the previous sketch as shown.

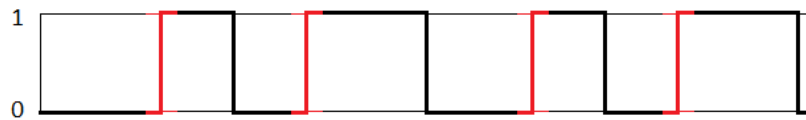


Figure 39.27. Rising edge overlay.

A proper *edge detector* is designed to convert rising edges into short pulses. Continuing the example on hand, this amounts to turning the red portion in the previous sketch into pulses of their own as depicted in Figure 39.28.



Figure 39.28. Pulses derived from rising edges.

5.1 Edge Detector Circuit

Astonishingly, it’s possible to build an edge detector circuit from a strange but simple configuration of logic gates. The idea is to couple both inputs to an AND gate, but to invert one of them as shown in Figure 39.29.

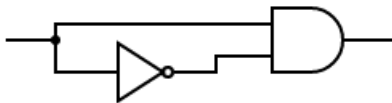


Figure 39.29. Edge detector circuit.

The key is to realize that the ‘evaluation time’ of a given logic gate is not instant, and the presence of the INV gate

momentarily lags down the signal reaching the second AND gate input. From this we see suddenly switching the input to 1 causes a brief moment where both inputs on the AND gate are 1, and a quick high pulse is emitted.

Extended Edge Detector

The particular make and model of the INV gate chosen for an edge detector circuit, along with the quality of the wires involved, has direct influence over its proper functionality. That is, some chips evaluate very quickly, sending a very narrow pulse downstream that is too slender to use.

This is compensated for by adding an additional INV gates to the circuit in even pairs. as shown in Figure 39.30.

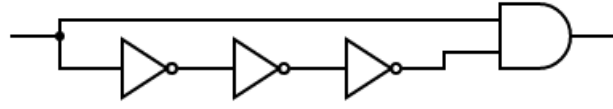


Figure 39.30. Extended edge detector circuit.

The extra pair of INV gates helps to slow the signal reaching the low AND gate input, thus making the detected edge last longer. As it turns out, this idea must also be exploited in simulation because the order in which components are defined has an effect on internal state. In order to guarantee a ‘proper’ pulse from the AND gate, the safer option is the extended edge detector. Our version of the extended edge detector takes the form:

Components:

```
("SPL", 1, 1001, 2002, 3001, 2, 3)
("INV", 2, 2002, 4001, 0, 4, -1)
("AND", 3, 3001, 3002, 3010, -1, -1)
("INV", 4, 4001, 5001, 0, 5, -1)
("INV", 5, 5001, 3002, 0, 3, -1)
```

Interacts:

```
(1, "PSH", 1001, 1)
(1, "LED", 3010, 3)
```

It’s worth having a precise look at the state of the extended edge detector. Starting with the ground state, we first have

$$S_{off} = \{[1] 000 [2] 000 [3] 000 [4] 000 [5] 000\} ,$$

which is in fact unstable, and the circuit jumps quickly to the rest state:

$$S_{rest} = \{[1] 000 [2] 010 [3] 010 [4] 100 [5] 010\}$$

From here, let us examine what happens by pressing and holding button 1. Doing so causes the intermediate state

$$S_{med} = \{[1] 100 [2] 010 [3] 010 [4] 100 [5] 010\} .$$

Immediately after this, the circuit jumps to the curious intermediate state

$$S_{edge} = \{[1] 111 [2] 100 [3] 101 [4] 010 [5] 100\} ,$$

which momentarily flashes the circuit’s only light. This indicates that the so-called edge has been successfully detected and any downstream components are now aware. Note the off-color behavior of the AND gate [3] at this moment: the inputs disagree, but the output is high.

Finally, with the button still pushed and held, the state becomes

$$S_{hold} = \{[1] 111 [2] 100 [3] 100 [4] 010 [5] 100\} .$$

Letting go of the button, the circuit returns to the rest state.

5.2 D Flip-Flop

Quickly reviewing the story of bit storage, we know that a clever use of feedback gives rise to the SR latch, and then the inputs to the circuit can be edited to give the ‘enabled’ SR latch. Modifying the input further produces the elegant D latch.

Here we continue the SR latch evolution story by incorporating edge detection. In particular, begin with the D latch and place an extended edge detector circuit immediately after the ‘enable’ button as shown in Figure 39.31.

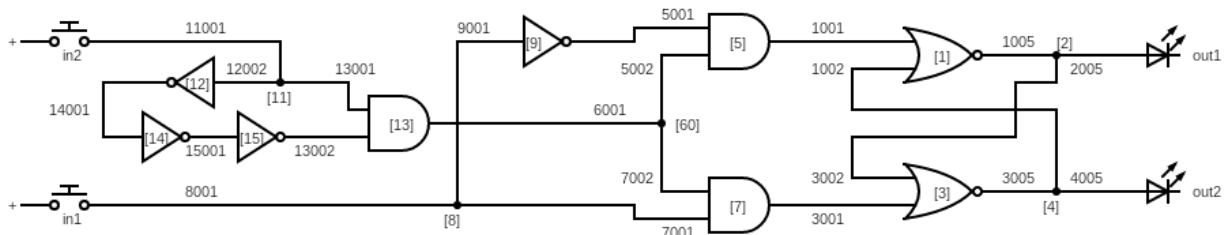


Figure 39.31. D flip-flop circuit.

This construct is called a *D flip-flop*. Different from the D latch, the ‘enable’ function only triggers on the leading edge of the enable button (in2). That is, the circuit is only enabled in pulses, not continuously. Internally, the D flip-flop does the same job as the D latch and obeys the same truth table.

The corresponding data for the D flip-flop circuit is as follows:

Components:

```
("NOR", 1, 1001, 1002, 1005, -1, 2)
("SPL", 2, 1005, 3002, 2005, 3, -1)
("NOR", 3, 3001, 3002, 3005, -1, 4)
("SPL", 4, 3005, 1002, 4005, 1, -1)
("AND", 5, 5001, 5002, 1001, -1, 1)
("SPL", 6, 6001, 5002, 7002, 5, 7)
("AND", 7, 7001, 7002, 3001, -1, 3)
("SPL", 8, 8001, 7001, 9001, 7, 9)
("INV", 9, 9001, 5001, 0, 5, -1)
("SPL", 11, 11001, 12002, 13001, 12, 13)
("INV", 12, 12002, 14001, 0, 14, -1)
("AND", 13, 13001, 13002, 6001, -1, 6)
("INV", 14, 14001, 15001, 0, 15, -1)
("INV", 15, 15001, 13002, 0, 13, -1)
```

Interacts:

```
(1, "PSH", 8001, 8) // D
(2, "PSH", 11001, 11) // E
(1, "LED", 2005, 2) // Q
(2, "LED", 4005, 4) // Q'
```

D Flip-Flop Symbol

As it turns out, the D flip-flop is popular enough to have its own symbol as shown in Figure 39.32. The input *D*, along with outputs *Q*, *Q'* are labeled accordingly (with *Q'* equivalent to $\overline{Q}$). The ‘set’ (*S*) and ‘reset’ (*R*) pins can be ignored for our purposes, which leaves the strange triangular symbol on the left for the ‘enable’ function. This input is by convention called the *clock*.

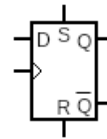


Figure 39.32. D flip-flop symbol.

6 Counters

The D flip-flop can be put to use in a variety of interesting ways, and we scratch the surface here by building a few counters.

6.1 One Bit Counter

Begin with the D flip-flop, and perform the strange surgery of removing the D -input, and re-routing its wire to the Q' output as shown in Figure 39.33. Also ignore the Q -output for a moment and only look at Q' .

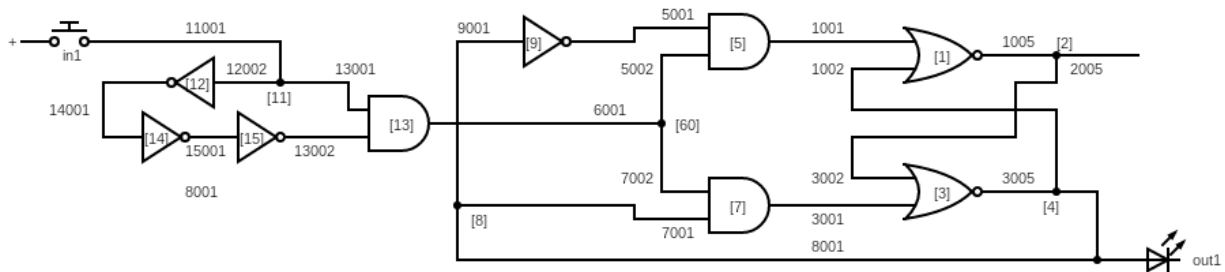


Figure 39.33. One bit counter circuit.

Such a modified D flip-flop is called a *one bit counter*. To ‘count’ a one bit number isn’t very exciting, as the circuit merely amounts to a button that turns a light on or off. The subtlety is, the circuit only responds to the moment the button is pressed, i.e. on the rising edge of the input.

represented in Figure 39.34. The equivalent as circuit data reads:

Evolution of State

Slightly less interesting than the states allowed in circuit is the order in which they occur. From the ground state S_{00} , pushing and holding the input leads to the state S_{11} . Releasing the button gives S_{01} . From there, pressing the button again gives S_{10} , and releasing it a second time gets back to S_{00} . In summary, we find the sequence of states to be:

$$S_{00} \rightarrow S_{11} \rightarrow S_{01} \rightarrow S_{10} \rightarrow S_{00} \rightarrow \dots$$

In other words, the light stays on after the button is released. It takes a second push to turn the light off.

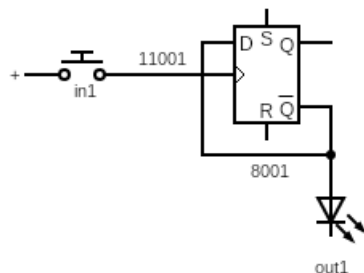


Figure 39.34. One bit counter circuit in terms of D flip-flop symbol.

In terms of the D flip-flop symbol, the one bit counter is

Components:

```
( "NOR", 1, 1001, 1002, 1005, -1, 2)
( "SPL", 2, 1005, 3002, 2005, 3, -1)
( "NOR", 3, 3001, 3002, 3005, -1, 4)
( "SPL", 4, 3005, 1002, 8001, 1, 8)
( "AND", 5, 5001, 5002, 1001, -1, 1)
( "SPL", 6, 6001, 5002, 7002, 5, 7)
( "AND", 7, 7001, 7002, 3001, -1, 3)
( "SPL", 8, 8001, 7001, 9001, 7, 9)
( "INV", 9, 9001, 5001, 0, 5, -1)
( "SPL", 11, 11001, 12002, 13001, 12, 13)
( "INV", 12, 12002, 14001, 0, 14, -1)
( "AND", 13, 13001, 13002, 6001, -1, 6)
( "INV", 14, 14001, 15001, 0, 15, -1)
( "INV", 15, 15001, 13002, 0, 13, -1)
```

Interacts:

```
(1, "PSH", 11001, 11) // Clock
(1, "LED", 8001, 8) // Q'
```

6.2 Two Bit Counter

One can straightforwardly build a two bit counter by linking a pair of one bit counters in series as shown in Figure 39.35.

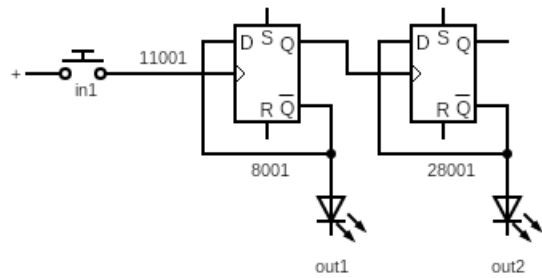


Figure 39.35. Two bit counter circuit in terms of D flip-flop symbol.

For clarity we have employed the symbolic notation for the D flip-flop, but keep in mind the circuit being described isn't precisely trivial. For completeness, the corresponding circuit data reads:

Components:

```
("NOR", 1, 1001, 1002, 1005, -1, 2)
("SPL", 2, 1005, 3002, 31001, 3, 31)
("NOR", 3, 3001, 3002, 3005, -1, 4)
("SPL", 4, 3005, 1002, 8001, 1, 8)
("AND", 5, 5001, 5002, 1001, -1, 1)
("SPL", 6, 6001, 5002, 7002, 5, 7)
("AND", 7, 7001, 7002, 3001, -1, 3)
("SPL", 8, 8001, 7001, 9001, 7, 9)
("INV", 9, 9001, 5001, 0, 5, -1)
("SPL", 11, 11001, 12002, 13001, 12, 13)
("INV", 12, 12002, 14001, 0, 14, -1)
("AND", 13, 13001, 13002, 6001, -1, 6)
("INV", 14, 14001, 15001, 0, 15, -1)
("INV", 15, 15001, 13002, 0, 13, -1)
("NOR", 21, 21001, 21002, 21005, -1, 22)
("SPL", 22, 21005, 23002, 22005, 23, -1)
("NOR", 23, 23001, 23002, 23005, -1, 24)
("SPL", 24, 23005, 21002, 28001, 21, 28)
("AND", 25, 25001, 25002, 21001, -1, 21)
("SPL", 26, 26001, 25002, 27002, 25, 27)
```

```
("AND", 27, 27001, 27002, 23001, -1, 23)
("SPL", 28, 28001, 27001, 29001, 27, 29)
("INV", 29, 29001, 25001, 0, 25, -1)
("SPL", 31, 31001, 32002, 33001, 32, 33)
("INV", 32, 32002, 34001, 0, 34, -1)
("AND", 33, 33001, 33002, 26001, -1, 26)
("INV", 34, 34001, 35001, 0, 35, -1)
("INV", 35, 35001, 33002, 0, 33, -1)
```

Interacts:

```
(1, "PSH", 11001, 11)
(1, "LED", 8001, 8)
(2, "LED", 28001, 28)
```

The two bit counter still operates on a single input. The so-far unused Q output of the first counter is sent as the input to the next, and the state of the counter is read from the pair Q' outputs.

The two bit counter's qualitative behavior extends that of the one bit, namely that the output state changes only when the input is initially pressed. Starting from the ground state and listing only the output values, the two bit counter yields the output sequence

$$00 \rightarrow 01 \rightarrow 10 \rightarrow 11 \rightarrow 00 \rightarrow \dots,$$

or in base ten,

$$0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0 \rightarrow \dots$$

That is, the circuit counts from 0 to 3 and repeats forever.

6.3 Four Bit Counter

Given the plug-and-play flexibility the one bit adder circuit, it's very easy to extend the two bit counter to add three, four, or any number of bits. Granted, it becomes more difficult to manage the such circuits, which is why D flip-flop circuits are often printed to chips.

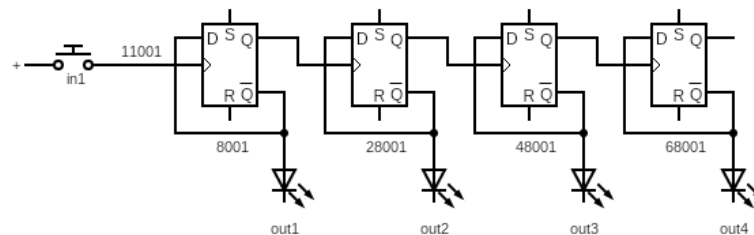


Figure 39.36. Four bit counter circuit.

A four bit counter has been prepared and is shown in Figure 39.36, which looks rather tame in terms of the D flip-flop symbol. Such a device uses four outputs to cycle through the integers 0000 to 1111, also known as zero to fifteen. As circuit data, this one is a monster:

Components:

```
("NOR", 1, 1001, 1002, 1005, -1, 2)
("SPL", 2, 1005, 3002, 31001, 3, 31)
("NOR", 3, 3001, 3002, 3005, -1, 4)
("SPL", 4, 3005, 1002, 8001, 1, 8)
("AND", 5, 5001, 5002, 1001, -1, 1)
("SPL", 6, 6001, 5002, 7002, 5, 7)
("AND", 7, 7001, 7002, 3001, -1, 3)
("SPL", 8, 8001, 7001, 9001, 7, 9)
("INV", 9, 9001, 5001, 0, 5, -1)
("SPL", 11, 11001, 12002, 13001, 12, 13)
("INV", 12, 12002, 14001, 0, 14, -1)
("AND", 13, 13001, 13002, 6001, -1, 6)
("INV", 14, 14001, 15001, 0, 15, -1)
("INV", 15, 15001, 13002, 0, 13, -1)
("NOR", 21, 21001, 21002, 21005, -1, 22)
("SPL", 22, 21005, 23002, 51001, 23, 51)
("NOR", 23, 23001, 23002, 23005, -1, 24)
("SPL", 24, 23005, 21002, 28001, 21, 28)
("AND", 25, 25001, 25002, 21001, -1, 21)
("SPL", 26, 26001, 25002, 27002, 25, 27)
("AND", 27, 27001, 27002, 23001, -1, 23)
("SPL", 28, 28001, 27001, 29001, 27, 29)
("INV", 29, 29001, 25001, 0, 25, -1)
("SPL", 31, 31001, 32002, 33001, 32, 33)
("INV", 32, 32002, 34001, 0, 34, -1)
("AND", 33, 33001, 33002, 26001, -1, 26)
("INV", 34, 34001, 35001, 0, 35, -1)
("INV", 35, 35001, 33002, 0, 33, -1)
("NOR", 41, 41001, 41002, 41005, -1, 42)
("SPL", 42, 41005, 43002, 71001, 43, 71)
("NOR", 43, 43001, 43002, 43005, -1, 44)
("SPL", 44, 43005, 41002, 48001, 41, 48)
("AND", 45, 45001, 45002, 41001, -1, 41)
("SPL", 46, 46001, 45002, 47002, 45, 47)
("AND", 47, 47001, 47002, 43001, -1, 43)
("SPL", 48, 48001, 47001, 49001, 47, 49)
("INV", 49, 49001, 45001, 0, 45, -1)
("SPL", 51, 51001, 52002, 53001, 52, 53)
("INV", 52, 52002, 54001, 0, 54, -1)
("AND", 53, 53001, 53002, 46001, -1, 46)
("INV", 54, 54001, 55001, 0, 55, -1)
("INV", 55, 55001, 53002, 0, 53, -1)
("NOR", 61, 61001, 61002, 61005, -1, 62)
```

```
("SPL", 62, 61005, 63002, 62005, 63, -1)
("NOR", 63, 63001, 63002, 63005, -1, 64)
("SPL", 64, 63005, 61002, 68001, 61, 68)
("AND", 65, 65001, 65002, 61001, -1, 61)
("SPL", 66, 66001, 65002, 67002, 65, 67)
("AND", 67, 67001, 67002, 63001, -1, 63)
("SPL", 68, 68001, 67001, 69001, 67, 69)
("INV", 69, 69001, 65001, 0, 65, -1)
("SPL", 71, 71001, 72002, 73001, 72, 73)
("INV", 72, 72002, 74001, 0, 74, -1)
("AND", 73, 73001, 73002, 66001, -1, 66)
("INV", 74, 74001, 75001, 0, 75, -1)
("INV", 75, 75001, 73002, 0, 73, -1)
```

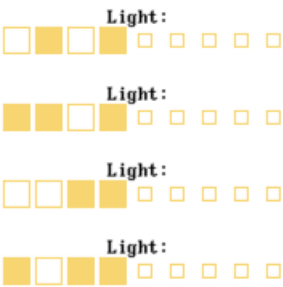
Interacts:

```
(1, "PLS", 11001, 11)
(1, "LED", 8001, 8)
(2, "LED", 28001, 28)
(3, "LED", 48001, 48)
(4, "LED", 68001, 68)
```

Output

As a reward for patience, the PSH button has been upgraded to species PLS, freeing the user's hand. Letting the circuit run, the output section cycles through the numbers 0 to 15 as follows:





7 And-Inv Completeness

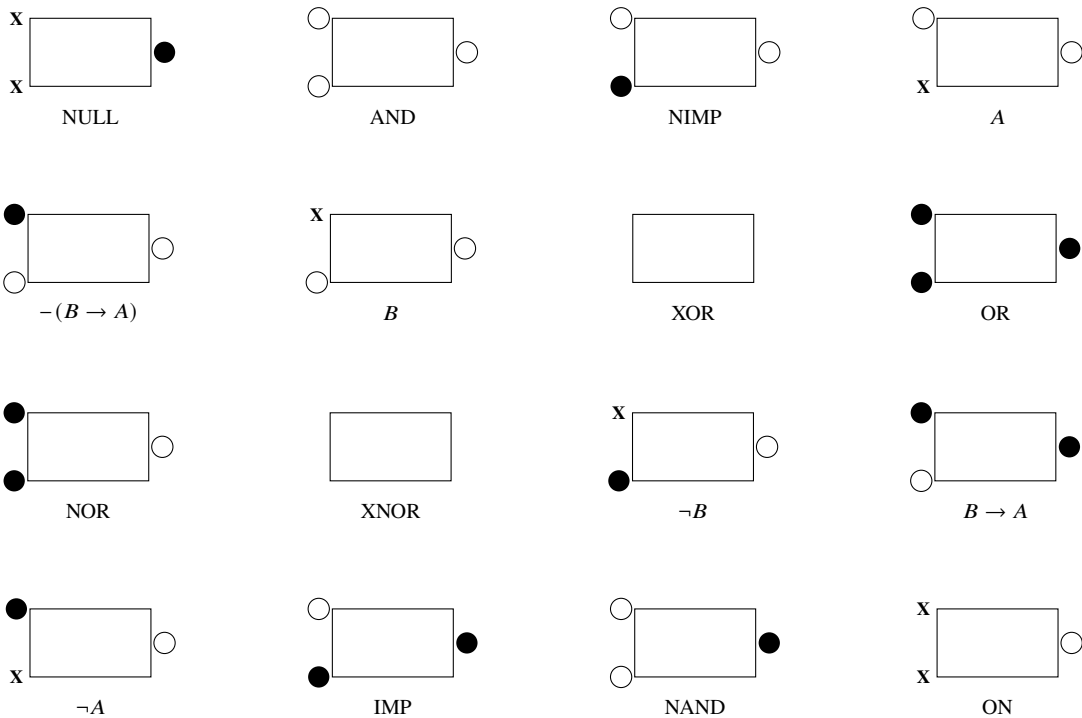
This note presents a compact diagrammatic notation for common two-input Boolean gates using a single rectangular box with small markers placed just outside its frame: the top-left marker denotes input A , the bottom-left marker denotes input B , and a right-edge marker denotes the output. Each

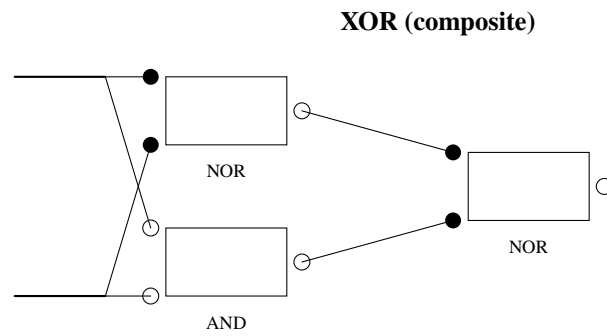
marker's style encodes how that signal is treated: open circle $\circ$ means the signal is used as-is (no inversion), a filled circle $\bullet$ means the signal is inverted, and an uppercase **X** means the input is unused. With these decorations, most two-input gates can be expressed directly; parity gates (XOR/XNOR) appear as simple composites.

Key / Legend

1. $\circ$ open circle = direct (no inversion)
2. $\bullet$ filled circle = inverted signal
3. **X** (uppercase) = unused input (no circle)
4. $\square$ empty box = AND gate (default box semantics)

Marker positions: A at the top of the left edge; B at the bottom of the left edge; output at the right edge center.





Note: XNOR is the logical negation of XOR — invert the final output marker.

Chapter 40

Sequence Analysis

1 Sequences

In the broadest sense, a *sequence* is any particular ordering of things or symbols. A sequence could be digits of a phone number, a list of strategic moves, the order of cards in a deck, etc.

1.1 Members and Elements

The N total constituents of a sequence are called *members*. The unique (non-repeated) members in a sequence are called *elements*.

Introducing notation to represent an example sequence, consider the statement

$$S_{16}(\phi_3) = (1\ 2\ 1\ 1\ 3\ 2\ 1\ 3\ 3\ 2\ 3\ 1\ 2\ 1\ 3\ 1) .$$

On the right, we have an ordered list of digits 1, 2, 1, 1, etc., and there are $N = 16$ members in total. To represent this symbolically, on the left we first write S_{16} , which means ‘a sequence with sixteen members’. Following this we represent the *element vector* ϕ_3 , which for the example sequence means

$$\phi_3 = \{1, 2, 3\} .$$

The members of the sequence are each drawn from one list of three elements.

Individual Items

The j th item in a sequence containing N members is written

$$S_N^j(\phi) ,$$

and similarly the j th item in an element vector of size M is written

$$\phi_M^j .$$

Either of these could be used to represent the members of a sequence.

1.2 Element Regrouping

Have another look at the example sequence above, namely

$$1\ 2\ 1\ 1\ 3\ 2\ 1\ 3\ 3\ 2\ 3\ 1\ 2\ 1\ 3\ 1 ,$$

and delete every second space to write

$$S_8(\phi_{3,2}) = (12\ 11\ 32\ 13\ 32\ 31\ 21\ 31) .$$

Representing the same information as $S_{16}(\phi_3)$, we have a new sequence with $16/2 = 8$ members, and the element vector is no longer ϕ_3 . In particular, the element vector becomes

$$\phi_{3,2} = \{11, 12, 13, 21, 22, 23, 31, 32, 33\} ,$$

which lists every possible pairing of the digits 1, 2, 3.

Order-Two Regrouping

In more symbolic terms, the move we just made amounts to the *order-two regrouping*

$$S_N(\phi_M) = S_{N/2}(\phi_{M,2}) + R_2(\phi_M) ,$$

where the sequence $S_{N/2}(\phi_{M,2})$ reads

$$S_{N/2}(\phi_{M,2}) = \left(S_{N/2}^1(\phi_{M,2})\ S_{N/2}^2(\phi_{M,2}) \cdots S_{N/2}^{N/2}(\phi_{M,2}) \right) ,$$

and the remainder term $R_2(\phi_M)$ is included for when N is not an even number.

Generalized Regrouping

A regrouped sequence contains the same information as the original while making the trade of having fewer members with a larger basis vector.

Continuing the example on hand, one may imagine turning the ϕ_3 basis vector into one containing $3^3 = 27$ elements:

$$\phi_{3,3} = (111, 112, 113, \dots, 331, 332, 333)$$

With this, a sequence $S_N(\phi_M)$ regroupes as

$$S_N(\phi_M) = S_{N/3}(\phi_{M,3}) + R_3(\phi_M) ,$$

where the remainder term contains the tail end of the original sequence, one or two digits in this case.

In the general case, one may perform an order- g regrouping, so long as g never exceeds N . To express this we write

$$S_N(\phi_M) = S_{N/g}(\phi_{M,g}) + R_g(\phi_M) , \quad (40.1)$$

and the number of members in $R_g(\phi_M)$ is N modulo g .

1.3 Worked Example

To demonstrate the utility of element regrouping, and to motivate further analysis, consider the binary sequence

$$S_{24}(\phi_2) = (010011001110010011001110) .$$

One may argue, at a glance, that the sequence given has no discernible order, or was perhaps generated by a random process. We shall use element regrouping to find out.

Proceed by conceiving the basis vector

$$\phi_{2,2} = \{00, 01, 10, 11\} ,$$

or equivalently, just changing symbols,

$$\phi_{2,2} = \{A, B, C, D\} .$$

The given sequence becomes

$$S_{12}(\phi_{2,2}) = (BADADC BADADC) .$$

No reason to stop there, of course. One can now imagine pairs of elements such that

$$\phi_{2,4} = \{AA, AB, AC, \dots, DB, DC, DD\} ,$$

or equivalently, switching to lowercase letters,

$$\phi_{2,4} = \{a, b, c, \dots, n, o, p\} .$$

In terms of the vector $\phi_{2,4}$, the given sequence boils down to six symbols:

$$S_{24}(\phi_2) = S_6(\phi_{2,4}) = (e m o e m o)$$

From the above we see the original sequence is anything but random, as the regrouped sequence is two instances of the same string *emo*.

Further Analysis

It's worth going one more step. Consider yet another element vector that is an order-three regrouping of the previous one, namely

$$\psi_{16,3} = (aaa, aab, aac, \dots, ppn, ppo, ppp) ,$$

having $16^3 = 4096$ elements, of which *emo* is just one of them, where for some unique k , we have

$$\psi_{16,3}^k = emo .$$

The equivalence

$$S_{24}(\phi_2) = S_2(\psi_{16,3}) = (\psi_{16,3}^k \psi_{16,3}^k)$$

tells us that either sequence - the original or the regrouped - embeds a probability

$$P = \frac{1}{4096}$$

that the second half is equal to the first half. Finally, we can rewrite the original sequence in a way that betrays its (now obvious) repetition:

$$S_{24}(\phi_2) = S_2(\phi_{2,12}) = (010011001110 010011001110)$$

Relevance to Birthday Problem

There is a famous question in probability courses that asks 'if there are N people in the room, what is the probability $P(N)$ that any two people share a birthday?'

The solution to this problem is a fun exercise, and the answer happens to be

$$P(N) = 1 - \frac{365!}{365^N (365 - N)!} ,$$

which makes use of the factorial ($!$) operator.

It just so happens that we just puzzled out the $N = 2$ case of the above, where the term 4096 replaces the number 365 in the formula.

For fun, suppose the example binary sequence was instead $N = 12 \times 3 = 36$ digits in length, consisting of three copies of the same sub-sequence, i.e.

$$S_{36}(\phi_2) = S_{12}(\psi_{16,3}) = (\psi_{16,3}^k \psi_{16,3}^k \psi_{16,3}^k) .$$

Like the $N = 2$ case, we can wonder about the chances of such a sequence occurring, which amounts to calculating

$$P(3) = 1 - \frac{4096!}{4096^3 (4096 - 3)!} ,$$

resulting in a number slightly larger than $1/4096$.

One might wonder why $P(3)$ is not significantly smaller than $1/4096$, because there's a feeling that 'twins are more common than triplets'. Keep in mind though that $P(N)$ calculates whether *any* of the N members are the same, not an original specific member.

1.4 Element Translation

In any basis ϕ , suppose we have a sample sequence

$$S_7(\phi) = (A B C D E F G) .$$

One useful manipulation involves demoting the first member of the sequence (A) to the last position (after G), thereby 'promoting' all other members. This is called *element translation*, and to denote this we attach a superscripted digit on S as follows:

$$^1S_7(\phi) = (B C D E F G A) .$$

The same sequence can be translated more than once by adjusting the new superscript:

$$^2S_7(\phi) = (C D E F G A B)$$

$$^3S_7(\phi) = (D E F G A B C)$$

Moreover, translations work in reverse when using a negative subscript:

$$^{-1}S_7(\phi) = (G A B C D E F)$$

$$^{-2}S_7(\phi) = (F G A B C D E)$$

In the general case, for any sequence (in shorthand notation)

$$S_N(\phi) = (S_N^1 S_N^2 \cdots S_N^N) ,$$

a translation of t members is represented by:

$$^tS_N(\phi) = (S_N^{1+t} S_N^{2+t} \cdots S_N^N \cdots S_N^1 S_N^2 \cdots S_N^t) \quad (40.2)$$

Regrouping after Translation

Translation of a sequence has significant consequence when regrouping its members. Using the same sequence from the worked example above, take the $t = 1$ translation to write

$$\begin{aligned} {}^1S_{24}(\phi_2) &= \tilde{S}_{24}(\phi_2) \\ &= (100110011100100110011100) \end{aligned}$$

In terms of the basis vector $\phi_{2,2}$, the above reads

$$\tilde{S}_{12}(\phi_{2,2}) = (CBCBDACBCBDA) ,$$

and furthermore using $\phi_{2,4}$ we have

$$\tilde{S}_6(\phi_{2,4}) = (j j m j j m) .$$

Notice right away that $j j m$ is different from *emo* that corresponds to the non-translated sequence. Unsurprisingly though, the translated sequence still resolves to two copies of *something*, namely

$$\psi_{16,3}^q = j j m$$

for some unique q .

2 Random Number Analysis

Now we use sequence analysis to solve a ‘real’ problem. Namely, suppose you are given a sequence of 10^4 of integers in the range $[0, \dots, 9]$ with the claim that the digits in the sequence are randomly distributed.

2.1 Histograms

One may visualize a given sequence using a *histogram*, which using our terms, is a 2D bar graph vertically representing the total occurrence count of each element in a given sequence.

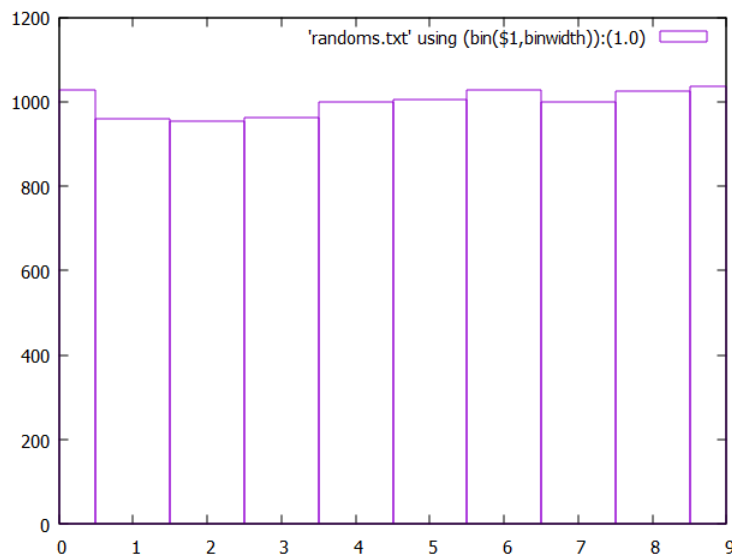


Figure 40.1. Histogram of $N = 10000$ random integers ranging 0 to 9.

If the incoming sequence is *truly* random, there should, at the very least, be no discernible structure visible in the corresponding histogram as shown in Figure 40.1. The horizontal axis lists the elements in the sequence, i.e. the integers 0 to 9. The vertical axis represents the number of occurrences of each integer, and we see an average of about 1000 per element. This makes sense, as the given sequence contains ten total elements and is ten thousand members large. The slight ‘bumpiness’ along the line through 1000 gives a sense of the standard deviation around this average.

Histograms can be blind to certain structures within the data they represent. For this case, we don’t have a way

of knowing, for instance, whether the pattern 123 occurs as often as 321. This, if it were the case, would alone be enough to taint the randomness of the sample sequence while the histogram remains innocent. The sequence analysis methods we develop here are a remedy to this.

Since computer automation will be essential as we trudge forward, it’s worth having a reliable system for manipulating large lists of data and generating histograms. To this end we employ the famous *gnuplot*, freely available for most workstations. The above histogram was generated using three *gnuplot* statements:

```
binwidth=1
bin(x,width)=width*floor(x/width)
plot 'randoms.txt' using (bin($1,binwidth)):(1.0) smooth freq with boxes
```

In `randoms.txt`, members of the sequence are separated by the ‘newline’ character.

2.2 Tainted Sample

Now, suppose we are handed a new sequence stored in `seq_10_0.txt` of comparable length to the above (at least 10^4 members). Running this sequence through the same

gnuplot program used above, we get a starkly different histogram as shown in Figure 40.2.

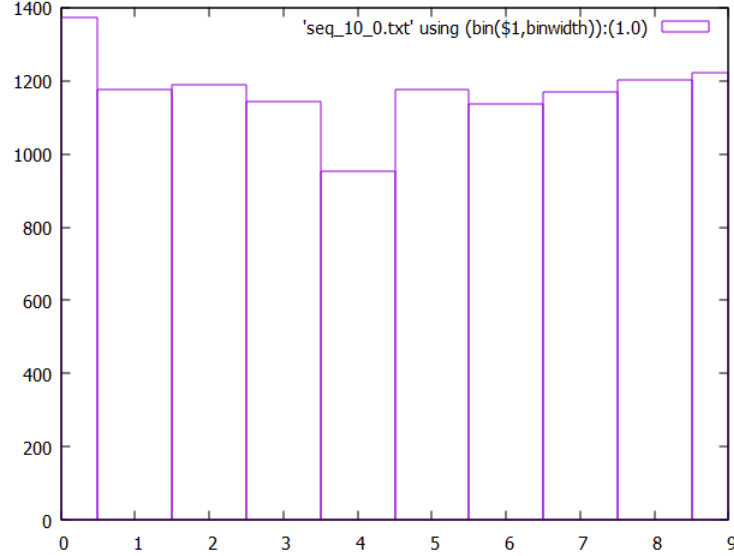


Figure 40.2. Histogram of $N \approx 11000$ non-random integers ranging 0 to 9.

Significant structure is visible in the histogram, thus the incoming sequence is tainted somehow. At minimum, we see a over-representation of the element 0 while 4 is strikingly under-represented. The remaining elements seem to share an average, at least qualitatively, but that's the best we can say so far.

2.3 Base-Ten Element Vectors

Now things get gritty. First, observe that the element vector behind any sequence containing the integers 0 to 9 is

$$\phi_{10} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} .$$

In the spirit of element regrouping, we can re-imagine the incoming sequence as being half as long while consisting of members between the numbers 00 to 99, suggesting the element vector

$$\phi_{10,2} = \{00, 01, \dots, 98, 99\} .$$

Of course, we can keep going with this to write

$$\phi_{10,3} = \{000, 001, \dots, 998, 999\} ,$$

and so on for a general grouping number g as suggested in Equation (40.1).

Without knowing what causes the non-random quality of the given sequence, it is most prudent to include as many

regrouping of the sequence as is reasonable. For the sake of demonstration, we choose a cutoff at $g = 8$ while including all values beneath this. Also, since the incoming sequence is sufficiently large, we ignore anything to do with remainder terms.

2.4 Translation and Regrouping

Starting with the incoming sequence $S^N(\phi_{10})$, suppose we produce the $t = 1$ translation ${}^1S^N(\phi_{10})$. This move has no effect whatsoever on the corresponding histogram, i.e. Figure 40.2 remains the same under this change.

This isn't always the case if we translate before regrouping. To illustrate, suppose the first six members of the incoming sequence are

$$S_N(\phi_{10}) = (0\ 0\ 6\ 0\ 4\ 0 \dots) .$$

From this, the $t = 1$ translation reads

$${}^1S_N(\phi_{10}) = (0\ 6\ 0\ 4\ 0 \dots 0) ,$$

and the $t = 2$ translation is

$${}^2S_N(\phi_{10}) = (6\ 0\ 4\ 0 \dots 0\ 0) .$$

Looking closely, we see that $S^N(\phi_{10})$ and ${}^1S_N(\phi_{10})$ will yield different results under a $\phi_{10,2}$ regrouping. To convince yourself quickly, let $A = 00$, $B = 60$, $C = 40$

and see what happens to these after translation. By the same token though, ${}^2S_N(\phi_{10})$ does not yield new useful information, as it simply kicks 00 to the back of the line. The corresponding histogram would be blind to this order.

What we learn from this is, if we are concerned with the $g = 2$ regrouping of the incoming sequence, then only $S_N(\phi_{10})$ and ${}^1S_N(\phi_{10})$ need to be brought into consideration. This argument extends upward for increasing g .

That is, if we are concerned with the $g = 3$ regrouping of the incoming sequence, then $S_N(\phi_{10})$, ${}^1S_N(\phi_{10})$, ${}^2S_N(\phi_{10})$ are relevant, and there is no need for ${}^3S_N(\phi_{10})$, for it just kicks the first member to the back.

An in-house script has been prepared to generate all of the translated and regrouped sequences required. Starting from the incoming sequence stored in `seq_10_0.txt`, the secondary data files required are suggested below.

- Order-two regrouping $\phi_{10,2}$ requires:
 - `seq_10_2_0.txt`
 - `seq_10_2_1.txt`
- Order-three regrouping $\phi_{10,3}$ requires:
 - `seq_10_3_0.txt`

```
set key off
binwidth=1
bin(x,width)=width*floor(x/width)
g=2
plot for[i = 0:g-1] "seq_10_".g."_".i.".txt" _
    using (bin($1,binwidth)):(1.0) smooth freq with boxes
```

The program uses a loop to stack histograms set by variable g that we'll tweak for other regroupings. For the

- `seq_10_3_1.txt`
- `seq_10_3_2.txt`

- Order- g regrouping $\phi_{10,g}$ requires:

- `seq_10_g_0.txt`
- `seq_10_g_1.txt`
- ...
- `seq_10_g_g-1.txt`

As mentioned, and for no good reason, we cut the analysis after $g = 8$. This can be expanded as needed.

2.5 Stacked Histograms

Despite having different structure, regrouped translations of a sequence share an element vector, and thus have the same number of horizontal bins in a corresponding histogram. For this reason it makes sense to plot each regrouping on the same graph.

Double-Stacked Histogram

For the $g = 2$ case, we use the following `gnuplot` program:

$g = 2$ case, we get the resulting histogram shown in Figure 40.3.

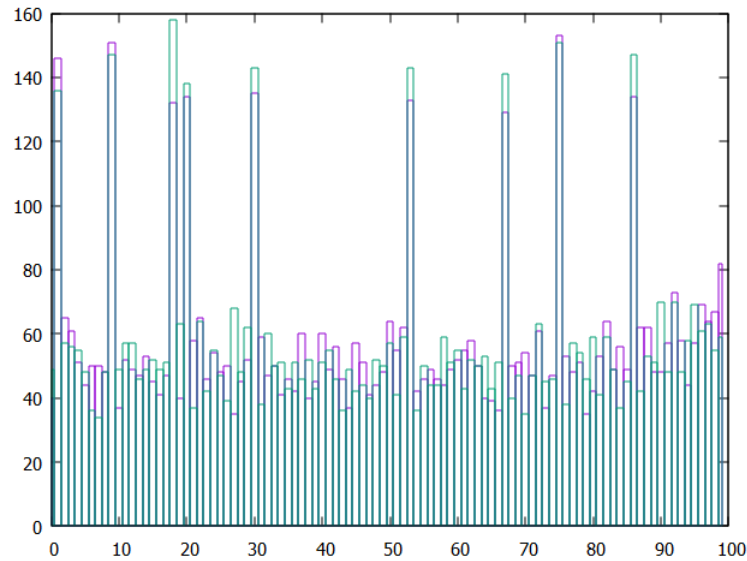


Figure 40.3. Stacked histogram containing the $g = 2$ regrouping.

The stacked histogram showing the regroupe sequences shows highly discernible structure. From lowest to greatest, there are peaks at

$$\omega_2 = (01, 09, 18, 20, 30, 53, 67, 75, 86) .$$

Interestingly, as hinted in the original histogram in Figure 40.2, the digit 4 is not involved in the above list, while the

number 0 occurs more than once.

Triple-Stacked Histogram

The above analysis is now repeated with everything left in place except we now set $g = 3$, and the result is a triple-stacked histogram shown in Figure 40.4.

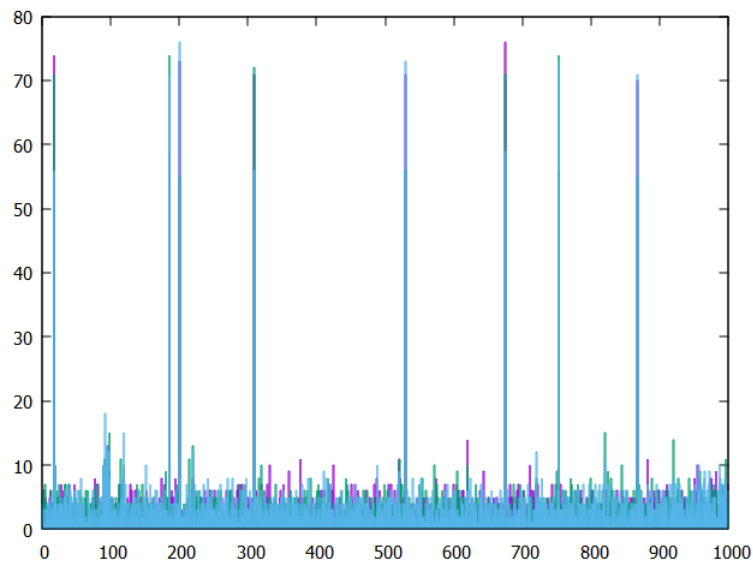


Figure 40.4. Stacked histogram containing the $g = 3$ regrouping.

With the triple-stacked histogram, the peaks are much higher and sharper than the background noise. Looking very closely (preferably in software at this point), we identify each peak:

$$\omega_3 = (018, 186, 201, 309, 530, 675, 753, 867) .$$

Quad-Stacked Histogram

Before the stacking game becomes too repetitive, let us hit the $g = 4$ case and count the number of peaks, which seems to decrease by one with every increase in g . Doing the now-familiar exercise leads to Figure 40.5.

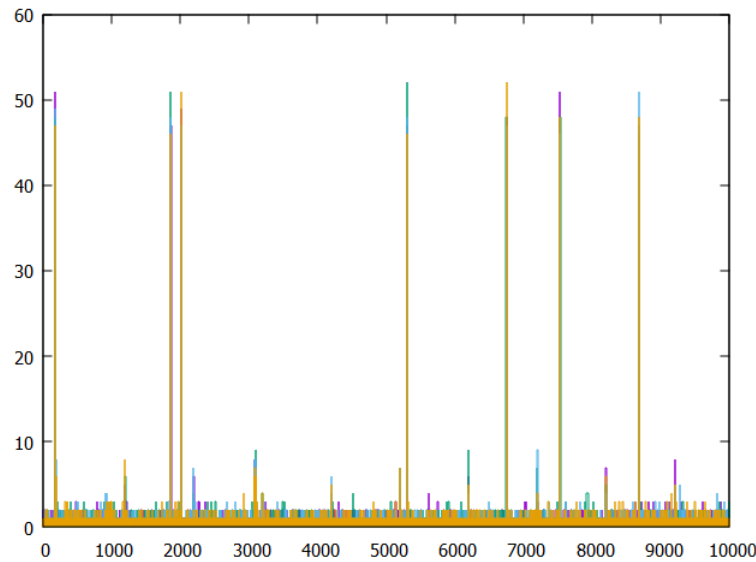


Figure 40.5. Stacked histogram containing the $g = 4$ regrouping.

Listing all peaks carefully, we find

$$\omega_4 = (0186, 1867, 2018, 5309, 6753, 7530, 8675) ,$$

totaling to 7. This is one fewer peaks than listed in ω_3 , continuing the pattern of decreasing by one for every increase in g . Running this forward in imagination, one can estimate that somewhere around ω_{10} will there be just one peak in the histogram. Surely by then the culprit is found.

2.6 Higher Regroupings

With increasing g , the histogram becomes harder to discern at scale and it's better to use software to dig out the corresponding peaks. Showing only the results, one should

find

$$\begin{aligned} \omega_5 &= (01867, 18675, 20186, 67530, 75309, 86753) \\ \omega_6 &= (018675, 186753, 201867, 675309, 867530) \\ \omega_7 &= (0186753, 1867530, 2018675, 8675309) \\ \omega_8 &= (01867530, 18675309, 20186753) \end{aligned}$$

2.7 Results

In case it was not obvious yet, it seems that the (formerly famous) telephone number

2018675309

is embedded a handful of times in the otherwise random incoming sequence.

3 Alphanumeric Analysis

Here we tackle another example that is similar in nature but greater in scale than the previous one.

