Offline Mathematics
MANUSCRIPT

William F. Barnes
[1]

April 12, 2024

Contents

Part I

Algebra

Chapter 1

Numbers and Arithmetic

1 Numbers

Numbers are symbols used for counting, measuring, or labeling a certain quantity. Each number is unique, meaning no two numbers represent the same value.

1.1 Number Line

When two numbers are compared, one number must have a greater value, and the other will have the lesser value. It follows that all numbers can be arranged on a number line, from lesser (left) to greater (right) as shown in Figure 1.1.

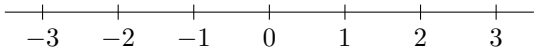


Figure 1.1: Number line (partial).

The number line is divided into two regions, positive and negative, with the special number 0 being neither of these. It extends infinitely in both directions, to $+\infty$ on the right and $-\infty$ on the left, and the whole number line is *continuous*.

Continuity

On the number line,'continuous' means that every conceivable decimal and fraction occurs somewhere on the line. This implies if we choose some number A and add or subtract from A to 'move over' to number B, we do so by strolling over all possible numbers between the two.

1.2 Real Numbers

The continuous set of all numbers on the number line is called the *real numbers*, and has the symbol $\mathbb{R}$. The values ∞ and $-\infty$ are not considered real numbers, nonetheless we can loosely capture all this by writing

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

Infinity

Briefly, one reason that infinity is not considered a real number is for the lack of compatibility with ordinary arithmetic. Questions like 'what is infinity plus one?' are answered by 'infinity again':

$$\begin{aligned}\infty + 1 &= \infty\\ \infty + \infty &= \infty\\ \infty \times \infty &= \infty\end{aligned}$$

While things like the above may be true in a certain intuitive way, these are not really algebraic statements. Carrying away one example on hand, one might want to conclude

$$1 = \infty - \infty = 0$$

and claim to have broken mathematics. This is ultimately nothing but abuse of notation.

Rational vs. Irrational

Real numbers can be divided into two categories, *rational numbers* and *irrational numbers*. A rational number, coming from the word 'ratio', is a number that can be expressed as a non-repeating decimal. Rational numbers are things like 7, 0.25, or 33/3.

On the other hand, numbers like $1/3 = 0.333\ldots$ that require an infinite trail of 3's after the decimal are irrational. Perhaps the most frequently used irrational numbers are π, e, and $\sqrt{2}$, but it's straightforward to reason that there are *many* more irrational numbers than there are rational numbers.

Set Notation

The symbol assigned to rational numbers is $\mathbb{Q}$, and the symbol assigned to irrational numbers is $\mathbb{Q}'$. Each of these qualifies as a *subset* of the real numbers, and to denote this we write:

$$\begin{aligned}\{\text{rational numbers}\} &= \mathbb{Q} \subset \mathbb{R}\\ \{\text{irrational numbers}\} &= \mathbb{Q}' \subset \mathbb{R}\end{aligned}$$

The *union* of $\mathbb{Q}$ and $\mathbb{Q}'$ reconstitute the real numbers:

$$\mathbb{Q} \cup \mathbb{Q}' = \mathbb{R}$$

1.3 Integers

Rational numbers that have no decimal component belong to the *integers*, denoted $\mathbb{Z}$. The set of integers is still infinite in size but is certainly 'smaller' than $\mathbb{Q}$ and $\mathbb{R}$:

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

Whole and Natural

Next there are *whole numbers*, which is the subset of $\mathbb{Z}$ that excludes negatives but includes zero:

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

Finally, we have *natural numbers*, denoted $\mathbb{N}$, which simply shaves the zero from the set of whole numbers:

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

1.4 Prime Numbers

Prime numbers are whole numbers that cannot be divided into smaller whole 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 Decomposition

For a whole number N *prime decomposition* of N is defined as a list of prime numbers whose product is N. To find the prime decomposition of an odd number, try dividing by 2. The result can be one of three things:

- If $N/2$ is a prime number, stop.
- If $N/2 = M$ is a whole number, then 2 is a factor. Repeat using M.
- If $N/2$ is fraction or decimal, replace 2 by the next prime number (3).
- If $N/3$ is fraction or decimal, replace 3 by the next prime number (5).
- If $N/5$ is fraction or decimal, replace 5 by the next prime number (7), and so on.
- When the prime being tested exceeds $N/2$, stop.

Problem 1

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

$$12 = 2 \times 2 \times 3 = 2^2\, 3$$
$$231 = 3 \times 7 \times 11$$
$$150 = 2 \times 3 \times 5 \times 5 = 2\,(3)\,5^2$$
$$225 = 3 \times 3 \times 5 \times 5 = 3^2\, 5^2$$

2 Expressions

The *mathematical operators* are symbols that are situated among numbers. A valid *mathematical expression* usually has at least two numbers and one operation, such as:

$$1 + 2\,.$$

In the above, the 'plus' operator (+) combines the pair of 'input numbers' 1 and 2.