3.1 ASCII Basis

Consider a sequence consisting of capital letters, lowercase letters, numbers, punctuation marks, and other symbols visible on a standard U.S. keyboard. In particular, we take the ASCII characters within from 32 (a blank space) to 126 (the ‘tilde’ $\sim$ symbol). As an element vector, this amounts to

$$\phi_{95} = \{a, b, A, B, 1, 2, !, @, \dots\},$$

having 95 total elements.

3.2 Incoming Sequence

Now, suppose we are handed a file `seq_abc_0.txt` containing an ASCII sequence of $N \approx 10^5$ members claimed to be randomly chosen from ϕ_{95} . The incoming sequence could look like the following 100 characters indicative of the rest:

```
FC19T8YSOW~q{rSF0'u%}+<t,B&a{,%HB3910;5c
GWTeJ4J_ii2#tk\P-9eQ2.{ 'Pnz:x^op)JtY|{v|
CEn&Gw)6%/<OW'U(+P(y
```

The task is to determine if the sequence is truly random versus having embedded some kind of order, and if so, discovering its cause.

3.3 Data Preparation

Preparation for analysis begins by admitting that even modest regroupings $\phi_{95,2}$ with $95^2 = 9025$ elements, let alone $\phi_{95,3}$ with $95^3 = 857375$ elements, are unsuitable for viewing on a histogram. Foregoing visualization of any kind, instead we employ the classic Unix toolkit, namely those available in a *Bash shell*.

For the sake of efficiency, let us proceed using only the regrouping values $g = 12, g = 16, g = 20, g = 24$. If the incoming sequence contains N members, this means we need, in addition to ${}^N S(\phi_{95})$:

$${}^{N/12} S(\phi_{95,12})$$

$${}^{N/16} S(\phi_{95,16})$$

$${}^{N/20} S(\phi_{95,20})$$

$${}^{N/24} S(\phi_{95,24})$$

As with the previous case, an in-house script has been prepared to generate all of the translated and regrouped sequences required. For each g involved, secondary data files occur as follows:

- Order- g regrouping $\phi_{95,g}$ requires:

- `seq_10_g_0.txt`
- `seq_10_g_1.txt`
- ...
- `seq_10_g_g-1.txt`

3.4 Processing

To review, an incoming sequence is given in the form `seq_abc_0.txt` containing $\approx 10^5$ members. Translated regroupings of this sequence are generated and stored in separate files `seq_10_12_0.txt`, `seq_10_12_1.txt`, and so on.

What remains is to churn an answer from the incoming sequence and secondary data by preparing a script as follows:

main.sh

```
#!/bin/bash

# Combine like-width datasets.
for g in 12 16 20 24; do
    ./combine.sh $g
done

# Core analysis.
for filename in data/*.dat; do
    echo "working..."
    ./core.sh $filename
done

# Produce the result
./analyze.sh
```

In `main.sh`, the first salient line of code erects a ‘for’ loop that churns the values 12, 16, 20, 24. One at a time, these are sent as arguments to another script file `combine.sh`, detailed as follows:

combine.sh

```
#!/bin/bash

touch data/data.tmp

j=$1
for filename in data/seq_abc_"$j"_*.txt; do
    cat data/data.tmp $filename > data/tmpf && cat data/tmpf > data/data.tmp
done

rm data/tmpf

newfile=data/seq_abc_"$j"_all.dat
mv data/data.tmp $newfile
```

The script `combine.sh` requires all sequence files to occur in a directory called `data`. In analogy to building a stacked histogram, all files corresponding to the same element vector (same `g`), are joined and stored in a temporary file `data.tmp`. The temporary file is then named

appropriately and stored with the `dat` extension. The above will lead to four `dat` files, as `g` takes on four values 12, 16, 20, 24.

Next, each `dat` file produced is sent to another script called `core.sh` as shown:

core.sh

```
#!/bin/bash

cp "$1" data_t.tmp

# Sort and count unique entries.
while read -r
do
    echo "$REPLY"
done < data_t.tmp > tmpf
sort tmpf | uniq -c | sort -nr > data_t.tmp

rm tmpf

mv data_t.tmp $(echo "$1" | cut -f 1 -d '.').res
```

The job of `core.sh`, somewhat analogous to analyzing a histogram, is to take an incoming `dat` file to tally and sort its members by occurrence count from top to bottom. The most-popular elements are reported along with their respective occurrence counts in a file with a similar name having extension `res`.

Finally, the script `analyze.sh` takes each `res` file and keeps only the top 20 records as detailed:

```
#!/bin/bash

for i in data/*.res; do
```

```
    awk 'NR>20{exit} {print $0}' $i
done > result.txt
```

The output of `analyze.sh` is stored in another file, `result.txt`.

3.5 Results

Executing `main.sh` delivers `results.txt`, which is the only output the user need interact with. However, it's enlightening to do the job manually using results written with

the `res` extension. Sampling the top five records from each of these, we have:

`seq_abc_12_all.res`

```
13 ~play m@ke$.
13 y m@ke$.j@ck
13 w0rk_&n0~pl
13 rk_&n0~play
13 play m@ke$.j
```

`seq_abc_16_all.res`

```
13 ~play m@ke$.j@ck
13 y m@ke$.j@ck^@^D
13 w0rk_&n0~play m
13 rk_&n0~play m@k
13 play m@ke$.j@ck^
```

`seq_abc_20_all.res`

```
13 ~play m@ke$.j@ck^@^D
13 y m@ke$.j@ck^@^Dull
13 w0rk_&n0~play m@ke$
13 rk_&n0~play m@ke$.j
13 play m@ke$.j@ck^@^Du
```

`seq_abc_24_all.res`

```
13 ~play m@ke$.j@ck^@^Dull
13 w0rk_&n0~play m@ke$.j@c
13 rk_&n0~play m@ke$.j@ck^
13 play m@ke$.j@ck^@^Dull b
13 n0~play m@ke$.j@ck^@^Dul
```

Evidently, there is an ordered *something* that occurs 13 times in the original sequence. Since the list corresponding to $g = 24$ contains the same motifs as those with lower g -numbers, we proceed with the 24-case only.

Taking ten more records from the file `seq_abc_24_all.res` yields:

```
13 ll w0rk_&n0~play m@ke$.
13 l@y m@ke$.j@ck^@^Dull b0
13 l w0rk_&n0~play m@ke$.j
13 k_&n0~play m@ke$.j@ck^@
13 _n0~play m@ke$.j@ck^@^Du
13 _&n0~play m@ke$.j@ck^@^
13 @y m@ke$.j@ck^@^Dull b0y
13 @ll w0rk_&n0~play m@ke$
13 0~play m@ke$.j@ck^@^Dull
13 0rk_&n0~play m@ke$.j@ck
```

3.6 Final Analysis

While the final step could be automated, it's mostly harmless to work with the output of `seq_abc_24_all.res` directly to discern the hidden message in the original sequence. Playing a game akin to tree ring matching, we arrange the above information as follows:

```
~play m@ke$.j@ck^@^Dull
w0rk_&n0~play m@ke$.j@c
rk_&n0~play m@ke$.j@ck^
play m@ke$.j@ck^@^Dull b
n0~play m@ke$.j@ck^@^Dul
11 w0rk_&n0~play m@ke$.
l@y m@ke$.j@ck^@^Dull b0
1 w0rk_&n0~play m@ke$.j
k_&n0~play m@ke$.j@ck^@
_n0~play m@ke$.j@ck^@^Du
_&n0~play m@ke$.j@ck^@^
@y m@ke$.j@ck^@^Dull b0y
@ll w0rk_&n0~play m@ke$
0~play m@ke$.j@ck^@^Dull
0rk_&n0~play m@ke$.j@ck
-----
@ll w0rk_&n0~play m@ke$.j@ck^@^Dull b0y
```

Finally, we unambiguously see the phrase 'All work and no play makes Jack a dull boy', heavily stylized, occurs in 13 locations throughout the incoming sequence.

4 Binary Sequence Prediction

Formally, a *binary* sequence is one that consists of just two symbols, whether they be 0/1, or Heads/Tails, etc. The world is full of binary sequences whether they occur naturally via *stochastic* processes, or artificially in telephone lines and computer circuits.

A stochastic process is usually physical in nature, and could be as simple as a game of coin flipping, or as disorderly as the drops falling from a rainstorm. While stochastic events may have deterministic causes, the mechanism responsible is never evident in the data collected from the system. In effect, then, sequences arising from stochastic processes are randomized.

Given the ‘random’ versus ‘artificial’ dichotomy in binary sequences, we shall pursue two goals using sequence analysis:

1. Develop a process for discerning whether a binary sequence is random versus artificial.
2. Develop a tool that predicts the next digit in a growing binary sequence given the previous digits.

Review

To bite into the issue, consider the example sequence

$$S_{48}(\phi_2) = 101101001010101101100110101100010001110101011010.$$

As a binary sequence, we may invoke the the element vector

$$\begin{aligned}\phi_{2,2} &= \{00, 01, 10, 11\} \\ &= \{A, B, C, D\},\end{aligned}$$

and the sequence is written

$$\begin{aligned}S_{24}(\phi_{2,2}) &= CDBACCCDBCBC \\ &\quad CDABABDBBBCC.\end{aligned}$$

Going further with the basis vector

$$\begin{aligned}\phi_{2,4} &= \{0000, 0001, \dots, 1110, 1111\} \\ &= \{AA, AB, \dots, DC, DD\} \\ &= \{a, b, \dots, o, p\},\end{aligned}$$

the sequence is written

$$S_{12}(\phi_{2,4}) = (leklgglbbnfk).$$

Hopefully the above rings familiar as review. One new case previously skipped is $\phi_{2,3}$, for which we have

$$\begin{aligned}\phi_{2,3} &= \{000, 001, \dots, 110, 111\} \\ &= \{a, b, c, d, e, f, g, h\},\end{aligned}$$

and the corresponding sequence is

$$S_{16}(\phi_{2,3}) = (ffbcffegfecbgfcd).$$

4.1 Population Vector

Now we introduce the *population vector* ρ_M that, for a given sequence $^N S(\phi_M)$, tallies the total occurrences of each element in the sequence. The population vector is not a new concept - it’s essentially a histogram converted to a line of data.

Illustrating with the example on hand, the sequence $S_{24}(\phi_2)$ listed above has

$$\rho_2 = \{23_0, 25_1\},$$

which means there are 23 zeros and 25 ones in the sequence. For the next regrouping, $S_{24}(\phi_{2,2})$ is characterized by

$$\rho_{2,2} = \{3_A, 8_B, 9_C, 4_D\}.$$

Going further, the sequence $S_8(\phi_{2,3})$ is characterized by

$$\rho_{2,3} = \{0_a, 2_b, 3_c, 1_d, 2_e, 6_f, 2_g, 0_h\}.$$

Often, it’s more useful to organize ρ with the most popular entries first, which for the previous example means

$$\rho_{2,3} = \{6_f, 3_c, 2_e, 2_g, 1_b, 1_d, 0_a, 0_h\}.$$

Sorting by occurrence count, we can express more lengthy population vectors in abbreviated form, for instance

$$\rho_{2,4} = \{3_l, 2_b, 2_g, 2_k, 1_e, 1_f, 1_n, 0_a, 0_o, \dots\}.$$

Population after Translation

In light of element translation, it would make sense that a population vector $\rho_{x,y}$ will be different depending on the translation applied. To track this in the spirit of Equation (40.2), we include a superscript before the ρ symbol i.e. $^t \rho_{x,y}$. Illustrating using the running example, we have, for the $t = 1$ translations of the sequence:

$$\begin{aligned}^1 S_{48}(\phi_2) &= 011010010101011011001101 \\ &\quad 011000100011101010110101\end{aligned}$$

$${}^1S_{24}(\phi_{2,2}) = \text{BCCBBBBCDADB} \\ \text{BCACADCCDDBB}$$

$${}^1S_{16}(\phi_{2,3}) = (\text{dccfddbfaedfcgfg})$$

$${}^1S_{12}(\phi_{2,4}) = (\text{gjfgmngcdklf})$$

From these, we write the translated population vectors (the ρ_2 case remains the same). While sharing some resemblance with the non-translated population vectors, the fine details are varied from the originals:

$$\begin{aligned} {}^1\rho_2 &= \{23_0, 25_1\} \\ {}^1\rho_{2,2} &= \{3_A, 9_B, 8_C, 4_D\} \\ {}^1\rho_{2,3} &= \{5_d, 4_f, 3_c, 1_a, 1_b, 1_e, 1_g, 0_h\} \\ {}^1\rho_{2,4} &= \{3_g, 2_f, 1_c, 1_d, 1_i, 1_k, 1_l, 1_m, 1_n, 0_a, \dots\} . \end{aligned}$$

Grand Population Vector

The population vector allows us to symbolize a process used twice previously, particularly in (i) histogram stacking, and (ii) creation of a *res* file.

For a given regrouping number g , we reasoned that the translation number t , starting from $t = 0$, should never exceed $t = g - 1$. In other words, a sequence $S_N(\phi_M)$ regrouped as $S_{N/g}(\phi_{M,G})$ means we must also bring into consideration:

$$\begin{aligned} &{}^1S_N(\phi_{M,g}) \\ &{}^2S_N(\phi_{M,g}) \\ &\dots \\ &{}^{g-1}S_N(\phi_{M,g}) \end{aligned}$$

Meanwhile, each of the above implies a unique population vector ${}^t\rho_{M,g}$, namely:

$$\begin{aligned} &{}^1\rho_{M,g} \\ &{}^2\rho_{M,g} \\ &\dots \\ &{}^{g-1}\rho_{M,g} \end{aligned}$$

From a given $\phi_{M,g}$, we define the *grand population vector* to be

$$\sigma_{M,g} = \sum_{t=0}^{g-1} ({}^t\rho_{M,g}) .$$

If the vectors ${}^t\rho_{M,g}$ are *not* sorted from greatest to least occurrence, then $\sigma_{M,g}$ corresponds to a g -stacked histogram. If, on the other hand the vectors ${}^t\rho_{M,g}$ are arranged from greatest occurrence to the least, then $\sigma_{M,g}$ corresponds to what we saw in any *res* file in previous work.

To demonstrate, consider the population vectors $\rho_{2,2}$, ${}^1\rho_{2,2}$ written above, namely

$$\begin{aligned} \rho_{2,2} &= \{3_A, 8_B, 9_C, 4_D\} \\ {}^1\rho_{2,2} &= \{3_A, 9_B, 8_C, 4_D\} , \end{aligned}$$

that correspond to $S_{24}(\phi_{2,2})$, ${}^1S_{24}(\phi_{2,2})$ respectively. Summing across like elements, we easily write the grand population vector for this case:

$$\sigma_{2,2} = \{6_A, 17_B, 17_C, 8_D\}$$

A more illustrative case explores the vectors $\rho_{2,3}$, ${}^1\rho_{2,2}$, ${}^2\rho_{2,3}$, which requires knowing $S_{16}(\phi_{2,3})$, ${}^1S_{16}(\phi_{2,3})$, ${}^2S_{16}(\phi_{2,3})$. Two of these we have written already, and the third is

$${}^2S_{16}(\phi_{2,3}) = (\text{gef cggdcgbahc ffc}) .$$

From these, we derive three population vectors (two are review):

$$\begin{aligned} \rho_{2,3} &= \{0_a, 2_b, 3_c, 1_d, 2_e, 6_f, 2_g, 0_h\} \\ {}^1\rho_{2,3} &= \{1_a, 1_b, 3_c, 5_d, 1_e, 4_f, 1_g, 0_h\} \\ {}^2\rho_{2,3} &= \{1_a, 1_b, 4_c, 1_d, 1_e, 3_f, 4_g, 1_h\} \end{aligned}$$

Summing by column, the grand population vector is

$$\sigma_{2,3} = \{2_a, 4_b, 10_c, 7_d, 4_e, 13_f, 8_g, 1_h\} ,$$

or equivalently,

$$\sigma_{2,3} = \{13_f, 10_c, 8_g, 7_d, 4_b, 4_e, 2_a, 1_h\} .$$

We conclude that motifs like $f = 101$ and $c = 010$ are much more common in the original sequence than things like $a = 000$ and $h = 111$.

4.2 Partial Guess

Now we're at the heart of the problem. Keeping the same working example, consider again the sequence

$$\begin{aligned} S_{48}(\phi_2) &= 101101001010101101100110 \\ &\quad 101100010001110101011010 . \end{aligned}$$

If such a sequence were to grow by adding one digit on the right, thereby becoming, $S_{49}(\phi_2)$, can we predict what the digit will be?

The answer is reliably ‘no’ if the sequence is truly random. However, if there are prevailing patterns in the sequence, it should be possible to detect them, and then to look at the rightmost digits and guess intelligently what digit would continue the sequence.

Partial Guess

The ‘final guess’ X of the new digit is calculated using the original sequence and its various regroupings. For each regrouping $S_{N/g}(\phi_{M,g})$ of the sequence, there is a notion of *partial guess*, x_g . The ensemble of partial guesses constitutes the final guess. Symbolically, this means

$$S_{N+1}(\phi) = (S_N(\phi) X), \quad (40.3)$$

and X depends on each x_g in a way we’ll cover soon:

$$X = f(x_1, x_2, x_3, \dots, x_g)$$

Worked Example

Digging into the sequence on hand, and starting with the trivial regrouping $g = 1$ (no regrouping at all), we found

$$\sigma_2 = \rho_2 = \{23_0, 25_1\},$$

in which there is no sense of pattern beyond there being slightly more ones in the sequence than there are zeros. With this, we write

$$x_1 = 1.$$

For $g = 2$, look at the end of the original, non-translated sequence, $S_{48}(\phi_2)$. We reason the last digit in the sequence, namely 0, could be the first digit of some element in $\phi_{2,2}$ whose second digit is x_2 . Given the choices 00, 01, 10, 11, we see any element starting with 1 is ruled out, and we must decide between 00 and 01. The judge of the matter shall be the grand population vector $\sigma_{2,2}$, namely

$$\sigma_{2,2} = \{6_A, 17_B, 17_8, 4_D\} = \{6_{00}, 17_{01}, 17_{10}, 8_{11}\}.$$

The ratio 6:17 says the pattern 00 is less likely to occur than 01, thus the sequence ought to end with the pattern 01, and we have the partial guess

$$x_2 = 1.$$

For $g = 3$, consider the last two digits in the original sequence, namely 10. We reason that these could be the

first two digits in some element from $\phi_{2,3}$ whose third digit is x_3 . There are eight elements for $g = 3$, but the items we need begin with 10, and these are 100, 101. The population grand vector $\sigma_{2,3}$ for this situation was found to be:

$$\sigma_{2,3} = \{13_f, 10_c, 8_g, \dots\} = \{13_{101}, 10_{010}, 6_{110}, \dots\}$$

The pattern 101 is more common in the sequence than is 100, so we have

$$x_3 = 1.$$

For $g = 4$, we take the last three digits in the original sequence, namely 010. We reason that these could be the first three digits in some element from $\phi_{2,4}$ whose fourth digit is x_4 . There are sixteen elements for $g = 4$, but only $e = 0100$, $f = 0101$ start with 010. The partial guess x_4 is determined by whichever of e or f is most common in the sequence, i.e. which has a higher value in $\sigma_{2,4}$.

For brevity we’ll stop the manual analysis here. In practice, it’s useful to extend beyond $g = 4$, best left to something automated up to some cutoff $g_{\max}$.

Undecided Cases

In the event that the contenders for a partial guess x_g have the same population number, then assigning $x_g = 0$ or $x_g = 1$ doesn’t quite do justice. In such a case, we instead let

$$x_g = \frac{1}{2},$$

the average of the two.

Randomness

If it turns out that most (or all) of the partial guess values are undecided, it could be that (i) an insufficient range of g -values are being tested, or (ii) the sequence is ‘random’ as far as we can tell.

4.3 Final Guess

With all partial guess values $\{x_g\}$ in hand, the job is to assemble the final guess first symbolized in Equation (40.3). For this, we choose a weighted average scheme to combine partial guesses, namely

$$Y = \frac{1}{T} \sum_{g=1}^{g_{\max}} w_g x_g. \quad (40.4)$$

The unknown terms w_g are called *weights*, obeying normalization condition

$$T = \sum_{g=1}^{g_{\max}} w_g.$$

A given weight w_g tells much how much the corresponding partial guess x_g contributes to the final guess X .

The average Y is most generally a decimal value somewhere between 0 and 1. To finish the job, we must round Y to either 0 or 1, giving the final guess:

$$X = \text{Round}(Y) = \begin{cases} 0 & \text{if } Y < .5 \\ 1 & \text{if } Y \geq .5 \end{cases} \quad (40.5)$$

In this sense we've transformed the question of how to predict the next digit in a growing sequence into a question of choosing a proper right hand side for Y .

Weighting Models

A specific choice of coefficients $\{w_g\}$ is a *weighting model*, easily represented as a vector

$$W_q = (w_1, w_2, \dots, w_{g_{\max}}) .$$

The subscript q is unique to the specific choices w_1/T , w_2/T , etc. Only one weighting model is required to calculate the final guess $X(Y)$, but choosing the right model is an art unto its own.

4.4 Guess-Reveal Algorithm

Reviewing binary sequence prediction in one sentence, we take an incoming binary sequence $S_N(\phi_2)$ and, based on any discernible order within the sequence, calculate the form of $S_{N+1}(\phi_2)$ given a weight model W_q .

Given an incoming sequence, consider now a truncated version

$$S_{n < N}(\phi_2) = (S^1 S^2 \dots S^n)$$

whose members match the first n members of the original sequence. Now devise a 'game' where, given the truncated sequence, predict the next digit using a fixed model W_q . With the guess on hand, reveal the next digit in the actual sequence and compare the guess to reality. From this we keep score in the form of win/loss ratio, winning streak, etc. The longer the incoming sequence, the more data we generate.

When the sequence is finished, i.e. when the guessing game has been played on each truncation of the incoming sequence, the model's performance is summarized.

4.5 Model Space

While there are technically an infinite number of weighting models one could sift through, we tame the problem by having each w_g be limited to 0 or 1, i.e. a 'binary' weighting

model. With this, every W_q -vector takes the form of a binary sequence, and for convenience we convert the sequence to base ten and assign that number to q :

$$\begin{aligned} W_0 &= (0\ 0\ 0\ 0\ \dots) \\ W_1 &= (1\ 0\ 0\ 0\ \dots) \\ W_2 &= (0\ 1\ 0\ 0\ \dots) \\ W_3 &= (1\ 1\ 0\ 0\ \dots) \\ W_4 &= (0\ 0\ 1\ 0\ \dots) \\ &\dots \end{aligned}$$

The total number of weighting models is determined by $g_{\max}$, and for the sake of demonstration we choose $g_{\max} = 10$. This invokes a total of $2^{10} = 1024$ models in total, with the last being

$$W_{1023} = (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1) .$$

The term *model space* refers to the set of all models $\{W_q\}$ that could be used to calculate $X(Y)$, where for completeness, Y is given by Equation (40.4), and X is given by Equation (40.3).

By choosing a binary model system, the infinitude of available models reduces to $2^{g_{\max}}$ discrete points with the argument that enough of model space is covered. Moreover, a systematic stepping through model space is as easy as moving among numbers from 0 to $g_{\max}$ in binary.

4.6 Automation

All of the above-described processes can be abstracted and packaged into a suitable program that operates on any incoming binary sequence. For this study, a *QB64* source file called *Binary-Analyzer.BAS*, which compiles to *Binary-Analyzer.exe*, has been prepared to do so. All exact results will come from this one program.

A precise implementation of the so-called 'automated guess-reveal algorithm' is attached in the aforementioned *BAS* file.

Model Hunt Mode

For a given sequence $S_N(\phi_2)$, the program must (i) open an 'outer loop' for stepping through model space (in total or in part), and also (ii) open an 'inner loop' that steps through each truncation of the given sequence.

Once we're two loops deep, the guess-reveal algorithm gets to play. That is, a 'final' guess is calculated for whichever truncation we're on, i.e.

$$S_{n+1}(\phi) = (S_n(\phi)\ X) ,$$

and then X is compared with the *actual* $S_{n+1}(\phi)$ as it's given. From here we gather whether the model was right or wrong, along with other statistics, and the final output should be a single model number q that best matches the sequence.

A key output from so-called 'model hunt mode' is the winning model's *correctness rating*, denoted C_R , which is a simple ratio of the number of times the guess-reveal algorithm yields correct results throughout the total number of attempts:

$$C_q = \frac{\text{number correct}}{\text{total attempts}} \quad (40.6)$$

Oracle Mode

A second running mode can also be prepared for interaction with a human (or similar stochastic system) in real time. In particular, the user is asked to manually input the ones and zeros of a handmade binary sequence, and the guess-reveal algorithm is applied and reported with each keystroke.

Like an oracle, the computer should be able to attain a higher correctness rating in real time than would be achieved by randomly guessing.

5 Binary Case Studies

5.1 Random Binary Sequences

Suppose now that we take a binary sequence known to be random. Like a fair coin-tossing game, we know that there should, on average, be an equal number of heads and tails as the game goes on. For N total trials of such a system, we can write the *standard deviation*

$$\sigma_N = \frac{1}{2}\sqrt{N},$$

giving a sense of what's reasonable for off-average trends.

For example, in a case of $N = 100$ trials, i.e. a binary sequence of length 100, we write $\sigma_{100} = \sqrt{100}/2 = 5$. That is, after a game of 100 tosses, we can often expect off-average outcomes to be biased only by about 5 units. Or, taking $N = 1000$, the standard deviation comes out to approximately 16.

Introducing another definition, the so-called *coefficient of variation*, denoted C_V , divides σ_N by the sample size:

$$C_V = \frac{\sigma_N}{N} = \frac{1}{2\sqrt{N}}$$

That is, for $N = 100$ we have $C_V = 5\%$, and for $n = 1000$ we have $C_V \approx 2\%$.

If a sequence is to be classified as 'random', then the correctness rating C_q and C_V relate by

$$C_q \approx 50\% + C_V.$$

Just to be sure, let's use $C_q > 55\%$ as a clear indicator of order in a sequence that is at or more than 100 members in length.

QB64 RND Function

A binary sequence of $N = 2000$ digits has been prepared using the RND function built into QB64. Running this sequence through model hunt mode with q running from 0 to 1023 yields the following information:

- The best model is $W_{66} = (100010\dots)$, rated 52%, with a longest streak of 13.
- Comparable to best-performing are W_{71} , W_{83} , W_{85} , W_{130} .
- The worst model is $W_1 = (100000\dots)$, rated 48%.

With the best and worst correctness ratings being 52% and 48% respectively, it's reasonable to conclude that sequence analysis finds no discernible order in the 2000-digit binary sequence generated by QB64. (A relief for both parties.)

Wolfram Rule 30

In the 1980s, researcher Stephen Wolfram was studying cellular automata using sets of rules encoded in binary, and was surprised by the result of so-called Rule 30, depicted in Figure 40.6.

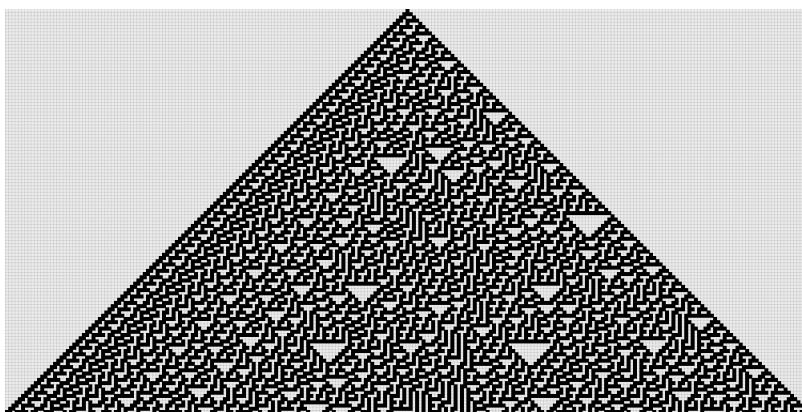


Figure 40.6. Wolfram's Rule 30.

Wolfram contends that the central column of alternating black/white squares constitutes a random binary sequence. Using an in-house utility to generate and extract $N = 2000$ digits down the middle column, we apply sequence analysis to find:

- The best model is $W_2 = (010000\dots)$, rated 51%, with a longest streak of 11.
- Comparable to best-performing are $W_2, W_4, W_5, W_7, W_{14}, W_{22}, W_{66}, W_{68}, W_{70}, W_{195}, W_{262}, W_{323}, W_{387}$.
- The worst model is W_{416} , rated 46%.

There may or may not be a prize (still) available for anyone who can prove that the central column in the Rule 30 image is not a random sequence. One way or the other, we find no order by sequence analysis.

Stock Markets

It is widely claimed that the world stock market is generally unpredictable, so this ought to be evident by sequence analysis (or get rich if not). For a source of data, it turns out that *Yahoo Finance* freely avails historical information for participating entities. A query was made for DIS, short for ‘Disney’, which returned a table of opening and closing prices for business dates spanning 08/18/2022 to 08/18/2023.

Converting opening and closing prices to binary data goes as follows: if today’s price is higher than yesterday’s price, write 1. Otherwise, write 0. Doing precisely this for the DIS data produces a sequence with 252 members. For completeness, this sequence reads:

```

 $S_{252}(\phi_2) = 000011010010011110100100$ 
 $000010011100001101001101$ 
 $101100000001100000001100$ 
 $111000111000001011001101$ 
 $111111100111111101110010$ 
 $000110101110001100100101$ 
 $101101010101001110010110$ 
 $110001100001100001011010$ 
 $000011101010100100011110$ 
 $111011011000110100101100$ 
 $011101000001$ 

```

Running the above through the sequence analysis meat grinder leads to the following summary:

- The best model is W_{546} , rated 55%, with a longest streak of 7.
- Comparable to best-performing are $W_{562}, W_{564}, W_{630}$.
- The worst model is W_8 , rated 47%.

In other words, a stock’s rising and falling is as random as it gets.

5.2 Percussive Sequences

A *percussive* binary sequence is one that has obvious regularity, as if generated with intent by a drummer. In light of sequence regrouping, percussive sequences are easily spotted as one or a few repeated elements from some element vector $\phi_{2,g}$.

A percussive sequence may embed several aspects of repetition, and the sequence as a whole eventually repeats itself. The period of repetition, i.e. the number of digits that pass by before the sequence repeats, is called the *period*. To denote a percussive binary sequence we’ll use the nomenclature $P_{T,N}$ with P replacing the letter S , and the subscript also carries the period number T .

Uniqueness

Only two period-1 sequences are possible in binary, and these are:

$$P_{1,N}(\phi_2) = 000000000000000000000000\dots$$

$$P_{1,N}(\phi_2) = 1111111111111111111111111111\dots$$

As for $P_{2,N}$, there are four total percussive sequences, but two of them are redundant to the above and are excluded:

$$P_{2,N}(\phi_2) = 01010101010101010101010101\dots$$

$$P_{2,N}(\phi_2) = 10101010101010101010101010\dots$$

However, note the pair of sequences $P_{2,N}(\phi_2)$ are identical up to a $t = 1$ translation. A proper sequence analysis would treat each version roughly the same.

By similar reasoning, there two unique percussive sequences explicit to period three:

$$P_{3,N}(\phi_2) = 011011011011011011011011011\dots$$

$$P_{3,N}(\phi_2) = 001001001001001001001001001\dots$$

Seed

Percussive sequences can be expressed more concisely by only stating the minimum number of digits needed to reproduce the sequence. The span of digits is called the *seed*, denoted D_T . The length of the seed is exactly one period T , and the seed can be sampled from anywhere in the sequence. For this reason, translations of a seed are still considered the same seed.

Listing the seed for the above cases plus one more, we have

$$\begin{aligned} D_1 &= \{0, 1\} \\ D_2 &= \{01\} \\ D_3 &= \{001, 011\} \\ D_4 &= \{0001, 0011, 0111\} . \end{aligned}$$

For the D_5 seeds, we not only have

$$D_5 = \{00001, 00011, 00111, 01111, \dots\} ,$$

but also the concatenations of D_2 with D_3 , namely

$$00101, 01011 .$$

Convergence Time

Percussive sequences with a period $T \leq g_{\max}$ are picked up swiftly by sequence analysis, and all nontrivial models in $\{W_q\}$ (not all zeros or not all ones) will lead the guess-reveal algorithm toward a perfect correctness rating.

Given a percussive sequence $S_N(\phi_2)$ with any model W_q , define the *convergence time* τ_q as a count of how many iterations, i.e. how many guess-and-reveal rounds it takes for the sequence analyzer to completely stop making mistakes. This is a rather high bar for any prediction tool, but for low-period percussive sequences we can almost always shoot for 100%.

For long-period cases, i.e. $T \geq g_{\max}$, the sequence still exhibits plenty of order, yielding a high correctness rating, however convergence time may not be evident.

Measurement: D_5

Consider a percussive binary sequence $S_{240}(\phi_2)$ that is based on a D_5 seed, particularly

$$S_{240}(\phi_2) = (01011 \dots)$$

with 240 total members. Running the sequence through so-called model hunt mode in the sequence analyzer gives

information on the performance on each model W_q tested, where q runs from 0 to 1023.

Reporting results in a way analogous to the random sequence study above, we have:

- The best model is W_{16} , rated 97%, with a longest streak of 226.
- Comparable to best-performing are W_{17} , W_{19} , W_{21} , W_{49} , W_{50} , W_{51} , W_{53} , W_{56} , W_{57} , W_{59} , W_{61} , W_{65} , W_{67} , W_{80} to W_{85} , W_{87} , W_{88} , W_{89} , W_{91} , W_{93} , W_{96} , W_{98} and plenty more.
- The worst model is W_1 , rated 58%.

Evidently, it's hard to *not* find the clear order in the given sequence, as even the worst model performs well above 50%. The best model W_{16} is able to correctly predict the next digit via the guess-reveal algorithm in 97% of cases, with a longest continuous 'win' streak of 226.

Subtracting the 'longest streak' number from the total sequence length indicates the convergence time of the best model:

$$\tau_{16} \approx 240 - 226 = 14$$

Note that we define the 'best' model by that which has the highest correctness rating C_q . In certain cases, it's possible for a model without the highest C_q have the longest correctness streak.

Measurement: D_{10}

Upping the ante slightly, consider another percussive binary sequence $S_{240}(\phi_2)$ that is based on a D_{10} seed, particularly

$$S_{240}(\phi_2) = (0001110011 \dots) .$$

Running the sequence through model hunt mode gives information on the performance on each model W_q tested, where q runs from 0 to 1023.

Reporting results in a way analogous to the random sequence study above, we have:

- The best model is W_{162} , rated 97%, with a longest streak of 221.
- A second-best model is W_{38} , rated 96%, with a longest streak of 206. Comparable to this are W_{98} , W_{102} .
- The worst model is W_3 , rated 30%.

Such a sequence clearly ordered as evident by the performance of the best model W_{162} along with its close contenders W_{38} , W_{98} , etc. From these we see the convergence time to be roughly

$$\tau_{162} \approx 240 - 221 = 19.$$

Interestingly, the worst-performing model W_3 gives a correctness rating of 30%, which, despite being an inverted answer, still tells us the sequence is ordered.

Measurement: D_{30}

A convenient pattern that seems to recur throughout the percussive example above is that, in both cases, the best model W_q is found long before the full range of q is traversed. From this we glean that low-order analysis with $g_{\max}$ set modestly to 10 can perform enough to achieve finite convergence time. Going for an extreme example, consider another percussive binary sequence $S_{240}(\phi_2)$ that is based on a D_{30} seed, particularly

$$S_{240}(\phi_2) = 010011000111000011110000011111 \dots$$

Despite having a seed length of $T = 30$, which is three times $g_{\max}$, the analysis yields:

- The best model is W_{770} , rated 89%, with a longest streak of 175.
- Comparable to best-performing is W_{772} .
- The worst model is W_3 , rated 34%.

Almost astonishingly, model W_{770} manages a correctness ratio of 89%, and achieves a convergence time after approximately

$$\tau_{770} \approx 240 - 175 = 65$$

digits.

5.3 Human Samples

Now we're ready for a fun and gratifying use of sequence analysis. A handful of individuals have volunteered hand-made binary sequences by pressing 0 and 1 on a keyboard 'as randomly as possible'. Samples submitted vary in length.

Data

A total of ten samples were collected from several individuals. Labeling the samples alphabetically, we have:

$A = S_{1023} = 1010101010101010101001 \dots$
 $B = S_{511} = 010111010110001101010010 \dots$
 $C = S_{439} = 101100101001010010010011 \dots$
 $D = S_{311} = 011001010010100011000011 \dots$
 $E = S_{299} = 101010101010111010000111 \dots$
 $F = S_{299} = 101110101010101010101010 \dots$
 $G = S_{196} = 111110111101001010111111 \dots$
 $H = S_{314} = 010101000110010100101010 \dots$
 $I = S_{299} = 110110110111010110100010 \dots$
 $J = S_{343} = 101010001010101011001100 \dots$

Testing

Each of the human-provided samples A to J is exposed to the full sequence analysis apparatus: translation, regrouping, grand population vector - the whole bit. For each sample, all we're shooting for is to find a model W_q that delivers a correctness rating of more than 55%, and then we'll know for sure that the sequence isn't inherently random.

It turns out our task is relatively easy. Alas, instead of running through all weight numbers $q = 0$ to $q = 1023$, it suffices to test a range much narrower, and for this we test models from $q = 15$ to $q = 64$.

Results

Letting a program do the hard work, the following results emerge:

Sample	Length	Best q	Correct	Streak
A	1023	18	86%	24
B	511	27	65%	11
C	439	18	78%	17
D	311	26	60%	12
E	299	24	69%	16
F	299	15	90%	56
G	196	19	70%	12
H	314	18	70%	14
I	299	20	61%	9
J	343	18	87%	38

From the table, we observe that, by far, no sample provided is 'random enough' to fool the machine, despite having sampled a small subset of model space. Keep in

mind the number to beat is 55% for a sample length of 100, so the bar is already high.

The most challenging cases are arguably D and I , and perhaps there are models beyond $q = 64$ that better fit these. Whoever typed out sequence F probably wasn't trying.

5.4 Pathological Sequences

Let us play with a strange idea. Take any weighting model W_q , and consider a sequence of length one, containing just a single zero:

$$S_1(\phi_2) = (0)$$

Now, apply half of the guess-reveal algorithm, just the 'guess' part, to this trivial sequence. Based on the existing zero in the sequence, the prediction will be a second zero, i.e.

$$S_2(\phi_2) = (0 \ X^2),$$

with $X^2 = 0$.

However, let's do the opposite of what the algorithm predicts, and assign $X^2 = 1$ instead. Then, starting with the sequence

$$S_2(\phi_2) = (0 \ 1),$$

let the algorithm predict the third digit, and then disobey the prediction. Whatever the prediction X^3 is, append the sequence with its opposite Z^3 :

$$S_3(\phi_2) = (0 \ 1 \ Z^3)$$

Carrying this process on, we end up with a sequence that is tuned to perfectly outsmart the model W_q . Such a sequence is *pathological* to the model, and when tested against W_q , yields a correctness rate of 0%. Once a sequence is known to be pathological, we change nomenclature to Z , replacing S :

$$S \rightarrow Z$$

Seed and Evolution

The first members of a pathological sequence determine the evolution of the rest. For this reason the first members are called the *seed*. For simplicity we choose a one-member seed, namely 0. Importantly, choosing a seed of 1 instead, let alone a more complicated one, alters all results that follow.

Also fundamental to a pathological sequence is the range of regroupings used to analyze the seed characterized by $g_{\max}$.

Examples: Low- q

Pathological sequences can take surprising form, even for low- q weighting models. Going through the first few non-trivial cases, we find the following:

$$Z_1 = 01101010101010101010101010101010 \dots$$

$$Z_2 = 011011100010111000101110001011100010 \dots$$

$$Z_3 = 011010011001100110011001100110011001 \dots$$

$$Z_4 = 011110110100111100010110100010111100 \dots$$

$$Z_5 = 011100101110100010110010111100001110 \dots$$

$$Z_6 = 011100101110100011100010111010001110 \dots$$

$$Z_7 = 011010011110001010111000110010111000 \dots$$

$$Z_8 = 011111011110110011101010110011111010 \dots$$

$$Z_9 = 011011100110101000101111100000101110 \dots$$

Examples: Larger q

Higher q -numbers lead to equally mysterious pathological sequences:

$$Z_{254} = 011110111111010110110011100011101101 \dots$$

$$Z_{255} = 011011111010111001111100011010010100 \dots$$

$$Z_{256} = 01111111111011111111011110111011111 \dots$$

Ring and Resonance

Peeling apart a given pathological sequence Z_q , there are evidently two kinds of behavior, one we'll call *ring*, and the other, *resonance*. The term 'ring' refers to qualitatively unordered behavior, exhibiting randomness or something without order. On the other hand, 'resonance' means the sequence takes on a percussive nature.

For instance, look again at Z_2 , and write the sequence as

$$Z_2 = (011) + (01110001) + (01110001) + \dots$$

In the new terminology, the first three members of Z_2 grouped on their own as 'ring' terms, and all other members occurring as resonating chunks of period $T = 8$.

The transition from ring to resonance, and perhaps back to ring, and perhaps a blend of the two, is not obvious in the general case.

```

1111001101011010111011010101011011001011
1110111110110110100110110111000110111010
0100111100100111110000111110101011100101
0011101101011111001011111110001111110011
011111111101111110110111110111011110111
0111011000111101000111111001001111001010
1001110011010110011011010011011110011101
1111110101010110101011110101101110101110
0101100101001100101101000101101101001001
0011000111001000111010000111011001101010
0000011110010011110100011110110101010010
1001011101101100001101110001101100100101
1011101111011101011011010010011010011011
001011111010011111011111100011111100111
10100011010101001010101111010111100000
1101111010000111010001110010001111010100
0111110001100110001101110001110110001111
100110101101110101110110101111100101111

```

Figure 40.7. Text editor view of Z_{256} (partial sequence).

To illustrate, the sequence Z_{256} started above was generated to about 3000 members and inspected with a text editor availing a spectacle as seen in Figure 40.7. A glance at the arrangement of digits shows *some* kind of pattern. It just happens that model W_1 , along with many others, can manage a correctness rating of 57% for the Z_{256} sequence.

5.5 Taxonomy of Correctness

Having tested a whole zoo of binary sequences by now, we've seen, for each case, the correctness rating for any given model falling somewhere between 0% and 100%.

Here we list five distinct 'correctness zones' based on Equation (40.6):

- $C_q = 100\% \rightarrow$ **Percussive** zone.
- $55\% < C_q \leq 99\% \rightarrow$ **Bias** zone.
- $45\% \leq C_q \leq 55\% \rightarrow$ **Random** zone.
- $1\% \leq C_q < 45\% \rightarrow$ **Transcendental** zone.
- $C_q = 0\% \rightarrow$ **Pathological** zone.

Any binary sequence incoming from the wild is, depending on its discernible order, classified as either percussive, biased, or random.

Transcendental Zone

It's mighty peculiar that we are able to speak of correctness ratings significantly lower than 50%, like a broken clock being correct more than twice a day. There are (at least) two ways of getting into the so-called *transcendental* zone.

- Send a pathological sequence Z_q through guess-reveal based on a different associated model $W_{r \neq q}$.
- Manually edit the first (few) digit(s) of a pathological sequence Z_q and re-evaluate using W_q .

That is, only sequences that come *from* the guess-reveal algorithm are able to enter the transcendental zone. Funny enough, any correctness rating significantly away from 50%, above or below, is enough to know the sequence is ordered.

6 Creating a PRNG

We learned from pathological sequences that the guess-reveal algorithm can be used in reverse to generate sequences instead of analyze them.

Here we engage in a new question: Can the same toolkit be used to create a *pseudo random number generator* (in binary, of course)? This is abbreviated PRNG, and the prefix ‘pseudo’ means, for our purposes, that the output is completely deterministic, i.e. running the same program twice gives the same results twice.

Multiplicity of Output

The output of the PRNG is determined by its seed value. For our purposes, a binary seed value D_{10} of 10 digits, relating to $g_{\max} = 10$, shall kick-start the PRNG. The seed is treated as an internal variable, which is to say that the seed does not automatically appear in the resulting sequence. There are 2^{10} total seeds. Right away, we see that there are 2^{10} possible ways to begin a pseudo-random sequence.

6.1 Initializing the Sequence

To prepare a pseudo-random binary sequence, first imagine two new variables:

K = temporary working sequence
 R = return value

The variable K is initialized to store the incoming seed, and the variable R is initialized as empty:

$K = D_{10}$
 $R = (\text{empty})$

6.2 Main Loop

Now the main loop is ready to go:

1. Isolate the last (rightmost) ten members in K . Convert this base-two number to base-ten, storing in q .
2. Initialize the weight model W_q .
3. Run the guess-reveal algorithm on K , storing the result in X .
4. Invert X (0 becomes 1 or vice versa) and store in Z .
5. Append Z to the right side of K and R .
6. Right-truncate K to main fixed length.

On a technical note, it helps to keep K limited to a few dozen characters, keeping perhaps the latest 40. This keeps the guess-reveal algorithm on its heels by denying the full sequence history from influencing later evolution.

Repeating the above steps N times places N members into the return sequence R_N .

6.3 Testing

To check the proposed PRNG’s quality, we use the same apparatus used for testing the ‘natural’ cases, i.e. sequences known already to be orderless. Checking various issues of R_N using different seeds and different sequence lengths tends to yield results characterized by

$$C_q \approx 48\% \pm 5\% ,$$

hovering nicely near the so-called ‘random’ zone.

Now, due to a possible a rubber-ruler argument, we won’t read *too* much into whether the proposed PRNG produces quality output, particularly because we’re testing with the same tool that forms the PRNG itself. Nonetheless it seems to work.

Chapter 41

Epicycles

1 Introduction

Intellectuals have been fascinated by idea of the *circle* for centuries. The cliché *invention of the wheel*, originating somewhere in Europe or China, sparks the image of rolling transportation as the first useful implementation of the circle. Equally unclear is precisely where and when circles were first used for grinding tools, milling equipment, architecture, writing systems - the list goes on and on. In any case, most cultures seem to discover the ‘circle’ for themselves and put it to various uses.

This study contains a brief survey of the historical use of circles for scientific or practical purposes. We then apply modern methods, namely the discrete Fourier transform, to develop a general-use curve-sketching apparatus by means of epicycles.

1.1 Ismail Al-Jazari

An abstraction of the wheel, namely the *cogwheel*, or the *gear*, is at the center of virtually every invention since the eighteenth century. Before then, the first conceptualization of using cogwheels in machines is credited to Arab engineer Ismail Al-Jazari, whose career peaked around the year 1206.

Quite ahead of his time, Al-Jazari mastered a cross section of what we’d now call Newtonian mechanics, namely the ideas of leverage, mechanical advantage, gear ratios, and hydrostatics to name a few. Figure 41.1 shows one of Al-Jazari’s devices depicted in his book, *The Book of Knowledge of Ingenious Mechanical Devices*. From his works, Al-Jazari clearly understood why inter-moving wheels can give rise to complex and precise outputs in a given machine.

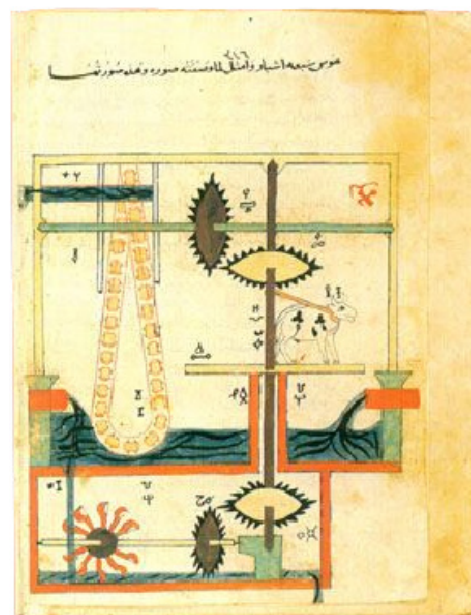


Figure 41.1. Hydro-powered saqiya chain pump device. Illustrated in Al-Jazari’s *The Book of Knowledge of Ingenious Mechanical Devices*, 1206.