Equations

The expression $1 + 2$ is equivalent to the *sum* of the two input numbers, namely 3. The expression and its result can be written together as a *mathematical equation* with an 'equality' (=) symbol balancing each side:

$$1 + 2 = 3$$

Simplifying Expressions

Using operators to reduce the complexity of an expression (toward a number) without introducing error is called *simplifying*, or *evaluating* the expression. The 'simplified' expression must be essentially equal to the original, and is given the term 'equivalent'. By convention, the equivalent expression is placed on the right side of the equality symbol (=).

Nested Expressions

Bracketing symbols called *parentheses* ( ) are used to embed an expression within an expression. There must always be parenthetical balance in an equation, 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:

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

2.1 Operators

There are six mathematical operators for basic arithmetic. Listing in a particular order, these are:

Operator	Symbol	Result
Parentheses	(N)	N
Exponent	$\hat{}$	product
Multiplication	$\times$ or $\cdot$	product
Division	$\div$ or $/$	ratio or quotient
Addition	$+$	sum
Subtraction	$-$	difference

At the top of the list is the set of *parentheses* ( ), which tell us to ignore everything else and solve whatever is inside the parentheses first. Next on the list is the exponent ($\hat{}$) operator, followed by multiplication ($\times$ or $\cdot$), and so on down to subtraction ($-$).

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 **P**lease **E**xcuse **M**y **D**ear **A**unt **S**ally. This means to look for parentheses first, then exponents, multiplication, division, addition, and subtraction follow in order.

The '(P)EMDAS' operators take two or more numbers as input and return one number as output.

Precedence and Binding

For some more terminology, it is said that 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*. In practice, one could read that the multiplication operator binds more tightly (to a number) than an addition operator.

Problem 2

Use the order of operations to simplify

$$4 + 5 \times 9$$

and make sure the following gives a different answer:

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

2.2 Special Operators

There is a class of special operators that take one number instead of two, 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 shows up 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:

$$20! = (5 + 4)! = 2432902008176640000$$

Floor and Ceiling

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

Implied Multiplication

An abuse of notation that students of mathematics quickly accept is the omission of any symbol for multiplication. When a number 'butts 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:

$$6 \div (2 \times (2+1)) = 1$$
$$(6 \div 2) \times (2+1) = 9$$

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, we could play games like this forever.

When dealing with equations in a serious way, it's best to disambiguate as much possible, which means to make liberal use of parentheses if any expression risks misinterpretation by human or machine.

2.3 Inequality

We've seen that the equality (=) sign is the 'balance' between two equivalent expressions. When two expressions are not equivalent, or conditionally equivalent, there are special symbols to denote the nature of imbalance. 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	$4 \geq \lvert -4 \rvert$

2.4 Zero and One

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:
$$N \times 0 = 0$$
- Division by zero produces no useful information[1]:
$$\frac{N}{0} = \text{Undefined}$$
- 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:
$$N \times 1 = \frac{N}{1} = N$$
- Raising one to any real power results in one:
$$1 \times 1 \times 1 \times 1 \times \cdots = 1^0 = 1^{-1.75} = 1$$
- Fractions with the same numerator and denominator are equivalent to one:
$$\frac{N}{N} = 1$$
- The infinite repeating decimal $0.999\overline{9}$ is equivalent to one:
$$0.999\overline{9} = 1$$
(Proven later with geometric series.)

3 Using Operators

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

[1] Unfortunately, the number of instructors who have been caught conveying the grave error $N/0 = 0$ is itself nonzero.

3.1 Properties of Addition

Commutative Property

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:

$$N = A + B$$
$$N = B + A$$

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

Associative Property

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:

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

3.2 Properties of Multiplication

Commutative Property

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:

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

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

Associative Property

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. That is:

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

...

Chapter 2

Quadratics and Cubics

1 Quadratic Expressions

Factoring quadratic expressions is among the most tortured exercises in mathematics education. It is often the student's first encounter with something presented as not *purely* formulaic, which is to mean factoring can still go wrong - to not turn out an answer - without making a technical mistake. It's the student's first real brush with the algebraic abyss.

Un-Foiling an Equation

It's worthwhile to pause on why factoring is difficult in the first place. Factoring is introduced to the student after topics like 'distribution' and polynomial multiplication are thoroughly gained, so statements such as

$$(2x+3)(4x-1) = 8x^2 + 10x - 3$$

are easily understood reading left-to-right.

On the left are *four* terms grouped by signs and parentheses, but on the right, after carrying out a swift FOIL operation, are *three* terms. Despite the algebraic balance of the equation above, there's still a sense that some information is lost in going left to right.

Factoring, of course, is the reverse job of polynomial multiplication. For the example on hand, we must begin with three pieces of information, namely $8x^2$, $10x$, -3, and use this to generate the whole cluster to the left of the $=$ sign.

As if un-scrambling eggs, it's not obvious how to un-multiply a polynomial into separate products. This topic is therefore not only difficult to learn, but inevitably difficult to teach. The student-teacher relationship won't be comparably strained until integral calculus.