1.2 Antikythera Mechanism

As if Al-Jazari wasn’t a historical enigma, leave it to the ancient Greeks to destroy all ancient stereotypes. A ship that sank sometime around 200 B.C. was carrying something now known as the *Antikythera mechanism*, depicted in Figure 41.2, named after a neighboring island. The Antikythera mechanism is nothing short of a full-blown *analog computer* based on connected cogwheels that ‘calculates’ the goings of the Sun, Moon, and other celestial objects of ancient Greek astronomy. At least a thousand years ahead of its

time, and more complex than a wristwatch, the Antikythera mechanism demonstrates the Greeks' mastery of the wheel not just for easing repetitive farming tasks, but for handling *information* in mechanical form.



Figure 41.2. Antikythera Mechanism. Photo released by the National Archaeological Museum in Athens, 2005.

1.3 Archimedes and Pi

Suspiciously contemporary (or nearly so) to the Antikythera mechanism was the career of Archimedes (BC 287 - 212), famous Greek mathematician and inventor. With regard to circles, ancient philosophers were already aware the circumference of a circle relates to its diameter in a fixed ratio, but Archimedes was the first to actually calculate this ratio rigorously, naming his result π .

While it is unclear whether Archimedes had anything to do with the Antikythera mechanism, it's undeniable that circular motion was of enormous technical and theoretical interest to Greek intellectuals. It wouldn't be long before the circle was hailed as the 'divine' shape. This idea spread, perhaps *too* thoroughly, all through the elementary sciences used by the Greeks.

2 Universe of Ptolemy

To accompany the mechanical and mathematical achievements of the ancient Greeks, efforts were made to crack the problem of the mysterious motions of heavenly bodies. Unlike today, the ancients didn't (rather, couldn't) differentiate the workings of the local solar system - planets, moons, comets - from the wider universe of stars. To the Greeks, the Earth was the center of *all* things, and every motion in the night sky was set for the sake of Earthly drama. This, of course, is known as the *geocentric model* of the universe.

2.1 Ptolemaic System

Seeking an explanation for the apparent 'perfect' motions observed in the heavens, Greek philosopher Ptolemy (AD 127-145) continued the assumption that the Earth remain fixed in space as the center of the universe. Then, measuring up from the ground, he supposed that the known celestial bodies - Moon, Mercury, Venus, Sun, Mars, Jupiter, Saturn - in that order, each moved on its own sphere concentrically surrounding the Earth. Depicted in Figure 41.3 is an adaptation of the Ptolemaic model as 'perfected' for theological purposes.

As a refinement of the geocentric model, Ptolemy's 'concentric spheres' upgrade also takes care of the stars: since the stars in the night sky appear mutually fixed, Ptolemy reasoned the stars to all be embedded the same outermost sphere with Earth precisely at the center. This notion is still of practical use today, known by modern astronomers as the so-called *celestial sphere*.

Ancient astronomers also knew the apparent arc of a celestial object's motion can be deduced over time by measuring its position at the same moment each night against the background of fixed stars. This is undoubtedly how the first few planets were discovered, and leads to why the standard year is partitioned into *months*, corresponding to the rough 30-day period of the Moon. The Sun is no exception to this pattern, for its motion too is apparently circular against the background of fixed stars, centered on Earth.

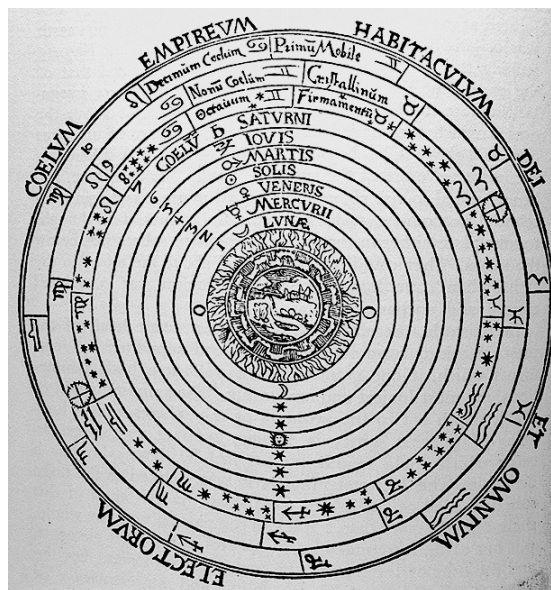


Figure 41.3. Engraving from Peter Apian's *Cosmographia*, depicting the *Christian Aristotelian cosmos*, 1524.

2.2 Planetary Symbols

Embedded in Figure 41.3 are the contemporary astronomical symbols used for the celestial bodies. For completeness, these are:

- ☾ = Moon
- ☿ = Mercury
- ♀ = Venus
- ☼ = Sun
- ♂ = Mars
- ♃ = Jupiter
- ♄ = Saturn

Interestingly, the symbol for the Sun includes a central dot, which is to denote the existence of *sunspots*, a well-known solar phenomena to the ancients.

2.3 Retrograde Motion

The 'measure-each-night' trick does not work out so beautifully for planets, with Mars being a particularly striking example. Every two or so years, Mars seems to deviate from its proper arc as seen from Earth. Strangely, Mars goes *backward* against the night sky for several months, and then course-corrects to resume its normal path, completing a kind of figure-eight, or Z-like curve. The apparent backtracking

of Mars, also exhibited by the other planets, became known as *retrograde motion*.

Since the original data used by Ptolemy isn't forthcoming, we can use present-day data to approximate the view shared by the ancient Greeks. Depicted in Figure 41.4 is the retrograde motion of Mars from Winter 1964 to Summer 1965 as seen from Earth by NASA. Faced with a similar observation, Ptolemy knew that retrograde motion would be the bitter end for a sphere-based a model of the cosmos, but this just *couldn't* be the case, because circles are supposed to be perfect. It would have been considered barbaric to propose new orbital shapes, such as ellipses, or to introduce any new element beside the circle. Ptolemy insisted that it be circles all the way down.

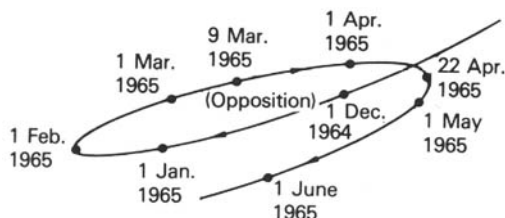


Figure 41.4. Retrograde motion of Mars. Originally in the *Book of Mars*, NASA SP-179 (1968).

2.4 Epicycles

To wiggle his way out of the retrograde motion problem, Ptolemy supposed that Mars may not be 'locked in place' on its respective concentric sphere, but perhaps Mars also moves along a smaller circle embedded on the greater circle. The 'smaller circle' is called an *epicycle*, and the 'greater circle' is called the *deferent*. It turns out that adding epicycles to the geocentric model is the perfect ingredient to account for all celestial motions, including retrograde motion of the known planets. This refinement to the geocentric theory became known as the *Ptolemaic model*.

Shown in Figure 41.5 is an 'early' reproduction of Ptolemy's work on epicycles, published in the *Almagest*, published in roughly 150 A.D. Inspired by the Greeks, many astronomers took the Ptolemaic model as a starting point to work out the epicyclic motions of Mars, Mercury, and so on. Figure 41.6 depicts one such study that tracks the evident position of Mars over a sixteen-year interval.

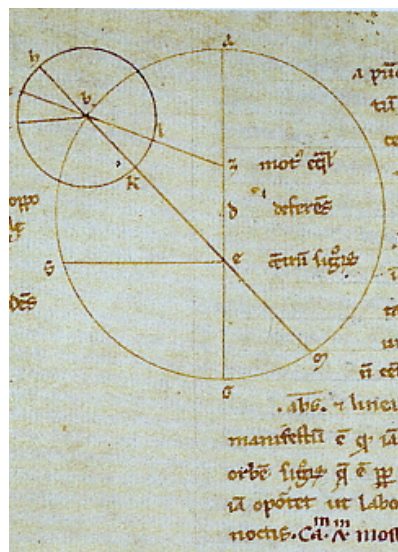


Figure 41.5. Claude Ptolemy, *Almagest*, translated into Latin by Gerard of Cremona, 13th century. Vatican, Biblioteca Apostolica Vaticana, Lat. 2057.

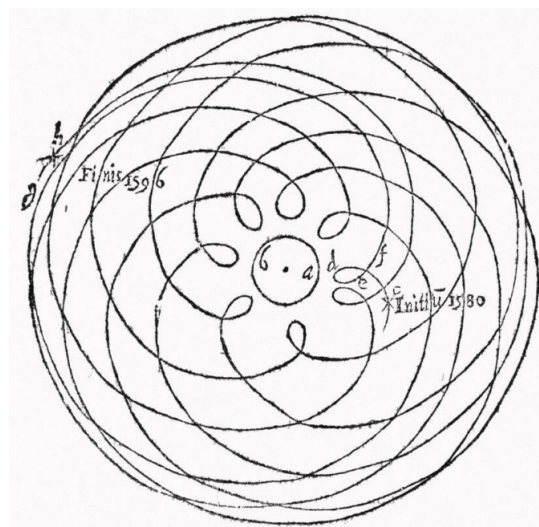


Figure 41.6. Epicycles of Mars as observed from Earth between 1580 and 1596. From *Sphaera mundi* (1635) by Italian astronomer Giuseppe Biancani (1566-1624). Released by Middle Temple Library (2016).

The Ptolemaic model also took the interest of Arab astronomer Ibn al-Shatir (1304-1375). Working in a looser spirit than the Greeks, Ibn al-Shatir wasn't so worried about the divinity of circles, and freely added an extra epicycle called the *Tusi-couple*, simplifying the theory. Depicted in Figure 41.7 is Shatir's explanation of the orbit of Mercury using epicycles. Among achievements similar to this, his

works were collected in *kitab nihayat as-sul fi tashih al-usul*, reanslating to ‘A Final Inquiry Concerning the Rectification of Planetary Theory’.

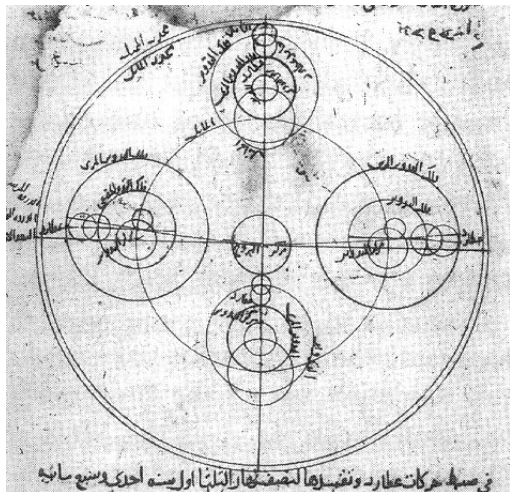


Figure 41.7. Ibn al-Shatir’s explanation of the orbit of Mercury using epicycles. Wikimedia Commons.

Remarks

To get the most use from the Greek’s new-found the ability to account for celestial motions, with epicycles affirming that the circle really *is* the divine shape at the heart of cosmic motion, the ‘circle bias’ drilled a little too deeply into the ancient worldview, and would eventually act to hinder progress in astronomy and other pursuits.

The Ptolemaic model wouldn’t be successfully challenged - and only barely so - until the days of Copernicus (1543) and Galileo (1610), who dared to suggest that the Sun was the center of the solar system. Zooming forward a bit more, it’s been clear since the days of Newton (1642-1727) that circles are *not* the key to understanding orbits in the solar system.

The dismissal of epicycles from astronomy, however, does not mean the subject be condemned as a historical relic. It turns out that the intermediate discussion behind epicycles requires a tour of duty through early college-level mathematics. From the comforts of modern notation and understand can we uniquely appreciate the mystery that kept the Greeks so intrigued.

3 Drawing Continua

Epicycles are naturally suited to sketching curves, usually closed, clover-like shapes with a fair degree of symmetry. It turns out that, with enough cleverness, epicycles can be used to trace many families of curves.

3.1 Index Notation

At any time t , the so-called ‘drawing tip’ in the epicycle construction is somewhere in the Cartesian plane. Using standard trigonometry, it’s easily seen that the curve traced out by N epicycles is given by:

$$x(t) = \sum_{j=1}^N r_j \cos(\phi_j + \omega_j t)$$

$$y(t) = \sum_{j=1}^N r_j \sin(\phi_j + \omega_j t)$$

In the above, r_j is each radius, ω_j is each angular frequency, and ϕ_j is each initial phase. For a curve represented by N epicycles, all of the information is contained in the table:

$$\{r_j, \omega_j, \phi_j\},$$

where

$$j = 1, 2, 3, \dots, N.$$

3.2 Horizontal Lines

Straight lines through the origin are exceedingly simple to draw with circles. The recipe requires two epicycles having equal radius and equal-and-opposite rotation speeds. Visualizing this in steps, Figure 41.8 shows the so-called *one-circle* approximation to a given horizontal straight line. This clearly can’t be the final answer, the circle only touches the line twice. Note too that this first-order circle goes clockwise for increasing time.

By introducing the second epicycle and tracing a point on its rim, we come up with the *two-circle* approximation. It happens that this case perfectly produces a horizontal line through the origin as shown in Figure 41.9. The ‘drawing tip’ on the rim of the second circle has a constant height.

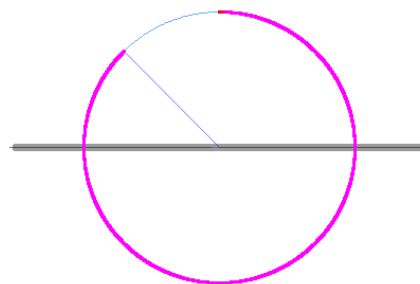


Figure 41.8. One-circle approximation of horizontal line through the origin, only correct at two points.

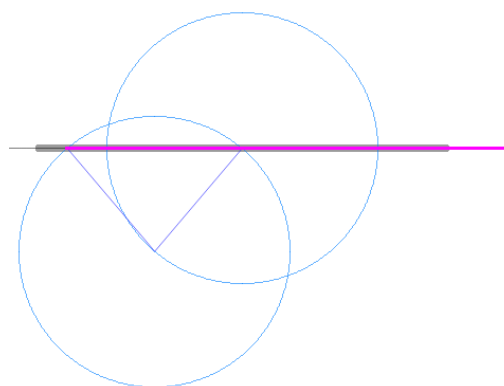


Figure 41.9. Two-circle approximation of horizontal line through the origin, correct at all points.

3.3 Arbitrary Straight Lines

Three epicycles are required to draw an arbitrary straight line in the Cartesian plane. Figure 41.10 illustrates the order-two approximation of such a line, but the approximation is clearly offset (in position and rotation). We correct for this by choosing just the right r_0, ϕ_0 for a non-rotating base-zero epicycle, leading to Figure 41.11.

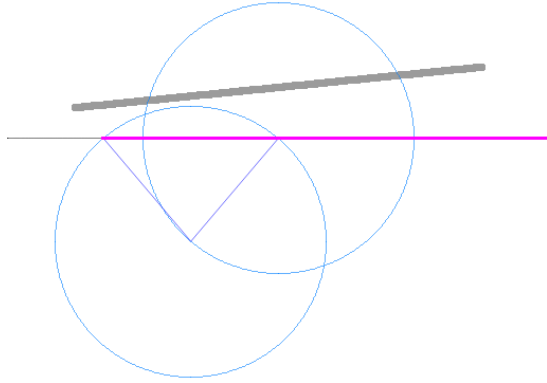


Figure 41.10. Two-circle approximation of horizontal line, incorrect up to a constant offset.

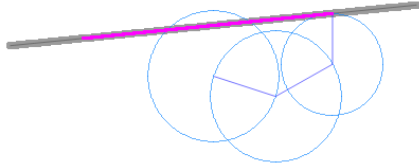


Figure 41.11. Three-circle approximation of horizontal line, corrected with a non-rotating epicycle.

3.4 Ellipses

An ellipse can be generated using two circles having unequal radii, and equally-opposite rotation rates. If the respective radii are $r_1 > r_2$, then the semi-major axis of the ellipse is the sum

$$a = r_1 + r_2 ,$$

and the semi-minor axis is the difference

$$b = r_1 - r_2 .$$

With this information, it's straightforward to show that the distance from the center of the ellipse to either focus is given by

$$c = ae = 2\sqrt{r_1 r_2} ,$$

where e is the eccentricity.

Demonstrated in Figure 41.12 is the case with $r_1 = 2r_2$. For completeness, the Figure is represented entirely by:

$$\begin{aligned} r_j &= \{128, 64\} \\ \omega_j &= 2\pi \cdot \{1, 1\} \\ \phi_j &= \{-\pi/2, \pi/2\} \end{aligned}$$

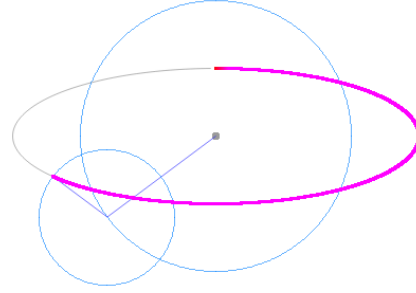


Figure 41.12. Tracing an ellipse using two circles.

3.5 Stars

A surprising comfort zone for epicycles is the family of 'stars'. To construct any given star, we choose a pair of counter-rotating circles with non-equal radii and non-equal rotation rates. Supposing one circle rotates at positive rate ω_1 , and the other has negative rate $-\omega_2$, the number of points in the resulting star is equal to $\omega_1 - \omega_2$.

Two example cases are detailed in Figures 41.13, 41.14. The first case is generated with the following data:

$$\begin{aligned} r_j &= \{128, 64\} \\ \omega_j &= 2\pi \cdot \{2, 3\} \\ \phi_j &= \{-\pi, \pi/4\} \end{aligned}$$

Likewise, for a second star:

$$\begin{aligned} r_j &= \{128, 96\} \\ \omega_j &= 2\pi \cdot \{3, 4\} \\ \phi_j &= \{-\pi, \pi/4\} \end{aligned}$$

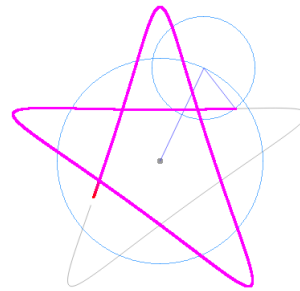


Figure 41.13. Five-vertex star using two circles.

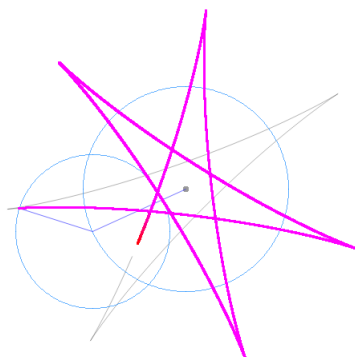


Figure 41.14. Seven-vertex star using two circles.

3.6 Lissajous Curves

Cartesian Formulation

A *Lissajous curve*, also known as a Bowditch curve, is the image of the set of parametric equations:

$$\begin{aligned}x(t) &= A \sin(2\pi at + \delta) \\ y(t) &= B \sin(2\pi bt)\end{aligned}$$

The constants A and B respectively control the horizontal and vertical scales. Inside the sine function, the constants a and b are angular frequencies, where δ , conventionally associated with the x -equation, is an overall phase shift. The ratio a/b determines the overall ‘fingerprint’ of the curve. A Lissajous curve is *closed* when a/b is rational.

With so many variables in play, it’s prudent to explore Lissajous curves by limiting $A = B$, and also restricting a , b to integer values. Plotted in Figure 41.15 is a table of Lissajous curves for various choices of a , b , and δ .

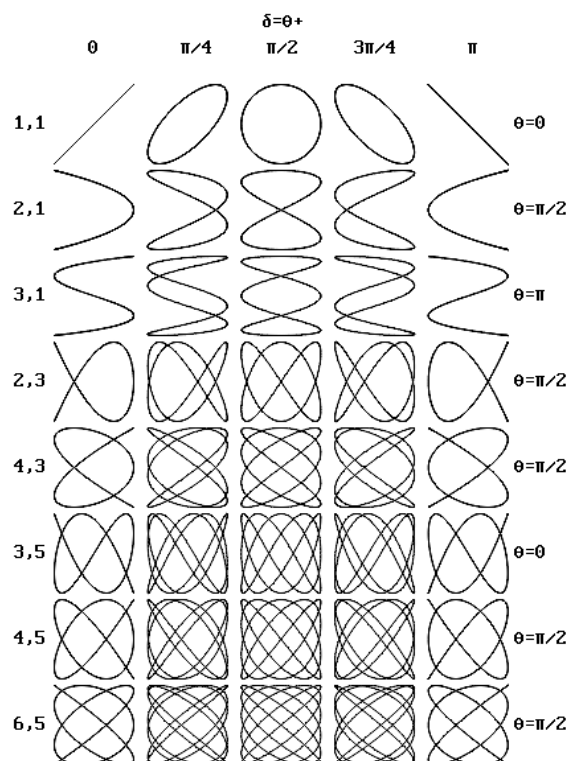


Figure 41.15. Table of Lissajous curves. Indexed vertically are respective angular frequencies a , b (scaled by 2π). The phase shift δ is indexed horizontally.

Epicycle Formulation

The epicyclic formula for Lissajous curves consists of two sets of counter-rolling circles. To illustrate, take a Lissajous curve such as any from Figure 41.15. Choosing the curve with $a = 3$, $b = 1$, $\delta = \pi + \pi/4$ for the sake of argument. (Note too that a , b are scaled by 2π .) From this, we need an equivalent representation in the form

$$\begin{aligned}\{r_j, \omega_j, \phi_j\} \\ j = 1, 2, 3, 4,\end{aligned}$$

where the zero-case can be ignored (the Lissajous curve is already centered). We also won’t worry much about the r_j -variable, for this is one and the same number across each j :

$$r = r_1 = r_2 = r_3 = r_4$$

With some fiddling, one can easily-enough work out the proper mapping from a , b , δ , onto ω_j , ϕ_j . For our case, this turns out to be:

$$\omega_1 = b = -\omega_3 \qquad \omega_2 = a = -\omega_4$$

$$\phi_j = \{-\pi, 3\pi/4, 0, -3\pi/4\}$$

Note that ω_1 is assigned to the lesser of the given angular frequencies, and ω_2 always takes the greater.

Having specified a given Lissajous curve in terms of epicycles, we can witness the curve being gradually constructed. Continuing the example on hand, Figures 41.16-41.19 illustrate increasing orders of approximation to the Lissajous curve, settling on fourth-order, giving the exact result. Positioned below each curve is a plot representing the radius (grey) and phase (red) of each epicycle used.

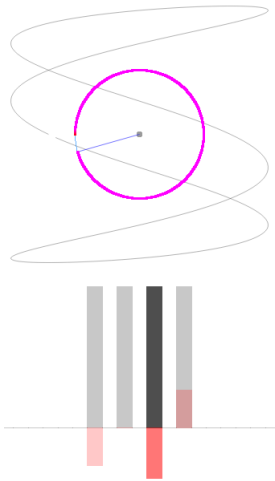


Figure 41.16. One-circle approximation of a Lissajous curve.

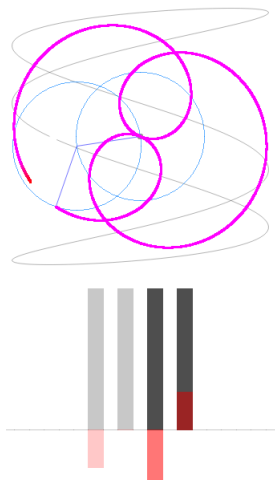


Figure 41.17. Two-circle approximation of a Lissajous curve.

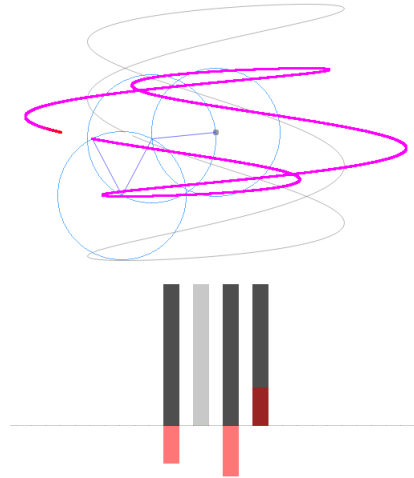


Figure 41.18. Three-circle approximation of a Lissajous curve.

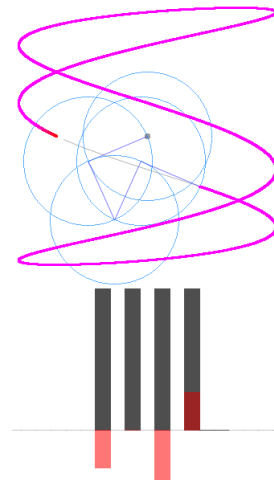


Figure 41.19. Four-circle (final) approximation of a Lissajous curve.

It turns out that any Lissajous curve from the family depicted in Fig. 41.15 can be generated by the same process demonstrated above. The angular frequencies a, b are straightforwardly assigned to each ω_j as discussed, however the setting the initial phase ϕ_j for each case can be quite a chore to do manually.

4 Epicycle Clock

An interesting tangent worth exploring is the *epicycle clock*, which is a re-design of a standard analog clock.

4.1 Standard Clock

Imagine a standard analog clock with the usual ‘hour’, ‘minute’, and ‘second’ hands. Placing the origin of a Cartesian system at the center of the clock face, the ‘tip’ of the hour hand can be written as a vector:

$$\mathbf{H} = \langle R_H \cos(\theta_H), R_H \sin(\theta_H) \rangle$$

In the above, R_H is the length of the hour hand, and the variable θ_H is an angular parameter that represents each hour and all moments in between. By convention $\theta = 0$ corresponds to 3:00, and θ_H increases in the *counterclockwise* direction, if that weren’t confusing enough. This means $\pi/2$ corresponds to 12:00, π corresponds to 9:00, and so on.

On a standard clock, the minute hand and second hands are also pinned to the origin, having respective vector representations:

$$\begin{aligned} \mathbf{M} &= \langle R_M \cos(\theta_M), R_M \sin(\theta_M) \rangle \\ \mathbf{S} &= \langle R_S \cos(\theta_S), R_S \sin(\theta_S) \rangle \end{aligned}$$

To be more precise, each angular parameter θ_j is a function of time, equal to

$$\begin{aligned} \theta_j(t) &= -\omega_j t + \phi_j \\ j &= H, M, S, \end{aligned}$$

where the angular velocity of each hand is captured by ω_j , and t is the varying parameter. The phase ϕ_j is tuned at $\pi/2$ so all clock hands overlap at 12:00 when $t = 0$. The way we capture one hour having 60 minutes, and one minute having 60 seconds, is by restricting the angular velocities:

$$\omega_S = 60 \omega_M = 60 (60 \omega_H)$$

4.2 Epicycle Clock

With the analog clock framed in terms of vectors and angles, we now make it more epicycle-like. The hour hand shall be based at the origin as before. As for the minute hand, we choose its base as the tip of the hour hand. Similarly, the second hand is placed at the tip of the minute hand. All together, this means that the ‘time vector’

$$\mathbf{T}(t) = \mathbf{H}(t) + \mathbf{M}(t) + \mathbf{S}(t)$$

lands at a specific place in the Cartesian plane for each value of t .

Shown in Fig. 41.20 is such a clock displaying the time 10:10:23. The shaded backdrop image is comprised of $60 \times 60 \times 60$ total lines, generated by tracing the second hand once for all possible times. As the clock runs, the hour hand traces out its ‘usual’ circle. The minute hand though, because it’s not pinned to the center, must swing around the hour hand, and likewise for the second hand. For this reason, we call this scheme the *epicycle clock*.

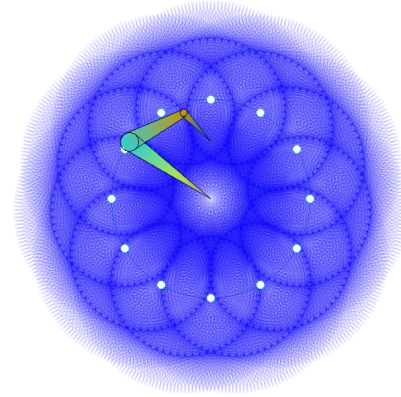


Figure 41.20. Twelve-hour epicycle clock displaying time 10:10:23.

4.3 Non-Twelve Hour Clocks

A modification to the epicycle clock changes the total period from 12 hours to some other number. That minute and second hands still occupy values from 0 to 60 to complete a revolution. When the ‘minutes’ and ‘seconds’ are the same number, the hands become momentarily parallel. Shown in Figure 41.21 is a *six-hour* clock displaying the time 04:21:21.

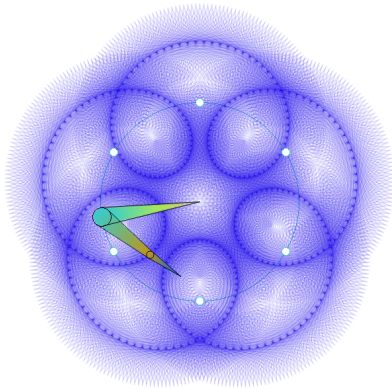


Figure 41.21. Six-hour epicycle clock displaying time 04:21:21. The minute- and second-hands are momentarily parallel when ‘minutes’ and ‘seconds’ are the same number.

4.4 Sidereal Time

One interesting detail of the epicycle clock is that the number of complete ‘loops’ traced out by the minute hand is *one less* than the number of hours on the clock. For instance, the minute hand completes 11 loops on the 12-hour clock. On the 6-hour clock, there are 5 loops. No matter how many hours are on the clock, there is always an ‘off-by-one’ discrepancy for the number of complete revolutions carried out by the minute hand.

Applying the same analysis to *modern* (heliocentric, Newtonian) planetary motion, particularly Earth, it’s possi-

ble to raise doubt about the number of days in one year. For a while now, we’ve agreed that there are about 365.2422 days in one year, but on *which* side of the off-by-one error is this number based on?

It turns out that modern astronomers have framed this problem by introducing *sidereal time* and *anit-sidereal time*. Consider the rotation of Earth *only*, ignoring the effect of going around the sun. Based on the apparent motion of the stars across the sky as measured from a fixed point on Earth, one would measure the ‘day’ to be not 24:00 hours (88512 seconds), but instead 23:56:04 (86164.0905 seconds). This time is called the *sidereal day*, i.e., the ‘pure’ rotation period of the Earth with respect to the fixed stars.

On the other hand, suppose instead that Earth did *not* rotate on its axis, but still revolves around the sun (ignoring the stars). In this situation, there is still a notion of sunrise and sunset on Earth, with a period of exactly one year. This effectively means that a complete year results a ‘free day’. It’s in this way that *solar time* is defined. Since the actual Earth rotates *and* revolves, and it’s convenient for Earthly affairs to base local clocks on the apparent solar time. The apparent solar time is of course modified by Earth’s angular velocity.

To summarize, we acknowledge that if Earth’s year is precisely N days, only $N - 1$ of those days correspond to rotation about the axis, and one day comes about by revolution around the Sun.

5 Connecting Discrete Points

Epicycles can be used to approximate to ‘connect’ a series of discrete points, somewhat like connect-the-dots.

5.1 Resolution

Resolution, in the context of epicycles, refers to the number of circles N required to draw a smooth curve through P particular points in the plane. Since epicycles trace continuous arcs by definition, there will be more ‘curve on the page’ than would arise from a linear connect-the-dots approach. For $N < P$, the resulting curve attempts to trace out the shape implied by all P points, but is only ‘exact’ for the case $N = P$.

5.2 The Frequency Trick

Consider any ordered list of points, where for simplicity let us take $(-30, 15)$, $(20, 65)$, and $(-20, -50)$ as an example. With three points given, one immediately knows (at most) three epicycles are required to connect the points in order:

$$\{r_j, \omega_j, \phi_j\}$$

$$j = 1, 2, 3$$

In total, there are six known quantities - two for each point - however there are *nine* unknown quantities. This is three unknowns too many to specify a solution, so an additional insight is needed.

One eventually notices that the applicable ω_j for small N should also be small, which is to say fast-rotating epicycles aren’t required to trace only three points. In fact, the *only* frequencies in play ought to be

$$\omega_j = 2\pi \cdot \{-1, 0, 1\} ,$$

which we shall call the ‘frequency trick’. Effectively, we have removed some freedom from the problem by insisting on the rotation speed ω_j of each circle. With this restriction on ω_j , the epicycle representation for this problem can be written as a system with only six unknowns:

$$\{r_j, j, \phi_j\}$$

$$j = -1, 0, 1$$

How such a system is solved is slightly nontrivial, but at least we know there probably is a solution.

Leaving the details aside for now, it happens that the three points $(-30, 15)$, $(20, 65)$, and $(-20, -50)$ are approximated by:

$$r_j = \{14.14, 45.42, 24.90\}$$

$$\omega_j = 2\pi \cdot \{0, 1, -1\}$$

$$\phi_j = \{-2.356, -2.827, 0.3719, \}$$

Figures. 41.22, 41.23 show a case representative of the example on hand.

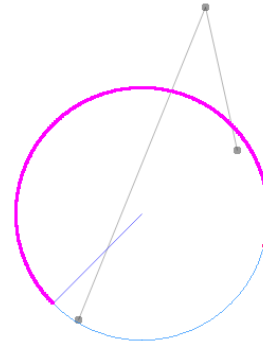


Figure 41.22. One-circle approximation of three given points using frequency number 1, ignoring $-1, 0$.

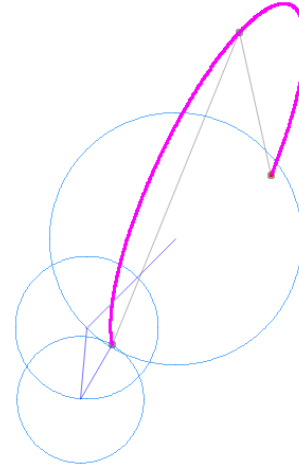


Figure 41.23. Three-circle approximation of three given points using frequency numbers $-1, 0, 1$.

Generalized Frequency Trick

It should make sense that the so-called frequency trick works in general for systems comprising of $2N + 1$ total points, in which case we tack on a new frequency number for each

new point added:

$$\omega_j = 2\pi \cdot \{-N, \dots, -2, -1, 0, 1, 2, \dots, N\}$$

The case $\omega_j = 0$ corresponds to the epicycle that does not rotate, but instead offsets the entire curve by a constant vector.

5.3 Squares

Unlike lines, simple arcs, and Lissajous curves, it should follow that shapes featuring rigid corners are troublesome for epicycles. For this reason, it's necessary to approximate things like squares, particularly because of the corners, using many circles.

In effort to 'guess' the correct epicycle representation for a square, let's first assume the square is represented by P discrete points, and thus it requires N epicycles to trace through each point. Using the so-called frequency trick, the angular frequencies can safely be assumed as

$$\omega_j = 2\pi \cdot \{0, \pm 1, \pm 2, \dots\}.$$

Of course, not all of these frequencies may be represented in the square's final approximation.

Speculating on the radius values r_j and phase values ϕ_j , begin with the observation that the *greatest* distance from the origin traced by an epicycle is any corner of the square. To reach the top-right corner, every phase must be momentarily in agreement, having value $\phi_j = \pi/4$. This occurs at three more instants, namely $\phi_j = -\pi/4$ and $\phi_j = \pm 3\pi/4$. When all of the phases are in sync, the length of the epicycle 'drawing arm' is the sum of all r_j . If the total width of the square is $2A$, then we may write

$$A\sqrt{2} \approx \sum_{j=1}^N r_j.$$

By similar arguments, the *shortest* distance from the origin occurs at four points on the square, and we must have

$$A \approx \sum_{j=1}^N a_j r_j$$

for some set of coefficients $\{a_j\}$.

At this stage, we have two equations for r_j that are vaguely suggestive of ϕ_j . Solving for r_j , ϕ_j this way is a very hard problem, so let us (once again) leave certain details aside, and simply list the correct answer. Choosing

$A = 1$, it turns out that, approximately:

$$r_j = \{1.146, 0.1274, 0.04585, 0.02339, \dots\}$$

$$\omega_j = \{-1, 3, -5, 7, -9, 11, -13, \dots\}$$

$$\phi_j = \{0, \pi, -\pi, 0, 0, \pi, -\pi, 0, 0, \dots\}$$

Reconciling this with the sums above, we can write

$$\sqrt{2} \approx 1.146 + 0.1274 + 0.04585 + 0.02339 + \dots$$

$$1 \approx 1.146 - 0.1274 - 0.04585 + 0.02339 + \dots$$

Fig. 41.24 illustrates the epicycle approximation of a square based on $N = 8000$ points. Plotted under the curve in grey is $\ln(1 + r_j)$ versus ω_j , centered at $\omega_0 = 0$. The corresponding phase $\pm\pi$ is superimposed in red. Interestingly, the sequence of r_j seems to stabilize around the values listed, with increased resolution N mattering little beyond $N \approx 400$.

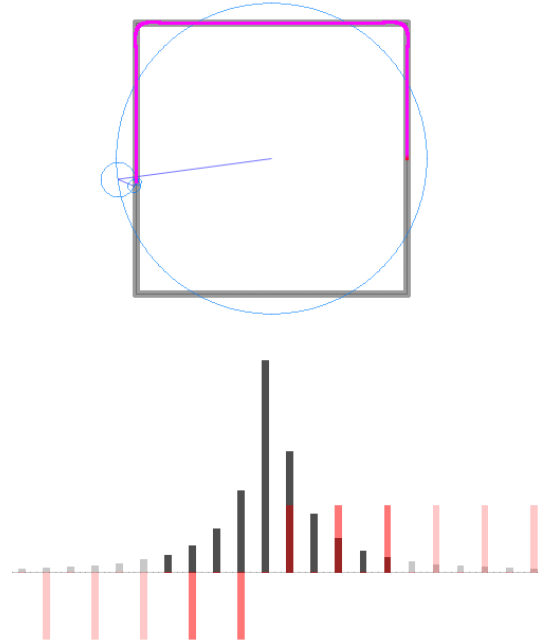


Figure 41.24. Approximation of a square based on $N = 8000$ points. Under the curve in grey is $\ln(1 + r_j)$ versus ω_j , centered at $\omega_0 = 0$. The corresponding phase $\pm\pi$ is superimposed in red.

Rounded Squares

By truncating the epicycle approximation of a square to $N = 2$ circles only, the resulting image is a square with rounded corners as shown in Figure 41.25.

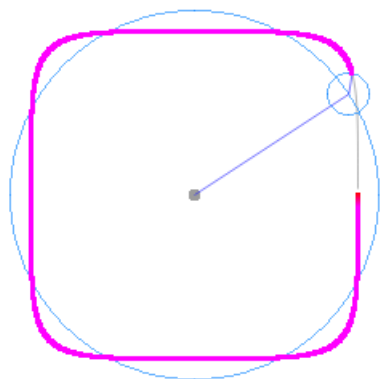


Figure 41.25. Rounded square using a two-circle approximation.

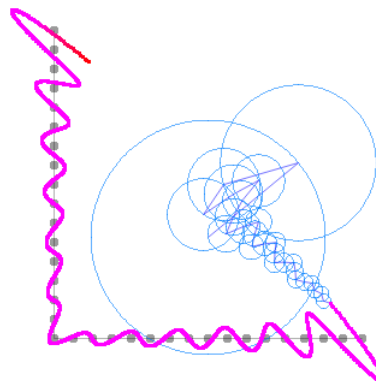


Figure 41.26. Approximation of open L-shape curve. $N/P < 1$.

5.4 Open Curves

Epicycles are most natural for drawing closed curves, but it's instructive to try approximating open curves. Checking this out for an L-shaped example, we see in Figs. 41.26-41.27 that the approximate curve does manage to connect all given points, but the intermediate behavior near the boundaries is a bit 'jumpy', exhibiting Gibbs phenomenon.

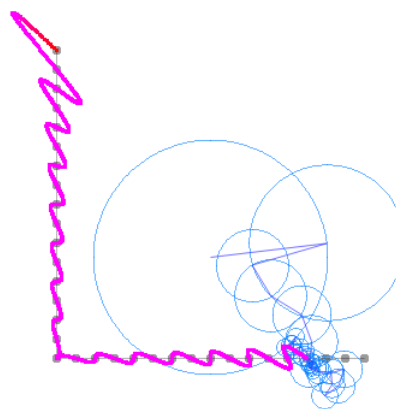


Figure 41.27. Approximation of open L-shape curve. $N = P$.

6 Drawing Arbitrary Curves

Here we develop a means for using epicycles to draw a curve through an arbitrary sequence of points. To show what this means, presented below is the *Piscis Vincentii*, depicted in Figures 41.28-41.29. The ‘raw’ fish consists of 168 ordered pairs in the Cartesian plane. The fully general trick for calculating all r_j and ϕ_j from P given points, leading to the fully cooked version, is called the *discrete Fourier transform*, or DFT for short.

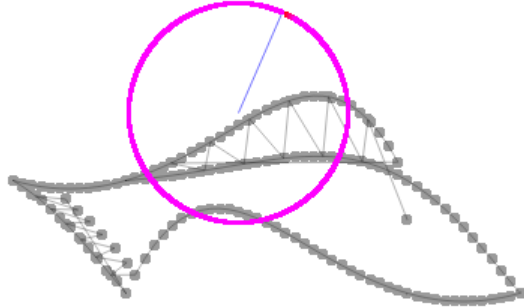


Figure 41.28. *Piscis Vincentii*, raw.

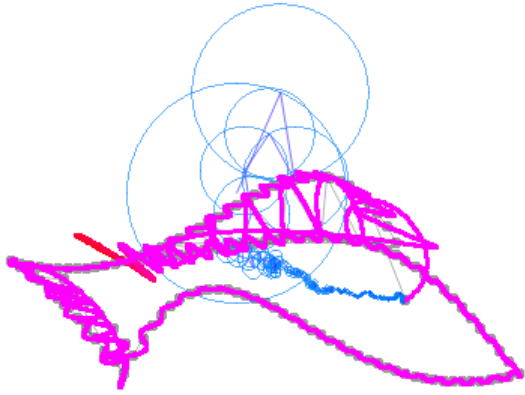


Figure 41.29. *Piscis Vincentii* approximated by 168 epicycles (fully cooked).

6.1 Motivation for DFT

Consider a given ordered list of P ordered pairs (x, y) on the Cartesian plane:

$$\{x_p, y_p\}$$

$$p = 1, 2, 3, \dots, P$$

The task on hand is to represent the p th point somewhere on the perimeter of the N th epicycle:

$$x_p = \sum_{j=1}^{N \leq P} r_j \cos(\phi_j + \omega_j t_p) \quad (41.1)$$

$$y_p = \sum_{j=1}^{N \leq P} r_j \sin(\phi_j + \omega_j t_p) \quad (41.2)$$

Note that when $N = P$, the above equations are guaranteed to send the resulting curve directly through each point.

6.2 Complex Epicycles

With the polar representation of each x_p, y_p , it makes sense to combine these into a complex number such that

$$z_p = x_p + i y_p \quad (41.3)$$

for each p . Then Equations (41.1), (41.2) join to become

$$z_p = \sum_{j=1}^N (r_j \cos(\phi_j + \omega_j t_p) + i r_j \sin(\phi_j + \omega_j t_p)) ,$$

which, by Euler’s formula, is equivalent to:

$$z_p = \sum_{j=1}^N r_j e^{i\phi_j + i\omega_j t_p} \quad (41.4)$$

To get a grip on the time dependence, let us first associate each point p with a unique value on the interval $[0 : 2\pi)$ such that

$$t_p = \frac{2\pi p}{N}$$

$$p = 0, 1, 2, \dots, N - 1 . \quad (41.5)$$

By doing so, the angular frequencies ω_j become dimensionless integers:

$$\omega_j = j$$

$$j = 0, 1, 2, \dots, N - 1 \quad (41.6)$$

You may trade factors like 2π from the t_p -variable to the ω_j variable without harm.

An equivalent but more convenient choice for the angular frequencies ω_j allows for negative values:

$$\omega_j = \left\{ 0, \pm 1, \pm 2, \pm 3, \dots, \frac{N}{2} \right\}$$

To get this from Equation (6.2), we write the conditional substitution:

$$\omega_j = \begin{cases} j & \text{if } j \leq N/2 \\ j - N & \text{if } j > N/2 \end{cases} \quad (41.7)$$

It also makes sense to separate time-dependent terms from the constant terms, so let's define a complex number Q_j such that

$$Q_j = r_j e^{i\phi_j}, \quad (41.8)$$

which combines the radius and the initial phase of each epicycle. Putting everything together, we so far have

$$z_p = \sum_{j=0}^{N-1} Q_j e^{i2\pi j p/N}. \quad (41.9)$$

6.3 Discrete Fourier Transform

The big job now is to solve Equation (41.9) for each complex coefficient Q_j . This is done by a method called *Fourier's trick*, which begins by introducing a similar exponential term with a negative argument, particularly

$$e^{-i2\pi k p/N},$$

where the 'frequency index' j is changed to k . Multiply this quantity into Equation (41.9) and sum over p :

$$\begin{aligned} \sum_{p=0}^{N-1} z_p e^{-i2\pi k p/N} &= \sum_{p=0}^{N-1} \sum_{j=0}^{N-1} Q_j e^{i2\pi j p/N} e^{-i2\pi k p/N} \\ &= \sum_{j=0}^{N-1} Q_j \sum_{p=0}^{N-1} e^{i2\pi(j-k)p/N} \end{aligned}$$

It's possible to prove (see below) that the right-most sum evaluates to a Kronecker delta relation

$$\sum_{p=0}^{N-1} e^{i2\pi(j-k)p/N} = N \delta(j-k),$$

and the above reduces to

$$\sum_{p=0}^{N-1} z_p e^{-i2\pi k p/N} = N \sum_{j=0}^{N-1} Q_j \delta(j-k) = N Q_k.$$

The effect of $\delta(j-k)$ is to 'pluck out' Q_k from the remaining sum, while flattening all other terms to zero. After realizing this, we have a formula for Q_j as it appears in Equation (41.9):

$$Q_j = \frac{1}{N} \sum_{p=0}^{N-1} z_p e^{-i2\pi j p/N} \quad (41.10)$$

For a consistency check, it's worth plugging Equation (41.10) back into Equation (41.9) to make sure z_p comes back out:

$$\begin{aligned} z_p &= \sum_{j=0}^{N-1} Q_j e^{i2\pi j p/N} \\ &= \sum_{j=0}^{N-1} \frac{1}{N} \sum_{p'=0}^{N-1} z_{p'} e^{-i2\pi j p'/N} e^{i2\pi j p/N} \\ &= \sum_{p'=0}^{N-1} z_{p'} \left(\frac{1}{N} \sum_{j=0}^{N-1} e^{i2\pi j(p-p')/N} \right) \\ &= \sum_{p'=0}^{N-1} z_{p'} \delta(p-p') = z_p \end{aligned}$$

Note that a similar Kronecker delta relation has been used.

6.4 Kronecker Delta Relations

The discrete Fourier transform relies on two Kronecker delta relations

$$\frac{1}{N} \sum_{p=0}^{N-1} e^{i2\pi(j-k)p/N} = \delta(j-k) \quad (41.11)$$

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{i2\pi j(p-p')/N} = \delta(p-p'), \quad (41.12)$$

where j, k, p, p' are integers. To really understand these, consider the 'pure' claim

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{i2\pi j p/N} = \delta(p), \quad (41.13)$$

and expand out the sum manually:

$$\begin{aligned} \sum_{j=0}^{N-1} e^{i2\pi j p/N} &= \\ &1 + e^{i2\pi p/N} + e^{i2\pi 2p/N} + \dots + e^{i2\pi(N-1)p/N} \end{aligned}$$

In this form, the sum looks especially like a geometric series, which in general obeys

$$\frac{1 - z^n}{1 - z} = 1 + z + z^2 + \dots + z^{n-1},$$

thus the above sum is

$$\sum_{j=0}^{N-1} e^{i2\pi j p/N} = \frac{1 - e^{i2\pi p}}{1 - e^{i2\pi p/N}}.$$

The above can only take one of two values. The interesting case is $p = 0$, leading to

$$\sum_{j=0}^{N-1} e^{i2\pi j0/N} = \sum_{j=0}^{N-1} 1 = N.$$

For *all* other integer values of p , the sum resolves to *zero*, as

$$1 - e^{i2\pi p} = 0$$

$$p = 1, 2, 3, \dots$$

In summary, we have found

$$\frac{1}{N} \sum_{j=0}^{N-1} e^{i2\pi jp/N} = \begin{cases} 1 & \text{if } p = 0 \\ 0 & \text{if } p \neq 0 \end{cases},$$

which is precisely (41.13). Note the argument p can be easily substituted and/or linearly shifted to recover (41.11)-(41.12).

6.5 Implementing the DFT

The star result of the DFT derivation is Equation (41.10), namely

$$Q_j = \frac{1}{N} \sum_{p=0}^{N-1} z_p e^{-i2\pi jp/N}.$$

As pseudo-code, the implementation Equation (41.10) looks like:

```
DFT (Given(), j0)
N = Size(Given)
re = 0
im = 0
For k = 0 TO N - 1
    arg = 2 * Pi * k * j0 / N
    (u, v) = Mult (Cos(arg), Sin(arg),
                  Given(k).x, Given(k).y)
    re = re + u
    im = im - v
Next k
re = re / N
im = im / N
Return (re, im)
```

The array *Given()* refers to the list of P given points represented by z_p . The output of the *DFT* function is the j th pair of Fourier coefficients Q_j . Translating each Q_j back into radius and phase is done as follows:

```
For j = 0 TO N - 1
    omega(j) = j
    If (j > N / 2) Then omega(j) = j - N
    (Q(j).x, Q(j).y) = DFT (Given(), j0)
    rad(j) = Sqr(Q(j).x^2 + Q(j).y^2)
    phase(j) = Atan2(Q(j).y, Q(j).x)
```

Chapter 42

Collatz Fractal

1 Introduction

In 1937, Lothar Collatz pointed out a pattern that has been bugging mathematicians and hobbyists ever since. Start with any natural number $n > 1$, and then:

- If n is even, change n to $n/2$.
- If n is odd, change n to $3n + 1$.
- Repeat until n changes to 1.

The so-called *Collatz conjecture* states that *any* natural number $n > 1$ will eventually reduce down to 1. So far so good, but where's the proof?

Any acceptable proof the Collatz conjecture has remained elusive to the most accomplished mathematicians. This problem, while gaining a slew of nicknames along the way, namely (but not limited to) the $3n + 1$ problem, the hailstone sequence, the hailstone numbers, and the wondrous numbers, has nonetheless been been 'proved' by a number of optimists. We will make no attempt to wrap our heads around the any proofs here, or to debunk any crackpot analysis. Rather, the plan is to further deepen the mystery by involving complex numbers.

1.1 Collatz Function

The Collatz conjecture can be contained in a rather ugly equation:

$$f(x) = \begin{cases} x/2 & x \text{ even} \\ 3x + 1 & x \text{ odd} \end{cases} \quad (42.1)$$

Note that there is no 'escape condition' for Equation (42.1), thus the infinite cycle $1 - 4 - 2 - 1 - 4 - 2 - \dots$ awaits

any natural number that satisfies the conjecture. A more utilitarian version of the above would contain a way to escape from such a cycle when $x = 1$ first occurs.

1.2 Continuous Collatz Function

In order to study the Collatz conjecture on the complex plane, it makes sense to generalize $f(x)$ into a new function $c(z)$ that is (i) smooth and differentiable, and (ii) reducible to $f(x)$ for real natural numbers. In the most general case then, we can write

$$c(z) = \left(\frac{z}{2}\right) p_{\text{even}}(z) + (3z + 1) p_{\text{odd}}(z), \quad (42.2)$$

where the functions $p_{\text{even}}(z)$, $p_{\text{odd}}(z)$ obey:

$$p_{\text{even}}(z) = \begin{cases} 1 & z \text{ even} \\ 0 & z \text{ odd} \end{cases}$$
$$p_{\text{odd}}(z) = \begin{cases} 0 & z \text{ even} \\ 1 & z \text{ odd} \end{cases}$$

Common literature¹ on the subject suggests trying

$$p_{\text{even}}(z) = \cos^2\left(\frac{\pi z}{2}\right)$$
$$p_{\text{odd}}(z) = \sin^2\left(\frac{\pi z}{2}\right),$$

and inserting these into Equation (42.2) gives, after simplifying,

$$c(z) = \frac{7z + 2}{4} - \frac{5z + 2}{4} \cos(\pi z). \quad (42.3)$$

One can easily check that Equation (42.3) satisfies Equation (42.2) by checking a few real natural numbers. Moreover, $c_1(z)$ is a smooth and differentiable function, which

¹Collatz conjecture'. *Wikipedia*. December 22 2022.
https://en.wikipedia.org/wiki/Collatz_conjecture

translates to ‘much more appealing to mathematicians than $f(x)$ ’.

It is instructive to go through a second derivation of Equation (42.3) by choosing² $p_{\text{even}}(z)$, $p_{\text{odd}}(z)$ such that

$$p_{\text{even}}(z) = \frac{(-1)^z + 1}{2}$$

$$p_{\text{odd}}(z) = -\frac{(-1)^z - 1}{2}.$$

As required, one function ‘turns off’ as the other ‘turns on’ for any given real natural number z . For non-whole z , the above become complex numbers, and may be reconciled via Euler’s formula

$$(-1)^z = e^{i\pi z} = \cos(\pi z) + i \sin(\pi z).$$

We see the term $\cos(\pi z)$ does the heavy lifting, as the imaginary term $\sin(\pi z)$ is precisely zero for real natural number z and can be omitted.

The provisional formula for $c(z)$ so far reads

$$c(z) = \frac{z}{2} \left(\frac{\cos(\pi z) + 1}{2} \right) + (3z + 1) \left(\frac{1 - \cos(\pi z)}{2} \right),$$

and simplifying this leads to Equation (42.3). Interestingly, note that either g -function can be raised to an integer power without overstaying its welcome in $c(z)$. For integers q_e , q_o , this means

$$c(z) = \frac{z}{2} \left(\frac{\cos(\pi z) + 1}{2} \right)^{q_e} + (3z + 1) \left(\frac{1 - \cos(\pi z)}{2} \right)^{q_o} \quad (42.4)$$

is another valid Collatz function.

1.3 Time Analysis

Convergence Time

If we take a real natural number z and recursively apply $z \rightarrow c(z)$, the number of iterations required to reach $z = 1$ is

the *convergence time* of z . For instance, starting with $z = 3$ we would have the sequence $3 - 10 - 5 - 16 - 8 - 4 - 2 - 1$, having convergence time $t = 7$. The special case $z = 1$ has convergence time $t = 0$. Certain numbers, especially odds, can have unexpectedly long convergence times.

Following are the natural numbers up to $z = 40$ along with corresponding convergence times:

1(0)	2(1)	3(7)	4(2)	5(5)
6(8)	7(16)	8(3)	9(19)	10(6)
11(14)	12(9)	13(9)	14(17)	15(17)
16(4)	17(12)	18(20)	19(20)	20(7)
21(7)	22(15)	23(15)	24(10)	25(23)
26(10)	27(111)	28(18)	29(18)	30(18)
31(106)	32(5)	33(26)	34(13)	35(13)
36(21)	37(21)	38(21)	39(34)	40(8)

Take note that starting with $3^3 = 27$ requires 111 iterations to converge.

Escape Time

While it seems that all natural numbers *probably* satisfy the Collatz conjecture, we still need to account for what happens when z is a decimal, a complex number, or something else that fails to satisfy the conjecture. It will turn out that most of the complex plane must be treated this way, and the converging cases are more rare than common.

When either (i) the magnitude of z exceeds some hard-coded maximum, or (ii) the number of iterations exceeds a cutoff number, then z has reached an *escape time*. In the former case, z is striding toward infinity, and the escape time indicates ‘how quickly’ this occurs. In the latter case, z is wandering around the complex plane with no apparent destiny. These are considered to converge infinitely slowly.

²Xander. ‘The Collatz Fractal’. Rhapsody in Numbers. January 12 2012.
https://yozh.org/2012/01/12/the_collatz_fractal/

2 Collatz Set

The *Collatz set* is comprised of numbers z_0 in the complex plane that reach a convergence time in a finite number of $z \rightarrow c(z)$ applications, where $c(z)$ is the continuous Collatz function (42.2) or a generalization thereof. Points characterized by an escape time are outside of the set.

2.1 Discrete Plot

Plotting Set Members

For visualizing the Collatz set, we can start by assigning a distinct value - namely color - to each convergence time. Any complex numbers z_0 that converge in the same number t steps get the same color, and different numbers z' with different convergence times t' get a different color, and so on. While the particular choice of colors does not quite matter, it helps to have a diversity of shades with noticeable contrast between boundaries.

Going right to some results, Figure 42.1 shows a heat

map of convergence times over the complex plane centered at $x = 15$. On the positive real axis are the (randomly chosen) shades assigned to the convergence time for the number at that position.

Plotting Set Non-Members

To represent points outside the Collatz set, we choose a two-color gradient to indicate escape time. Inputs that escape slower than others are shaded black, whereas points that jump right of the picture are white-shifted toward beige. Figure 42.1 shows such a gradient hugging the real axis. Note that the vertical stripe in the Figure corresponds to inputs $z = (1, y)$, and these inputs automatically satisfy the Collatz conjecture without iteration. The origin is centered one notch leftward of the vertical stripe.

The last category corresponds to complex inputs that fail to converge and diverge, eventually exceeding the cutoff number. These are assigned a flat color (green). As you may imagine, the negative integers fit this description.

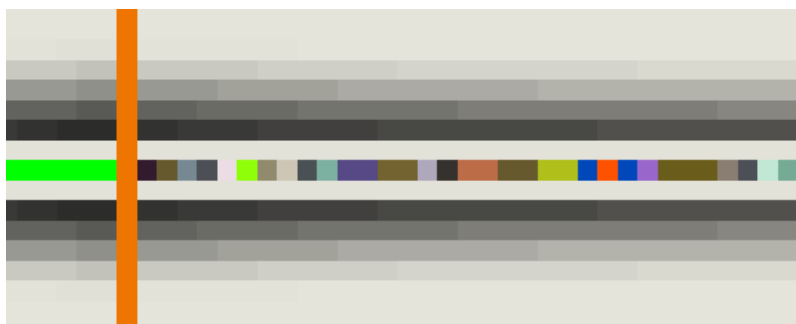


Figure 42.1. Collatz set over complex integer inputs. Vertical stripe corresponds to $x = 1$. (Center: $z = 15$, Zoom: 0.125)

2.2 Continuous Plot

With an idea of how the Collatz set appears for integers, now we construct the set over the continuous complex plane. Shown in Figure 42.2 is the continuous analog to Figure

42.1, and some stark differences are visible. Most notably (i) the natural numbers with a convergence time seem to be *gone*, (ii) the gradient has a ‘spiky’ appearance, undoubtedly due to the cosine term in $c(z)$, and (iii) inputs near the origin seemingly don’t converge.

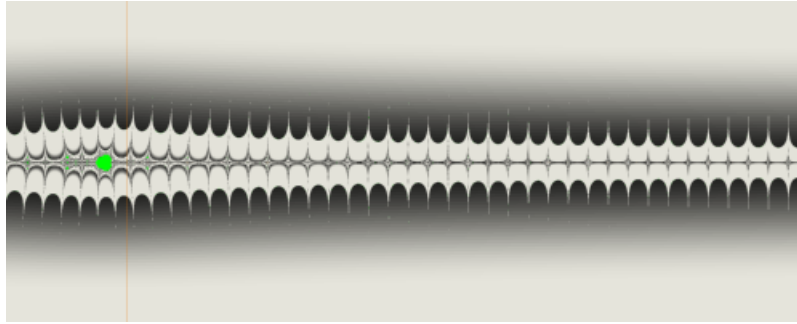


Figure 42.2. Collatz set over complex continuous inputs. Vertical stripe corresponds to $x = 1$. (Center: $z = 15$, Zoom: 0.125)

3 Collatz Fractal

The barely-visible boundary separating members of the Collatz set from non-members is called the *Collatz fractal*. Shown in Figure 42.3 shows a detailed view of the Collatz

fractal generated using Equation (42.3) near the origin. Note that the subtle vertical line at $x = 1$ is not magnified by zooming in - this line has zero thickness and renders as one pixel at best.

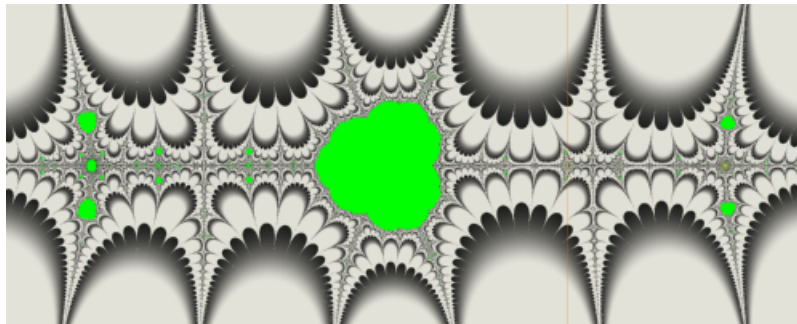


Figure 42.3. Collatz set over complex continuous inputs centered on $z = 0$ in a window of width ≈ 5 . The boundary of the green regions is the Collatz fractal. (Zoom: 1)

3.1 Self Similarity

Part of the definition of ‘fractal’ is the property of *self-similarity*, which means a piece of the set reflects the whole set. This means, for instance, that we can center near any ‘green blob’ and magnify, and the resulting image should contain elements from Figure 42.3. The Collatz fractal surely demonstrates this, as shown in Figures 42.4, 42.5. The former Figure, zoomed in at the point $z = (2.01, 0.368)$, shows a near-ninety degree rotation base fractal. Choosing another point on which to center and zoom, particularly $z = (2.04, 0.415)$, we produce the latter Figure, containing a varied view of the base fractal.

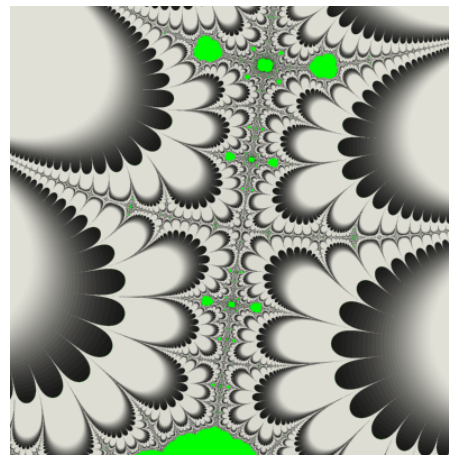


Figure 42.4. Self-similarity of the Collatz set. Centered at $(2.01, 0.368)$. (Zoom: 32)

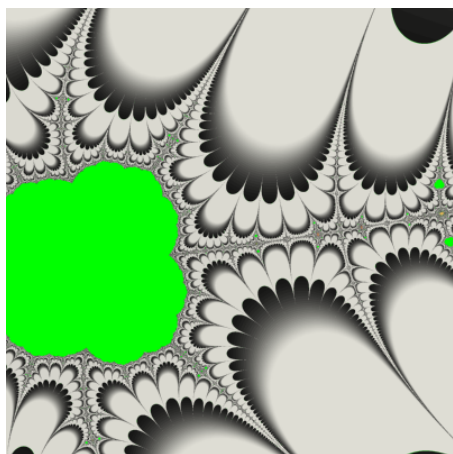


Figure 42.5. Self-similarity of the Collatz set. Centered at (2.04, 0.415). (Zoom: 256)

4 Collatz Gems

The natural numbers, which are the testing ammunition for the Collatz conjecture on the real line, bear minuscule representation in Collatz fractal as depicted in Figure 42.3. Such numbers are surely in the set, but are they reduced to infinitesimal points? The answer is a resounding ‘no’, and we must zoom in a bit to appreciate what happens on and near natural numbers.

4.1 Anatomy of a Gem

Shown in Figures 42.6, 42.7 are two images of the Collatz ‘gem’ (a word we’ll keep) located at $z = 1$. The boundary of the gem is part of the Collatz fractal, and the interior is colored according to the convergence time of the point tested. Note the subtle presence of the infinitesimal line $z = (1, y)$ streaking down the center of each image. Outside of the gem is an infinitude of self-similarity of the fractal, as evident by the green speckles scattered about.

From what we know of the discrete Collatz set, $z = 1$ should be associated with the same shade of orange that occupies the vertical stripe in Figure 42.1. In the Collatz gem, we have to magnify profoundly to verify this as shown in Figures 42.8, 42.9. Many narrow bands of color are passed on the way there.

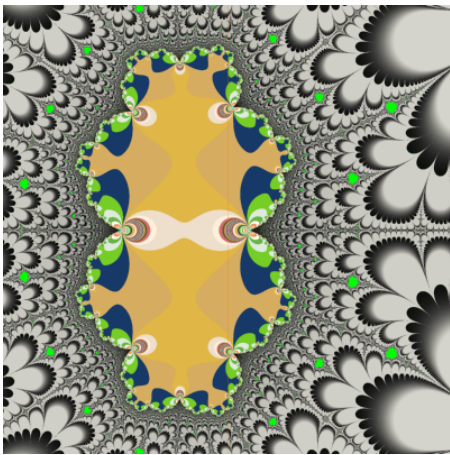


Figure 42.6. Collatz gem at $z = 1$. (Zoom: 128)

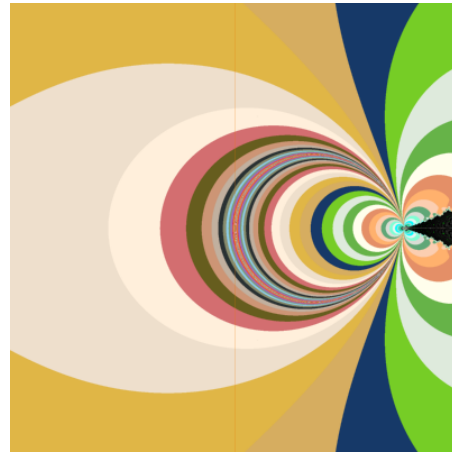


Figure 42.7. Collatz gem at $z = 1$. (Zoom: 1024)



Figure 42.8. Collatz gem at $z = 1$. (Zoom: 4194304)



Figure 42.9. Collatz gem at $z = 1$. (Zoom: 536870912)

4.2 Collatz Gems

Collatz gems occur at (seemingly) all natural number positions $z > 1$, and can be studied in similar fashion to the $z = 1$ case. The color of the gem varies throughout its shape, but the shade pinpointed on the exact natural number always corresponds to the discrete case. A gem's interior shading near the boundaries accentuates the Collatz fractal occurring there. Interestingly too, the gem is never (or rarely) centered on the natural number value itself. Figures 42.10-42.15 survey the Collatz gems from $z = 2$ to $z = 7$.

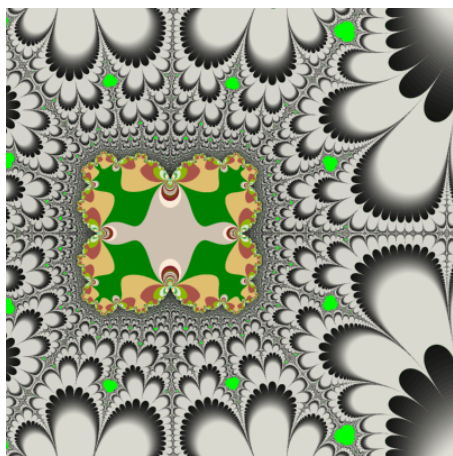


Figure 42.10. Collatz gem at $z = 2$. (Zoom: 32)

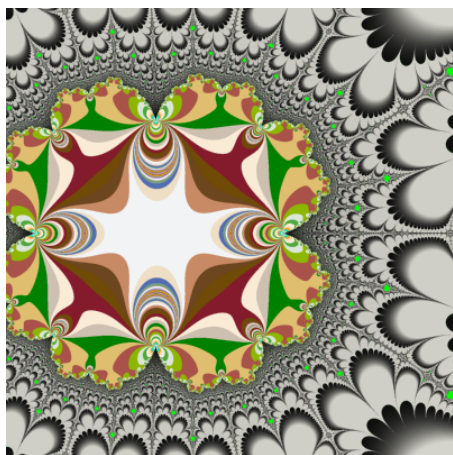


Figure 42.11. Collatz gem at $z = 3$. (Zoom: 512)

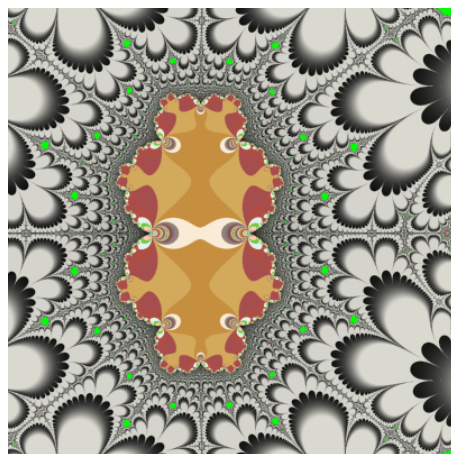


Figure 42.12. Collatz gem at $z = 4$. (Zoom: 32)

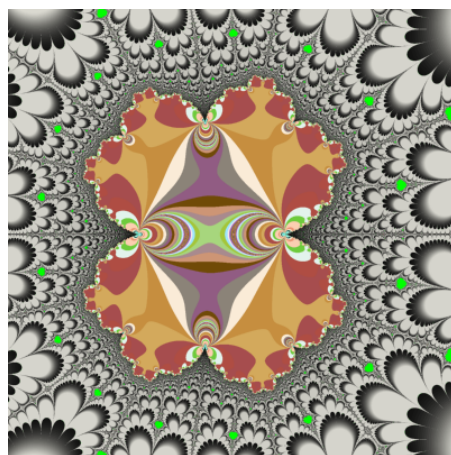


Figure 42.13. Collatz gem at $z = 5$. (Zoom: 256)

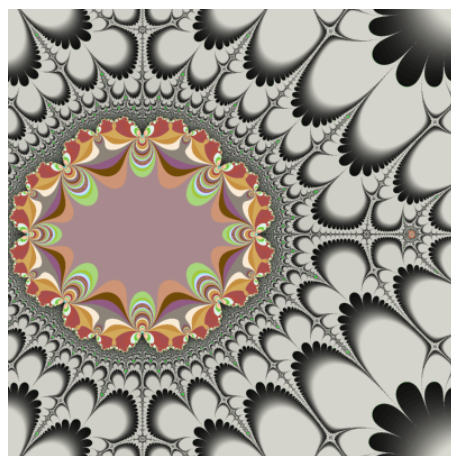


Figure 42.14. Collatz gem at $z = 6$. (Zoom: 128)

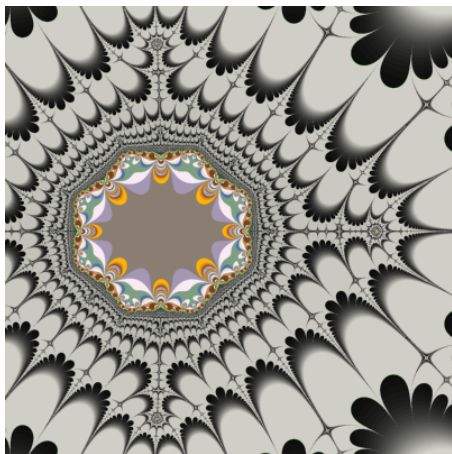


Figure 42.15. Collatz gem at $z = 7$. (Zoom: 1024)

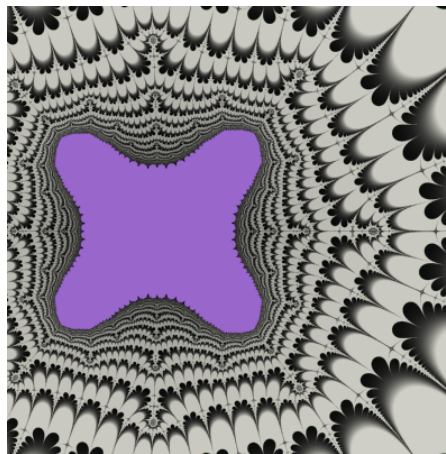


Figure 42.16. Collatz gem at $z = 27$. (Zoom: 4096)

4.3 Pathological Gem $z=27$

The number 27 is arguably the most quirky input one can send to the Collatz function. As mentioned, it takes 111 iterations to finally converge, begging the question of whether the Collatz gem at $z = 27$ tells a related story. Recursing on values near c (27) and looking up the corresponding gem, we get the rather uninteresting shape shown in Figure 42.16. It is of uniform shade with index 111 corresponding to the convergence time. The boundary is still a Collatz fractal, but the overall appearance is unlike other gems.

Let us chase the progression of 27 through the Collatz pipeline, i.e. $c(27) = 82$, and then $c(82) = 41$, etc., and view the corresponding gems. For the 82-case we produce Figure 42.17. The shading is again flat and few features are discernible. This pattern continues to the next number in the sequence, i.e. 41, as shown in Figure 42.18.

Going the other direction, we can check power-of-two multiples of 27 and see the same ‘flat color’ effect. For instance, trying $z = 27 \times 2^5 = 864$ leads to Figure 42.16, and the interior gem shading is still flat. Finally when we try $z = 27 \times 2^{10} = 27648$ and magnify like crazy, we see the alternating color bands in a typical Collatz gem shown in Figure 42.21. The view had to be slightly offset to capture the image.

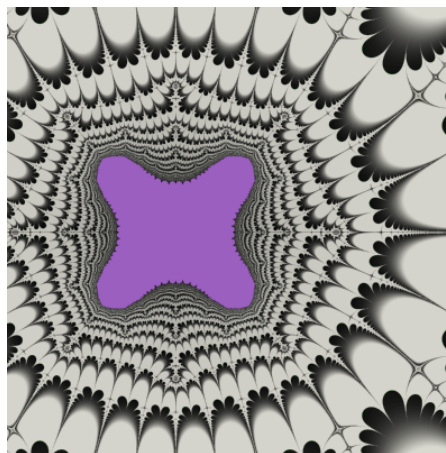


Figure 42.17. Collatz gem at $z = 82$. (Zoom: 1024)

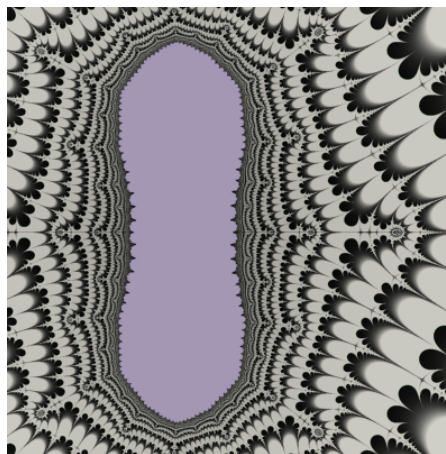


Figure 42.18. Collatz gem at $z = 41$. (Zoom: 4096)

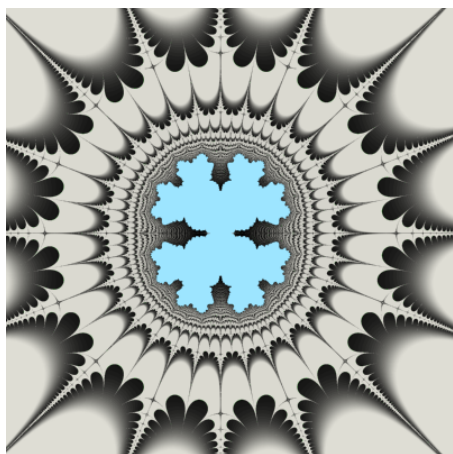


Figure 42.19. Collatz gem at $z = 864$. (Zoom: 2048)

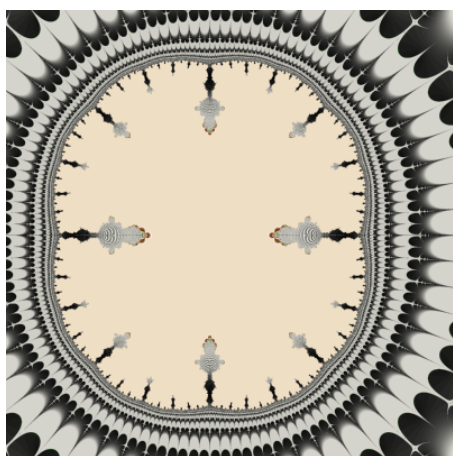


Figure 42.20. Collatz gem at $z = 27648$. (Zoom: 131072)

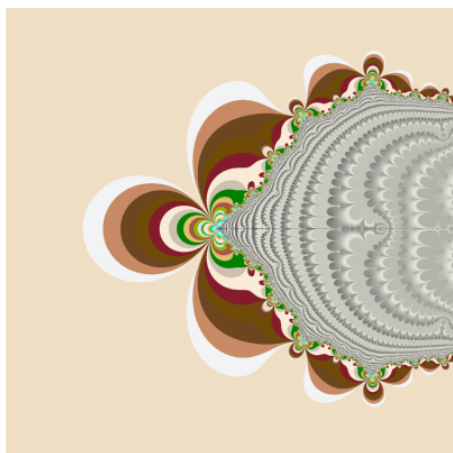


Figure 42.21. Collatz gem at $z = 27648$. (Zoom: 2097152, Off-centered)

4.4 Non-natural numbers

The Collatz set, and any gem that helps fill it, is not limited to the positive real axis. We find interesting structure at decimal z and at negative points on the real axis as shown in Figures 42.22, 42.23. We see another Collatz gem in the former case, as evident by the alternating bands of color representing convergence time. In the latter case we find with a clover-like glyph, yet another face of the Collatz fractal.

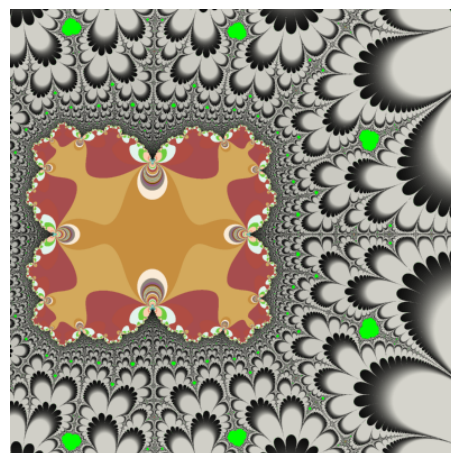


Figure 42.22. Collatz gem. Center: $z = (0.61082, 0)$. (Zoom: 256)



Figure 42.23. Collatz set. Center: $z \approx (-1.2656, 0)$. (Zoom: 8192)

4.5 Further Study

All views of the Collatz fractal shown above, along with all embeded gems, have been generated using Equation (42.3),

an order-one generalization of Equation (42.4). The study could be repeated using Equation (42.4) using any set of integers q_e, q_o , which will undoubtedly have some bearing on the appearance of the fractal and it's gems.

Shown in Figures 42.24, 42.25 are simplified views of the familiar 'order-one' Collatz fractal, followed by a pure order-two representation. The qualitative views are different in each, and the gem at $z = 2$ is visible in the latter case.

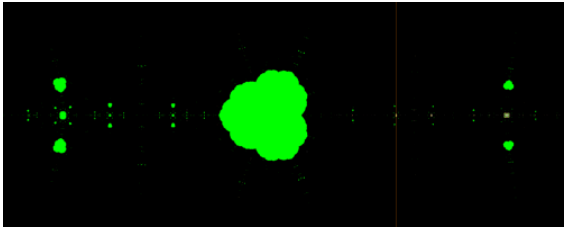


Figure 42.24. Simplified view of order-one Collatz fractal.

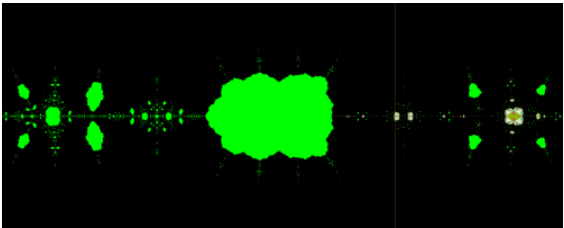


Figure 42.25. Simplified view of order-two Collatz fractal.

5 Futility

Regarding the Collatz conjecture, Paul Erdős himself conceded that *mathematics may not be ready for such problems*. Despite this, let us try the problem head-on.

5.1 Collatz Sequence

Let us process the first few natural numbers $\{n\} = 1, 2, 3, \dots$, via Equation (42.1), namely

$$f(n) = \begin{cases} n/2 & n \text{ even} \\ 3n + 1 & n \text{ odd} \end{cases},$$

until $f(n) = 1$. Stopping at $n = 15$, a table of sequences

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	10	2	16	3	22	4	28	5	34	6	40	7	46
		5	1	8	10	11	2	14	16	17	3	20	22	23
		16		4	5	34	1	7	8	52	10	10	11	70
		8		2	16	17		22	4	26	5	5	34	35
		4		1	8	52		11	2	13	16	16	17	106
		2			4	26		34	1	40	8	8	52	53
		1				2	13	17		20	4	4	26	160
						1	40	52		10	2	2	13	80
							20	26		5	1	1	40	40
							10	13		16			20	20
							5	40		8			10	10
							.	.		.			.	.

can be made, where columns that end with a dot (.) continue to a sub-sequence written elsewhere in the table.

In the table above, each column is a sequence $\phi(n, k)$ that starts at number n and ends at number k . The simplest meaningful case starts with $n = 2$, meaning $\phi(2, 1) = (2, 1)$, whereas for $n = 3$ we have $\phi(3, 1) = (3, 10, 5, 16, 8, 4, 2, 1)$. Let us denote the ‘size’ of a sequence, i.e. the number of elements, as $\tilde{n} + 1$, where $\tilde{n}$ is the convergence time of n .

For certain base numbers, we observe that the resulting sequence may start repeating that of a previous base number. For instance, take $n = 6$ to write

$$\begin{aligned} \phi(6, 1) &= (6, 3, 10, 5, 16, 8, 4, 2, 1) \\ &= (6, 3, 10) (5, 16, 8, 4, 2, 1) \\ &= \phi(6, 10) \phi(5, 1) . \end{aligned}$$

Progress Operator

The Collatz function written as Equation (42.1) is meant to be used recursively. However, let us coin a second term for the process

$$n \rightarrow f(n)$$

by calling f the *progress operator*. Referenced in this way, we note that $f(n)$ is applied not recursively, but rather once to the input number n .

Up to $n = 45$, we use the progress operator to write the following table:

$f(1) = 4$	$f(16) = 8$	$f(31) = 94$
$f(2) = 1$	$f(17) = 52$	$f(32) = 16$
$f(3) = 10$	$f(18) = 9$	$f(33) = 100$
$f(4) = 2$	$f(19) = 58$	$f(34) = 17$
$f(5) = 16$	$f(20) = 10$	$f(35) = 106$
$f(6) = 3$	$f(21) = 44$	$f(36) = 18$
$f(7) = 22$	$f(22) = 11$	$f(37) = 112$
$f(8) = 4$	$f(23) = 70$	$f(38) = 19$
$f(9) = 28$	$f(24) = 12$	$f(39) = 118$
$f(10) = 5$	$f(25) = 76$	$f(40) = 20$
$f(11) = 34$	$f(26) = 13$	$f(41) = 124$
$f(12) = 6$	$f(27) = 82$	$f(42) = 21$
$f(13) = 40$	$f(28) = 14$	$f(43) = 130$
$f(14) = 7$	$f(29) = 88$	$f(44) = 22$
$f(15) = 46$	$f(30) = 15$	$f(45) = 136$

Additionally, define the *compound progress operator*

$f_k(n)$ as the progress operator applied $k > 1$ times to a base number n , for instance:

$$\begin{aligned} f_2(n) &= f(f(n)) \\ f_k(n) &= f(f_{k-1}(n)) \end{aligned}$$

Let $\tilde{n}$ denote the convergence time for a given input n . In terms of the progress operator and convergence time, we see that any number n satisfying the Collatz conjecture must also satisfy

$$f_{\tilde{n}}(n) = 1,$$

where presumably $\tilde{n}$ is finite for finite n . This is called the *Collatz condition*.

5.2 Regress Operator

Let $\{n\}$ represent the set of all n that lead to the same $f(n)$ such that

$$f(\{n\}) = f(n).$$

The output of the progress operator is not unique, easily seen with $f(3) = 10$ and $f(20) = 10$ and many others. For this reason, let us define the *regress operator* $g(n)$, whose output is multi-valued, as

$$g(f(n)) = \{n\}.$$

For a consistency check, apply f to each side to find

$$f(g(f(n))) = f(\{n\}) = f(n),$$

showing that f, g are mutually inverse to each other.

Up to $n = 45$, we use the regress operator to write the following table:

$g(1) = 2$	$g(16) = 32, 5$	$g(31) = 62$
$g(2) = 4$	$g(17) = 34$	$g(32) = 64$
$g(3) = 6$	$g(18) = 36$	$g(33) = 66$
$g(4) = 8, 1$	$g(19) = 38$	$g(34) = 68, 11$
$g(5) = 10$	$g(20) = 40$	$g(35) = 70$
$g(6) = 12$	$g(21) = 42$	$g(36) = 72$
$g(7) = 14$	$g(22) = 44, 7$	$g(37) = 74$
$g(8) = 16$	$g(23) = 46$	$g(38) = 76$
$g(9) = 18$	$g(24) = 48$	$g(39) = 78$
$g(10) = 20, 3$	$g(25) = 50$	$g(40) = 80, 13$
$g(11) = 22$	$g(26) = 52$	$g(41) = 82$
$g(12) = 24$	$g(27) = 54$	$g(42) = 84$
$g(13) = 26$	$g(28) = 56, 9$	$g(43) = 86$
$g(14) = 28$	$g(29) = 58$	$g(44) = 88$
$g(15) = 30$	$g(30) = 60$	$g(45) = 90$

The $g(n)$ table above exhibits clear order. For each $n > 0$, there is at least one solution $g(n) = 2n$, and this accounts for all even outputs from $g(n)$.

The odd outputs from $g(n)$ occur as second solutions to the cases 4, 10, 16, 22, 28, 34, and so on. To quantify these, define an integer index q , and by inspection of the above, write

$$g(4 + 6q) = \begin{cases} 2(4 + 6q) \\ 1 + 2q \end{cases}$$

for $q = 0, 1, 2, 3, \dots$

Regress Tree

The regress operator can be used recursively, necessitating the *compound regress operator* applied k times to a number n :

$$g_k(n) = g(g_{k-1}(n))$$

Of course, since the regress operator may return multiple values, we expect the total output of $g_k(n)$ is a tree-like structure of numbers.

To develop a so-called *regress tree*, begin with the Collatz condition

$$f_{\tilde{n}}(n) = 1$$

and apply the g_1 operator to both sides:

$$g_1(f_{\tilde{n}}(n)) = g_1(1) = g(1) = 2$$

On the left, the presence of g_1 effectively knocks $\tilde{n}$ down by one, resulting in

$$f_{\tilde{n}-1}(n) = 2.$$

Start again with the Collatz condition and apply g_2 to find

$$f_{\tilde{n}-2}(n) = g_2(1) = g(g(1)) = g(2) = 4,$$

and continue the pattern to get

$$f_{\tilde{n}-3}(n) = g_3(1) = g(g(g(1))) = g(4) = 8,$$

and furthermore

$$f_{\tilde{n}-4}(n) = g_4(1) = g(8) = 16.$$

The first multi-valued case is $\tilde{n} - 5$, particularly:

$$f_{\tilde{n}-5}(n) = g_5(1) = g(16) = \begin{cases} 32 \\ 5 \end{cases}$$

With two branches to tend, all subsequent cases develop on one or the other branch. For $\tilde{n} - 6$, apply $g(n)$ to each side to have:

$$f_{\tilde{n}-6}(n) = g_6(1) = \begin{cases} g(32) \\ g(5) \end{cases} = \begin{cases} 64 \\ 10 \end{cases}$$

The pattern started here can be repeated indefinitely. Continuing beyond $g_6(1)$, we produce the following regress trees:

$$g_7(1) = \begin{cases} \begin{cases} g(64) \\ g(10) \end{cases} \\ \begin{cases} 128 \\ 21 \\ 20 \\ 3 \end{cases} \end{cases}$$

$$g_8(1) = \begin{cases} \begin{cases} g(128) \\ g(21) \end{cases} \\ \begin{cases} g(20) \\ g(3) \end{cases} \end{cases} = \begin{cases} \begin{cases} 256 \\ 42 \end{cases} \\ \begin{cases} 40 \\ 6 \end{cases} \end{cases}$$

$$g_9(1) = \begin{cases} \begin{cases} g(256) \\ g(42) \end{cases} \\ \begin{cases} g(40) \\ g(6) \end{cases} \end{cases} = \begin{cases} \begin{cases} \begin{cases} 512 \\ 85 \end{cases} \\ 84 \end{cases} \\ \begin{cases} \begin{cases} 80 \\ 13 \end{cases} \\ 12 \end{cases} \end{cases}$$

$$g_{10}(1) = \begin{cases} \begin{cases} \begin{cases} g(512) \\ g(85) \end{cases} \\ g(84) \end{cases} \\ \begin{cases} \begin{cases} g(80) \\ g(13) \end{cases} \\ g(12) \end{cases} \end{cases} = \begin{cases} \begin{cases} \begin{cases} 1024 \\ 170 \end{cases} \\ 168 \end{cases} \\ \begin{cases} \begin{cases} 160 \\ 26 \end{cases} \\ 24 \end{cases} \end{cases}$$

$$g_{11}(1) = \begin{cases} \begin{cases} \begin{cases} \begin{cases} g(1024) \\ g(170) \end{cases} \\ g(168) \end{cases} \\ \begin{cases} \begin{cases} g(160) \\ g(26) \end{cases} \\ g(24) \end{cases} \end{cases} = \begin{cases} \begin{cases} \begin{cases} \begin{cases} 2048 \\ 341 \end{cases} \\ 340 \end{cases} \\ 336 \end{cases} \\ \begin{cases} \begin{cases} \begin{cases} 320 \\ 53 \end{cases} \\ 52 \end{cases} \\ 48 \end{cases} \end{cases}$$

$$g_{12}(1) = \begin{cases} \begin{cases} \begin{cases} \begin{cases} g(2048) \\ g(341) \end{cases} \\ g(340) \end{cases} \\ g(336) \end{cases} \\ \begin{cases} \begin{cases} g(320) \\ g(53) \end{cases} \\ g(52) \\ g(48) \end{cases} \end{cases} = \begin{cases} \begin{cases} \begin{cases} \begin{cases} 4096 \\ 682 \end{cases} \\ 680 \\ 113 \end{cases} \\ 672 \end{cases} \\ \begin{cases} \begin{cases} \begin{cases} 640 \\ 106 \end{cases} \\ 104 \\ 17 \end{cases} \\ 96 \end{cases} \end{cases}$$

$$g_{13}(1) = \begin{cases} \begin{cases} \begin{cases} \begin{cases} g(4096) \\ g(682) \end{cases} \\ g(680) \\ g(113) \end{cases} \\ g(672) \end{cases} \\ \begin{cases} \begin{cases} g(640) \\ g(106) \end{cases} \\ g(104) \\ g(17) \\ g(96) \end{cases} \end{cases} = \begin{cases} \begin{cases} \begin{cases} \begin{cases} 8192 \\ 1365 \end{cases} \\ 1364 \\ 227 \end{cases} \\ 1344 \end{cases} \\ \begin{cases} \begin{cases} \begin{cases} 1280 \\ 213 \end{cases} \\ 212 \\ 35 \end{cases} \\ 208 \\ 34 \\ 192 \end{cases} \end{cases}$$

$$g_{14}(1) = \left(\left(\left(\left(\begin{array}{l} g(8192) \\ g(1365) \end{array} \right) \right) \left(\begin{array}{l} g(1364) \\ g(227) \end{array} \right) \right) \left(\begin{array}{l} g(1360) \\ g(226) \end{array} \right) \right) g(1344) = \left(\left(\left(\left(\begin{array}{l} 16384 \\ 2730 \end{array} \right) \right) \left(\begin{array}{l} 2728 \\ 454 \end{array} \right) \right) \left(\begin{array}{l} 2720 \\ 453 \end{array} \right) \right) \left(\begin{array}{l} 452 \\ 75 \end{array} \right) 2688$$

$$\left(\left(\left(\left(\begin{array}{l} g(1280) \\ g(213) \end{array} \right) \right) \left(\begin{array}{l} g(212) \\ g(35) \end{array} \right) \right) \left(\begin{array}{l} g(208) \\ g(34) \end{array} \right) \right) g(192) = \left(\left(\left(\left(\begin{array}{l} 2560 \\ 426 \end{array} \right) \right) \left(\begin{array}{l} 424 \\ 70 \end{array} \right) \right) \left(\begin{array}{l} 416 \\ 69 \end{array} \right) \right) \left(\begin{array}{l} 68 \\ 11 \end{array} \right) 384$$

By flattening each regress tree to a sequence of numbers $h(\tilde{n})$, we can represent all of the above in a condensed table:

$$\begin{aligned} h(0) &= (1) \\ h(1) &= (2) \\ h(2) &= (4) \\ h(3) &= (8) \\ h(4) &= (16) \\ h(5) &= (32, 5) \\ h(6) &= (64, 10) \\ h(7) &= (128, 21, 20, 3) \\ h(8) &= (256, 42, 40, 6) \\ h(9) &= (512, 85, 84, 80, 13, 12) \\ h(10) &= (1024, 170, 168, 160, 26, 24) \\ h(11) &= (2048, 341, 340, 336, 320, 53, 52, 48) \\ h(12) &= (4096, 682, 680, 113, 672, 640, 106, 104, 17, 96) \\ h(13) &= (8192, 1365, 1364, 227, 1360, 226, 1344, 1280, 213, 212, 35, 208, 34, 192) \\ &\dots \end{aligned}$$

By construction, any given sequence $h(\tilde{n})$ contains all members with convergence time $\tilde{n}$. No number occurs more than once throughout the table.

Tree Growth

Looking again at the regress operator, it follows that $g(n)$ is single valued for five of every six instances, and is multi-valued for one of every six. From this, it should follow that each member in a given $h(\tilde{n})$ has a rough $1/6$ chance to introduce a new number in the preceding $h(\tilde{n} + 1)$.

Let $N_{\tilde{n}}$ denote the number of members in any given $h(\tilde{n})$. From the above, we write

$$N(\tilde{n} + 1) \approx N(\tilde{n}) \left(1 + \frac{1}{6}\right),$$

which implies exponential growth in N :

$$N(\tilde{n}) \approx e^{\tilde{n}/6}$$

As a matter of curiosity, recall that the number 27 has a convergence time of $\tilde{n} = 111$. In light of the above, there are approximately

$$N(111) \approx e^{111/6} \approx 10^8$$

numbers with the same convergence time.

Let $T(\tilde{n})$ denote the sum of all $N(k)$ for all $k \leq \tilde{n}$, i.e.

$$T(\tilde{n}) = \sum_{k=0}^{\tilde{n}} N(k) \approx \sum_{k=0}^{\tilde{n}} e^{k/6}.$$

In the large- $\tilde{n}$ regime, we may further approximate the sum of exponential terms as a continuous integral in k having step size dk :

$$T(\tilde{n}) \approx \int_0^{\tilde{n}} e^{k/6} dk = 6 e^{k/6} \Big|_0^{\tilde{n}} \approx 6 e^{\tilde{n}/6}$$

5.3 Base-Odd Lattice

It's easy to see that odd numbers are the star players in the problem, as all even numbers eventually collapse down to an odd number m . Rewriting the set of natural numbers in this light, we can write out a *base-odd lattice*, where a column starts with m and grows downward in multiples of 2^λ , where $\lambda = 1, 2, 3, \dots$. Doing so for 10 rows, 9 columns, we have:

<i>1</i>	<i>3</i>	<i>5</i>	<i>7</i>	<i>9</i>	<i>11</i>	<i>13</i>	<i>15</i>	<i>17</i>
2	6	10	14	18	22	26	30	34
4	12	20	28	36	44	52	60	68
8	24	40	56	72	88	104	120	136
16	48	80	112	144	176	208	240	272
32	96	160	224	288	352	416	480	544
64	192	320	448	576	704	832	960	1088
128	384	640	896	1152	1408	1664	1920	2176
256	768	1280	1792	2304	2816	3328	3840	4352
512	1536	2560	3584	4608	5632	6656	7680	8704

Collatz sequences can be easily visualized using the base-odd lattice: any process $n \rightarrow 3n + 1$ will land a new n in the field of even numbers. By $n \rightarrow n/2$, n effectively ‘boils upward’ to an odd number in the top row. In the above, italicized odd integers lead to boldface evens in nearby columns.

As it turns out, a complete base-odd lattice has (nearly)

every second bold entry corresponding to *some* odd number. For instance, jumping down the leftmost column to 16, 64, 256, 1024, 4096, etc., it follows that each original $n \rightarrow 3n + 1$ had to be 5, 31, 85, 341, 1365, etc. Continuing in the space allotted, we find a pattern in the boldface even numbers that is conspicuously absent in any column whose elements are divisible by three.

1	3	5	7	9	11	13	15	17
2	6	10	14	18	22	26	30	34
4	12	20	28	36	44	52	60	68
8	24	40	56	72	88	104	120	136
16	48	80	112	144	176	208	240	272
32	96	160	224	288	352	416	480	544
64	192	320	448	576	704	832	960	1088
128	384	640	896	1152	1408	1664	1920	2176
256	768	1280	1792	2304	2816	3328	3840	4352
512	1536	2560	3584	4608	5632	6656	7680	8704

Any odd number m in the top row sits on an infinite stack of even numbers $\{m \cdot 2^l\}$. It follows that if m satisfies the Collatz conjecture, so too does the stack of evens under m . Occurring in columns divisible by three are boldface even numbers, each of which points to a different odd number $\neq m$ that would satisfy the Collatz conjecture by relationship to m .

Denote $\beta(m)$ as the set of odd numbers that come from boldface evens in the m th column of the base-odd lattice. Also let $\beta_k(m)$ denote the k th odd number in the set $\beta(m)$, where the index $k = 1, 2, 3, \dots$ is always positive. Reading down the $m = 1$ column, the odd numbers $\beta(1)$ implied by the bold evens are given by

$$\beta_k(1) = \frac{1 \cdot 2^{2k} - 1}{3}.$$

The set $\beta(3)$ contains nothing. The next nontrivial column is $m = 5$ and we correspondingly have

$$\beta_k(5) = \frac{5 \cdot 2^{2k-1} - 1}{3}.$$

The pattern in $\beta_k(m)$ repeats every three columns. Defining an index $j = 0, 1, 2, \dots$ such that $m = 1 + 6j$ and so on, we may write

$$\beta_k(1 + 6j) = \frac{(1 + 6j) \cdot 2^{2k} - 1}{3}$$

$$\beta_k(3 + 6j) = \emptyset$$

$$\beta_k(5 + 6j) = \frac{(5 + 6j) \cdot 2^{2k-1} - 1}{3}.$$

Expanding these for the first few j, k leads to the following:

$$\begin{aligned} \beta(1) &= (1, 5, \mathbf{21}, 85, 341, \mathbf{1365}, 5461, 21845, \mathbf{87381}, \dots) \\ \beta(5) &= (3, 13, 53, \mathbf{213}, 853, 3413, \mathbf{13653}, 54613, \dots) \\ \beta(7) &= (\mathbf{9}, 37, 149, \mathbf{597}, 2389, 9557, \mathbf{38229}, 152917, \dots) \\ \beta(11) &= (7, 29, \mathbf{117}, 469, 1877, \mathbf{7509}, 30037, 120149, \dots) \\ \beta(13) &= (17, \mathbf{69}, 277, 1109, \mathbf{4437}, 11749, 70997, \mathbf{283989}, \dots) \\ \beta(17) &= (11, \mathbf{45}, 181, 725, \mathbf{2901}, 11605, 46421, \mathbf{185685}, \dots) \\ \beta(19) &= (25, 101, \mathbf{405}, 1621, 6485, \mathbf{25941}, 103765, 415061, \dots) \\ \beta(23) &= (\mathbf{15}, 61, 245, \mathbf{981}, 3925, 15701, \mathbf{62805}, 251221, \dots) \\ \beta(25) &= (\mathbf{33}, 133, 533, \mathbf{2133}, 8533, 34133, \mathbf{136533}, 546133, \dots) \\ \beta(29) &= (19, 77, \mathbf{309}, 1237, 4949, \mathbf{19797}, 79189, 316757, \dots) \\ \beta(31) &= (41, \mathbf{165}, 661, 2645, \mathbf{10581}, 42325, 169301, \mathbf{677205}, \dots) \end{aligned}$$

Explicitly, we see if $m = 1$ satisfies the Collatz conjecture, then so does $(5, 21, 85, \dots)$. Similarly, if $m = 5$ satisfies the Collatz conjecture, then so does $(3, 13, 53, \dots)$, and so on.

Odd numbers divisible by three are denoted in boldface, corresponding to the empty sequences $\beta(3), \beta(9)$, etc.

Self-Similarity

The sequences $\{\beta_k(m)\}$ exhibit some self-similarity. For instance, one can easily puzzle out that

$$\begin{aligned}\beta_{k+1}(5) &= 10 \cdot \beta_k(1) + 3 \\ \beta_{k+1}(25) &= 100 \cdot \beta_k(1) + 33,\end{aligned}$$

along with several other patterns the reader is invited to find.

Odd Number Generator

The equations $\beta_k(m)$ above are equipped to generate a table of odd numbers as follows:

j	k	$\beta_k(1 + 6j)$	j	k	$\beta_k(5 + 6j)$
0	1	1	0	1	3
0	2	5	1	1	7
1	1	9	2	1	11
2	1	17	0	2	13
0	3	21	3	1	15
3	1	25	4	1	19
4	1	33	5	1	23
1	2	37	6	1	27
5	1	41	1	2	29
6	1	49	7	1	31
			8	1	35
			9	1	39
			10	1	43
			2	2	45
			11	1	47
			12	1	51
			0	3	53
			13	1	55

Forbidding the odd number m from being divisible by three, i.e.

$$\frac{m}{3} \notin \mathbb{N},$$

the β -equations above can be generalized into a single equation

$$\gamma_\lambda(m) = \frac{m \cdot 2^\lambda - 1}{3}.$$

Generating a similar table of odd numbers from γ leads to:

m	λ	$\gamma_\lambda(m)$	m	λ	$\gamma_\lambda(m)$
1	2	1	11	3	29
5	1	3	47	1	31
1	4	5	25	2	33
11	1	7	29	2	35
7	2	9	7	4	37
17	1	11	59	1	39
5	3	13	31	2	41
23	1	15	65	1	43
13	2	17	17	3	45
29	1	19	71	1	47
1	6	21	37	2	49
35	1	23	77	1	51
19	2	25	5	5	53
41	1	27	83	1	55

For a sanity check, note that the progress operator applied to γ , which is always odd, reads

$$f(\gamma_\lambda(m)) = m \cdot 2^\lambda.$$

The right side is always even by construction. Moreover, we see that λ additional applications of the progress operator brings the working number down to m :

$$f_{\lambda+1}(\gamma_\lambda(m)) = m$$

Since γ is arbitrary, we now see that any odd number is related by the progress operator to another odd number not divisible by three. Meanwhile, any odd number divisible by three can only occur as the first odd number in a Collatz sequence ϕ .

Integer Oddness

In the formula $\gamma_\lambda(m)$, we can think of $1/\lambda$ as the ‘oddness’ of any given odd number. The limit case $\lambda = 1$ corresponds to ‘very odd numbers’, which arise when the input m is greater than the result $\gamma_1(m)$.

Crossout Analysis

The sequences $\{\beta_k(m)\}$ avail a method for testing a range of odd integers starting from one and ending at a cutoff N . Choosing $N = 29$ for the sake of demonstration, we start by listing the odd numbers between 1 and N , inclusive.

Treating $n = 1$ as the only ‘tested’ case, cross out from the list any numbers less than N that occur in the sequence $\beta(1)$:

~~1~~	3	~~5~~	7	9
11	13	15	17	19
~~21~~	23	25	27	29

Since $n = 5$ has been crossed out, we may read from $\beta(5)$ and cross out any numbers less than N :

$$\begin{array}{ccccc} \cancel{1} & \cancel{3} & \cancel{5} & 7 & 9 \\ 11 & \cancel{13} & 15 & 17 & 19 \\ \cancel{21} & 23 & 25 & 27 & 29 \end{array}$$

This time $n = 13$ has been crossed out, which means we may read across $\beta(13)$ and cross out $n = 17$, which then gives us $n = 11$, and immediately leading to 7, 29 from $\beta(11)$:

$$\begin{array}{ccccc} \cancel{1} & \cancel{3} & \cancel{5} & \cancel{7} & 9 \\ \cancel{11} & \cancel{13} & 15 & \cancel{17} & 19 \\ \cancel{21} & 23 & 25 & 27 & \cancel{29} \end{array}$$

Continuing this pattern, one finds that 9, 19, 25 can also be crossed out:

$$\begin{array}{ccccc} \cancel{1} & \cancel{3} & \cancel{5} & \cancel{7} & \cancel{9} \\ \cancel{11} & \cancel{13} & 15 & \cancel{17} & \cancel{19} \\ \cancel{21} & 23 & \cancel{25} & 27 & \cancel{29} \end{array}$$

So far then, only 15, 23, 27 don't follow automatically from $\beta(1)$ and its branches. Going after the 15-case manually, one writes

$$15 = \frac{(5 + 6j) \cdot 2^{2k-1} - 1}{3},$$

solved by $j = 3$, $k = 1$, which corresponds to the first entry in the list $\beta(23)$. Therefore we may cross out 15, whose fate depends on $n = 23$.

To proceed, we can show that $n = 23$ is generated by $j = 5$, $k = 1$ using the same equation, which means 23 is the first entry in $\beta(35)$. We're stuck verifying either 23 or $n = 35$ explicitly.

Finally, we have that $n = 27$ is generated by $j = 6$, $k = 1$, which means 27 is the first entry in $\beta(41)$. We're stuck verifying either 27 or $n = 41$ explicitly.

Curiously, note that the numbers not trivially covered by crossout analysis - namely 15, 23, 27 - each qualify as very odd integers with $k = 1$.

5.4 Cycles

A cycle occurs if a given number n is the result of q applications of the progress operator:

$$f_q(n) = n$$

The number q is identified as the period of the cycle.

Denote a cycle of period q as a sequence $\theta(q)$. For instance, a cycle representing the English alphabet would read

$$\theta(26) = (a, b, c, \dots, x, y, z, a)$$

Note there are $26 + 1$ members in the sequence, as the last entry is an extra copy of the first. If a number n occurs in a cycle, it should make sense that its immediate neighbor $f(n)$, and at least one solution of $g(n)$, also occur in the (same) cycle.

The numbers in a cycle cannot be monotonically increasing or decreasing, thus there must be at least one odd number in a cycle.

Single-Odd Cycle

Suppose a cycle of period q contains a single odd number m at position j in the sequence. To represent this, begin with:

$$\theta(q) = \left(2^{j-1}m, \dots, 8m, 4m, 2m, m, 3m+1, \frac{3m+1}{2}, \frac{3m+1}{4}, \dots, \frac{3m+1}{2^{q-j}} \right)$$

Since the first and last terms are equal, we eliminate j to find

$$2^{q-1}m = 3m + 1,$$

and furthermore

$$m = \frac{1}{2^{q-1} - 3}.$$

This system is solved by $m = 1$, $q = 3$, telling us the cycle has period three and contains the odd number 1. From

this we finally write the single-odd cycle:

$$\theta(3) = (1, 4, 2, 1)$$

Double-Odd Cycle

Suppose there exists a cycle $\theta(q)$ that contains exactly two odd numbers m, n :

$$\theta(q) = (a, \dots, 2m, m, 3m+1, \dots, 2n, n, 3n+1, \dots, a)$$

Also suppose that n is λ positions to the right of m , and similarly that m is λ' positions to the right of n . That is

$$\begin{aligned} f_\lambda(m) &= n \\ f_{\lambda'}(n) &= m, \end{aligned}$$

where the period q of the cycle is given by

$$q = \lambda + \lambda'.$$

Since all members in θ are even besides m, n , we can explicitly write

$$\begin{aligned} \frac{3m+1}{2^{\lambda-1}} &= n \\ \frac{3n+1}{2^{\lambda'-1}} &= m. \end{aligned}$$

Solve the system for m, n explicitly to find

$$\begin{aligned} m &= \frac{2^{\lambda-1} + 3}{2^{q-2} - 9} \\ n &= \frac{2^{\lambda'-1} + 3}{2^{q-2} - 9}. \end{aligned}$$

The above can be used to check for odd solutions for m, n with any given λ, λ' .

In the special case that $\lambda = \lambda'$, we must also have $n = m$, and suddenly we're back to a one-odd cycle, contradicting the starting assumption. Thus we must have $\lambda \neq \lambda'$ for a double-odd cycle.

The common denominator in the equations for m, n can never be negative, telling us

$$q > 2 + \log_2(9),$$

therefore the minimum q is 6.

For another special case, note if λ, λ' are large numbers, the above system becomes

$$\begin{aligned} m &\approx 2^{\lambda-q} = \frac{1}{2^{\lambda'}}, \\ n &\approx 2^{\lambda'-q} = \frac{1}{2^\lambda}, \end{aligned}$$

which has no relevant solution. This gives assurance that very large λ, λ' need not be checked explicitly.

Analyzing the m, n equations once more, note that solutions only exist if the numerator sufficiently outweighs the denominator so as to produce odd numbers on the left. This has no way of occurring for low q or high q , and we conclude no double-odd cycle exists.

Arguments similar to this can be used to rule out triple-odd cycles and all others.

5.5 Conclusions

There is enough evidence afoot to believe the Collatz conjecture to be *true*, despite not writing a formal proof.

Getting the easy conclusions out of the way first, we can be sure that there are no Collatz cycles $\theta(q)$ except for the trivial case $1 \rightarrow 4 \rightarrow 2 \rightarrow 1$. Also, if one is interested in checking the Collatz conjecture by brute force, we've shown that the computational burden could be reduced via crossout analysis.

Going a little deeper, recall that the odd number generators $\beta_k(m)$ are able to emit any odd natural number. Doing so led to the sequences $\beta(1), \beta(5)$, etc. listed above. Comparing the sequences $\beta(m)$ to the flattened regress tree $h(\tilde{n})$, we can write:

$$\begin{aligned}
h(0) &= (\beta_1(1)) \\
h(1) &= (2) \\
h(2) &= (4) \\
h(3) &= (8) \\
h(4) &= (16) \\
h(5) &= (32, \beta_2(1)) \\
h(6) &= (64, 10) \\
h(7) &= (128, \beta_3(1), 20, \beta_1(5)) \\
h(8) &= (256, 42, 40, 6) \\
h(9) &= (512, \beta_4(1), 84, 80, \beta_2(5), 12) \\
h(10) &= (1024, 170, 168, 160, 26, 24) \\
h(11) &= (2048, \beta_5(1), 340, 336, 320, \beta_3(5), 52, 48) \\
h(12) &= (4096, 682, 680, \beta_1(85), 672, 640, 106, 104, \beta_1(13), 96) \\
h(13) &= (8192, \beta_6(1), 1364, \beta_1(341), 1360, 226, 1344, 1280, \beta_4(5), 212, \beta_1(53), 208, 34, 192) \\
&\dots
\end{aligned}$$

Observe that all odd numbers in the flattened regress tree occur in the sequences $\beta(q)$. Each starts with a very odd integer $\beta_1(q)$ and goes in order with increasing k . This result gives assurance that no odd number is excluded from the regress tree.

Additionally, we can see that the columns of the base-odd lattice are encoded in the flattened regress tree, and vice versa. Denote m_x as the number in the x th row in the m th column of the base-odd lattice. Then, we can write $h(\tilde{n})$ as:

$$\begin{aligned}
h(0) &= (1_1) \\
h(1) &= (1_2) \\
h(2) &= (1_3) \\
h(3) &= (1_4) \\
h(4) &= (1_5) \\
h(5) &= (1_6, 5_1) \\
h(6) &= (1_7, 5_2) \\
h(7) &= (1_8, 21_1, 5_3, 3_1) \\
h(8) &= (1_9, 21_2, 5_4, 3_2) \\
h(9) &= (1_{10}, 85_1, 21_3, 5_5, 13_1, 3_3) \\
h(10) &= (1_{11}, 85_2, 21_4, 5_6, 13_2, 3_4) \\
h(11) &= (1_{12}, 341_1, 85_3, 21_5, 5_7, 53_1, 13_3, 3_5) \\
h(12) &= (1_{13}, 341_2, 85_4, 113_1, 21_6, 5_8, 53_2, 13_4, 17_1, 3_6) \\
h(13) &= (1_{14}, 1365_1, 341_3, 227_1, 85_5, 113_2, 21_7, 5_9, 213_1, 53_3, 35_1, 13_5, 17_2, 3_7) \\
&\dots
\end{aligned}$$

From the above, we affirm that the flattened regress tree contains not only all odd numbers, but also all even numbers. Any number that violates the Collatz conjecture has nowhere to hide, so we can probably stop now.

Chapter 43

Sxript Programming Language

```
.
+@aaa,      @a
@aaaaaaa'    @aa:
@aaaaaaa    +@aaaa:
#   @aaa#    @aaaaa
, @aaa      .++aaa+
  @aaa#    a   @
; @aaa    ,a
  @aaa#    a
# @aaa' @
  @aaa,
'#####
@aaaaaaa.
. @aaaaaaa
' '... @aaa#
, @aaa
a ' @aaa#
' @ @aaa
a; @aaa'
a @aaa
@; #+ ' @aaa'
, @aa' @ @aaa
@aaaaaa . @aaa; +
@aaaaaa @aaaaa+
+ @aaa . @aaa#
@a . #aaa
' '+
```

Sxript is an interpreted functional/procedural programming language designed for numerical evaluation and problem solving with high-order functions, variable storage, scoped subprograms, and arbitrary-precision numbers, along with file I/O, graphical output, and more. Sxript may be used as a stand-alone tool on most devices, or included as a component of any BASIC, JavaScript, or C++ application. All source code is written in a strict subset of the QB45 language, and is auto-translated by the BAS-TO-X utility into target languages.

1 Hello World

Successfully booting a stand-alone instance of `sxript.exe` in the console should produce something like the following:

```
*** Welcome to Sxript ***
      (QB64 Build)
```

```
[1]:
```

REPL Prompt

The core of any stand-alone instance is a typical *REPL* prompt, which means ‘read-eval-print-loop’. In the same way you type commands into a terminal and press Enter to continue, the above prompt awaits a string you prepare, which is sent through the `SxriptEval` function when you press Enter. God willing, the output emerges on the next line.

Jokes aside, this means `Sxript` is an *expression-oriented* language, which means every piece of code return a value. There are no statements for toggling modes, settings, or otherwise ‘doing something’ in the computer. This is a text-in, text-out language.

First Session

Typing a few test statements into the prompt, a typical first session could be appear as in the code box that follows. Right away there are a few things that stand out.

```
[1]: hello to the world
[1]: hellototheworld

[2]: 'hello to the world'
[2]: hello to the world

[3]: 1 + 1
[3]: +2.0

[4]: 3
[4]: 3
```

Whitespace

Sxript condenses and therefore ignores all unclaimed spaces, tabs, and line breaks. For instance, the input `hello to the world` is condensed down to a single word `hellototheworld` with all spaces removed. In isolation, a piece of ambiguous text such as `helloworld` or `sqrt` is formally called a *word*.

The same whitespace condensation, although hidden, turns `1 + 1` (five characters) into `1+1` (three characters) before attempting the addition.

Quotes

One way to preserve the integrity of text is to enclose the string in bracketing symbols as indicated on input [2] above. The left bracketing symbol is the back-tick key (to the left of 1, under Esc). The right bracketing character is the single quote (apostrophe).

As a proper quote, the input `'hello to the world'` results in `hello to the world`, which is at the level of human readability. The bracketing symbols are stripped as the output reaches the console.

Double Quotes

The double quote symbol is not used or reserved in Sxript whatsoever. The symbol (`"`) can be used freely.

Numbers

For the third input we have supplied the arithmetic statement `1 + 1`, resulting in the third output `+2.0`. The result of a numerical operation is always formatted as a *formatted* number, which means to always include a sign prefix, either plus or minus, along with a decimal component, even for whole numbers.

Scientific notation is used for numbers large or small enough. Under the hood, numbers are represented by a floating-point variable type that behaves similarly among JavaScript and C-family languages.

Plain numbers like 3 are called *unformatted*, which work as regular numbers in an arithmetic sense.

len() Primitive

The `len(x)` function, also known as the `len(x)` *primitive*, is a built-in function that takes an argument `x` and returns the length in characters of `x`.

When used on a quote, the length excludes enclosing quote symbols. When used on a number, the length includes all sign and decimal information.

```
[4]: len('hello world')
[4]: 11

[5]: len(+5.0)
[5]: 4

[6]: len(a b c d e)
[6]: 5
```

2 Numerical Operations

2.1 Arithmetic

The *arithmetic* operators are one-character symbols for exponents (^), multiplication (*), division (/), modulus (%), addition (+), subtraction (-) as follows:

```
^ * / % + -
```

Evaluation

A numerical evaluation is triggered when an operator is situated between two numbers, either formatted or non-formatted.

The result of a numerical evaluation is always a formatted number, i.e. includes a sign and a decimal. Expressions containing multiple operators are evaluated internally without showing each step. For instance:

```
[1]: 4 * 2 + 3
[1]: +11.0
```

Precedence

The order in which the operators are listed above also denotes their precedence of evaluation, reading left-to-right. No two operators have the same precedence.

When multiple instances of a given operator occur in an expression, evaluation proceeds in *right-to-left* order.

Since multiplication takes precedence over addition, the previous result can be recovered via:

```
[2]: 3 + 4 * 2
[2]: +11.0
```

Embedding

A way to subvert operator precedence is to enclose part(s) of an expression with parentheses. Priority is passed to the most-embedded, most-left set of parentheses. The 'evaluation edge' passes recursively upward through layers of parentheses until finished.

```
[4]: 2*(3+6)/3
[4]: +6.0
```

```
[5]: 24/(3*(1+1))
[5]: +4.0
```

Precision and Magnitude

Depending on the host language's implementation of floating-point numerals, there are usually about fifteen digits of precision per number. Exact fractions like 1/3 are approximated via:

```
[6]: 1/3
[6]: +0.3333333333333333
```

For numbers requiring scientific notation, the suffix D (QB64) or e (JavaScript) occurs after the decimal portion:

```
[7]: 256^16 / QB64
[7]: +3.402823669209385D+38
```

```
[7]: 256^16 / JavaScript
[7]: +3.402823669209385e+38
```

The largest number allowed in a double-precision variable is around 10^{308} .

Negative Sign Duality

One unfortunate accident of history is the dual-use of the negative sign. On one hand, we've already reserved the dash symbol (-) for the subtraction operator. On the other hand, the same dash symbol is always used for negative numbers such as -3.5 that do not invoke subtraction at all.

With this dilemma, Sxript adopts a rule where the negative sign prefers to bind to numbers (like an electron) to make them negative, *before* going between numbers to act as a subtraction operator.

Binding of the negative sign is demonstrated in the box below. By operator precedence, the exponent operator is treated first, which triggers a scan for candidate numbers surrounding the operator. On the right, we have 2, whereas on the left, the scan captures -3. Since $(-3)^2$ is solved by +9.0, the expression simplifies to 0+9.0 or +9.0.

```
[8]: 0-3^2
[8]: +9.0
```

This example begs the question of how the result -9.0 is attained. As a note of caution, one way that *won't* work is demonstrated box below. The result is $+9.0$ again due to the parentheses taking precedence over any operators. Internally, the quantity (3) is evaluated to 3 , and then we're back on the path to getting $0+9.0$.

```
[9]: 0-(3)^2
[9]: +9.0
```

The result -9.0 is attained by enclosing the term 3^2 in parentheses before subtracting from zero:

```
[10]: 0-(3^2)
[10]: -9.0
```

Operator Reduction

All input passes through an operator reduction filter prior to evaluation, which reduces things like $+++$ to a single $+$ and so on. Occurrences of $-+$ are interpreted as $-$, making inputs such as $3-+3$ evaluate to $+0.0$. Occurrences of a 'double negative' $--$ are converted to a single positive $+$. For a mental exercise, check that the following expressions are consistent with your expectations:

```
[11]: 0---(-3^2)
[11]: -9.0

[12]: 0----(-3^2)
[12]: +9.0
```

2.2 Numeric Primitives

Recall that a primitive function is one that is hard-wired into the Sxript lexicon. Letting x represent an arbitrary number, we have the following primitives, in alphabetical order:

abs(x)

Removes the $\pm$ sign from x .

atan(x)

Trigonometric arctangent of x .

atan2(x,y)

Cartesian arctangent of the polar angle formed between the positive x -axis and the point x, y .

cos(x)

Trigonometric cosine of x .

eul(x)

Euler exponential e^x .

greater(x,y)

Returns $+1.0$ if x is greater than y . Returns $+0.0$ otherwise.

int(x)

Convert x to integer.

ln(x)

Natural logarithm of $x > 0$. Returns an error otherwise.

rand(x)

Random number between zero and x .

sgn(x)

Returns $+1.0$ if x is greater than zero. Returns $+0.0$ otherwise.

cos(x)

Trigonometric sine of x .

sqrt(x)

Square root of $x > 0$. Returns an error otherwise.

cos(x)

Trigonometric tangent of x .

unf(x)

Removes number formatting from x .

3 Logical Operators

Following the numerical operators in precedence are the *logical* operators for inversion (!), equality (=), the logical and (&), and the logical or (|):

```
! = & |
```

Like the arithmetic operators, the logical operators situate between any two numbers *or other types*. However, the logical operators only emit binary output, which is to say 0 or 1.

Inversion

The **inversion** operation takes a single argument to the left of the (!) symbol. If the argument is not zero, then 0 is returned. Otherwise, 1 is returned as shown:

```
[1]: -3!  
[1]: 0  
  
[2]: 0!  
[2]: 1  
  
[3]: (2+3)!  
[3]: 0
```

Equality

As the name suggests, the logical equality operator returns 1 when situated between two equivalent numbers. Otherwise, the result is 0. Use of the equality operator is straightforward, however things get interesting when all available floating-point digits are used as hinted below:

```
[4]: 1/3 = +0.333  
[4]: 0  
  
[5]: 1/3 = +0.3333333333333333  
[5]: 1
```

The fraction 1/3 is equivalent to $0.33\overline{3}$ where the string of 3's runs infinitely. The best a floating-point number can do is store about fifteen of the trailing digits, and then no real distinction exists between the fraction and its decimal equivalent.

And, Or

The logical **and** returns 1 if both inputs are greater than zero. The logical **or** returns 1 when at least one of the inputs is greater than zero. Otherwise, each output is 0. Some trivial use cases are illustrated below:

```
[6]: 1 & (3-1)  
[6]: 1  
  
[7]: (4-2*2) | 0  
[7]: 0  
  
[8]: 4 + 3|0  
[8]: 1
```

Numeric Comparison

There is no operator notation for numeric comparison with the angle brackets <, >, as these symbols are reserved for enclosing vectors. Instead the **greater** primitive (function gets) the job done by taking the two numbers as function parameters. Then, the **greater** primitive only returns +1.0 if the first argument is greater than the second.

```
[9]: greater(5,6)  
[9]: 0  
  
[10]: greater(5,0)  
[10]: 1  
  
[11]: greater(5,5)  
[11]: 0
```

Of course, we have an effective **lesser** by tacking on the inversion operator at the end, reversing each output:

```
[12]: (greater(5,6))!  
[12]: 1  
  
[13]: (greater(5,0))!  
[13]: 0  
  
[14]: (greater(5,5))!  
[14]: 1
```


4 Quote Operations

In `Sxript`, the ‘string’ data type is called a *quote*. A quote is a sequence of characters enclosed on the left by the back-tick symbol (```), and enclosed on the right by an apostrophe (`'`).

Formatted Output

Before reaching the terminal or output file, the enclosing characters (```) are erased from any quotes. For example, the qquote ‘Hello’ is seven characters in the string space of the host language, however we want the length to resolve to five characters, which is indeed the case when using the `len()` primitive:

```
[1]: 'Hello'
[1]: Hello

[2]: len('Hello')
[2]: 5
```

Note that when two quotes are juxtaposed, as in the example ‘Hello’ ‘World’, the internal behavior is *undefined*. Such an input formats to `HelloWorld` as expected, with all enclosing characters hidden, however the length of the resulting quote is not ten, but instead twelve. The back-to-back enclosing characters are still counted. There is no reason to do this.

```
[3]: len('Hello' 'World')
[3]: 12
```

Operator Overloading

Quotes can be manipulated using the same operators `*`, `/`, `+`, `-` that occur in numerical evaluation, a typical move called *operator overloading*.

The way an operator is used should be evident from context, but this isn’t a guarantee. For instance, we know that the expression `5+6` is undoubtedly mathematical, whereas `‘hello ’ + ‘world’` is the concatenation of quotes.

Concatenation

Proper quote concatenation is done with the addition operator placed between quotes, which evaluates to a single quote.

```
[4]: 'Hello ' + 'World'
[4]: Hello World

[5]: len('Hello ' + 'World')
[5]: 11

[6]: 'do ' + 're ' + 'mi'
[6]: do re mi
```

Subtraction

The tail end of a quote can be subtracted off using the negation operator. If the second argument matches the right-most characters of the first argument, then the first quote minus truncated before the second quote is returned.

```
[7]: 'file.txt' - '.txt'
[7]: file

[8]: len(' ' - 'xyz')
[8]: 0
```

Multiplication

When the multiplication operator occurs between quotes, the result has a copy of the second argument inserted after each character of the first argument.

```
[9]: 'abc' * 'xy'
[9]: axybxycxy

[10]: 'abcde...' * ' '
[10]: a b c d e . . .
```

Division

When the division operator occurs between quotes, the result has the first argument with each instance of the second argument removed.

```
[11]: 'axybxycxy' / 'xy'
[11]: abc

[12]: 'abc' / 'xy' * 'xy'
[12]: abc
```

Binary Logic

Binary logical operator overloading extends to quotes. Any non-empty quote (`' '`) is analogous to a zero-input, and any non-empty quote is considered non-zero.

```
[13]: 'race' + 'car' = 'racecar'
[13]: 1

[14]: 'race' & 'car'
[14]: 1

[15]: ' ' | 'car'
[15]: 1

[16]: ' ' & ' '
[16]: 0
```

4.1 Special Characters

Several special ASCII characters, namely the `tab` (9), `line feed` (10), and `carriage return` (13) are given special treatment because Sxript normally ignores or eats such whitespace entities. It's also necessary to have a way to include the backtick (```) and apostrophe (`'`) characters in the body of a formatted quote.

Tab

To have a `tab` inside a quote, use the `backslash-t` combination as shown below. Note this also works outside of the quote environment as well.

```
[1]: 'Hello\tWorld'
[1]: Hello      World

[2]: Hello\tWorld
[2]: Hello      World
```

Backtick and Apostrophe

In order to have a backtick or apostrophe appear within the body of a quote, precede these characters with a backslash.

```
[3]: 'Here\'s a \'quote\''
[3]: Here's a 'quote'
```

Line Feed, Carriage Return

The 'line ending characters' LF and CR, respectively, are produced with `backslash-n` and `backslash-r`. The behavior of these can vary based on implementation. Using a QB64 implementation, the line feed and carriage return characters operate as shown:

```
[1]: 'Hello\nWorld'
[1]: Hello
World

[2]: 'Hello\rWorld'
WorldHello
```

Observe how the carriage return in line [2] moves the cursor to the beginning of the quote and bumps all characters to the right to accomodate `World` on the left.

Note also that the matching [2] notation on the output line has vanished. This is a case where Sxript has an effect on its hosting environment, and certain information has been invisible 'scrolled upward' before hitting the terminal.

Different behaviors emerge from JavaScript and C++ implementations of Sxript under the same test.

Escape Character

Formally, the backslash is called the *escape character*. When encountered, Sxript treats the symbol immediately following the escape character as a special case.

In order to have a backslash show up on purpose, use two of them, unless at the end of a quote. For this, a special construct is used, as the enclosing apostrophe needs to be shielded.

```
[1]: 'Careful with \\ slashes'
[1]: Careful with \ slashes

[2]: 'Careful with \\ slashes\\\{\}'
[2]: Careful with \ slashes\
```

Null

For some situations, there is reason to return nothing - which really means nothing - not an empty quote, not an empty set of parentheses. For this we have that when the internal result of an expression is `{null}`, such is erased.

```
[3]: {null}
[3]:
```

4.2 Textual Primitives

Let *c* represent any single-character quote, and let *x* represent a quote of arbitrary length. Also let any instance of *n* denote an integer. The primitives for dealing with quotes are as follows:

asc(*c*)

Integer ASCII code of *c*.

chr(*n*)

Character matching the ASCII code of *n*.

instr(*x*,*y*)

Return *n* if quote *x* contains quote *y* at starting at position *n*. Returns +0.0 otherwise.

lcase(*x*)

Replace all uppercase characters in *x* with lowercase.

left(*x*,*n*)

Return the first *n* characters in quote *x*.

len(*x*)

Return the number of characters in quote *x*.

mid(*x*,*n1*,*n2*)

Return a subquote of *x* with length *n2* starting at position *n1*.

quote(*x*)

Enclose *x* in bracketing symbols, i.e. '*x*'.

right(*x*,*n*)

Return the last *n* characters in quote *x*.

lcase(*x*)

Replace all lowercase characters in *x* with uppercase.

unquote(*x*)

Remove bracketing symbols from *x*.

5 Vectors

A *vector* is a data structure for keeping numbers, quotes, or other structures in a linear list, somewhat like an array.

Vector Anatomy

A vector is enclosed by left- (<) and right-angle brackets (>). Each member in the vector or list is called an *element*, and each element is separated by a comma: (,)

```
[1]: <1,2,1+2,'Computer'>
[1]: <1,2,+3.0,Computer>

[2]: <a, b, c, d, e>
[2]: <a,b,c,d,e>

[3]: <'hello','world'>
[3]: <hello,world>
```

Vectors can contain data of mixed types, including other vectors.

```
[4]: <1,2,3,<a,b,c,d,e,f>,<4,5,6>>
```

Evaluation

By default, the elements in a vector are evaluated in left-to-right order. However, the 'deepest and left-most sub-expression' rule always applies, meaning the most-embedded content in the whole vector is always simplified first.

Note that outside the console environment where pressing **Enter** doesn't trigger an evaluation, as in a text file, the above vector can be written with any configuration of whitespace for added clarity. For instance:

```
<1,2,3,
  <a,b,c,d,e,f>,
  <4,5,6>>
```

Vector Length

The `len()` primitive operating on a vector returns the number of elements in the vector. The 'empty' vector `<>` contains one element, and it is undefined for a vector to have zero elements.

```
[5]: len(<a, b, c, d, e>)
[5]: 5

[6]: len(<1,2,3,<a,b,c>,<4,5,6>>)
[6]: 5

[7]: len(<>)
[7]: 1
```

Matrix

If each element of a vector is itself a vector with a fixed number of elements, such a structure is a *matrix*. A matrix is a rectangular list of lists.

```
<
  <a,b,c>,
  <d,e,f>,
  <g,h,i>
>
```

5.1 Vector Operations

The overloading of operators `*`, `/`, `+`, `-` extends to vectors of the same length. Internally, the operator is situated between corresponding vector components and the result is a vector of the same length as the inputs.

For instance, the expression `<1,2,3>+<4,5,6>` is first transformed to `<1+4,2+5,3+6>`, which becomes `<+5.0,+7.0,+9.0>`. Had the operator been multiplication, division, etc., we would get:

```
[1]: <1,2,3> * <4,5,6>
[1]: <+4.0,+10.0,+18.0>

[2]: <1,2,3> / <4,5,6>
[2]: <+0.25,+0.4,+0.5>

[3]: <1,2,3> ^ <4,5,6>
[3]: <+1.0,+32.0,+729.0>

[4]: <'hello ', 'world'> * <'!','>
[4]: <h!e!l!l!o! !,w#o#r#l#d#>
```

Vector operations apply recursively into embedded vectors. For a less trivial example, one could have:

```
[5]:
<<1,2>,
  <3,<9,0>,'Toyota Camry'>>
-
<<6,7>,
  <8,<9,10>,'Camry'>>

[5]:
<<-5.0,-5.0>,
  <-5.0,<+0.0,-10.0>,Toyota >>
```

5.2 Vector Primitives

The primitives for dealing with vectors are as follows:

column(x,n)

Return the *n*th column (as a vector) of matrix *x*.

elem(x,n)

Return the *n*th element of vector *x*.

left(x,n)

Return a subvector of the first *n* elements of vector *x*.

len(x)

Return the number of elements in vector *x*.

mid(x,n1,n2)

Return the subvector with *n2* elements starting at element number *n1* in vector *x*.

replace(x,n,y)

Replace the *n*th element in vector *x* with *y*.

right(x,n)

Return a subvector of the last *n* elements of vector *x*.

smooth(x)

Remove all empty elements from vector *x*.

stack(x,y)

Append vector *y* to the right of vector *x*.

unvector(x)

Remove the bracketing symbols (<>) from vector *x*.

vector(x)

Enclose *x* in bracketing symbols (<>) to return a vector.

6 Mixed-Type Operations

As we've seen with quote operations and vector operations, the arithmetic and comparison syntax

```
^ * / % + - ! = & |
```

also extends to *mixed-type* operations. This is to say that numbers, quotes, and vectors can be musually added, multiplied, etc. to yield new structues.

Quote-Number Addition

The simplest mixed-type operation is number-quote addition (+), which works much like concatenation. The result of such an operation always yields a proper quote, never a number. The same rules hold for the commuted scenario of quote-number addition.

```
[1]: (1 + 2) + ' Three'
[1]: +3.0 Three

[2]: 1 + (2 + ' Three')
[2]: 12 Three

[3]: 1 + 2 + ' Three'
[3]: 1+2 Three

[4]: 'Four ' + 2 + 2
[4]: Four 4.0
```

To make sense of the above, recall that evaluation goes right to left when identical operators occur in the samne expression.

Quote-Number Multiplication

An exotic number-quote interaction, or to be more precise, whole number-quote interaction, is multiplication (*). Multiplying an integer *n* into a quote *c* results in a quote that is *n* copies of *c* juxtaposed.

```
[5]: 3 * 'Hello '
[5]: Hello Hello Hello

[6]: 'World ' * 2
[6]: World World
```

Quote-number multiplication is commutative, which means swapping the arguments leaves the result unchanged.

Note that multiplying any quote by zero results in an empty quote.

```
[7]: 0 * 'Hello'
[7]:

[8]: type(0 * 'Hello')
[8]: quote

[9]: len(0 * 'Hello')
[9]: 0
```

Distribution into Vector

An operator situated between a vector and a number or quote results in a modified vector of the same size where the outer content is distributed with the operator into each component. Distribution is non-commuting, which is to say the outer content will occur left- or right-distributed based on whether the operator occurs to the left or to the right of the vector.

```
[7]: <'file','notes','book'> + '.txt'
[7]: <file.txt,notes.txt,book.txt>

[8]: 2 * <1, 'Two ', <2, 'One '>>
[8]: <+2.0,Two Two ,<+4.0,One One >>
```

Undefined Operations

Certain combinations of data types and operators yield undefined results. For instance, if we try the equality (=) operator between a number and a quote, no operation takes place and we instead get a formatted version of the input.

```
[9]: 'Apple'=7
[9]: Apple=7

[10]: len(7='Apple')
[10]: 9
```

To reinforce that something like `7='Apple'` should not occur, observe that the `len` primitive around this structure includes the operator and the quote bracketing symbols.

Summary

A summary of allowed operations is listed in the Table below. The diagonal entries correspond to same-type operations with numbers, quotes, and vectors. Off-diagonal entries are

mixed-type, and indicate to which operations take place as described above.

Note that the word type doesn't interact with the oth-

ers or with itself, with the exception of the equality (=) operation.

		right			
		number	quote	vector	word
4*left	number	Numeric Operations	* +	right dist.	
	quote	* +	Quote Operations	right dist.	
	vector	left dist.	left dist.	Vector Operations	
	word				=

Table 43.1. Allowed type operations.

6.1 Clerical Primitives

Several language primitives that serve clerical purposes are reviewed here, namely `type`, `join`, and `identity`, from most- to least-complicated.

Type Primitive

The `type` primitive is a single-argument function that takes any data structure as an argument. The return from `type` is a quote containing one of: `number`, `quote`, `vector`, `word`, `occult`, `symbol`, `paren`.

Note that there is no type for 'operator', which is to say a type check on the plus sign (+) is different from a check on the asterisk (*).

The following expressions return `number`:

```
[1]: type(2*3)
[2]: type(+ -)
[3]: type(7Apple)
```

The following expressions return `quote`:

```
[4]: type('Hello' + ' World')
[5]: type('')
[6]: type(type(1))
```

The following expressions return `vector`:

```
[7]: type(<a,b,c,1,2,3>)
[8]: type(<<1,2>,<3,4>>)
[9]: type(<>)
```

The following expressions return `word`:

```
[10]: type>Hello)
[11]: type(len)
[12]: type(*)
```

For completeness, the three thus-far unmentioned outputs from `type` are extracted as follows:

```
[13]: type({})
[13]: occult

[14]: type([])
[14]: symbol

[15]: type()
[15]: paren
```

Join Primitive

The `join` primitive is a two-argument function whose purpose is to juxtapose the two arguments as if the programmer typed them without using `join`.

```
[1]: join(abc,123)
[1]: abc123

[2]: join(5,5)
[2]: 55

[3]: join(1+1, 2+2)
[3]: +6.0
```

Note that the juxtaposition of two formatted numbers induces addition or subtraction between the two.

Identity Primitive

The most benign primitive is the `identity`, which takes a single argument and returns the argument unchanged. Of course, the output of this function is still caught in the evaluation loop and is rendered down to a stable result as if `identity` was never there.

```
[1]: identity('hello world')
[1]: hello world

[2]: identity(1+1)
[2]: +2.0
```

Date and Time

Sxript embeds two special primitives for reporting the system date and the system time. Not surprisingly, these are called via `date()` and `time()`, respectively:

```
[1]: date()
[1]: 1-8-2025

[2]: time()
[2]: 22:4:15
```

Note the return from each is formally a quote, however the contents is a mixture of numbers and symbols.

7 Apply, Map, Reduce

Here we introduce the `apply`, `map`, `reduce` primitives, each being a collection-controlled loop maneuver for using functions with vectors.

Apply Primitive

Consider a vector v with components $v_1, v_2, \dots, v_N$, where N is the length of the vector. The `apply` primitive takes a function name f , along with the vector v and returns a new vector $\langle f(v_1), f(v_2), \dots, f(v_N) \rangle$.

```
[1]: apply(sqrt,<9,16,25,121>)
[1]: <3,4,5,11>

[2]: apply(len,<'dog','cat','mouse'>)
[2]: <3,3,5>

[3]: apply(type,<1+1,'Hello',type>)
[3]: <number,quote,word>
```

Map Primitive

Consider two vectors v and w of equal length N . For a given two-argument function f , the `map` primitive produces a new vector $\langle f(v_1, w_1), f(v_2, w_2), \dots, f(v_N, w_N) \rangle$.

```
[1]: map(<10,5,3>,greater,<6,5,4>)
[1]: <1,0,0>

[2]: map(<'dog','cat','mouse'>,
        instr,<'o','z','u'>)
[2]: <2,0,3>

[3]: map(<1,2,3>,join,<4,5,6>)
[3]: <14,25,36>

[4]: map(<1,2,3>,join,<4,5,6>+0)
[4]: <+5.0,+7.0,+9.0>
```

Reduce Primitive

The `reduce` primitive uses a two-argument function f to collapse a vector v of length N down to a single value via

$$f(\dots f(f(f(v_1, v_2), v_3), v_4), \dots, v_N).$$

```
[1]: reduce(join,<a,b,c,d,e>)
[1]: abcde

[2]: reduce(join,<1,2,3+0,4,5>)
[2]: +15.045

[3]: reduce(join,<5,*,4,*,3,*,2,*,1>)
[3]: +120.0

[4]: reduce(greater,<4,3,0,-2>)
[4]: 1
```

Application: Vector Dot Product

The vector 'dot product' i.e. scalar product, is the projection of one vector v onto another vector w of the same length N . In mathematical notation, one would write

$$v \cdot w = v_1 w_1 + v_2 w_2 + \dots + v_N w_N.$$

Using `sxript` apparatus, it turns out that the products $v_1 w_1, v_2 w_2$, etc. are naturally handled by vector operations. Taking a three-dimensional example, we already have:

```
[1]: <1,2,3> * <4,5,6>
[1]: <+4.0,+10.0,+18.0>
```

Explicitly, the dot product between the vectors $\langle 1, 2, 3 \rangle$ and $\langle 4, 5, 6 \rangle$ is the sum of the elements in $\langle +4.0, +10.0, +18.0 \rangle$. Given that each of these is a formatted number, it follows that all we effectively need to do is remove the commas and bracketing symbols to give $+4.0 + 10.0 + 18.0$, which is valid input for numerical evaluation. This is done using the `reduce` primitive with `join` as the working function.

```
[2]: reduce(join,<1,2,3> * <4,5,6>)
[2]: +32.0
```

8 Variables

Like most programming languages, a chunk of data can be stored in a *variable*. After being stored, the content of a variable is referenced through the corresponding variable name.

Let Primitive

The syntax for variable storage is `let(name,x)`, where `name` is a word representing the variable name and `x` stands for any valid data structure. The return from the `let` primitive is `x` fully simplified.

For instance, the expression `let(a,1+2)` is first reduced to `let(a,+3.0)`, thus the variable `a` holds `+3.0`. Note that variable storage follows *eager evaluation*, which is to say `1+2` is not stored, rather `+3.0` is computed and stored.

A variable name can be any word, with a short list of exceptions. To prevent interference with user-defined functions, the letters `x`, `y`, `z`, `t`, `u`, `v` are reserved.

Restoration

A stored variable is recalled by enclosing its name in square brackets (`[]`). When embedded in an expression, variable restoration is given priority over function evaluation and mathematical evaluation. By the time the information in a variable ‘hits’ a function or operator, the data is already in literal form, as if never stored in a variable.

Mutability

A variable cannot be ‘edited’ in the sense that there are no specialized operators `++`, `+=`, etc., as found in other languages. What is allowed however is to use `let` to overwrite new content into a previously-used variable name.

Symbol Type

Formally, any structure enclosed in square brackets (`[]`) is a *symbol*. There is a `symbol` primitive designed for this in case manually enclosing a word in square brackets is disadvantageous.

```
[1]: let(a,1+2)
[1]: +3.0

[2]: [a] + cos([a])
[2]: +2.010007503399555

[3]: let(a,[a]-1)
[3]: +2.0

[4]: symbol(a)
[4]: +2.0
```

Corollaries

Sine the output of the `let` primitive is its stored content, it follows that variable storage can be used in-line with compatible operations. For instance, `4+let(a,2)` outputs `+6.0` after storing 3 in the name `a`.

As an ordinary primitive, `let` can be embedded within other primitives, including another `let`. For instance, `let(b,4+let(a,3))` will first store 3 in `a`, simplifying to `let(b,+7.0)`.

```
[1]: 4+let(a,2)
[1]: +6.0

[2]: let(b,4+let(a,3))
[2]: +7.0

[3]: <[a],[b]>
[3]: <3,+7.0>
```

Toward Functional Programming

It should be emphasized that variable storage works not only on the standard data structures, but also *words*, including the names of functions.

```
[1]: let(f,cos)
[1]: cos

[2]: let(g,sin)
[2]: sin

[3]: [f]([g](3))
[3]: .9900590857598653

[4]: cos(sin(3))
[4]: .9900590857598653
```

9 Code Blocks

A *code block* provides a way to contain several lines of Sxript code in a single entity.

A code block begins with the word `block`, followed by a set of curled braces enclosed in parentheses (`{}`). Inside this structure, we write a sequence of expressions with each separated by the colon symbol (`:`). At least one of the expressions must be prefixed by `print_`, which is considered the output of the code block.

Code blocks do not have private scope, which is to say that any variables defined inside the block are accessible after the block is used. Similarly, any variables defined outside the block can be used within.

```
[1]: block({let(a,5):print_[a]+1})
[1]: +6.0

[2]: [a]
[2]: 5

[3]: block({let(b,3):print_[a]+[b]})
[3]: +8.0
```

Occult Type

Any structure enclosed in curled braces (`{}`) is called *occult*, which is yet another creature in the zoo of data types.

Occulted code is ‘protected’ from evaluation until triggered by `block` or similar primitive. This means occulted code can be stored in a variable and recalled for later use.

```
[4]: let(b,{let(a,5):print_[a]+1})
[4]: {let(a,5):print_[a]+1}

[5]: [b]
[5]: {let(a,5):print_[a]+1}

[6]: block([b])
[6]: +6.0
```

The language primitives `elem`, `left`, `len`, `mid`, `replace`, `right`, `smooth`, and `stack` defined for vectors also work on occulted structures. There are two new primitives, namely `occult` and `unoccult`, whose purpose is self-explanatory.

Print

Observing that `block` is itself a language primitive, anything following `print_` is considered part of the output. There can be more than one `print_` statement per code block, but there should never be zero. The simplest return from a code block is a set of empty parentheses:

```
[7]: block({print_()})
[7]: ()
```

Some caution must be used when using more than one `print_` statement, as each output is joined with the next without disgression. You are not, for instance, automatically handed a vector.

To illustrate, puzzle out why the following code block returns `+8.0`:

```
block({
  print_1+1:
  2+2:
  print_3+3
})
```

Evaluation

Occulted code is evaluated in strictly linear order, which is to say an expression is completely processed before evaluation moves beyond the colon (`:`) separator.

Some subtleties still lurk, however. Consider a code block that references a variable `a` before it is stored:

```
block({
  let(b,[a]+1):print_[b],:
  let(a,5)      :print_[b]
})
```

In the above, note the liberal use of whitespace (the evaluator doesn’t care) with the particular use of punctuation. This block is designed to emit two numbers separated by a comma, which is by itself bad form. It would more appropriate to enclose the whole structure in angle brackets to make a vector:

```
<block({
  let(b,[a]+1):print_[b],:
  let(a,5)      :print_[b]
})>
```

Since the symbol `[a]` is referenced before it is defined via `let`, it can't possibly happen that `b` stores the number `+6.0`, rather we expect `[a]+1` to occur at least once. However, on evaluation, the above gives:

```
<+6.0,+6.0>
```

We could have predicted the pair of `+6.0`'s by realizing that the variable `a` is defined outside the scope of the code block, i.e. a *global* variable. The output of the code block is truly `[a]+1,+6.0`, however such a structure is still caught in the evaluation cycle, and 5 is filled in for `[a]` before the final answer appears.

One can check this by printing occulted code instead of the 'raw' symbol for `b`:

```
<block({
  let(b,[a]+1):print_occult([b]),:
  let(a,5)      :print_occult([b])
})>
```

Wrapping each instance of `[b]` accordingly we see a modified output there the first reference to `[a]` is protected as occulted code:

```
<{[a]+1},{+6.0}>
```

The `unoccult` primitive can be used to dig out the vector contents and get the pair of `+6.0`'s again:

```
apply(unoccult,<{[a]+1},{+6.0}>)
```

Embedding

The `block` primitive can be used to return occulted code. For a minimal example, consider the following structure:

```
block({
  print_{print_'Hello World'}
})
```

Such a code block returns the occulted structure `{print_Hello World}`. Note that the output is formatted, which means there are no enclosing symbols around the embedded quote. Let this be a caution that copying code directly from the terminal, as it often does, comes with hazard.

In order to access the embedded occulted code, surround the entire code block with another block primitive as shown:

```
block(
  block({
    print_{print_'Hello World'}
  })
)
```

Note that the outer `block` does not enclose its content directly in curled braces (`{}`). The formatted result is `Hello World`, and running the entire structure above through the type primitive returns quote as expected.

9.1 Subprograms

There is a variation on the code block called the *subprogram* that makes use of privately-scoped variables. A subprogram has access to all variables and data in its environment, but can only send information back to the environment via the `print_` statement. All else is forgotten when the subprogram ends.

A subprogram is initiated with the `sub` primitive. Right away, we can test the tricky case reviewed previously with `sub` replacing `block`. Fully simplified, the following structure reduces to `<[a]+1,+6.0>`.

```
<sub({
  let(b,[a]+1):print_[b],:
  let(a,5)      :print_[b]
})>
```

For a less trivial example, consider the following code block with three variables `a`, `b`, `c`.

```
block({
  let(a,1):
  let(c,
    sub({
      let(b,2):
      let(a,[a]+[b]):
      print_[a]
    })
  ):
  print_<[a],[b],[c]>
})
```

Working from the inside out, we see a subprogram that emits whatever value `[a]` represents, and this is stored in the variable `c` via the enclosing `let`. Within the subprogram

the variable `b` is defined, but notice there is no definition for `b` outside the subprogram. When the final `print_` is encountered, we have variables `a` and `c` properly filled, but `b` remains empty. The final output of the code block is `<1, [b], +3.0>`.

10 Functions

Subroutines

Combining the ability to store variables via `let` with the notion of storing occulted code that is activated using `block` or `sub`, one can imagine building custom subroutines or functions.

For a trivial example, consider the ‘function’ that adds one to the variable `a` and returns the result:

```
let(plusone,{print_[a]+1})
```

Executing this mentally, we have a variable `plusone` storing one line of occulted code. This means any instance of `[plusone]` is replaced accordingly, which needs either `block` or `sub` to finish evaluating. For this it’s convenient to define another variable `b` to save some typing, where any instance of `[b]` is replaced by the contents of a plus one as defined:

```
let(b,sub([plusone]))
```

Of course, this is a clunky way to program, particularly because the variable `a` needs to be hard-coded into both the environment and the occulted line(s). What we’ve got so far is a subroutine at best.

10.1 Func Primitive

A user-defined function, or UDF, is constructed with the `func` primitive. Analogous to variables, any function consists of a name associated with stored data. The function body must occur in the form of occulted code:

```
func(newUDF,{...})
```

As a language primitive, the return from `func` is the function name itself, which occurs as a *word*. Like variables, any UDF defined in a subprogram is forgotten when the subprogram closes.

Once defined, a UDF is referenced like any other function or language primitive, which is to say we write the name of the function immediately followed by any required arguments enclosed in parentheses.

Arguments sent to a UDF are referenced within the body as parameters using the symbols `[x]`, `[y]`, `[z]`, `[t]`, `[u]`, `[v]`, and these are the only symbols allowed. If there is need to send more than six arguments, one must pack some

of them into a vector or comparable data structure, all of which fits into a single symbol such as `[x]`. (One could get by using *only* `[x]`.)

Single-Line Functions

The simplest UDF has a single line as its function body. For instance, one could write and test a set of functions for converting temperature units between Fahrenheit and Celsius:

```
[1]: func(degC,{(5/9)*([x]-32)})
[1]: degC

[2]: func(degF,{(9/5)*[x]+32})
[2]: degF

[3]: degC(degF(0))
[3]: +0.0

[4]: degC(-40) = degF(-40)
[4]: 1
```

Immediate Use

Any UDF can be used immediately after being defined, even on the same line. Generically, for a two-argument function, this looks like:

```
func(newUDF,{...[x]...[y]...})(x,y)
```

For a trivial example, take a function that we use immediately for multiplying two numbers:

```
[1]: func(mult,{[x]*[y]})(4,5)
[1]: +20.0
```

Recursion

User-defined functions that reference themselves are allowed. While there are numerous ways to implement recursion, we can make clever use of mixed-type operations to have a UDF call itself if a certain condition is met. In the example that follows, the `fac` function calculates the factorial of an integer greater than one.

```
func(fac,{
  [x]*(unquote(
    greater([x],2)*'fac'
  ))([x]-1)
})
```

A quick test of the above might look like:

```
[1]: fac(6)
[1]: +720.0

[2]: fac(fac(3))
[2]: +720.0
```

10.2 Multi-Line Functions

Technically, any given UDF stores just one line of code. However, we can pack more than one line into a **block**, or even better, for privately-scoped multi-line functions, a **sub**.

Privacy and Scope

Because **sub** carries occulted code, it follows that the parameters `[x]`, `[y]`, etc. are buried in at least two layers of curled braces (`{}`) within the body of a function.

There is an eventual limit to how ‘deep’ the parameter references work, thus it may be prudent to store `[x]`, `[y]`, etc. as local variables in the **sub**. To reinforce this, consider the following session in which a variable `a` is defined and then referenced in a variety of ways:

```
[1]: let(a,1)
[1]: 1

[2]: {a}
[2]: 1

[3]: {{a}}
[3]: {1}

[4]: {{{a}}}
[4]: {{1}}
```

```
[5]: {{{{a}}}}
[5]: {{{1}}}

[6]: {{{{{{a}}}}}
[6]: {{{{{1}}}}}

[7]: {[a]}
[7]: {[a]}
```

We see structures such as `{a}` being replaced as if equivalent to `[a]`, and this pattern penetrates through three additional layers of curled braces (`{}`). Any more than four total layers and the variable isn’t seen by the evaluator.

The rule is more strict for function parameters `[x]`, `[y]`, etc. In the following example, we witness the technique of storing the incoming parameter `[x]` in a new variable `a`. Next we open an embedded **sub** which makes reference to `a` and `x`.

```
func(test,{
  sub({
    let(a,[x]):
    print_sub({
      print_<[a],[x]>
    })
  })
})(1)
```

Running the above example produces a vector `<1,[x]>`. That is, we see `[a]` properly restored, whereas the inner **sub** doesn’t know what to do with `[x]`.

Application: Polish Notation Calculator

Jumping into an almost-practical example of a multi-line function, you may know about *Polish* notation, also known as *prefix* notation, which is a way of writing mathematical expressions by placing the operator before any numbers. For instance, the infix expression `3 + 4` is written `+ 3 4` in Polish notation. The advantage to prefix notation is the list can grow on the right. If we write `× 3 4 9`, the infix translation is `3 × 4 × 9`.

The plan is to create a function that can take a Polish arithmetic expression and return the simplified result, which requires converting to infix notation familiar to **Sxript**. The most suitable format for Polish notation is a vector, with the operator as the first element and each number separated by a comma, for instance `<*,3,4,9>`. This is the kind of structure we expect to send to the function.

Converting to infix notation is done by plucking off the

operator from the front of the vector, and then inserting the operator between each number. As a final step we use `reduce` and `join` to distill a vector to a number. The function `Polish` does exactly this as shown:

```
func(Polish,{
  sub({
    let(op,unvector(left([x],1))):
    let(dat,right([x],len([x])-2)):
    let(nfx,
      stack(mid([x],2,1),
        apply([op],[dat]))
  ):
  print_reduce(join,[nfx])
})
```

The use case would be, for instance:

```
[2]: Polish(<*,3,4,9>)
[2]: +108.0
```

To help trace what happens to the input, we can sketch (in fake notation) what happens internally:

```
op = *
dat = <4,9>
nfx = <3,*4,*9>
print_ 3*4*9
```

10.3 Anonymous Functions

There is another species of function, called an *anonymous* function, also known as a *lambda*, which has instructions but no name. To motivate this, recall the example used to highlight the immediate use of a UDF:

```
[1]: func(mult,{[x]*[y]})(4,5)
[1]: +20.0
```

What we have is the name, specification, and use of a function all in one line. Supposing `mult` is used nowhere else in the program, one could argue that the name `mult` didn't play a material role in multiplying two numbers. That is, only the body of the function and the arguments sent to it are significant to the program and the programmer.

Lambda Primitive

Anonymous functions are initiated with the `lambda` primitive, which takes one argument as occulted code. This is very much like `func` without the leading name argument.

Recasting the previous example as an anonymous function, one would write:

```
[2]: lambda({[x]*[y]})(4,5)
[2]: +20.0
```

Note that any `lambda` must be accompanied on the right by the list of arguments enclosed in parentheses. If we exclude this, the return from `lambda` is the *internal* name of the anonymous function. If you see this, you're probably making a stylistic error.

```
[3]: lambda({[x]*[y]})
[3]: lambda4115
```

The \$ Shortcut

The word `lambda` can be replaced by a dollar sign (`$`). This is merely for saving keystrokes and reducing clutter.

```
[4]: $({[x]*[y]})(4,5)
[4]: +20.0
```

For a nontrivial example, the following construction is used for calculating the average number in a list:

```
${(reduce(join,(0+[x])))/len([x]))}
(<1,2,3,4,5>)
```

10.4 Report Primitive

A peculiar primitive called `report` is designed to take the name of a UDF as input and return the body of the function as occulted code.

For a trivial example of this, we could set up the following session:

```
[1]: func(mult,{[x]*[y]})
[1]: mult

[2]: report(mult)
[2]: {[x]*[y]}
```


Since the output of `report` is occulted code, we should be able to wrap the whole expression with `lambda` to use the UDF in a roundabout way.

```
[3]: $(report(mult))(3,4)
[3]: +12.0
```

Note, however, that the *visible* return from the `report` primitive is formatted for the terminal. If the function contains special characters, particularly pertaining to quotes, these are ‘formatted away’ before reaching the screen. Internally, of course, the `report` primitive does not go through the formatting step.

10.5 Special Operators

There are two ‘operators’, namely the question mark (?) and the tilde (~), that may ease the burden of using many parentheses when calling functions.

Question Operator

Whenever the pattern `x?y` is encountered, where `x` and `y` stand for anything, the interpretation `x(y)` is returned. A typical use case could go as follows:

```
[5]: cos ? 3
[5]: -.9899924966004454

[6]: cos ? cos ? 3
[6]: .5486961336030971

[7]: cos(cos(3))
[7]: .5486961336030971
```

Tilde Operator

Consider any two-argument function `f(x,y)`. When the structure `a ~ b` is encountered, where `a` and `b` stand for anything, the interpretation `f(a,b)` is returned. A typical use case could go as follows:

```
[1]: func(tmult,{cos([x])*sin([y])})
[1]: tmult

[2]: 3 ~tmult~ 4
[2]: +0.7492287917633427

[3]: 3 ~tmult~ 4 ~tmult~ 5
[3]: -0.5806818746830253

[4]: cos(3)*sin(cos(4)*sin(5))
[4]: -0.5806818746830253
```

11 Flow and Procedure

Procedural control pertains to the flow of a program taking one channel or another based on a condition. Sxript offers a standard implementation of *procedural* statements, namely the ‘if’ structure present in most languages, along with ‘for’ loops and ‘do’ loops, and even the notorious ‘goto’. Several of these are quite unlike the standard language primitives.

11.1 Iff Primitive

The ‘if’ primitive is written `iff`, which is a special function that takes three arguments. The first argument is always a binary condition, which must numerically reduce to either one or zero. If the condition resolves to one, control is passed to the second argument sent to `iff`. If the condition resolves to zero, control is passed to the third argument sent to `iff`. The second and third arguments must occur as a single line of occulted code.

```
[1]: iff(1=1,{‘True’},{‘False’})
[1]: True

[2]: iff(0=1,{‘True’},{‘False’})
[2]: False
```

For a less trivial example, a typical use case of `iff` would involve multiple lines per eventuality, achieved with code blocks or subprograms. Needless to mention, the `iff` primitive can be embedded within an enclosing code block or subprogram.

```
iff([condition],{
  block({
    a:b:print_c
  })
},{
  block({
    d:e:print_f
  })
})
```

The `iff` primitive makes quick work of a recursive factorial function:

```
func(fac,{
  iff([x]=1,{+1.0},{[x]*fac([x]-1)})
})
```

11.2 For Primitive

There is a primitive designated to the classic ‘for’ loop, denoted `for`. The `for` primitive takes two bulky arguments: (i) a vector specifying the looping variable, the lower bound, the upper bound, and increment, and (ii) occulted code to be looped over in which the looping variable is referenced. For a simple example, a for loop could look like:

```
[1]: for(<k,-3,5,2>,{[k],})
[1]: -3.0,-1.0,+1.0,+3.0,+5.0,
```

Had it not been for the comma after `[k]`, the above would evaluate to `+5.0`. Also, it’s prudent to make sure a proper data structure emerges from the calculation, so we can enclose the whole thing in angle brackets (`<>`), and then use `smooth` to get rid of the extra comma.

```
[2]: for(<k,-3,5,2>,{[k],})
[2]: +5.0

[3]: smooth(<for(<k,-3,5,2>,{[k],})>)
[3]: <-3.0,-1.0,+1.0,+3.0,+5.0>
```

Fill UDF

The above is useful enough to adapt into a UDF for creating vectors with evenly-spaced numeric elements:

```
func(fill,{
  smooth(<for(
    <k,[x],[y],[z]>,{[k],}
  )>)
})
```

```
[5]: fill(-1,2,.5)
[5]: <-1.0,-0.5,+0.0,
      +0.5,+1.0,+1.5,+2.0>
```

Yet another factorial function can be made using `fill`. This one is non-recursive:

```
func(fac,{sub({print_
  reduce($([x]*[y])),fill(1,[x],1))
})})
```

Note that the sub apparatus would not be needed had it not been for the anonymous function containing the sym-

bol [x]. Instead we could have first defined a UDF for multiplication so fact becomes simpler:

```
func(mult,{[x]*[y]}):

func(fact,{
  reduce(mult,fill(1,[x],1))
}):
```

For a variation on the filled vector, we can also create an unfilled vector using a one-argument UDF:

```
func(emptyvec,{
  <for(<i,1,[x]-1,1>,{,})>
}):
```

Application: FizzBuzz

A classic problem in computer science and job interviews is the *FizzBuzz* problem, which goes like:

Write a program that scans the integers from 1 to 100. Print Fizz for multiples of three, and print Buzz for multiples of five. For multiples of both, print FizzBuzz.

As a *Sxrit* program, we can solve the FizzBuzz problem using a *for* loop:

```
print_
for(<i,1,100,1>,{sub({
  let(a,''):
  iff([i]%3=0,{let(a,[a]+'Fizz')}):
  iff([i]%5=0,{let(a,[a]+'Buzz')}):
  print_unf([i]) ' : ' iff(([a]='')!,{
    [a]'\n'},{[i]'\n'}):
  })))
```

For a sanity check, the output FizzBuzz emerges for the numbers 15, 30, 45, 60, 75, and 90.

Application: ASCII Table

Using a *for* loop and the *iff* primitive together, we can make a miniature ASCII table provided the special characters are either ignored or otherwise stepped around.

```
for(<i,32,126,1>,{block({let(out,''):
  iff([i]=39,{let(out,'\'')}):
  iff([i]=40,{let(out,'('')}):
  iff([i]=41,{let(out,')')}):
  iff([i]=44,{let(out,',')}):
  iff([i]=60,{let(out,'<')}):
  iff([i]=62,{let(out,'>')}):
  iff([i]=91,{let(out,'['')}):
  iff([i]=92,{let(out,'\\{\\')}):
  iff([i]=93,{let(out,']')}):
  iff([i]=96,{let(out,'`')}):
  iff([i]=123,{let(out,'{')}):
  iff([i]=125,{let(out,'}'')}):
  print_quote(unf([i])) + ' ':
  print_iff([out]='',{
    quote(chr([i]))},{[out]})\t:
  print_iff(([i]+1-32)%5=0,
    {\n},{'}))})))
```

Columnar output is made possible by the tab character (9).

32	33 !	34 "	35 #	36 \$
37 %	38 &	39 '	40 (	41)
42 *	43 +	44 ,	45 -	46 .
47 /	48 0	49 1	50 2	51 3
52 4	53 5	54 6	55 7	56 8
57 9	58 :	59 ;	60 <	61 =
62 >	63 ?	64 @	65 A	66 B
67 C	68 D	69 E	70 F	71 G
72 H	73 I	74 J	75 K	76 L
77 M	78 N	79 O	80 P	81 Q
82 R	83 S	84 T	85 U	86 V
87 W	88 X	89 Y	90 Z	91 [
92 \	93]	94 ^	95 _	96 `
97 a	98 b	99 c	100 d	101 e
102 f	103 g	104 h	105 i	106 j
107 k	108 l	109 m	110 n	111 o
112 p	113 q	114 r	115 s	116 t
117 u	118 v	119 w	120 x	121 y
122 z	123 {	124	125 }	126 ~

11.3 Do... Loop

The *do...loop* is a *count-controlled* loop, which means the contents of the loop is executed a fixed number of times. To set the number, simply let a number hang before the leading *do* as shown:

```
block({
  print_ 'Hello w' + :
  5:
  do:
    print_ 'o' + :
  loop:
    print_ 'rld!'
  })
```

The loop executes exactly five times, thus there are five o's repeated in the output:

```
Hello woooooorld!
```

The do...loop structure can be self-nested, and the total number of iterations need not be a hardcoded number. Puzzle out why the following code block returns +24.0:

```
block({
  3:do:
    1+1:do:
      2+2:do:
        print_1+0:
      loop:loop:loop})
```

Application: Quote Permutation

The do...loop structure also plays well with recursive functions. In the following we create a UDF called `perm`, which is designed to generate all permutations of an input quote. Initially, the argument `[x]` is empty, and the argument `[y]` takes the input. As evaluation goes on, the input is chopped up and sent back to `perm` in various chunks until all permutations are generated.

```
func(perm, {sub({
  let(a, [x]):let(b, [y]):
  print_ iff(len([b])=0, {[a]}, {
    sub({let(i, 1):
      len([b]):
      do:let(newpre,
        [a] + mid([b], [i], 1)):
      let(newsuf,
        left([b], [i]-1)
        + right([b], len([b])-[i])):
      print_perm([newpre], [newsuf]):
      let(i, [i]+1):
      loop }}))}}))})
```

A support function `permquote` is defined to serve as a front-end to `perm`.

```
func(permquote, {
  smooth(<perm('', [x])>):
  }):
```

```
[3]: permquote('abcd')
[3]: <abcd, abdc, acbd, acdb, adbc, adcb,
    bacd, badc, bcad, bcda, bdac, bdca,
    cabd, cadb, cbad, cbda, cdab, cdba,
    dabc, dacb, dbac, dbca, dcab, dcba>
```

Application: Riemann Sums

For approximating the area under a curve, we employ the notion of Riemann sums, particularly the trapezoid rule, the midpoint rule, and Simpson's rule. Any one of these suffices for an approximation, however it's instructive to show several running simultaneously. Moreover, one can show that the result for Simpson's rule is a weighted average of the others mentioned, helping to verify overall accuracy against a direct implementation.

To demonstrate, consider the function $f(x) = 4x - x^2$ in the interval $0 \leq x \leq 4$. Let us slice the interval into $n = 15$ discrete bins. In Sxript notation, one writes:

```
func(f, {(
  4*[x] - [x]*[x]
  }):
  let(x0, 0):let(xn, 4):let(n, 15):
```

It is necessary to specify the spatial interval $\Delta x = (x_n - x_0) / n$, and also initialize the accumulator variables for left sum, right sum, midpoint sum, trapezoid sum, and the first of two Simpson's rule sums.

```
let(dx, ([xn]-[x0])/[n]):
let(ls, 0):let(rs, 0):let(ms, 0):
let(ts, 0):let(si1, 0):
```

Next we set up a loop that makes n iterations to approximate the left sum, right sum, and midpoint sum. The trapezoid sum and the first Simpson's sum are calculated after the loop.

```

let(j,0):
[n]:
do:
  let(xj,[x0] + [j]*[dx]):
  let(x1,[xj] + [dx]):
  let(xm,([xj] + [x1])/2):
  let(ls,[ls] + [dx]*(f([xj]))):
  let(rs,[rs] + [dx]*(f([x1]))):
  let(ms,[ms] + [dx]*(f([xm]))):
  let(j,[j]+1):
loop:
let(ts,(1/2)*([ls]+[rs])):
let(si1,(1/3)*([ts]+2*[ms])):

```

In order to verify the contents of [si1], we construct a second loop that recalculates the Simpson's rule result.

```

let(h,([xn]-[x0])/(2*[n])):
let(si2,0):
let(j,0):
[n]:
do:
  let(xj, [x0] +
    [h] + ([j]/[n])*([xn]-[x0])):
  let(f1,f([xj]-[h])):
  let(f2,f([xj])):
  let(f3,f([xj]+[h])):
  let(si2, [si2] +
    [h]*(1/3)*([f1] + 4*[f2] + [f3])):
  let(j,[j]+1):
loop:

```

Check all outputs simultaneously using:

```

print_<[ls]\n,[rs]\n,[ms]\n,
[ts]\n,[si1]\n,[si2]>

```

```

<+10.619259259259259
,+10.619259259259259
,+10.690370370370369
,+10.619259259259259
,+10.666666666666664
,+10.666666666666666>

```

We see various approximations either under- or over-shooting the average result, with the last two being spot on. Ignoring the last floating-point digit in each result, the two approximations via Simpson's rule are in agreement. For completeness, the exact solution to the area under the curve

in this example is:

$$\int_0^4 (4x - x^2) dx = \frac{32}{3} = 10.666\bar{6}$$

11.4 Anchor... Goto

Not needing line numbers, Sxript instead has 'anchors', which are a way to bookmark a location in code to which control may be passed using goto_.

An anchor is initiated by the anchor_ keyword. The characters following the underscore constitute the name of the anchor, and these must be hard-coded. There cannot be a variable, function, or other structure used to 'calculate' an anchor name.

Control is passed to an anchor using the keyword goto_. Note that anchor_ and goto_ only apply locally to the code block or subprogram in which they're defined. Jumping to other code blocks isn't possible.

The information following goto_ must match an anchor name, but does not have to be hard-coded. The target of goto_ may be calculated from a variable or otherwise. In the following example we store in a variable a some word representing a randomly-selected anchor elsewhere in the code.

```

let(a,unquote(
'cat'+unf(int(rand(3)))))

print_'A cat says ': goto_[a]:
anchor_cat0:print_'meow.':goto_cat9:
anchor_cat1:print_'purr.':goto_cat9:
anchor_cat2:print_'hiss.':goto_cat9:
anchor_cat9

```

Across multiple trials, the output is any one of three possibilities:

```

[1]: A cat says purr.
[2]: A cat says meow.
[3]: A cat says meow.
[4]: A cat says purr.
[5]: A cat says hiss.

```

11.5 If... Goto

The if_ keyword initiates a conditional jump. Following the underscore is a condition that evaluates to one or zero. If the condition is zero, control is passed to the next line as if nothing happened. Otherwise, after the condition is

829

```
func(largefac,{
  iff([x]=0)|([x]=1),{1},{
    largemul([x],
      largefac(largesub([x],1)))}}})
```

thousand factorial, i.e.

$$1000! = 1000 \cdot 999 \cdot 998 \cdots 2 \cdot 1$$

is given by:

```
print_largescl(largefac(1000))
```

For a quick show of force, the above is readily applied to integers in the hundreds or thousands. For instance, one

The result is has more than 2500 decimals of precision:

```
<+4.0238726007709377354370243392300398571937486421071463254379991042993851239862
90205920442084869694048004799886101971960586316668729948085589013238296699445909
97424504087073759918823627727188732519779505950995276120874975462497043601418278
09464649629105639388743788648733711918104582578364784997701247663288983595573543
25131853239584630755574091142624174743493475534286465766116677973966688202912073
79143853719588249808126867838374559731746136085379534524221586593201928090878297
```

```
30843139284440328123155861103697680135730421616874760967587134831202547858932076
71691324484262361314125087802080002616831510273418279777047846358681701643650241
53691398281264810213092761244896359928705114964975419909342221566832572080821333
18611681155361583654698404670897560290095053761647584772842188967964624494516076
5353408198901385442487984959953319101723355566021394503997362807501378376153071
27761926849034352625200015888535147331611702103968175921510907788019393178114194
```

```
54525722386554146106289218796022383897147608850627686296714667469756291123408243
92081601537808898939645182632436716167621791689097799119037540312746222899880051
95444414282012187361745992642956581746628302955570299024324153181617210465832036
78690611726015878352075151628422554026517048330422614397428693306169089796848259
01254583271682264580665267699586526822728070757813918581788896522081643483448259
93266043367660176999612831860788386150279465955131156552036093988180612138558600
```

```
30143569452722420634463179746059468257310379008402443243846565724501440282188525
24709351906209290231364932734975655139587205596542287497740114133469627154228458
62377387538230483865688976461927383814900140767310446640259899490222221765904339
90188601856652648506179970235619389701786004081188972991831102117122984590164192
10688843871218556461249607987229085192968193723886426148396573822911231250241866
49353143970137428531926649875337218940694281434118520158014123344828015051399694
```

```
29015348307764456909907315243327828826986460278986432113908350621709500259738986
35542771967428222487575867657523442202075736305694988250879689281627538488633969
09959826280956121450994871701244516461260379029309120889086942028510640182154399
45715680594187274899809425474217358240106367740459574178516082923013535808184009
69963725242305608559037006242712434169090041536901059339838357779394109700277534
72,+2567.0>
```

Application: Binomial Coefficients

The binomial coefficients

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

tend to outgrow the standard floating-point number type when high precision is desired. Instead, one must use a large-number toolkit as we do here:

```
func(binom,{
  largediv(
    largefac([x]),
    largemul(largefac([y]),
      largefac(largesub([x],[y]))
    ))}):
```

13 ASCII Plotting

As advertised, Sxript is capable of generating graphical output. Not surprisingly though, this comes in the form of strings, also known as ASCII plotting.

13.1 Graphing $y=f(x)$

Mathematical functions of the form $y = f(x)$ can be graphed in the interval $x_{\min} \leq x \leq x_{\max}$ with step size Δx using the `plot` primitive, which takes strictly seven arguments:

- Function name as a word.
- $x_{\min}$ as a floating-point number.

- $x_{\max}$ as a floating-point number.
- Δx as a floating-point number.
- Horizontal window size as integer.
- Vertical window size as integer.
- Toggle axis numerals, 1 or -1.

For a demonstration, let us plot `cos` in a window appropriate to show its character:

```
plot(cos, -5, 5, 0.01, 40, 20, 1)
```

```
{cos, -5, 5, 0.01, 40, 20}
      * ^ ^ ^ ^ *
-----1
**      **      0.8947369088564453
*      *      0.7894738177128906
**      **      0.6842107265693359
*      *      0.5789476354257812
**      **      0.4736845442822264
<      *      *      **0.3684214531386718
<      *      *      * 0.2631583619951171
*      **      **      ** 0.1578952708515623
*      *      *      * 0.0526321797080076
*      **      **      * -0.05263091143554721
**      *      *      ** -0.15789400257910202
*      *      *      * -0.2631570937226564
**      **      **      ** -0.3684201848662112
*      *      *      * -0.4736832760097658
**      **      **      ** -0.5789463671533206
*      *      *      * -0.6842094582968754
**      *      **      ** -0.78947254944043
*****      ***** -0.8947356405839848
*      *      *      * -0.9999987317275394
```

```
-----+-----
5444333322221111000000001111222233334445
.....
0742974296419641863113681469146924792470
48371604937261594822849516273940617384
37047158259360371488417306395285174073
51739517384062840622604826048371593715
87654321098765432100123456789012345678
9998888777766665555555666677778888999
74297429641964186311368146914692479247
```

The plot is headed by the function name, along with relevant information sent to the `plot` primitive. Points in the plot are represented by an asterisk. Points slightly outside the boundary of the window are indicated with an arrow-like character.

Note that the function being plotted need not be a language primitive. Any UDF can go as the first argument in `plot`, including an anonymous function.

13.2 Scatter Plot

A second two-dimensional graph, namely the scatter plot, is invoked by the `scatter` primitive, which takes strictly four arguments:

- List of ordered pairs $\langle x, y \rangle$ contained in a vector.
- Horizontal window size as integer.
- Vertical window size as integer.
- Toggle axis numerals, 1 or -1.

For a trivial example, suppose we want to plot the four points $\langle 1, 4 \rangle$, $\langle 2, 7 \rangle$, $\langle 5, 3 \rangle$, $\langle 4, 5 \rangle$ in a window fifteen units across and then units tall. In a single line, we submit:

```
scatter(
  <<1,4>,<2,7>,<5,3>,<4,5>>,
  15,10,1)
```

Somewhat resembling a standard function plot, the scatter plot lives in the Cartesian plane in an automatically-fitted window:

```

      *              7
      6.5555555555555
      6.1111111111111
      5.6666666666667
          *          5.2222222222222
          4.7777777777778
      *          4.3333333333334
          3.8888888888889
          3.4444444444444
          *3

+++++
11112223334445
.....
025814702581470
087542108754210
051728405172840
074185207418520
012457801245780
048271504827150
025814702581470
```

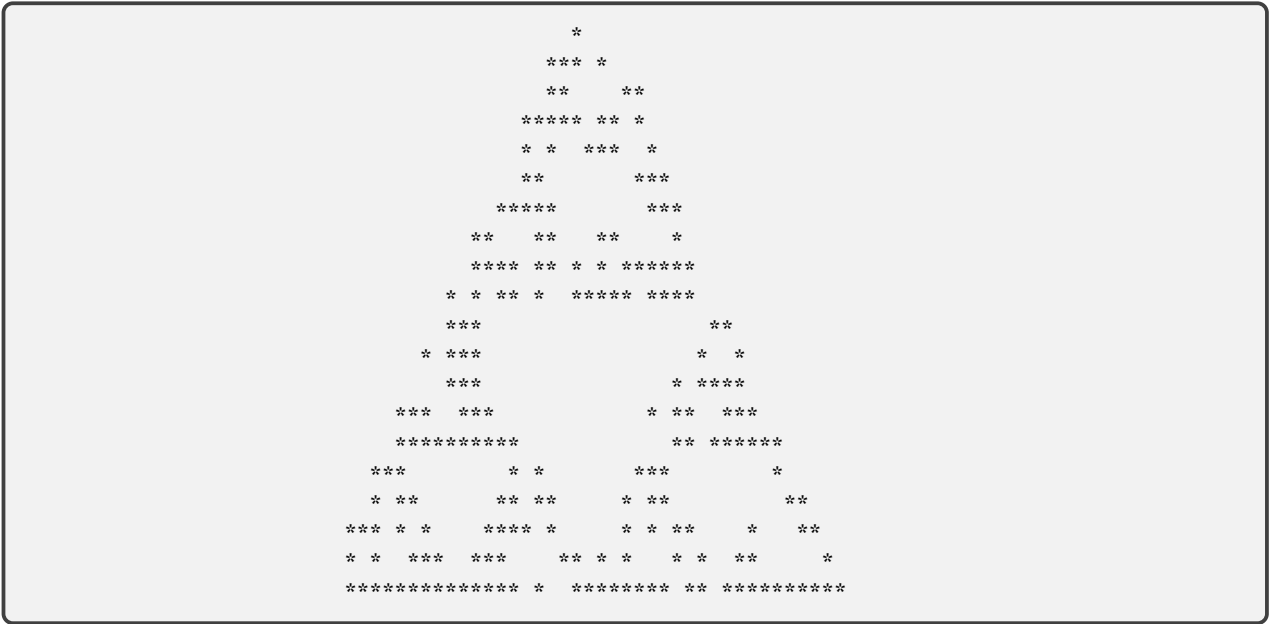
Application: Sierpinski Triangle

A scatter plot can be used for crudely visualizing a Sierpinski triangle, which is an infinite nesting of equilateral triangles. Setting up a well-known algorithm for plotting points within the Sierpinski triangle, we produce the following program:

```
let(xx1,0):let(yy1,0):
let(xx2,2.4/(3^.5)):let(yy2,0):
let(xx3,1.2/(3^.5)):let(yy3,1.2):
let(xx,0):let(yy,0):

print_scatter(smooth(<sub({
350:do:
  let(ii,int(rand(3))+1):
  iff([ii]=1,{block({
    let(xx,.5*([xx]+[xx1])):
    let(yy,.5*([yy]+[yy1])):
    print_(){}},{}):
  iff([ii]=2,{block({
    let(xx,.5*([xx]+[xx2])):
    let(yy,.5*([yy]+[yy2])):
    print_(){}},{}):
  iff([ii]=3,{block({
    let(xx,.5*([xx]+[xx3])):
    let(yy,.5*([yy]+[yy3])):
    print_(){}},{}):
  print_<[xx],[yy]>,:loop}>))
,40,20,-1)
```

Note that all code is sent to the `scatter` primitive. The axis numerals are disabled via the `-1` sent as the last argument. Despite receiving decimal-value inputs, each point is placed in an appropriate bin determined by the size of the window.



14 Functional Programming

Despite being limited to the string space of the given host language, **Sxript** encourages a coding discipline classified as *functional programming*. Right off the bat, functional programming generally means two things: (i) programming in a ‘functional’ language such as **Lisp** or **Haskell**, and (ii) programming in an ordinary language, such as **C++** or **Python** while under a self-imposed discipline that functional languages would enforce.

Functional programming looks and smells differently than the *imperative* paradigm, which is the usual step-by-step approach to coding. Rather than setting up the state of the computer to get ready for a calculation, as one might have in a **for** loop, one instead performs iterations directly using functions. One reason for this, apart from having to define extra variables, is to not trick the computer into a state that interferes unexpectedly with the program or its neighbors.

Some of the key tenets of functional programming are reviewed in no particular order:

14.1 Purity

A function is ‘pure’ if the function always returns the same output for the same input. In addition, a pure function causes no impact on any part of the program outside the function, which is to say a pure function has no *side effects*.

Nearly all of the **Sxript** primitives, and thereby most possible user-defined functions, are compatible with the rule of purity. We say ‘nearly’ because things like **date** and **time** are not pure by definition.

Sloppy use of the **block** primitive leads to side effects if variables and functions in the parent scope are overwritten. To erase this concern, the **sub/return_** combination is the safest way to perform calculations that can’t escape their container.

Immutability

Following the rule of purity is the rule of immutability, which refers to the stability of data. Once created, immutable data is never deleted or edited. This includes variables and functions.

Of course, it’s difficult to fathom programming without being able to assign a new value to an existing variable, at least certainly for complex tasks. Nonetheless, immutability is something to shoot for when writing clean code.

14.2 High-Order Functions

A functional programming language allows functions to be passed around as if they’re variables or data, and functions are thus known as *first class citizens*. This is to say functions can be stored, sent to functions, returned from other functions, etc.

A function *high order* if it takes a function as input and/or returns a function as output. Sometimes these are called ‘generators’ or ‘factories’, depending on the resource.

It was mentioned previously that **Sxript** primitives and user-defined functions are straightforwardly stored as variables, for instance:

```
[1]: let(f,cos)
[1]: cos

[2]: let(g,sin)
[2]: sin

[3]: [f]([g](3))
[3]: .9900590857598653
```

One may also note that the primitives **apply**, **map**, **reduce** are each high-order functions by virtue of taking a function as an argument.

Callbacks

When a function is passed to another function as an argument, the function being passed is called a *callback*. The so-called callback function can be used within the function to which it is sent.

For an example, let us define two arithmetic user-defined functions **mult**, **add**. The **mult** UDF multiplies its arguments **x**, **y** and returns the product. For simplicity assume **x**, **y** are positive integers. Similar applies to **add**.

For a ‘workhorse’, define a UDF called **calculate**, taking three arguments, namely two numbers, and a function name, i.e. a callback.

```
func(mult,{[x]*[y]}):
func(add,{[x]+[y]}):

func(calculate,{[z]([x],[y]))):
```

Demonstrating on a trivial example, one applies the above via:

```
[4]: calculate(10,7,mult)
[4]: +70.0

[5]: calculate(10,7,add)
[5]: +17.0
```

Currying

As a way to manage complexity in a program, *currying* refers to condensing a multi-argument function to one with fewer arguments. Along the way we demonstrate the use of high-order functions.

To proceed, create a function to wrap around `mult` called `currymult`, whose job is to take a single integer argument `a` and then lay out a new function that multiplies its single argument by `a`. The return from `currymult` is a chunk of occulted code.

To help with calls to `currymult`, another UDF called `newmult` is defined to ‘unlock’ the new function from the occult brackets. Any call to `newmult` creates a new curried function named after the hard-coded number sent to `newmult`. When executed, the following returns `+18.0`, `+30.0`.

```
func(mult,{[x]*[y]}):

func(currymult,{sub({
  let(a,unf([x])):
  print_{func(mult{a},{mult({a},[x]))}}
}}):

func(newmult,{
  unoccult(currymult([x]))}):

newmult(3):
newmult(5):

print_mult3(6) ,:
print_mult5(6)  :
```

For an example that is more useful in the field, we can transform a natural logarithm function `ln()` of base e into a more general function `logn()` of base n .

```
func(logn,{ln([x])/ln([y])}):

func(currylog,{sub({
  let(n,unf([x])):
  print_{func(log{n},{logn([x],[n]))}}
}}):

func(newlog,{
  unoccult(currylog([x]))}):

newlog(10):
newlog(5):

print_log10(1000) ,:
print_log5(1000)  :
```

14.3 Composition

Another way to build and manage complexity in a program is using function *composition*. This could mean something as simple as rolling two functions into one using an imperative practice, however the use of high-order functions makes composition a more powerful asset.

To illustrate, consider the `+1` operation that simply adds one to any number, and let’s always take `0` as the starting point. Naturally, one could express the integer `4` using `0 + 1 + 1 + 1 + 1`, which requires the programmer to invoke the `+1` operation four times. Similar would apply for any positive integer.

To make something better, start with a function `f` that applies the `+1` operator to the argument sent to `f`. Also, define a two-argument high-order function `p` such that

$$p(x,y) = x(x(y)) ,$$

where x, y could stand for anything so far. As user-defined functions, these are:

```
func(f,{[x]+1}):

func(p,{
  [x]([x]([y]))
}):
```

For a sanity check, one could test `f(0)` to get `+1.0`, and `p(f,0)` to get `+2.0`. Note that the `p`-function takes the function `f` as its first argument.

The construction $p(f,x)$ at $x = 0$ attains the integer `+2.0`, thus it’s worth abstracting this into a function `g` as shown. While we’re at it, exploit the same pattern to define

a new function h that takes g as the first argument, and so on for a few layers:

```
func(g, {p(f, [x])}):
func(h, {p(g, [x])}):
func(i, {p(h, [x])}):
func(j, {p(i, [x])}):
```

Choosing any for illustration, we have that the function h is defined such that

$$h(x) = p(g, x) ,$$

which simplifies via:

$$\begin{aligned} h(x) &= g(g(x)) \\ h(x) &= g(p(f, x)) \\ h(x) &= g(f(f(x))) \\ h(x) &= p(f, f(f(x))) \\ h(x) &= f(f(f(f(x)))) \end{aligned}$$

Clearly then, we see $h(0)$ is equivalent to four applications of the f -function, therefore

$$h(0) = 4 .$$

As a matter of notational convenience, a shorthand way to write $h(0)$ takes advantage of the question operator (?). The following are equivalent representations of the integer four:

```
f(f(f(f(0)))) = f? f? f? f? 0 = h?0
```

In terms of the functions f, g, h, i , and remembering that **Script** ignores whitespace, we prepare a vector expressing the first sixteen positive integers:

```
print_apply(unf, <
      f? 0 ,
      g? 0 ,
      g? f? 0 ,
      h? 0 ,
      h? f? 0 ,
      h? g? 0 ,
      h? g? f? 0 ,
      i? 0 ,
      i? f? 0 ,
      i? g? 0 ,
      i? g? f? 0 ,
      i? h? 0 ,
      i? h? f? 0 ,
      i? h? g? 0 ,
      i? h? g? f? 0 ,
      j? 0 >)
```

If you haven't spotted it yet, the 'digits' f, g, h, i are like binary switches counting upward in the binary number system.

```
<1,2,3,4,5,6,7,8,
  9,10,11,12,13,14,15,16>
```

Note that the above can be used for a few more purposes. For instance, we can replace the function p with a version that more-deeply embeds the y -argument:

$$p(x, y) = x(x(y))$$

With this, a call to $j(0)$ returns the number $+81.0$. The reason why is left as an exercise.

Or, leaving $p(x, y)$ in its original form, one may replace f with $f(x) = \cos(x)$. Then, the expression $j ? 0$ is equivalent to sixteen layers of cosine applied to zero:

```
[8]: j ? 0
[8]: 0.7383692041223232

[9]: cos(cos(cos(cos(cos(cos(cos(
      cos(cos(cos(cos(cos(cos(cos(
      cos(cos(0))))))))))))))
[9]: 0.7383692041223232
```

Application: Digits of Pi

If you can't remember the digits of π , start with $x = 3.14$ and calculate $x - \tan(x)$ to find

$$3.14 - \tan(3.14) = 3.14159265 \dots$$

Interestingly, the result gives eight correct digits after the decimal point with floating-point junk afterward. Feed this result back into the recursive formula $x = \tan(x)$ to get more correct digits.

To set up this calculation quickly, redefine the function `f` via:

```
func(f,{[x] - tan([x])}):
```

Then, using the composite functions defined above, it only requires `g(3.14)` to get a result that exceeds the floating-point precision of the host language.

```
[1]: g(3.14)
[1]: +3.141592653589793
```

14.4 Recursion

In functional programming, performing repetitive actions using recursion is preferred (or mandated) over other loops that introduce extra state to the program. We've seen this several times while defining various versions of the factorial UDF, with a tight functional-friendly expression being:

```
func(fac,{
  iff([x]=1,{+1.0},{[x]*fac([x]-1)})
})
```

Of course, sometimes it's a better to use a traditional flow control, even a `do...loop`, as long as other safe practices are also in play.

Application: Ackermann Function

The *Ackermann* function is an early example of a recursive function that is easy to write but takes a very nontrivial path to arrive at a final answer. For non-negative integers, the Ackermann function is specified by:

$$A(x, y) = \begin{cases} x = 0 : & y + 1 \\ x > 0, y = 0 : & A(x - 1, 1) \\ x > 0, y > 0 : & A(x - 1, A(x, y - 1)) \end{cases}$$

```
func(Ack,{sub({
  let(a,[x]):let(b,[y]):
  print_iff([a]=0,{
    [b]+1 },{
    iff(greater([a],0) & ([b]=0),
      {Ack([a]-1,1)},
      {Ack([a]-1,Ack([a],[b]-1))
    })})})})
```

The Ackermann function has an enormous convergence time but except when the most meager of arguments are sent.

```
[2]: Ack(0,0)
[2]: +1.0

[3]: Ack(3,1)
[3]: +13.0

[4]: Ack(3,2)
[4]: +29.0

[5]: Ack(3,3)
[5]: +61.0
```

Fixed-Point Combinator

Now we get into a somewhat high-minded idea. Consider first the standard factorial UDF, but switch the variable from `x` to `y`:

```
func(fac,{
  iff([y]=1,{+1.0},
    {[y]*fac([y]-1)})})
```

The next move is to turn the recursive call to `fac` into a variable, particularly `x`. This makes the factorial UDF into a two-argument function:

```
func(fac,{
  iff([y]=1,{+1.0},
    {[y]*([x]([x],[y]-1))})})
```

Note that `[x]` occurs twice in the function because the embedded function call must also have two arguments. As a sanity check, one may use such a function via:

```
[2]: fac(fac,6)
[2]: +720.0
```

To proceed, it would be nice to abstract away the multiple calls to `fac`, which means replacing the arguments `fac` and `6` with variables. For this we define a function for *self-application* as follows:

```
func(self,{
  [x]([x],[y])
})
```

```
[4]: self(fact,6)
[4]: +720.0
```

To make the above even more abstract, one notices that we didn't really need the `self` UDF in a language that has anonymous functions. This means we can write:

```
[5]: $({[x]([x],[y]))}(fact,6)
[5]: +720.0
```

Similarly, we can replace the two-argument `fact` UDF with an anonymous function to write the final answer.

```
$({[x]([x],[y]))}
  ($({
    iff([y]=1,{+1.0},
      {[y]*[x]([x],[y]-1)})
  }),6)
```

The above is called a *fixed-point combinator*, a class of computational objects that come in many flavors and have many purposes. In our case, we have used self-application to achieve recursion using the construction $x(x, y)$.

If the embedded argument `6` seems irritatingly imprisoned, we can cheat by setting the argument equal to a temporary constant, i.e., a variable that doesn't change in accordance with the rule of immutability. By doing so, one can fully separate the function body from its input.

```
$({sub({let(a,[x]):
  print_
  $({[x]([x],[y]))}
  ($({
    iff([y]=1,{+1.0},
      {[y]*[x]([x],[y]-1)})
  }),[a])
}))}(6)
```

Application: Newton's Method

Students of calculus learn of Newton's method for iteratively solving one-dimensional equations. Fundamental to this is the formula for the derivative of a continuous function, which as `Sxript` code is given by:

```
func(deriv,{
  (([x]([y]+[z]))
  -([x]([y]-[z]))) / (2*[z])
}):
```

In the above, `x` is the equation to be solved, `y` is the test point, and `z` is a small deviation from the test point. In terms of the same parameters, Newton's method applies a recursive formula captured by:

```
func(linapx,{
  [y] - ([x]([y]))/deriv([x],[y],[z])
}):
```

In a traditional scheme, one might implement Newton's method by defining a function to capture the problem, and then write a `for` loop to churn a fixed number of iterations. For instance, to solve the problem $x = \cos(x)$, one could write:

```
func(p,{[x] - cos([x]))):
```

Using an initial guess `y=1` and going for ten iterations, one gets a satisfactory result using:

```
func(newtonloop,{
  sub({
    let(fxn,[x]):
    let(ans,[y]):
    for(<k,1,[z],1>, {
      let(ans,linapx([fxn],[ans],0.001))
    }):
    print_[ans]
  }):
  print_newtonloop(p,1,10)
```

Of course, we'd rather do this with a pure approach that avoids the `for` loop. Using a recursive function, the same result is achieved with:

```
func(newton,{
  iff(
    (greater(0.000001,abs(p([x]))))
    |(greater(abs(p([x])),1000000)),
    {[x]},
    {newton(linapx(p,[x],0.001))})
}):

print_newton(1)
```

Next, turn the recursive call to `newton` into a variable, which will take slot `x`, bumping the numeric argument to `y`. This transforms the function's appearance and behavior via:

```
func(newton,{
  iff(
    (greater(0.000001,abs(p([y]))))
    |(greater(abs(p([y])),1000000)),
    {[y]},
    {[x]([x],linapx(p,[y],0.001))})
}):

print_newton(newton,1)
```

Furthermore, let us recast the hard-coded calls to function `p` into a variable sent to the function, taking slot `y`. This bumps the numeric argument to `z`:

```
func(newton,{
  iff(
    (greater(0.000001,abs([y]([z]))))
    |(greater(abs([y]([z])),1000000)),
    {[z]},
    {[x]([x],[y],linapx([y],[z],0.001))})
}):

print_newton(newton,p,1)
```

This example is looking much like the factorial example above. At this stage, we need a self-application function of the form:

```
func(self,{
  [x]([x],[y],[z])
})
```

Implementing this as an anonymous function, the calculation now looks like:

```
print_
$({[x]([x],[y],[z]))}
({(
  iff(
    (greater(0.000001,abs([y]([z]))))
    |(greater(abs([y]([z])),1000000)),
    {[z]},
    {[x]([x],[y],linapx
      ([y],[z],0.001))})}),
p,1)
```

Finally, we can abstract `linapx`, `deriv`, and `p` into anonymous functions to write the whole enchilada as a single line:

```
$({[x]([x],[y],[z]))}
({(
  iff(
    (greater(0.000001,abs([y]([z]))))
    |(greater(abs([y]([z])),1000000)),
    {[z]},
    {[x]([x],[y],
      $({[y] - ([x]([y]))/
        $({([x]([y]+[z]))
          -([x]([y]-[z])))/(2*[z])
        })([x],[y],[z])
      })([y],[z],0.0001))})}),
$({[x] - cos([x]))},
1)
```

Omega Combinator

Previously, we encountered the two-argument self-application function

$$T = x(x, y)$$

in the form of:

```
func(self,{[x]([x],[y]))}
```

In the mathematical notation, we choose the letter *T* after Alan Turing, where $x(x, y)$ is known as the *Turing* combinator, also known as the *Mockingbird* combinator.

Now comes a simpler creature called the *Omega* combinator, defined as

$$\Omega = x(x)$$

which is a one-argument self-application function. Denoting the omega combinator with the symbol `m`, the equivalent as a Sxript function reads:

```
func(m, {[x]([x])}):
```

In the lambda calculus notation, $x(x)$ is equivalent to xx , which is also true for this discussion. The body of function m could be changed to $[x][x]$ (no parentheses) without consequence.

The Omega combinator is as simple as it looks. For instance $m(5)$ returns 55. Of course, $m(+5)$ returns $+10.0$. Interestingly though, $m(5+)$ returns $5+5+$. This essentially tricks the language into lazy-evaluating the incomplete math expression, which is to say, skips the evaluating.

Putting Ω to greater use, also define two functions r , s such that

$$r = \Omega(\Omega(x))$$

$$s = r(r(x)) ,$$

or equivalently:

```
func(r, {m(m([x]))}):
func(s, {r(r([x]))}):
```

For lack of proper names, let's call r the *second-order* self-applicator, and correspondingly s the *fourth-order* self-applicator. While this toolset may seem meager at face value, maybe a term like 'fourth-order' may foreshadow what's to come. For a quick test of the above, one finds:

```
[4]: r(0)
[4]: 0000

[5]: s(0)
[5]: 0000000000000000
```

That is, $r(0)$ returns 0000, a total of four digits. However, $s(0)$ returns sixteen digits. Naturally, the functions m , r , s can be compounded by standard means, i.e. $m(r(0))$ returns 00000000, whereas $m(s(0))$ returns 32 digits.

Next, recall an obscure shortcut of syntax reserved by the question mark operator (?). The question operator allows a single-argument function to be called as if in-line with arithmetic. For instance, $\cos ? 0$ is equivalent to $\cos(0)$, or, $r?0$ resolves to 0000.

Trying the combinations $m(r?)(0)$, $r(m?)(0)$, we receive sixteen zeros in each case.

```
[6]: m(r?)(0)
[6]: 0000000000000000

[7]: r(m?)(0)
[7]: 0000000000000000
```

To understand these, it helps to take away the argument on the right to attain the lazy evaluation of the left.

```
[8]: m(r?)
[8]: r?r?

[9]: r(m?)
[9]: m?m?m?m?
```

We can insert any of m , r , s in place of the function calls above. For instance, each of $r(r?)(0)$, $m(s?)(0)$, return $4^4 = 256$ digits. Things get bigger from here, as $s(m?)(0)$, $r(s?)(0)$ each return $4^8 = 65536$ digits.

```
[10]: len(r(r?)(0))
[10]: 256

[11]: len(m(s?)(0))
[11]: 256

[12]: len(s(m?)(0))
[12]: 65536

[13]: len(r(s?)(0))
[13]: 65536
```

Although no Sxript implementation has finished the next calculation to date, the expression $s(s?)(0)$ is equivalent to 4^{16} zeros, where

$$4^{16} = 4294967296 .$$

At one byte per character, this is 4 GB of data!

More fun can be had with the Omega combinator. For this, let us define a few functions that are less abstract than the previous three:

```
func(e, {[x]+1}):
func(f, {[x]*2}):
```

To see the program and the output without having to execute anything, the following expressions have the anticipated result subtracted such that the return is all zeros.

```
(m(m(e?))(0)) - (2*2)
(m(r(e?))(0)) - (2*4)
(r(m(e?))(0)) - (2*4)
(r(r(e?))(0)) - (2*8)
(m(s(e?))(0)) - (2*16)
(s(m(e?))(0)) - (2*16)
(r(s(e?))(0)) - (2*32)
(s(s(e?))(0)) - (2*128)
```

```
(m(m(f?))(1)) - (4^2)
(m(r(f?))(1)) - (4^4)
(r(m(f?))(1)) - (4^4)
(r(r(f?))(1)) - (4^8)
(m(s(f?))(1)) - (4^16)
(s(m(f?))(1)) - (4^16)
(r(s(f?))(1)) - (4^32)
(s(s(f?))(1)) - (4^128)
```

We may also consider the rather ugly function `manycos`, which takes an initial number `x` along with an iteration number `y` and calculates

$$\text{manycos}(x) = \cos(\cos(\cos(\dots x \dots))) ,$$

with `y` total cosines.

```
func(manycos, {sub({
  let(a, cos([x])):
  [y]-1:do:
    let(a, cos([a])):
  loop:print_[a]
}))})
```

Interestingly, even for relatively small `y`, the return from `manycos` tends to converge to a single answer, regardless of the initial value `x`. The infinite-`y` return from `manycos` happens to be the solution to the transcendental equation $x = \cos(x)$, solved by

$$x = 0.7390851332151607 \dots$$

To evaluate the near-equivalent of `manycos` using recursion, we recycle the `m`, `r`, `s` functions defined above.

```
(m(m(cos?))(0)) - manycos(0,4)
(m(r(cos?))(0)) - manycos(0,8)
(r(m(cos?))(0)) - manycos(0,8)
(r(r(cos?))(0)) - manycos(0,16)
(m(s(cos?))(0)) - manycos(0,32)
(s(m(cos?))(0)) - manycos(0,32)
(r(s(cos?))(0)) - manycos(0,64)
(s(s(cos?))(0)) - manycos(0,256)
```

We can get more abstract by defining a function

$$w = x(x(y)) ,$$

which generalize the functions `r` and `s` above. Then, using the tilde (`~`) operator, we build expressions such as `r ~ w ~ 0`, which is equivalent to `r(r(0))`, returning sixteen zeros.

14.5 Impure Functions

Closure

In functional programming, a *closure* is a function that captures its surrounding environment, allowing it to access and, in some cases, modify variables and functions from its defining scope, even when called in a different context. For an example, consider the following ‘counter’ program implemented in JavaScript:

```
function makecounter() {
  let num = 0;
  return function() {
    num++;
    console.log(num);
  };
}

let counterA = makecounter();
counterA(); // returns 1
counterA(); // returns 2

let counterB = makecounter();
counterB(); // returns 1
counterB(); // returns 2
counterB(); // returns 3
```

The high-order function `makecounter` initializes a variable `num` to one and then returns a function whose job is to increment `num` by one.

Using `makecounter` to create a new function, namely `counterA`, it turns out that a copy of `num`, initialized to one, is ‘saved’ invisibly and is associated with `counterA`

in memory. Then, whenever `counterA` is invoked, the variable `num` is treated as mutable.

Using `makecounter` again to create another function `counterB`, this function gets its own fresh copy of the `num` variable, regardless of how `counterA` has treated its own copy.

Seemingly, closure threatens to breed confusion about hidden mutable state, breaking of scope, or other subtle dependencies. At face value, the behavior of the above counter functions appear to behave *impurely*, namely because the output always changes while the call to the function does not.

Mutable Functions

`Sxript` does not (need to) exhibit closure in the pure sense. To achieve a closure-like effect though, we should be able to define a function that embeds the data intended to be enclosed, which means building a function that is intentionally impure.

For an example comparable to the JavaScript case above, we write the following `Sxript` code:

```
func(makecounter, {sub({
  let(name, [x]):
  print_{
    func({name}, {
      0:
      func({name},
        replace(report({name}), 1,
          unoccult(elem(report({name}), 1))
            + 1 )})})})}):

unoccult(makecounter(counterA)):
unoccult(makecounter(counterB)):

func(tick, {block({
  unoccult(elem(report([x]), 2)):
  print_unf(unoccult(
    elem(report([x]), 1)))})}):

print_tick(counterA),:
print_tick(counterA),:
print_tick(counterB),:
print_tick(counterB),:
print_tick(counterB)
```

The `makecounter` UDF is a high-order function that takes the name of a new counter as the argument.

For a return, `makecounter` emits a piece of occulted code containing two items. First is the ‘enclosed’ data referenced by the function, a hard-coded zero in this case. Second is a `let` primitive containing an instruction to increment the enclosed data and overwrite the definition of the counter. That is, we have a high-order function giving rise to another high-order function that rewrites itself.

Initializing a new counter requires removing the occulting structure returned by `makecounter`. In order to use a counter, we define another function `tick` that executes the instruction in a given counter and then reports the respective enclosed data. The output of the above example is the same as the JavaScript case:

```
1,2,1,2,3
```

Note that the same program can be written, behaving identically, if we replace all calls to `func` with `let`, and all calls to `report` with `symbol`.

```
func(makecounter, {sub({
  let(name, [x]):
  print_{
    let({name}, {
      0:
      let({name},
        replace(symbol({name}), 1,
          unoccult(elem(symbol({name}), 1))
            + 1
        )
      )
    })
  })})}):

func(tick, {block({
  unoccult(elem(symbol([x]), 2)):
  print_unf(unoccult(
    elem(symbol([x]), 1)))})}):
```

15 Structured Programs

Strictly speaking, there is no such thing as a multi-line Sxript program. As we know, multi-line programs are allowed within the `block` primitive (public scope) or `sub` primitive (private scope).

As it turns out, multi-line programs are also allowed to be stored via text file and read during execution. The contents of the file does not need to be wrapped in a `block`-like primitive. For example, consider the file `vector.txt`, which contains:

```
func(last,{elem([x],len([x]))}):

func(fill,{
  smooth(<for(
    <k,[x],[y],[z]>,{[k],})>)):

func(dotprod,{reduce(join,[x]*[y]))}:

func(crossprod,{
  sub({
    let(r1,[x]):
    let(r2,[y]):
    print_<
      (elem([r1],2)*elem([r2],3)-
       elem([r1],3)*elem([r2],2)),
      -1*(elem([r1],1)*elem([r2],3)-
        elem([r1],3)*elem([r2],1)),
      (elem([r1],1)*elem([r2],2)-
        elem([r1],2)*elem([r2],1))
    >)}}):

print_ 'vector.txt':
```

The file `vector.txt` is stored in the `prg` directory (short for 'programs'), which is neighbor to `bin`, in which the Sxript executable is launched. (These details will vary per implementation but the in-program use will be the same throughout.)

Include Primitive

Clearly, the intent of `vector.txt` is to impart a handful of user-defined functions in a subsequent program, thus we want a `block`-like primitive having public scope to take on the file contents. For this, we use `include`, which takes the file location and name as a quote.

```
[1]: include('../prg/vector.txt')
[1]: vector.txt
```

With the above, we can test the new functions a number of ways, for instance:

```
[2]: report(last)
[2]: {elem([x],len([x]))}

[3]: dotprod(<1,2,3>,<4,5,6>)
[3]: +32.0

[4]: crossprod(<1,2,3>,<4,5,6>)
[4]: <-3.0,+6.0,-3.0>
```

Run Primitive

A second primitive designated to running external programs, called `run` does everything `include` does with the exception of having private scope. That is, variables and functions within `run()` do not touch the environment unless carried away by `print_`.

For example, the examples used to discuss `apply`, `map`, `reduce` are contained in the file `applymapreduce.txt`.

```
print_apply(sqrt,<9,16,25,121>),:
print_apply(len,
  <'dog','cat','mouse'>),:
print_apply(type,<1+1,'Hello',type>),:
print_map(<10,5,3>,greater,<6,5,4>),:
print_map(<'dog','cat','mouse'>,
  instr,<'o','z','u'>),:
print_map(<1,2,3>,join,<4,5,6>),:
print_map(<1,2,3>,join,<4,5,6>+0),:
print_reduce(join,<a,b,c,d,e>),:
print_reduce(join,<1,2,3+0,4,5>),:
print_reduce(
  join,<5,*,4,*,3,*,2,*,1>),:
print_reduce(greater,<4,3,0,-2>)
```

Running this program gives all familiar outputs. For a sanity check, we also verify that the present scope is empty.


```
[1]: run('../prg/applymapreduce.txt')
[1]: <3,4,5,11>,<3,3,5>,
      <number,quote,word>,
      <1,0,0>,<2,0,3>,<14,25,36>,
      <+5.0,+7.0,+9.0>,abcde,
      +15.045,+120.0,1

[2]: report()
[2]: {print_()}
```

Dependency

Using the `include` primitive, it's straightforward for external programs to depend on external programs, allowing the programmer to use the sum of all involved variables and functions.

For example, present in the `prg` directory is the file `cpmplexnum.txt`, which contains all of the user-defined functions for complex arithmetic. In addition, there exists a file `complexfunc.txt`, which contains:

```
include('../prg/complexnum.txt'):

func(cderiv,{
  (([x]([y][c+][z]))[c-]
   ([x]([y][c-][z]))[c/]
   (<2,0>[c*][z]))):

func(clinapx,{
  [y] [c-] (([x]([y])) [c/]
            cderiv([x],[y],[z]))):

func(cnewton,{
  iff(
    (greater(0.000001,
             abs(cmag([x]([y])))))
    | (greater(abs(cmag([x]([y]))),
               1000000)),
    {[y]},
    {cnewton([x],
              clinapx([x],[y],<0.001,0>))}):

print_ 'complexfunc.txt':
```

Executing the above, we load the whole complex number toolkit.

```
[1]: include('../prg/complexfunc.txt')
[1]: complexfunc.txt
```

For a test, consider the order-two equation

$$z + z^2 = 7 + 12i,$$

which has two solutions

$$z \approx \begin{cases} 2.7611 + 1.8398i \\ -3.7611 - 1.8398i \end{cases}.$$

The next job is to turn this into a `Sxript` function.

```
[2]: func(p,{
      ([x][c+]( [x][c^]2)[c-]<7,12>)
      })
[2]: p
```

Supplying the initial guesses (2,2), (-2,-2), respectively, we use the loaded `cnewton` UDF to find two solutions.

```
[3]: cnewton(p,<2,2>,5)
[3]: <+2.7611405131341713,
      +1.8398471258245837>

[4]: cnewton(p,<-2,-2>,5)
[4]: <-3.7611405259465363,
      -1.839847138247968>
```

16 Applications

16.1 Calculus 101

Differentiation

Some of our calculus toolkit was developed either before or during the study on high-order functions, thus some streamlining is due for the sake of programmer-friendliness. To begin we restate the functions for the derivative and for the linear approximation at the heart of Newton's method in one dimension:

```
func(deriv,{
  (([x]([y]+[z]))
  -([x]([y]-[z]))) / (2*[z])
}):

func(linapx,{
  [y] - ([x]([y]))/deriv([x],[y],[z])
}):
```

For completeness, the 'baseline' UDF for Newton's method is stated as follows - note the hardcoded function `p`, which represents the problem to be solved.

```
func(newton,{
  iff(
    (greater(0.000001,abs(p([x]))))
    |(greater(abs(p([x])),1000000)),
    {[x]},
    {newton(linapx(p,[x],0.001))})
}):
```

Of course, it would be cleaner to abstract the call to `p` into a variable, thus bumping the numerical argument `x` over to slot `y`.

```
func(newton,{
  iff(
    (greater(0.000001,abs([x]([y]))))
    |(greater(abs([x]([y])),1000000)),
    {[y]},
    {newton([x],
      linapx([x],[y],0.001))})}):
```

Then, we can send the problem and the initial guess directly to the `newton` UDF to receive an answer. For instance, we can set $p(x) = x - \cos(x)$ and supply an initial guess of $x_0 = 1$ to find:

```
[4]: newton$({[x]-cos([x])}),1)
[4]: +0.7390851333834115
```

Integration

Recall that one-dimensional integrals are numerically approximated using Riemann sums. The program is required to specify (i) the function to be integrated, (ii) the lower- and (iii) upper-bounds on the dependent variable, always `x`, and (iv) the number of integration bins.

Building a four-argument function to evaluate the Riemann sum, one can write:

```
func(rsum,{sub({
  let(dx,([z]-[y])/[t]):
  let(ls,0):let(rs,0):let(ms,0):
  let(ts,0):let(si,0):let(j,0):
  [t]:do:
    let(xj,[y] + [j]*[dx]):
    let(x1,[xj] + [dx]):
    let(xm,([xj] + [x1])/2):
    let(ls,[ls] + [dx]*([x]([xj]))):
    let(rs,[rs] + [dx]*([x]([x1]))):
    let(ms,[ms] + [dx]*([x]([xm]))):
    let(j,[j]+1):loop:
  let(ts,(1/2)*([ls]+[rs])):
  let(si,(1/3)*([ts]+2*[ms])):
  print_[si]})}):
```

The output of the function is contained in `[ls]` for the left-hand sum, `[rs]` for the right-hand sum, `[ms]` for the midpoint sum, `[ts]` for the trapezoid sum, and `[si]` for Simpson's rule. By default, the function returns `si`. For a test, let us calculate the area under the curve $y = 4x - x^2$ in the interval $0 \leq x \leq 4$ using $t = 15$ bins to reproduce the example used previously:

```
[7]: rsum$({4*[x]-[x]*[x]}),0,4,15)
[7]: +10.666666666666664
```

16.2 Complex Numbers

Any complex number z contains a real part a and an imaginary part b and is written

$$z = a + ib,$$

where i is the imaginary unit $\sqrt{-1}$.

For representing such a complex number in Sxript, we use a vector `<a,b>`. Due to this, functions for addition

and subtraction of complex numbers are straightforwardly written:

```
func(cadd, {[x]+[y]}):
func(csub, {[x]-[y]}):
```

For multiplication of two complex numbers z_1, z_2 , the product is written

$$z_1 \cdot z_2 = (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1) .$$

Churning this into a UDF, one writes:

```
func(cmul, {<
  elem([x],1)*elem([y],1)-
  elem([x],2)*elem([y],2),
  elem([x],1)*elem([y],2)+
  elem([x],2)*elem([y],1)>}):
```

For division of two complex numbers z_1, z_2 , the quotient is written

$$\frac{z_1}{z_2} = \frac{z_1 \cdot \bar{z}_2}{|z_2|^2} .$$

For this we need functions for the complex conjugate and the magnitude of a complex number.

```
func(conj, {
  <elem([x],1), -1*elem([x],2)>}):

func(cmag2, {
  elem([x]*[x],1)+
  elem([x]*[x],2)}):

func(cmag, {sqrt(cmag2([x]))}):

func(cdiv, {
  (cmul([x], conj([y])))/
  (cmag2([y]))}):
```

For navigation in the complex plane, we must easily convert complex numbers from the Cartesian form $z = (a, b)$ to the polar form

$$z = (r \cos(\phi), r \sin(\phi)) ,$$

where r is the magnitude and ϕ is the phase. For this we define a function `cpol` that takes a complex number z and returns (r, ϕ) . The `pcart` function implements the z -equation as written above.

```
func(cpol, {
  <cmag([x]),
  atan2(elem([x],2), elem([x],1))>}):

func(pcart, {
  elem([x],1)*
  <cos(elem([x],2)), sin(elem([x],2))>}):
```

Raising a complex number to a real-number (whole or decimal) power is achieved with the `cpow` UDF.

```
func(cpow, {
  sub({
    let(a, cpol([x])):
    let(rad, elem([a],1)):
    let(ang, elem([a],2)):
    let(pwr, [y]):
    let(rad, [rad]^[pwr]):
    let(ang, [ang]*[pwr]):
    print_pcart(<[rad], [ang]>)}))):
```

For complex exponents, we need something more intricate:

```
func(cpow2, {sub({
  let(ex, elem([x],1)):
  let(ey, elem([x],2)):
  let(ea, elem([y],1)):
  let(eb, elem([y],2)):
  let(lnz, [ex]*[ex] + [ey]*[ey]):
  print_iff([lnz]=0, {<0,0>}, {sub({
    let(lnz, 0.5*ln([lnz])):
    let(argz, atan2([ey], [ex])):
    let(mag,
      eul([ea]*[lnz] - [eb]*[argz])):
    let(ang, [ea]*[argz] + [eb]*[lnz]):
    print_pcart(<[mag], [ang]>)}))}):
```

All of the work for a complex logarithm has already been done, thus we assemble the `clog` UDF with the following:

```
func(clog, {sub({
  let(ex, elem([x],1)):
  let(ey, elem([x],2)):
  let(lnz, [ex]*[ex] + [ey]*[ey]):
  print_iff([lnz]=0, {<0,0>}, {sub({
    let(lnz, 0.5*ln([lnz])):
    let(argz, atan2([ey], [ex])):
    print_<[lnz], [argz]>)}))}):
```

Finally, it's convenient to recast the functions for complex arithmetic to inline operators via the tilde (~) operator.

```
let(c+,~cadd~):
let(c-,~csub~):
let(c*,~cmul~):
let(c/,~cdiv~):
let(c^,~cpow~):
let(c2^,~cpow2~):
let(I,<0,1>):
```

For a quick test, suppose we put all of the functions for complex arithmetic into a single block. Then, a typical session in the terminal could look like:

```
[2]: <3,0> [c-] [I]
[2]: <+3.0,-1.0>

[3]: <1,2> [c+] <3,4>
[3]: <+4.0,+6.0>

[4]: <1,2> [c*] <3,4>
[4]: <-5.0,+10.0>

[5]: <3,0>[c+]<4,0>[c*][I]
[5]: <+3.0,+4.0>

[6]: conj(<4,-3>)
[6]: <4,+3.0>
```

```
[7]: cmag(<4,-3>)
[7]: 5

[8]: cdiv(<4,-3>,<1,2>)
[8]: <-0.4,-2.2>

[9]: <1,2> [c/] <4,-3>
[9]: <-0.08,+0.44>

[10]: cpow2(<3,4>,<1,2>)
[10]: <-0.4198131755619574,
      -0.6604516942073322>

[11]: clog(<3,4>)
[11]: <+1.6094379124341003,
      0.9272952180016122>
```

Complex Newton's Method

Here we combine the tools of calculus and complex numbers. Exhibit one is the derivative of a complex function,

which takes the original `deriv` UDF and replaces any real numbers with complex numbers, along with any standard arithmetic operators with complex operators. Do the same with the `linapx` UDF.

```
func(cderiv,{
  ([x]([y][c+][z]))[c-]
  ([x]([y][c-][z]))[c/]
  (<2,0>[c*][z]))):

func(clinapx,{
  [y] [c-] ([x]([y])) [c/]
  cderiv([x],[y],[z]))):
```

Next we retool the function for Newton's method to handle complex numbers.

```
func(cnewton,{
  iff(
    (greater(0.000001,
      abs(cmag([x]([y]))))
    | (greater(abs(cmag([x]([y]))),
      1000000)),
    {[y]},
    {cnewton([x],
      clinapx([x],[y],<0.001,0>))}):
```

For a test, consider the order-three equation

$$z^3 + 2z^2 + 3z + 4 = 0,$$

which has one real solution and two complex solutions:

$$z \approx \begin{cases} -1.6506 \\ -0.1747 + 1.5468i \\ -0.1747 - 1.5469i \end{cases}$$

Encoding this as a Sxript function and testing, we write:

```
func(p,{(
  ([x][c^]3) [c+]
  (<2,0>[c*]([x][c^]2)) [c+]
  (<3,0>[c*][x]) [c+] <4,0>
)}) :

print_cnewton(p,<-1,0>,5) ,\n:
print_cnewton(p,<0,1>,5) ,\n:
print_cnewton(p,<0,-1>,5) :
```

The output of the above, as expected, consists of the three approximate solutions to the order-three equation.

17 Problems and Solutions

17.1 100 Doors

There are 100 doors in a row that are all initially closed. You make 100 passes by the doors.¹

- The first time through, visit every door and toggle the door (if the door is closed, open it; if it is open, close it).
- The second time, only visit every 2nd door (door 2, 4, 6, ...), and toggle it.
- The third time, visit every 3rd door (door 3, 6, 9, ...), etc, until you only visit the 100th door.

After the last pass, which doors are open?

Solution

As subprogram:

```
func(mult,{[x]*[y]}):

func(fill,{
  smooth(<for(
    <k,[x],[y],[z>,{[k],})>>)):

let(doors,apply$({-1.0}),fill(1,100,1)):

func(toggle,{
  smooth(<
    for(<i,1,100,1>,{
      iff([i]%[x]=0,{-1,},{1,})>>)):

for(<i,1,100,1>,{
  let(doors,map([doors],mult,toggle([i]))}):

print_smooth(<for(<i,1,100,1>,{
  iff(elem([doors],[i])=1,{[i],})>>
```

Output

```
<+1.0,+4.0,+9.0,+16.0,+25.0,+36.0,+49.0,+64.0,+81.0,+100.0>
```

¹Rosetta Code. 100 Doors. https://rosettacode.org/wiki/100_doors

17.2 Babbage Problem

What is the smallest positive integer whose square ends in the digits 269696?²

Solution

As single line:

```
func(babbage,{
  iff([x]([y]))=int([x]([y])),{
    [x]([y])
  },{
    babbage([x],[y]+1)
  })
})($({([x])*(10^6) + 269696}^.5),1)
```

Output

```
+25264.0
```

²Rosetta Code. Babbage Problem. https://rosettacode.org/wiki/Babbage_problem

17.3 Coconut Pyramids

Build a three sided pyramid and a four sided pyramid by stacking coconuts. Keep building them higher and higher. At some point, you will find you have used nearly the same number of coconuts for each pyramid. When the difference in the number of coconuts is 1, how big are the pyramids, i.e., how many coconuts used in each?

Solution

As subprogram:

```
let(iSq,1): let(nSq,1): let(iTri,1): let(nTri,1):

func(copy,{for(<i,1,[y],1>,{[x]})}):
func(tabfill,{copy(\ ,9-len([x]))[x]}):

print_\n'Triangle pyramid      Square pyramid\n':
print_(apply(tabfill,<'i','n','i','n'>)):

100:
do:
  iff(greater([nSq],[nTri]),{block({print_():
    let(iTri,[iTri] + 1):
    let(nTri,[nTri] + [iTri] * ([iTri] + 1) / 2)})
},{block({print_():
  let(iSq,[iSq] + 1):
  let(nSq,[nSq] + [iSq] ^ 2)})
}):
print_
iff(abs([nTri]-[nSq])=1,{
  \n (apply(tabfill,<[iTri],[nTri],[iSq],[nSq]>))
}):
loop:

print_\n'----- last checked -----'\n:
print_(apply(tabfill,<[iTri],[nTri],[iSq],[nSq]>))
```

Output

```
Triangle pyramid      Square pyramid
<      i,          n,      i,          n>
<    +2.0,    +4.0,    +2.0,    +5.0>
<    +6.0,    +56.0,    +5.0,    +55.0>
<   +11.0,   +286.0,    +9.0,   +285.0>
----- last checked -----
<   +57.0, +32509.0,   +45.0, +31395.0>
```

17.4 Golden Ratio

Write a program that computes the golden ratio.

Solution

As structured program:

```
include('../prg/sqrt.txt'):

let(acc,250):

print_
largetrim(
  (1 [1+] lsqrt(5,2.22,[acc])) [1/] 2
,[acc])
```

Output

```
+1.61803398874989484820458683436563811772030917980576286213544862270526046281890
24497072072041893911374847540880753868917521266338622235369317931800607667263544
33389086595939582905638322661319928290267880675208766892501711696207032221043216
2695486262963
```


17.5 Haskell Tutorial

One of the best tutorials on computer programming ever published was posted to the YouTube channel *FrungyKing* on August 18, 2013. The video is called *Programming - Why Haskell is Great - 10 minutes*. The full url to the video is: <https://www.youtube.com/watch?v=RqvCNb7fKsg>

The challenge is to reproduce the tutorial using Sxript in place of Haskell.

Solution

As subprogram:

```
print_ let(greeting, 'Hello') \n:
print_ func(swedish, {[x]*'f' - 'f'}) \n:
print_ swedish ? [greeting] \n:
print_ let(personname, 'Satoshi') \n:
print_ swedish ? [personname] \n:
print_ let(sentence, [greeting] + ' ' + [personname]) \n:
print_ swedish ? [sentence] \n:
print_ func(yell, {[x]+'!'}) \n:
print_ func(enthusiastically, {[x]([x]([x]([y]))})) \n:
print_ func(very, {enthusiastically([x], [y])}) \n:
print_ yell ? [greeting] \n:
print_ yell ~ enthusiastically ~ [greeting] \n:
print_ yell ? swedish ? [greeting] \n:
print_ yell ~ enthusiastically ~ (swedish ~ very ~ [greeting]) \n:
print_ (swedish~very~[sentence])($({very([x], ())})~enthusiastically~yell)
```

Output

```
Hello
swedish
Hfeflflfo
Satoshi
Sfaftfofsfhfi
Hello Satoshi
Hfeflflfof fSfaftfofsfhfi
yell
enthusiastically
very
Hello!
Hello!!!
Hfeflflfo!
Hffffffefffffffflfffffflfffffffo!!!
Hffffffefffffffflfffffflfffffffoffffff fffffffSfffffffafffffftfffffffoffffff
fsfffffffhffffffi!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

17.6 Koch Fractal

Plot the Koch fractal snowflake.

Solution

As subprogram:

```
'let(path,'F')':      'line':
let(path,'FRRFRRF'):'snowflake':
let(depth,3):

func(sweep,{sub({
  let(a,[x]):let(b,[y]):
  let(c,[z]):let(r,''):
  print_for(<i,1,len([a]),1>,{block({
    let(d,mid([a],[i],1)):
    iff([d]=[b],{
      let(r,[r]+[c]),{
      let(r,[r]+[d])}):
    print_()})}):print_[r]})}):

[depth]:do:let(path,
  sweep([path],'F','FLFRRFLF')):
loop:
```

```
let(turtle,<0,0,0>):
let(pi,4*atan(1)):
func(crawl,{
  <elem([x],1)+
    ([y])*cos(elem([x],3)*[pi]/180),
    elem([x],2)+
    ([y])*sin(elem([x],3)*[pi]/180),
    elem([x],3)>}):
func(swivel,{
  <elem([x],1),elem([x],2),
    elem([x],3) + [y]>}):
func(fwd,{let(turtle,
  crawl([turtle],[x]))}):
func(lft,{let(turtle,
  swivel([turtle],[x]))}):
func(rgt,{let(turtle,
  swivel([turtle],[x]))}):
```

```
print_scatter(smooth(<sub({
  print_for(<i,1,len([path]),1>,{
  block({
    iff(mid([path],[i],1)='F',{fwd(2)}):
    iff(mid([path],[i],1)='L',{lft(60)}):
    iff(mid([path],[i],1)='R',{rgt(-60)}):
    print_<reduce($({[x],[y]}),
      left([turtle],2))>,}>)),
  36,36,-1)
```

Output

As scatter plot (showing top half only):

```
      *** *
        * * * *
      *** *   *** *
        *           *
        **         * *
          *         *   *
      *** *   **   * *   *** *
    * * * * *   * * * * *
  ** *   *** *   *** *   *** *
 *
 *
 *
 *
 *
 *
***          *** *
* *          * *
* *          * *
```

17.7 Nth

Write a function/method/subroutine/... that when given an integer greater than or equal to zero returns a string of the number followed by an apostrophe then the ordinal suffix.³

Use your routine to show here the output for at least the following (inclusive) ranges of integer inputs: 0-25, 250-265, 1000-1025. Note: apostrophes are now optional to allow correct apostrophe-less English.

Solution

As structured program:

```
func(nth,{
  sub({
    let(a,right([x],2)):
    let(result,'th'):
    iff(right([a],1)=1,{let(result,'st')}):
    iff(right([a],1)=2,{let(result,'nd')}):
    iff(right([a],1)=3,{let(result,'rd')}):
    iff([a]-11=0,{let(result,'th')}):
    iff([a]-12=0,{let(result,'th')}):
    iff([a]-13=0,{let(result,'th')}):
    'print_[x]\'[result]':
    print_[x][result]
  })
}):

include('../prg/vector.txt'):

print_ apply(nth,apply(unf,fill(1,25,1)))      \n\n:
print_ apply(nth,apply(unf,fill(250,265,1)))   \n\n:
print_ apply(nth,apply(unf,fill(1000,1025,1))) :
```

Output

```
<1st,2nd,3rd,4th,5th,6th,7th,8th,9th,10th,11th,12th,
13th,14th,15th,16th,17th,18th,19th,20th,21st,22nd,
23rd,24th,25th>

<250th,251st,252nd,253rd,254th,255th,256th,257th,
258th,259th,260th,261st,262nd,263rd,264th,265th>

<1000th,1001st,1002nd,1003rd,1004th,1005th,1006th,
1007th,1008th,1009th,1010th,1011th,1012th,1013th,
1014th,1015th,1016th,1017th,1018th,1019th,1020th,
1021st,1022nd,1023rd,1024th,1025th>
```

³Rosetta Code. Nth. <https://rosettacode.org/wiki/N%27th>

17.8 Pascal Triangle

Write program that calculates the *n*th row of the Pascal triangle. Display the first sixteen rows. Display the row that begins with 1, 100, ...

Solution

As structured program:

```
include('..prg/factorial.txt'):

func(pascal,{
  iff(let(n,[x])=0,{<1>},{sub({
    iff([n]%2=0,{block({print_():
      let(p,smooth(<for(<i,0,[n]/2-1,1>,{binom([n],[i]),})>>)):
      let(q,stack([p],binom([n],[n]/2))))))
  },{block({print_():
    let(p,smooth(<for(<i,0,[n]/2,1>,{binom([n],[i]),})>>)):
    let(q,[p])
  }})):
  let(r,smooth(<for(<i,len([p]),0,-1>,{elem([p],[i]),})>>)):
  print_stack([q],[r])
}}))):

include('..prg/largenum.txt'):

print_
for(<i,0,15,1>,{
  apply(largen,pascal([i])) \n
}):

print_apply(largen,pascal(100))
```

Output

(1 of 2)

```
<1>
<1,1>
<1,2,1>
<1,3,3,1>
<1,4,6,4,1>
<1,5,10,10,5,1>
<1,6,15,20,15,6,1>
<1,7,21,35,35,21,7,1>
<1,8,28,56,70,56,28,8,1>
<1,9,36,84,126,126,84,36,9,1>
<1,10,45,120,210,252,210,120,45,10,1>
```

Output

(2 of 2, square wrapping)

<1,100,4950,161700,3921225,75287520,1192052400,16007560800,186087894300,1902231808400,17310309456440,141629804643600,1050421051106700,7110542499799200,44186942677323600,253338471349988640,1345860629046814650,6650134872937201800,30664510802988208300,132341572939212267400,535983370403809682970,2041841411062132125600,7332066885177656269200,24865270306254660391200,79776075565900368755100,242519269720337121015504,699574816500972464467800,1917353200780443050763600,4998813702034726525205100,12410847811948286545336800,29372339821610944823963760,66324638306863423796047200,143012501349174257560226775,294692427022540894366527900,580717429720889409486981450,1095067153187962886461165020,1977204582144932989443770175,3420029547493938143902737600,5670048986634686922786117600,9013924030034630492634340800,13746234145802811501267369720,20116440213369968050635175200,28258808871162574166368460400,38116532895986727945334202400,49378235797073715747364762200,61448471214136179596720592960,73470998190814997343905056800,84413487283064039501507937600,

93206558875049876949581681100,98913082887808032681188722800,100891344545564193334812497256,98913082887808032681188722800,93206558875049876949581681100,84413487283064039501507937600,73470998190814997343905056800,61448471214136179596720592960,49378235797073715747364762200,38116532895986727945334202400,28258808871162574166368460400,20116440213369968050635175200,13746234145802811501267369720,9013924030034630492634340800,5670048986634686922786117600,3420029547493938143902737600,1977204582144932989443770175,1095067153187962886461165020,580717429720889409486981450,294692427022540894366527900,143012501349174257560226775,66324638306863423796047200,29372339821610944823963760,12410847811948286545336800,4998813702034726525205100,1917353200780443050763600,699574816500972464467800,242519269720337121015504,79776075565900368755100,24865270306254660391200,7332066885177656269200,2041841411062132125600,535983370403809682970,132341572939212267400,30664510802988208300,6650134872937201800,1345860629046814650,253338471349988640,44186942677323600,7110542499799200,1050421051106700,141629804643600,17310309456440,1902231808400,186087894300,16007560800,1192052400,75287520,3921225,161700,4950,100,1>

17.9 Points on a Circle

Given three Cartesian points (x_1, y_1) , (x_2, y_2) , (x_3, y_3) , write a program that determines the center x -coordinate, center y -coordinate, and radius of a circle traced through the points. That is, find h , k , and r in the equation

$$(x - h)^2 + (y - k)^2 = r^2.$$

Demonstrate on $(4, 1)$, $(-3, 7)$, $(5, -3)$.

Solution

As single line:

```
func(circo,{
  sub({
    let(x1,elem([x],1)):let(x2,elem([y],1)):let(x3,elem([z],1))
    :let(y1,elem([x],2)):let(y2,elem([y],2)):let(y3,elem([z],2))
    :let(alpha,((([x2])^2)-((([x1])^2)+((([y2])^2)-((([y1])^2)))
    :let(delta,((([x3])^2)-((([x1])^2)+((([y3])^2)-((([y1])^2)))
    :let(beta,    -2*([x2])+2*([x1]))
    :let(gamma,   2*([y1])-2*([y2]))
    :let(epsilon,-2*([x3])+2*([x1]))
    :let(rho,     2*([y1])-2*([y3]))
    :let(machine,0.0001):
    :‘Some division-by-zero prevention...’
    :let(beta,    iff(greater([machine],abs([beta])),
                      {[beta]+[machine]},{[beta]}))
    :let(gamma,   iff(greater([machine],abs([gamma])),
                      {[gamma]+[machine]},{[gamma]}))
    :let(epsilon,iff(greater([machine],abs([epsilon])),
                      {[epsilon]+[machine]},{[epsilon]}))
    :let(rho,     iff(greater([machine],abs([rho])),
                      {[rho]+[machine]},{[rho]}))
    :let(k,([alpha]/[beta]-[delta]/[epsilon])
           /([rho]/[epsilon]-[gamma]/[beta]))
    :let(h,([alpha]/[gamma]-[delta]/[rho])
           /([epsilon]/[rho]-[beta]/[gamma]))
    :let(rad,sqrt((([x1]-[h])^2+([y1]-[k])^2))
    :let(result,<[h],[k],[rad]>)
    :print_[result]})(<4,1>,<-3,7>,<5,-3>)
```

Output

```
<-6.045454545454544,-3.6363636363636367,11.063770821622635>
```

17.10 Square Root Table

Write a program that generates the square root of each integer up to a cutoff value.

Solution

As structured program:

```
include('..prg/largecalc.txt'):

func(rtgen,{sub({
  let(a,unf([x])):
  print_{$({([x][1^]2)[1-]{a})}}
}))):

func(lsqrt,{
  lnewton(
    $({unoccult(rtgen([x]))})([y])
    ,1+[y]/2,[z]))):

print_
for(<i,1,25,1>,{<unf([i]),
  lsqrt([i],1+[i]/2,20)>,\n})
```

Output

```
<1,+1.0>,
<2,+1.4142135623730950488016887242096980785696718753769480731766797379907324784621>,
<3,+1.7320508075688772935274463415058723669428052538103806280558069794519330169088>,
<4,+2.0>,
<5,+2.2360679774997896964091736687312762354406183596115257242708972454105209256378>,
<6,+2.4494897427831780981972840747058913919659474806566701284326925672509603774573>,
<7,+2.6457513110645905905016157536392604257102591830824501803683344592010688232302>,
<8,+2.8284271247461900976033774484193961571393437507538961463533594759814649569242>,
<9,+3.0>,
<10,+3.1622776601683793319988935444327185337195551393252168268575048527925944386392>,
<11,+3.316624790355399849114932736670686683927088545589353597058682146116484642609>,
<12,+3.4641016151377545870548926830117447338856105076207612561116139589038660338176>,
<13,+3.605551275463989293119221267470495946251296573845246212710453056227166948293>,
<14,+3.7416573867739413855837487323165493017560198077787269463037454673200351563069>,
<15,+3.8729833462074168851792653997823996108329217052915908265875737661134830919369>,
<16,+4.0>,
```

```
<17,+4.1231056256176605498214098559740770251471992253736204343986335730949543463376>,  
<18,+4.2426406871192851464050661726290942357090156261308442195300392139721974353863>,  
<19,+4.3588989435406735522369819838596156591370039252324449368903441381595573282031>,  
<20,+4.4721359549995793928183473374625524708812367192230514485417944908210418512756>,  
<21,+4.5825756949558400065880471937280084889844565767679719026072421239068684255477>,  
<22,+4.6904157598234295545656301135444662805882283534117371536057018910170246327532>,  
<23,+4.7958315233127195415974380641626939199967070419041293464853091144482572359074>,  
<24,+4.8989794855663561963945681494117827839318949613133402568653851345019207549146>,  
<25,+5.0>
```

Index

0 pow 0, 173

optimization problem, cylinder, 282

aasymptotes, 177

absolute convergence, 188

absolute value, 14, 186

acceleration, 221, 345

acceleration vector, 349, 351

acceleration, freefall, 74

acceleration, inverse-square, 533

acceleration, polar coordinates, 345

acceleration, uniform, 71

acute triangle, 103

addition, associative property, 16

addition, commutative property, 16

addition, complex, 227

addition, fractions, 18

addition, matrix, 246

addition, rates, 33

addition, utilization, 27

addition, vector, 207

additive inverse, 207

adventitious angles, Langley's, 125

affine parameter, 487

air damping, kinematics, 296

algebraic identities, 60

algebraic operations, 16

algebraic properties, vectors, 209

analog clock face, 49

analytic geometry, 340

angle, negative, 110

angle-sum formulas, trigonometric, 111

angles, 102

angular coordinate, 132

angular frequency, 305

angular momentum, 533

anti-commutativity, vectors, 211

anti-symmetric matrix, 255

antiderivative, 299, 310

antiderivative, arccosecant, 303

antiderivative, arccosh, 304

antiderivative, arccosine, 302

antiderivative, arccotangent, 303

antiderivative, arccoth, 304

antiderivative, arccsch, 304

antiderivative, arcsecant, 303

antiderivative, arcsech, 304

antiderivative, arcsine, 302

antiderivative, arcsinh, 304

antiderivative, arctangent, 303

antiderivative, arctanh, 304

antiderivative, $\cos(x) \cdot \exp(x)$, 301

antiderivative, $\cos(x) \cdot \sin(x)$, 302

antiderivative, $\cos(x)^2$, 302

antiderivative, cosecant, 302

antiderivative, cotangent, 301

antiderivative, diminished natural logarithm, 300

antiderivative, exponential, 301

antiderivative, hyperbolic, 304

antiderivative, logarithmic, 300

antiderivative, modified natural logarithm, 301

antiderivative, natural logarithm, 299

antiderivative, nonlinear natural logarithm, 300

antiderivative, powers, roots, 300

antiderivative, reciprocal, 300

antiderivative, roots, powers, 300

antiderivative, secant, 302

antiderivative, shifted natural logarithm, 300

antiderivative, tangent, 301

antiderivative, trigonometric, 301

antiderivative, $x \cdot \cos(x)$, 299

antiderivative, $x \cdot \exp(x)$, 301

antiderivative, $x^2 \cdot \exp(x)$, 301

antiderivatives, inverse trigonometric, 302

apocalypse problem, 539

- apogee, 536
- applied differentiation, 280
- applied polynomial division, 89
- approximation, order, 291
- approximation, small-angle, 116
- arc length, circular, 107
- arc length, integral, 329
- arc length, parametric, 343
- arc length, polar coordinates, 344
- arccosecant, 114, 120
- arccosecant antiderivative, 303
- arccosecant derivative, 278
- arccosecant integral, 323
- arccosh antiderivative, 304
- arccosh derivative, 279
- arccosine, 113, 119
- arccosine antiderivative, 302
- arccosine derivative, 278
- arccotangent, 114, 120
- arccotangent antiderivative, 303
- arccotangent derivative, 278
- arccoth antiderivative, 304
- arccoth derivative, 279
- arccsch antiderivative, 304
- arccsch derivative, 279
- arcsecant, 114, 120
- arcsecant antiderivative, 303
- arcsecant derivative, 278
- arcsecant integral, 323
- arcsech antiderivative, 304
- arcsech derivative, 279
- arcsine, 113, 119
- arcsine antiderivative, 302
- arcsine derivative, 278
- arcsinh antiderivative, 304
- arcsinh derivative, 279
- arctangent, 113, 120
- arctangent antiderivative, 303
- arctangent at two, Taylor expansion, 294
- arctangent derivative, 278
- arctangent integral, 326
- arctangent near one, Taylor expansion, 294
- arctangent near two, Taylor expansion, 294
- arctangent near zero, Taylor expansion, 294
- arctanh antiderivative, 304
- arctanh derivative, 279
- area as displacement, 72
- area element, Cartesian, 347, 353
- area element, cylindrical, 355
- area element, differential, 347
- area element, polar coordinates, 347
- area element, spherical, 357
- area integral, polar coordinates, 343
- area, circle, 107
- area, ellipse, 157, 323
- area, Kepler's second law, 538
- area, parallelogram, 212
- area, polynomial, 162
- area, right triangle, 103
- area, triangle, 103, 212
- area, under curve, 307
- area, unit circle, 224, 317
- arithmetic expressions, 13
- arithmetic operators, 13
- arithmetic, complex, 227
- arrangements, 199
- arrow trick, 207
- associative property, 16
- associative property, addition, 16
- associative property, complex, 228
- associative property, multiplication, 16
- associativity, matrix, 247
- associativity, scalars, 209
- associativity, vector addition, 207
- asymptote, Taylor expansion, 296
- asymptotes, hyperbola, 144
- augmented matrix, 255
- average value, binomial distribution, 445
- average velocity, 72
- average, statistical, 202
- Avogadro's number, 47
- axioms, complex arithmetic, 227
- axis, imaginary, 231
- axis, real, 231
- Babylonian method, 666
- BAC-CAB formula, 213
- backward Euler's method, 669
- basepoint, change, 160
- basis vector derivatives, 345, 357
- basis vectors, 214
- basis vectors, Cartesian coordinates, 214
- basis vectors, change, 217, 218
- basis vectors, cylindrical, 355
- basis vectors, generalized, 487
- basis vectors, instantaneous, 349
- basis vectors, matrix representation, 357
- basis vectors, orthogonality, 356
- basis vectors, polar coordinates, 216
- basis vectors, spherical, 356

- basketball bouncing forever, 93
Bayes' theorem, 195
bijective, 177
binomial coefficients, 80, 293, 660
binomial distribution, 445
binomial distribution, average value, 445
binomial distribution, mode, 445
binomial distribution, standard deviation, 445
binomial expansion, 293
binomial theorem, 80, 82, 164
binomial theorem abuse, 82
binomials, FOIL method, 17
binormal vector, 351
birthday problem, 200
Bohr radius, 48
Bombelli, 68, 225
Bombelli's wild thought, 69
bound orbit, 543
bounded, 187
bounded above, 187
bounded below, 187
box method, 59
branch cuts, 240
branches, complex, 239
Bézier curves, 658
- Cartesian area element, 347, 353
Cartesian basis vectors, 214
Cartesian coordinate system, 353
Cartesian coordinates, 485
Cartesian coordinates, metric, 491
Cartesian functions, 176
Cartesian plane, 52
Cartesian position vector, 214
Cartesian volume element, 353
ceiling, 15
celsius, 30
center of mass, 532
central potential, 546
centroid, 331
cesium decay, 47
chain rule, derivataion, 274
chain rule, derivative, 274
chain rule, integration, 314
change of basis, 217, 218
change of basis, matrix, 247
characteristic equation, 259
characteristic polynomial, 259
checksum, Pascal triangle, 98
circle intersecting circle, 359
circle intersecting points, 360
circle line problem, 174
circle, area, 107
circle, circumference, 107
circle, diameter, 107
circle, offset, 134
circle, parameterized, 108
circle, taxonomy, 107
circle, unit, 108
circles, 107
circular arc length, 107
circular disk method, 329
circular inscription problem, 107
circular motion, 135
circular motion, uniform, 135, 346
circular orbit, 543
circumcircle, 124
circumference, circle, 107
classical probability, 192
classifying functions, 177
classifying general conics, 156
classifying rotated conics, 156
coefficient sum identity, binomial, 80
coefficients, 26, 159
coefficients, binomial, 80, 293, 660
collision of shperes, 530
combinations, 200
combinatorics, 199
common denominator, 18
commutative property, 16
commutative property, addition, 16
commutative property, multiplication, 16
commutativity, vector addition, 207
commutativity, vectors, 210
completing the square, 60
complex addition, 227
complex arithmetic, 227
complex arithmetic, axioms, 227
complex arithmetic, generalized, 229
complex components, 227
complex conjugate, 227
complex division, 229, 237
complex eigenvalues, 260
complex exponent, 239
complex functions, 240
complex inverse, 240
complex magnitude, 228
complex multiplication, 227, 237
complex natural logarithm, 239, 240
complex nth root, 239

- complex number, definition, 227
- complex numbers, 225, 227, 346
- complex numbers, embedded, 234
- complex numbers, matrices, 232
- complex numbers, polar, 237
- complex numbers, vectors, 231
- complex operators, 230
- complex phase, 230
- complex plane, 230
- complex radius, 230
- complex roots, 239
- complex rotations, 236
- complex square root, 239–241
- component subscripts, 205
- component, imaginary, 225
- components, complex, 227
- components, matrix, 244
- components, vector, 205, 215
- composite functions, 274
- compound events, 191
- Compton wavelength, 47
- concavity, 56, 286
- conditional convergence, 189
- conditional probability, 194
- conic sections, 137, 153
- conic sections, general, 155
- conic trajectory, 537
- conjugate, complex, 227
- conservation of energy, 530, 539
- conservation of momentum, 530
- conservation, energy, 169
- constant shift, integration, 315
- constant, Euler, 164
- constant, fine structure, 48
- constant, integration, 311
- continuity, 11, 181
- continuous distributions, 447
- continuous systems, 440
- contraction, 485, 490
- contravariant gradient, 489
- contravariant tensor, 483
- continuity, tensor type, 484
- convention, Einstein summation, 484
- convention, theta, 107
- convergence, 93, 173, 188
- convergence, absolute, 188
- convergence, conditional, 189
- convergent sequence, 185
- converging series, 188
- converging, infinite, 188
- coordinate rotation, 220
- coordinate system, Cartesian, 353, 491
- coordinate system, cylindrical, 355
- coordinate system, polar, 132, 216, 341, 491
- coordinate system, spherical, 356
- coordinate, angular, 132
- coordinate, radial, 132
- coordinates, Cartesian, 485
- coordinates, polar, 485
- coordinates, rotated, 155
- Copernicus method, 197
- cosecant, 109, 118
- cosecant antiderivative, 302
- cosecant derivative, 271
- cosh, 166
- cosh derivative, 279
- cosine, 104, 118, 168, 272
- cosine derivative, 270
- cosine superposition, 112
- cotangent, 109, 118
- cotangent antiderivative, 301
- cotangent derivative, 271
- cotangent near $\pi/2$, Taylor expansion, 296
- cotangent near zero, Taylor expansion, 296
- coth derivative, 279
- counting states, 192
- covariant gradient, 489
- covariant tensor, 483
- coworker problem, 193
- Cramer's rule, 248, 252
- critical point stability, 285
- critical points, 281
- critical vibrations, 338
- cross product, 211
- csch derivative, 279
- cube root, 667
- cube roots, 665
- cubic equations, 66
- cubic, depressed, 66
- curvature, 349, 351
- curvature, polar coordinates, 350
- curve fitting, 653
- curve, periodic, 167
- curved space, 483
- curves, Bézier, 658
- curves, quadratic, 56
- curves, three dimensions, 351
- cycloid, 333, 340
- cylinder, optimization problem, 282
- cylinder, volume, 39

- cylindrical area element, 355
- cylindrical basis vectors, 355
- cylindrical differential arc length, 355
- cylindrical differential interval, 355
- cylindrical line element, 355
- cylindrical position vector, 355
- cylindrical shell method, 330
- cylindrical volume element, 356
- cylindrical coordinate system, 355

- $d(\sin)$ shortcut, integration, 315
- dark integer, 165
- decimal, terminating, 19
- decimals, repeating, 99
- definite integral, 310
- definition, tensor, 485
- degenerate systems, 262
- degrees, 102
- del Ferro-Tartaglia, 225
- delta, Kronecker, 486
- denominator, change, 18
- density, 38
- density, variable, 541
- density, water, 38
- depressed cubic, 66
- derivative, 265
- derivative operator, 277
- derivative, applied, 280
- derivative, arcsecant, 278
- derivative, arccosh, 279
- derivative, arccosine, 278
- derivative, arccotangent, 278
- derivative, arccoth, 279
- derivative, arcsech, 279
- derivative, arcsecant, 278
- derivative, arcsech, 279
- derivative, arcsine, 278
- derivative, arcsinh, 279
- derivative, arctangent, 278
- derivative, arctanh, 279
- derivative, basis vector, 345, 357
- derivative, chain rule, 274, 292
- derivative, cosecant, 271
- derivative, cosh, 279
- derivative, cosine, 270
- derivative, cotangent, 271
- derivative, coth, 279
- derivative, csch, 279
- derivative, diminished natural logarithm, 269
- derivative, $\exp(x^x)$, 267
- derivative, implicit, 277
- derivative, inverse square, 266
- derivative, inverse trigonometric, 278
- derivative, logarithmic, 269
- derivative, mixed techniques, 278
- derivative, modified natural logarithm, 270
- derivative, natural exponential, 267
- derivative, natural logarithm, 269
- derivative, nonlinear natural logarithm, 269
- derivative, parabola, 265
- derivative, parametric, 342
- derivative, power rule, 276
- derivative, product rule, 273, 292
- derivative, quotient rule, 274, 292
- derivative, reciprocal, 266
- derivative, rules, vectors, 345
- derivative, secant, 271
- derivative, sech, 279
- derivative, second, 285, 293
- derivative, shifted natural logarithm, 269
- derivative, sine, 270
- derivative, sinh, 279
- derivative, square root, 266
- derivative, standard exponential, 267
- derivative, tangent, 271
- derivative, tanh, 279
- derivative, trigonometric, 270
- derivative, $x^{\text{pow } x}$, 268, 276
- derivative, $x \cdot \sin(x)$, 272
- derivatives, elementary, 265
- derivatives, exponential, 267
- derivative, exponential, squared argument, 276
- determinant, 248
- determinant multiplication, 251
- determinant notation, 211
- determinant transpose rule, 251
- devtor derivatives, 345
- diagonal form, matrix, 255
- diagonal matrix, 261
- diagonalization, matrix, 260
- diameter, circle, 107
- dice stacking, 197
- differentiable functions, 265
- differential arc length, cylindrical, 355
- differential area element, 347
- Differential geometry, 483
- differential interval, 346, 489
- differential interval, cylindrical, 355
- differential limit, 171
- differential line element, 346, 487

- digits of pi, 665
- dimension, 31
- dimensional analysis, 35
- dimensionless variables, 551
- diminished natural logarithm antiderivative, 300
- diminished natural logarithm derivative, 269
- direction, vector, 206
- discontinuity, endpoint, 181
- discrete distributions, 443
- discriminant, conic, 156, 157
- discriminant, 57
- dispersion, 442
- displacement as area, 72
- displacement, net, 72
- displacement, relative, 532
- distance below horizon, 129
- distance from rocket problem, 284
- distance to curve, 282
- distance to horizon, 129
- distribution function, probability, 440
- distribution, binomial, 445
- distribution, continuous, 447
- distribution, discrete, 443
- distribution, Gaussian, 447
- distribution, normalized probability, 443
- distribution, Poisson, 447
- distributive property, 16
- distributive property, complex, 228
- distributive property, vector, 209, 210
- divergence, 173
- divergent sequence, 185
- division, complex, 229, 237
- division, factoring, 87
- division, fractions, 18
- division, infinite sums, 84
- division, long, 82, 83
- division, polynomial, 82, 89
- domain, 176
- dot product, 210
- double-angle formulas, trigonometric, 111
- drawing, ellipse, 140
- dual line element, 487
- e vs. pi, 283
- Earth, local gravity, 541
- eccentricity, 154, 155, 535, 544
- eccentricity, ellipse, 137
- eigenvalues, 259
- eigenvalues, complex, 260
- eigenvalues, equal, 261
- eigenvalues, non-equal, 261
- eigenvectors, 259
- eigenvectors, basis, 261
- Einstein summation convention, 484
- elemental mass units, 47
- elementary derivatives, 265
- elimination, 257
- ellipse, 137, 340, 537
- ellipse intersecting ellipse, 360
- ellipse intersecting line, 358
- ellipse intersecting parabola, 361
- ellipse, area, 157, 323
- ellipse, tangent line, 277
- elliptic motion, 348
- elliptical orbit, 543
- embedded complex numbers, 234
- endpoint discontinuity, 181
- energy, 40, 670
- energy conservation, 539
- energy, conservation, 169
- energy, gravity potential, 539
- energy, kinetic, 170
- energy, potential, 169
- envelope of trajectories, 77
- envelope, velocity, 333
- equation of a plane, 353
- equation, characteristic, 259
- equation, linear, 53
- equation, manipulation, 26
- equation, quadratic, 56
- equation, transcendental, 30
- equations, 26
- equations, cubic, 66
- equations, deparameterized, 76
- equations, parametric, 340
- equations, transcendental, 664
- equatorial belt problem, 666
- equilateral triangle, 103
- equilibrium, mechanical, 170
- escape velocity, 542
- essential singularity, 181
- Euler constant, 164, 236
- Euler's formula, 236, 237
- Euler's method, 669
- Euler's method, backward, 669
- Euler's method, forward, 669
- Euler's method, mixed, 670
- even functions, 178
- events, 191
- events, compound, 191

- events, independent, 193
- events, mutually exclusive, 192
- expansion, binomial, 293
- expectation value, 202
- explicit vs. implicit, 681
- exponent, complex, 239
- exponent, utilization, 27
- exponential antiderivatives, 301
- exponential derivatives, 267
- exponential growth, 165
- exponential, integration, 313
- exponents, 13
- exponents, compound, 22
- exponents, multiplication, 22
- exponents, negative, 22
- exponents, negative base, 22
- exponents, special, 22
- exponents, zero, 23
- expressions, 13
- expressions, arithmetic, 13
- expressions, nested, 13
- expressions, parabolic, 149

- factorable numerator, 181
- factorial, 14, 41
- factoring, 19
- Factoring Cubic Equations, 66
- factoring cubics, 66
- factoring, division, 87
- factoring, quadratics, 59
- factorization, 19
- Fahrenheit, 30
- Ferro-Tartaglia formula, 67
- Fibonacci numbers, 91
- Fibonacci numbers, negative, 91
- Fibonacci sequence, 91
- field lines, 380
- fine structure constant, 48
- finite series, 188
- finite sums, 189
- fitting, curve, 653
- flat earth, 129
- flat space, 483
- flight time, 77
- floor, 15
- FOIL method, binomials, 17
- folding paper, 43
- formula, quadratic, 56
- formulas, Serret-Frenet, 352
- Forward Euler's method, 669

- fractions, 17
- fractions, addition, 18
- fractions, subtraction, 18
- frame, TNB, 352
- freefall acceleration, 74
- freefall motion, 76
- freefall, gravity, 170
- Frobenius method, 297
- FTC, 310
- function, composite, 274
- function, probability distribution, 440
- function, sinc, 182
- functions, 176
- functions, Cartesian, 176
- functions, classifying, 177
- functions, complex, 240
- functions, differentiable, 265
- functions, even, 178
- functions, hyperbolic, 238
- functions, inverse, 177
- functions, limits, 178
- functions, odd, 178
- functions, piecewise, 181
- functions, trigonometric, 237, 291
- fundamental theorem of calculus, 310
- fundamental trig identity, 105
- fundamental trigonometric identities, 111
- fundamental units, 46

- G-Shortcut, 95
- Gabriel's horn, 331
- Gaussian distribution, 447
- GCF, 19
- general conic sections, 155
- general conics, conics general, 155
- general form, 56
- generalized basis vectors, 487
- generalized complex arithmetic, 229
- generalized kinematics, 161, 289
- generalized Taylor expansion, 294
- geometric cubic, 66
- geometric sequence, 186
- geometric Series, 93
- geometric series, 291
- geometry, analytic, 340
- geometry, integral, 329
- geometry, limits, 174
- geometry, Riemannian, 483, 507
- geometry, trigonometry, 122
- Gerolamo Cardano, 66

- golden ratio, 91, 184
- gradient, 488
- gradient, contravariant, 489
- gradient, covariant, 489
- graph, one-dimensional, 49
- graph, two-dimensional, 49
- graphical solutions, 51
- graphing, 49
- gravitation, universal, 534
- gravitational potential energy, 539
- gravity near Earth, 541
- gravity, freefall, 170
- gravity, self energy, 541
- greatest common factor, 19
- growth, exponential, 165

- half-angle formulas, trigonometric, 112
- harmonic law, 538
- harmonic oscillator, simple, 305
- Hermitian matrix, 260, 261
- hole in a bucket, 283
- hole through Earth problem, 542
- Hooke's law, 170
- horizon, distance below, 129
- horizon, distance to, 129
- horizontal range, 77
- horizontal tank problem, 336
- hunter, 79
- hyperbola, 144, 340
- hyperbolic antiderivatives, 304
- hyperbolic arctangent near zero, expansion, 296
- hyperbolic curve, 166
- hyperbolic functions, 238
- hyperbolic integrals, 321
- hyperbolic orbit, 543
- hyperbolic, Taylor expansion, 296

- identities, kinematics, 73
- identities, trigonometric, 111
- identities, vector, 212
- identity operator, 244
- identity, binomial coefficient sum, 80
- identity, binomial row sum, 81
- identity, reflection, 80
- imaginary axis, 231
- imaginary component, 225
- imaginary numbers, 225
- imaginary unit, 230
- imperial system, 46
- implicit differentiation, 277
- implicit vs. explicit, 681
- indefinite integral, 311
- independent events, 193
- independent random variables, 441
- index notation, 483
- index, lowering, 490
- index, raising, 490
- inequality, 14
- inequality, triangle, 12, 125
- infinite limit, 172
- infinite sequences, 185
- infinite series, 188
- infinite sum, 172
- infinite sum analysis, 99
- infinite sums, division, 84
- infinity, 11
- inflection, 286
- injective, 177
- inner product, 485
- inscribed rectangle, 283
- inscribed triangle, 122
- instantaneous basis vectors, 349
- instantaneous velocity, 71
- integer, dark, 165
- integers, 12
- integral, 310
- integral calculus, 307
- integral notation, 310
- integral operator, 311
- integral, arccosecant, 323
- integral, arcsecant, 323
- integral, arctangent, 326
- integral, geometry, 329
- integral, hyperbolic, 321
- integral, interval, 310
- integral, natural logarithm, 317
- integral, parametric, 343
- integral, series expansion, 324
- integral, shifted natural logarithm, 317, 325
- integral, $\sin(x \cdot x)$, 326
- integral, symmetric domain, 311
- integral, trigonometric, 318
- integration by parts, 316
- integration constant, 311
- integration interval, 310
- integration limits, 310
- integration, applied, 333
- integration, chain rule, 314
- integration, constant shift, 315
- integration, $d(\sin)$ shortcut, 315

- integration, exponential, 313
- integration, exponents, 313
- integration, label trick, 317
- integration, mirror trick, 324
- integration, natural exponential, 314
- integration, reciprocal, 313
- integration, roots, 313, 316
- integration, techniques, 313
- integration, u-substitution, 314
- intercepts, 53, 56
- interior identities, ellipse, 139
- interior identities, hyperbola, 146
- internal relations, ellipse, 138
- internal relations, hyperbola, 145
- internal relations, parabola, 150
- interpolation, 116
- intersection, circles, 359
- intersection, ellipses, 360
- intersection, line segments, 358
- intersection, line, ellipse, 358
- intersection, line, parabola, 57
- intersection, parabola, ellipse, 361
- intersection, planes, 354
- intersection, point, triangle, 361
- intersection, points, circle, 360
- intersections, 358
- interval, 31
- interval, differential, 346, 489
- introduction to vectors, 205
- invariance, 486
- inverse cube attraction, 540
- inverse functions, 177
- inverse matrix, 252
- inverse reciprocal identities, trigonometric, 113
- inverse rotation matrix, 220
- inverse square derivative, 266
- inverse triangle identities, 113
- inverse trigonometric antiderivatives, 302
- inverse trigonometric derivatives, 278
- inverse trigonometric notation, 113
- inverse trigonometry, 113
- inverse, complex, 240
- inverse-square acceleration, 533
- inverted Pythagorean theorem, 104, 110
- involute, 133, 340
- irrational numbers, 12
- isosceles triangle, 103
- iterative methods, 649
- jerk, nonuniform, 158
- jerk, uniform, 158, 289
- Kepler's first law, 538
- Kepler's laws, 537
- Kepler's third law, 538
- kinematic identities, 73
- kinematics, 70, 289, 333, 340
- kinematics, air damping, 296
- kinematics, circle, 135
- kinematics, generalized, 161, 289
- kinematics, time-shifted, 158, 159
- kinetic energy, 170, 539
- Kronecker delta, 486
- label trick, integration, 317
- ladder touching cube problem, 64
- Langley's adventitious angles, 125
- laptop shop problem, 196
- large numbers, 41
- large-n systems, 440
- law of cosines, 122, 134, 210
- law of ellipses, 538
- law of equal areas, 538
- law of sines, 123
- laws of motion, 169
- LCM, 20
- least common multiple, 20
- left sum, 308
- Leibniz formula, 326
- length, 46
- light, wave, 590
- lightyears, 43
- like terms, 26
- limit, differential, 171
- limit, infinite, 172
- limits, functions, 178
- limits, geometry, 174
- limits, integral, 310
- limits, nonexistent, 185
- limits, vectors, 221
- limts, 171
- line, 53
- line element, cylindrical, 355
- line element, differential, 346, 487
- line element, dual, 487
- line intersecting ellipse, 358
- line intersecting parabola, 57
- line segments, intersection, 358
- line, horizontal, 53
- line, normal, 334

- line, number, 11
- line, perpendicular, 55
- line, tangent, 334
- line, vertical, 53
- linear combinations, 215, 219
- linear equation, 53
- linear motion, 100
- linear system, 30, 257, 484
- linear systems, 243
- linearity, 30
- lines, 132, 209
- lines, field, 380
- lines, perpendicular, 209
- Lissajous Curves, 136
- locating ships, 147
- logarithm operator, 23
- logarithm trick, 276
- logarithm, addition, 24
- logarithm, base, 23
- logarithm, base-ten, 23
- logarithm, change of base, 24
- logarithm, equals one, 23
- logarithm, inverse, 23
- logarithm, natural, 25, 166, 182
- logarithm, of one, 23
- logarithm, product, 24
- logarithm, subtraction, 24
- logarithm, utilization, 27
- logarithm, zero, 23
- logarithmic antiderivatives, 300
- logarithmic derivatives, 269
- logarithmic integral, tricky, 317
- logarithmic spiral, 342
- long division, 82, 83
- longitudinal vibrations, 337
- lottery game problem, 200
- lower triangular form, matrix, 255
- lowering index, 490
- Lucas numbers, 90
- Lucas sequence, 90
- Lucas-Fibonacci numbers, generalized, 92
- L'Hopital's rule, 280

- magnitude, 41, 306
- magnitude, complex, 228
- magnitude, vector, 206, 210
- manifold, 487
- mass, 46, 169
- mass between springs problem, 336
- mass units, 47
- mass units, elemental, 47
- mass versus weight, 38
- mass, center, 532
- mass, reduced, 532
- matrices, complex numbers, 232
- matrix addition, 246
- matrix associativity, 247
- matrix components, 244
- matrix diagonal form, 255
- matrix diagonalization, 260
- matrix formalism, 244
- matrix lower triangular form, 255
- matrix minor, 251
- matrix multiplication, 246
- matrix multiplication formula, 246
- matrix non-commutativity, 246
- matrix operations, 246
- matrix representation, basis vectors, 357
- matrix RREF, 256
- matrix scalar multiplication, 246
- matrix transpose, 255
- matrix triangular forms, 255
- matrix upper triangular form, 255
- matrix, anti-symmetric, 255
- matrix, augmented, 255
- matrix, change of basis, 247
- matrix, diagonal, 261
- matrix, Hermitian, 260, 261
- matrix, inverse, 252
- matrix, modal, 260
- matrix, orthogonal, 255
- matrix, symmetric, 255, 260
- matrix-operator equivalence, 244
- maximal range, 78
- mean, 202
- mean value theorem, 280
- mechanical equilibrium, 170
- median, 202
- melting ice sheet problem, 284
- methods, iterative, 649
- methods, Newton's, 664
- methods, numerical, 649
- metric tensor, 487
- metric tensor symmetry, 488
- metric, Minkowski, 492
- midpoint sum, 308
- Minkowski metric, 492
- Minkowski space, 590
- mirror trick, integration, 324
- misguided missile problem, 347

- missing face problem, 195
mixed differentiation techniques, 278
mixed Euler's method, 670
modal matrix, 260
mode, 202
mode, binomial distribution, 445
modified natural logarithm antiderivative, 301
modified natural logarithm derivative, 270
momentum, angular, 533
monotonic sequence, 187
monkey, 79
motion, circular, 135
motion, elliptic, 348
motion, freefall, 76
motion, laws, 169
motion, linear, 100
motion, one dimension, 70
motion, two-dimensional, 76
multi-state system, 444
multiplication, associative property, 16
multiplication, commutative property, 16
multiplication, complex, 227, 237
multiplication, determinant, 251
multiplication, distributive property, 16
multiplication, fractions, 17
multiplication, implied, 14
multiplication, matrix, 246
multiplication, rates, 35
multiplication, scalar, 209, 227, 341
multiplication, utilization, 27
multiplicity, 443
mutually exclusive events, 192

n-sided polygon, 224
NASA, 46
natural logarithm integral, 317
natural exponential derivative, 267
natural exponential, integration, 314
natural logarithm, 25, 166, 182
natural logarithm antiderivative, 299
natural logarithm derivative, 269
natural logarithm, complex, 239, 240
natural numbers, 12
negative angle, 110
nested expressions, 13
nested radicals, pi, 221
net displacement, 72
netball problem, 446
Newton's law, second, 531
Newton's law, third, 531

Newton's method, 664
Newton's method, second order, 666
Newton's second law, 169
non-Cartesian coordinates, 355
non-commutativity, matrix, 246
non-exclusivity, 193
nonexistent limits, 185
nonlinear natural logarithm antiderivative, 300
nonlinear natural logarithm derivative, 269
nonlinear system, 30
norm, tensor, 485
normal line, 334
normal line, ellipse, 142
normal line, hyperbola, 148
normal line, parabola, 150
normal vector, 348, 351, 361
normal vector, plane, 354
normalization, probability, 192
normalization, statistical, 202
normalizing, quadratic, 60
notation, determinant, 211
notation, exponent, 22
notation, index, 483
notation, integral, 310
notation, inverse trigonometric, 113
notation, scientific, 41
notation, Sigma, 42, 87, 172, 188
notation, sigma, 81
notation, variable, 26
nth root, complex, 239
number line, 11, 95
number, Avogadro's, 47
numbers, 11
numbers, complex, 225, 227
numbers, Fibonacci, 91
numbers, imaginary, 225
numbers, irrational, 12
numbers, large, 41
numbers, Lucas, 90
numbers, natural, 12
numbers, prime, 20
numbers, rational, 12
numbers, real, 11, 225
numbers, small, 171
numbers, special, 15
numbers, whole, 12
numerical methods, 649
numerical precision, 44

obtuse triangle, 103

- odd functions, 178
- offset circle, 134
- one, 15
- one-sided, 173
- opening direction, parabola, 149
- operations, algebraic, 16
- operations, matrix, 246
- operator, identity, 244
- operator, integral, 311
- operator, logarithm, 23
- operator-matrix equivalence, 244
- operators, 13
- operators, arithmetic, 13
- operators, complex, 230
- operators, order, 13
- operators, special, 14
- operators, unary, 14
- optimization problems, 282
- orbit, 542
- orbit, bound, 543
- orbit, circular, 543
- orbit, elliptical, 543
- orbit, hyperbolic, 543
- orbit, parabolic, 542
- orbit, unbound, 543
- order of approximation, 291
- order of magnitude, 42
- order of operations, 13
- ordered pairs, 50
- origin, 52
- orthogonal matrix, 255
- orthogonality, basis vectors, 214, 356
- orthogonality, trigonometry, 321
- orthogonality, vectors, 210
- oscillations, small, 170
- oscillator, simple harmonic, 305
- outer product, 484

- paper, folding, 43
- parabola, 149
- parabola intersecting ellipse, 361
- parabola, derivative, 265
- parabolic expressions, 149
- parabolic orbit, 542
- parallel vectors, 209
- parallelogram, area, 212
- parameter, affine, 487
- parameterized circle, 108
- parametric equations, 340
- parametric arc length, 343
- parametric derivatives, 342
- parametric integrals, 343
- parametric representation, ellipse, 139
- parametric representation, hyperbola, 146
- parametric second derivative, 342
- parametric surface of revolution, 343
- partial fractions, 85
- partial sum, 188
- parts, integration, 316
- Pascal triangle checksum, 98
- Pascal's triangle, 80
- Pascal's triangle alternating, 81
- Pascal's triangle, alternating, 97
- Pascal's triangle, geometric series, 97
- PEMDAS, 13
- perigee, 536
- periodic curve, 167
- periodic sequence, 185
- periodicity, 110, 176
- permutations, 199
- perpendicular lines, 209
- phase, 110, 306
- phase shift π , 110
- phase shift $\pi/2$, 110
- phase, complex, 230
- physical pendulum, 324
- Pi from nested radicals, 221
- π vs. e , 283
- π , Taylor expansion, 665
- piecewise functions, 181
- pipe, 39
- plane curve analysis, 348
- plane from three points, 362
- plane, complex, 230
- plane, equation, 353
- plane, normal vector, 354
- plane, vector representation, 354
- planes, 353
- planes intersecting, 354
- planetary motion, 531
- plot, position, 70, 73
- plot, velocity, 71
- plots, trigonometry, 118
- point intersecting triangle, 361
- point-slope form, 54
- points intersecting circle, 360
- Poisson distribution, 447
- polar area element, 347
- polar basis vectors, 216
- polar coordinate system, 132, 216, 341

- polar coordinates, 485
- polar coordinates arc length, 344
- polar coordinates, acceleration, 345
- polar coordinates, area integral, 343
- polar coordinates, curvature, 350
- polar coordinates, slope, 342
- polar coordinates, velocity, 345
- polar coordinates, metric, 491
- polar form, complex, 237
- polar position vector, 216
- polar representation, conics, 155
- polar representation, ellipse, 139
- polar representation, hyperbola, 146
- polar representation, vector, 216
- polygon, n-sided, 224
- polynomial division, 82, 95
- polynomial division, applied, 89
- polynomial, area, 162
- polynomial, characteristic, 259
- polynomial, slope, 162
- polynomial, Taylor, 158, 161, 290
- position plot, 70, 73
- position vector, 206
- position vector, Cartesian coordinates, 214
- position vector, cylindrical coordinates, 355
- position vector, polar coordinates, 216
- position vector, spherical coordinates, 356
- potential energy, 169
- potential energy, gravity, 539
- potential, power law, 550
- power law potential, 550
- power rule, derivative, 276
- power, whole number, 266
- powers, roots, antiderivative, 300
- precision, 41
- precision, numerical, 44
- precision, propagation, 45
- pressure, 39
- prime factorization, 20
- prime numbers, 20
- probability, 192
- probability distribution function, 440
- probability distribution, normalized, 443
- probability statistical, 192
- probability theory, 191
- probability, classical, 192
- probability, conditional, 194
- probability, normalization, 192
- problem, circular inscription, 107
- product formulas, trigonometric, 111
- product rule, derivative, 273
- product, independent random variables, 441
- product, inner, 485
- product, outer, 484
- product, scalar, 486
- product, tensor, 484
- products, vector, 210
- projector, 244
- properties of one, 15
- properties of zero, 15
- property, associative, 16
- property, commutative, 16
- Pythagorean theorem, 103
- Pythagorean theorem, inverted, 104, 110
- Pythagorean triplets, 106
- quadrants, 52
- quadratic curves, 56
- quadratic equation, 56
- quadratic equation, alternate, 57
- quadratic formula, 56, 61
- quadratic, normalizing, 60
- quotient rule, derivative, 274
- radial coordinate, 132
- radians, 102
- radicals, 23
- radicals, negative, 68
- radioactive decay, 195
- radius, complex, 230
- rail problem, 665
- raising index, 490
- random product problem, 448
- random sums problem, 449
- random variables, 441
- random variables, independent, 441
- random variables, product, 441
- random variables, sum, 441
- range, function, 176
- range, horizontal, 77
- range, maximal, 78
- rate, 31
- ratio, golden, 184
- rational numbers, 12
- ratios, 17
- real axis, 231
- real numbers, 11, 225
- reciprocal, 17
- reciprocal antiderivative, 300
- reciprocal binomial theorem, 98

reciprocal, derivative, 266
 reciprocal, integration, 313
 rectangle, inscribed, 283
 recursion, 90, 92
 recursion relations, 89
 recursive sequence, 185
 recursive sequences, 89
 reduced mass, 532
 reduction formulas, trigonometric, 320
 reflection identity, 80
 reflection property, ellipse, 141
 reflection property, hyperbola, 147
 reflection property, parabola, 151
 related rate problems, 283
 relative displacement, 532
 relativity, special, 590
 remainder, 83
 remainder, Taylor series, 332
 removable singularity, 181
 repeating decimals, 99
 revolution, surface, 331
 revolution, volume, 329
 Riemann sums, 308
 Riemann surface, 240
 Riemannian geometry, 507
 rifle problems, 79
 rifles, identical, 79
 right 3-4-5 triangle, 106
 right 30-60-90 triangle, 106
 right 5-12-13 triangle, 106
 right focal chord, 150
 right hand rule, 211
 right isosceles triangle, 105
 right $\pi/16$ triangle, 106
 right $\pi/8$ triangle, 105
 right sum, 308
 right triangles, special, 105
 right triangle, area, 103
 right triangles, 103
 root, cube, 665, 667
 root, integration, 316
 root, square, 666
 root, utilization, 27
 roots, 23
 roots, complex, 239
 roots, powers, antiderivative, 300
 rotated coordinates, 155, 220
 rotated vector, 220
 rotation, 133
 rotation matrix, inverse, 220

rotation operator, complex, 230
 rotations, 220
 rotations, complex, 236
 row operations, 249, 251, 255
 row sum identity, binomial, 81
 royal flush, 200
 RREF, matrix, 256
 rules of differentiation, 345
 Runge-Kutta techniques, 675
 Runge-Lorenz vector, 535

 scalar multiplication, 209, 227, 341
 scalar multiplication, matrix, 246
 scalar product, 486
 scalar, associativity, 209
 scale, 133
 scalene triangle, 103
 scientific notation, 41
 secant, 109, 118
 secant antiderivative, 302
 secant derivative, 271
 secant substitution, integration, 322
 sech derivative, 279
 second derivative, 285, 293
 second derivative, parametric, 342
 second-order Newton's method, 666
 seed, 92
 self energy, gravity, 541
 semilatus rectum, 138
 sequence, 185
 sequence, convergent, 185
 sequence, divergent, 185
 sequence, Fibonacci, 91
 sequence, geometric, 186
 sequence, Lucas, 90
 sequence, monotonic, 187
 sequence, periodic, 185
 sequence, recursive, 89, 185
 sequences, infinite, 185
 Series, 188
 series analysis, 332
 series expansion, integration, 324
 series, finite, 188
 series, geometric, 93
 series, infinite, 188
 Serret-Frenet formulas, 352
 shell theorem, 531, 544
 shift of index, 190
 shifted natural logarithm antiderivative, 300
 shifted natural logarithm derivative, 269

- shifted natural logarithm integral, 317, 325
shifted natural logarithm, Taylor expansion, 294
SHO, 305, 673
SI System, 46
sidedness, 173
Sigma notation, 42, 87, 172, 188
sigma notation, 81
significant figures, 44
simple harmonic oscillator, 305
sinc function, 182
sine, 104, 118, 168, 272
sine derivative, 270
sine substitution, integration, 322
sines superposition, 112
singularity, essential, 181
singularity, removable, 181
sinh, 166
sinh derivative, 279
slicing the cone, 153
slope, 54, 70
slope at a point, 265
slope of slope, 285
slope, polar coordinates, 342
slope, polynomial, 162
slope-intercept form, 54
small numbers, 171
small oscillations, 170
small vibrations, 337
small-angle approximation, 116, 272
smoothness, 181
snap, uniform, 159
SohCahToa, 105
solid sphere, 540
solving for a variable, 28
space, curved, 483
space, flat, 483
spanning, vector space, 215
special numbers, 15
special operators, 14
special relativity, 590
special right triangles, 105
speed, 71
sphere, visibility, 128
spheres, collision, 530
spherical area element, 357
spherical basis vectors, 356
spherical coordinate system, 356
spherical position vector, 356
spherical velocity vector, 357
spherical volume element, 357
spiral, logarithmic, 342
split-term method, 60
square disk method, 330
square root, 666
square root derivative, 266
square root Taylor expansion, 293
square root, complex, 239–241
squared argument, 271
squaring the geometric series, 97
squeeze theorem, 183, 186
stability, critical point, 285
standard deviation, 202
standard deviation, binomial distribution, 445
standard exponential derivative, 267
state, 191
statistical average, 202
statistical probability, 192
Stirling's approximation, 326
stochastic process, 191
strange beast, 183
strange function, 327
strange product, 327
strategy, 28, 44
stress, thermal, 556
sub-matrix, 251
subscripts, component, 205
subtraction, fractions, 18
subtraction, utilization, 27
subtraction, vector, 207
sum, infinite, 172
sum, partial, 188
sum, random variables, 441
summation, Einstein convention, 484
sums, finite, 189
superposition relationships, trigonometric, 112
superposition, cosines, 112
superposition, sines, 112
surface of revolution, 331
surface of revolution, parametric, 343
surjective, 177
symmetric integrationno domain, 311
symmetric matrix, 255, 260
symmetry, ellipse, 138
symmetry, functions, 177
symmetry, hyperbola, 145
symmetry, metric tensor, 488
symmetry, tensor, 491
system, continuous, 440
system, degenerate, 262
system, large-n, 440

- system, linear, 484
- system, multi-state, 444
- system, two-state, 443
- systems of equations, 29

- tables, trigonometry, 115
- tangent, 104, 118
- tangent antiderivative, 301
- tangent derivative, 271
- tangent line, 280, 334
- tangent line, circle, 108
- tangent line, ellipse, 140, 277
- tangent line, hyperbola, 147
- tangent line, parabola, 150
- tangent near $\pi/2$, expansion, 296
- tangent near $\pi/4$, Taylor expansion, 295
- tangent near zero, Taylor expansion, 295
- tangent substitution, integration, 322
- tangent vector, 348, 351
- tanh derivative, 279
- tapered candle problem, 284
- tautochrone, 335
- taxonomy of circles, 107
- Taylor expansion, arctangent at two, 294
- Taylor expansion, arctangent near one, 294
- Taylor expansion, arctangent near two, 294
- Taylor expansion, arctangent near zero, 294
- Taylor expansion, asymptotes, 296
- Taylor expansion, cotangent near $\pi/2$, 296
- Taylor expansion, cotangent near zero, 296
- Taylor expansion, generalized, 294
- Taylor expansion, hyperbolic, 296
- Taylor expansion, hyperbolic arctangent near zero, 296
- Taylor expansion, shifted natural logarithm, 294
- Taylor expansion, square root, 293
- Taylor expansion, tangent near $\pi/2$, 296
- Taylor expansion, tangent near $\pi/4$, 295
- Taylor expansion, tangent near zero, 295
- Taylor polynomial, 158, 161, 290
- Taylor series, 332
- Taylor series, remainder, 332
- Taylor's theorem, 289, 290
- Taylor's theorem, differentiation, 292
- Taylor's theorem, proof, 290
- Taylor's theorem, testing, 291
- techniques, Runge-Kutta, 675
- tensor norm, 485
- tensor product, 484
- tensor symmetry, 491
- tensor type, 483
- tensor type continuity, 484
- tensor, contravariant, 483
- tensor, covariant, 483
- tensor, definition, 485
- tensor, metric, 487
- tensors, 483
- testing Taylor's theorem, 291
- Thales' theorem, 122, 212
- theorem, Bayes', 195
- theorem, binomial, 80
- theorem, calculus, fundamental, 310
- theorem, mean value, 280
- theorem, Pythagorean, 103
- theorem, reciprocal binomial, 98
- theorem, shell, 531, 544
- theorem, squeeze, 183, 186
- theorem, Taylor's, 289, 290
- theorem, Thales', 122, 212
- thermal stress, 556
- theta convention, 107
- three points, plane, 362
- three-dimensional curves, 351
- time, 47
- time-shift analysis, 72, 158, 289
- time-shifted kinematics, 158, 159, 289
- TNB frame, 352
- torsion, 351
- trajectory, conic, 537
- trajectory, envelope, 77
- transcendental equations, 664
- transpose matrix, 255
- transpose rule, determinant, 251
- transverse vibrations, 337
- trapezoid rule, 309
- tree method, 20
- triangle inequality, 12, 125
- triangle, acute, 103
- triangle, area, 103, 212
- triangle, equilateral, 103
- triangle, inscribed, 122
- triangle, isosceles, 103
- triangle, obtuse, 103
- triangle, Pascal, 80
- triangle, right, 103
- triangle, right 3-4-5, 106
- triangle, right 30-60-90, 106
- triangle, right 5-12-13, 106
- triangle, right isosceles, 105
- triangle, right $\pi/16$, 106
- triangle, right $\pi/8$, 105

- triangle, scalene, 103
- triangles, 103
- triangular form, matrix, 255
- trig identity, fundamental, 105
- trigonometric antiderivatives, 301
- trigonometric derivatives, 270
- trigonometric functions, 237, 291
- trigonometric identities, 111
- trigonometric identities, fundamental, 111
- trigonometric integrals, 318
- trigonometric inverse reciprocal identities, 113
- trigonometric reduction formulas, 320
- trigonometric substitution, 303
- trigonometric substitution, integration, 321
- trigonometry, 102
- trigonometry plots, 118
- trigonometry tables, 115
- trigonometry, geometry, 122
- trigonometry, inverse, 113, 119
- trigonometry, polynomials, 116
- trinomial, 17
- triple product, 212
- triplets, Pythagorean, 106
- two balls and a building, 74
- two-body problem, 532
- two-sided, 174
- two-state system, 443
- type, tensor, 483
- type, tensor continuity, 484

- u-substitution, integration, 314
- unary operators, 14
- unbound orbit, 543
- uniform acceleration, 71
- uniform circular motion, 135, 346
- uniform jerk, 158, 289
- uniform snap, 159
- uniform velocity, 71
- uniqueness, 176
- unit circle, 108
- unit circle, area, 224, 317
- unit conversion, 35
- unit rate, 35
- unit vectors, 214
- unit, imaginary, 230
- units, 31
- units, area, 36
- units, fundamental, 46
- units, mass, 47
- units, mixed, 32

- units, physical, 46
- units, quantum mechanics, 47
- units, standard, 46
- units, volume, 36
- unity condition, 219
- universal gravitation, 534
- upper triangular form, matrix, 255

- variable assignment, 26
- variable density, 541
- variables, 26
- variables, dimensionless, 551
- variables, random, 441
- variables, random independent, 441
- variables, random sum, 441
- variance, 442
- vector addition, 207
- vector anti-commutativity, 211
- vector commutativity, 210
- vector components, 205
- vector direction, 206
- vector distributive property, 209, 210
- vector identities, 212
- vector magnitude, 206, 210
- vector orthogonality, 210
- vector polar representation, 216
- vector products, 210
- vector representation, plane, 354
- vector rotation, 220
- vector subtraction, 207
- vector, acceleration, 349, 351
- vector, binormal, 351
- vector, normal, 348, 351, 361
- vector, position, 206
- vector, Runge-Lorenz, 535
- vector, tangent, 348, 351
- vector, velocity, 348
- vector, zero, 207
- vectors, 205
- vectors, algebraic properties, 209
- vectors, basis, 214
- vectors, complex numbers, 231
- vectors, generalized basis, 487
- vectors, limits, 221
- vectors, parallel, 209
- vectors, unit, 214
- velocity, 70, 161, 221, 345
- velocity envelope, 333
- velocity plot, 71
- velocity vector, 348

- velocity vector, spherical, 357
- velocity, average, 72
- velocity, circle, 135
- velocity, escape, 542
- velocity, instantaneous, 71
- velocity, polar coordinates, 345
- velocity, uniform, 71
- vertex, 56, 76
- vibrations, critical, 338
- vibrations, longitudinal, 337
- vibrations, small, 337
- vibrations, transverse, 337
- viking problem, 446
- visibility on sphere, 128
- volume element, Cartesian, 353
- volume element, cylindrical, 356
- volume element, spherical, 357
- volume of revolution, 329
- volume, cylinder, 39
- washer method, 330
- water, 38
- water, density, 38
- wave, light, 590
- weight versus mass, 38
- whole number powers, 266
- whole numbers, 12
- wishing well problem, 74
- work, 31
- x equals $\cos(x)$, 664
- x^x , derivative, 268
- Zeno's paradox, 100
- zero, 15
- zero vector, 207