1.1 Quadratic Formula

Factoring a quadratic expression $Ax^2 + Bx + C$ is equivalent to seeking the x-intercepts in the graph

$$y = Ax^2 + Bx + C\,.$$

By setting $y = 0$, we can find two locations, each represented by the same letter x (confusingly enough), where the graph touches the x-axis:

$$0 = Ax^2 + Bx + C \tag{2.1}$$

$$0 = A(x-p)^2 + Aq \tag{2.2}$$

Introducing two new variables p, q as written, the problem translates to solving for these in terms of A, B, and C. Doing this exercise, one finds

$$p = \frac{-B}{2A} \tag{2.3}$$

$$q = \frac{C}{A} - \frac{B^2}{4A^2}\,, \tag{2.4}$$

so now x may be written

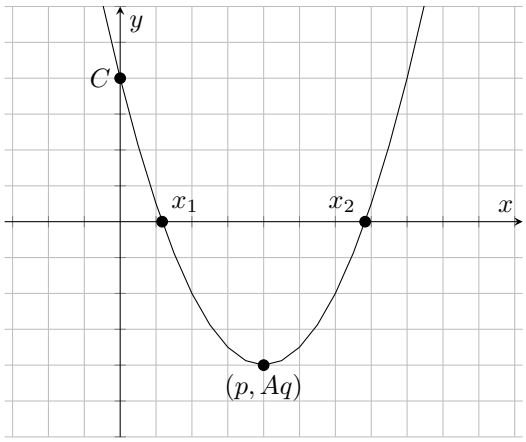
$$x = p \pm \sqrt{-q}\,. \tag{2.5}$$

Replacing p, q via Equations (2.3), (2.4) gives the *quadratic formula*:

$$x = \frac{-B}{2A} \pm \frac{\sqrt{B^2 - 4AC}}{2A} \tag{2.6}$$

Due to the plus-or-minus ($\pm$) symbol, there are two solutions to the quadratic formula, x_1, x_2. The factored answer can be written:

$$y = A(x - x_1)(x - x_2)$$



Factoring quadratics could begin and end at the quadratic formula if mathematics were taught a certain way. By understanding the above derivation,

the student hits all of the essential points required for factoring: solutions as x-intercepts, completing the square, substitution, and the quadratic formula.

Problem 1

Derive Equations (2.3), (2.4), (2.5) from Equations (2.1), (2.2).

Product of Quadratic Solutions

A seldom-utilized identity that pertains to factoring quardatic expressions relates the product of two the soltuions x_1, x_2. Carrying this out, one finds

$$x_1 \cdot x_2 = \left(\frac{-B}{2A} + \frac{\sqrt{B^2 - 4AC}}{2A}\right)\left(\frac{-B}{2A} - \frac{\sqrt{B^2 - 4AC}}{2A}\right) ,$$

simplifying to:

$$x_1 \cdot x_2 = \frac{C}{A}$$

2 Factoring Techniques

There are at least half a dozen 'common' methods for factoring quadratic expressions, some more useful than others in certain regimes. Following is a brief survey of some of these.

2.1 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 factor 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.

Problem 2

Use the box method to show:

$$6x^2 + 11x + 4 = (3x + 4)(2x + 1)$$

Hint:

$3x$	4	
$6x^2$	$8x$	$2x$
$3x$	4	1

Problem 3

Use the box method to show:

$$3x^2 - 2x - 8 = (x - 2)(3x + 4)$$

Hint:

x	-2	
$3x^2$	$-6x$	$3x$
$4x$	-8	4

2.2 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.

2.3 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

$$x^2 + 6x + 2 = \left(x + 3 - \sqrt{7}\right)\left(x + 3 + \sqrt{7}\right)\ .$$

Problem 4

Derive the quadratic formula by completing the square on:

$$ax^2 + bx + c = 0$$

Hint:

$$\left(\sqrt{a}x + \frac{b}{2\sqrt{a}}\right)^2 = \frac{b^2}{4a} - c$$

2.4 Algebraic Identities in Factoring

The limit case of a lucky factoring problem is one that maps perfectly onto a known algebraic identity. Committing identities to memory such as

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

quick work can be made of problems that fit them:

$$\begin{aligned} x^2 + 10x + 25 &= (x+5)^2 \\ x^2 - 6x + 9 &= (x-3)^2 \\ 9x^2 - 16 &= (3x+4)\,(3x-4) \end{aligned}$$

2.5 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.

2.6 Special Quadratic Coefficients

Unit Leading Coefficient

In the special case that $A = 1$, we have

$$x^2 + Bx + C = (x - x_1)\,(x - x_2)\ ,$$

implying

$$\begin{aligned} x_1 + x_2 &= -B \\ x_1 \cdot x_2 &= C\,, \end{aligned}$$

and furthermore

$$\begin{aligned} x_1 &= \frac{1}{2}\left(-B + \sqrt{B^2 - 4C}\right) \\ x_2 &= \frac{1}{2}\left(-B - \sqrt{B^2 - 4C}\right)\ . \end{aligned}$$

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 variable. Solving one equation for x solves the modified equation (should you encounter one) for z, and vice versa.

2.7 Discriminant

One can't encounter quadratic-anything without dealing with the *discriminant*:

$$\mathcal{D} = B^2 - 4AC$$

For solutions to a quadratic equation be real-valued, the discriminant must be positive. The special case $B^2 = 4A$ corresponds to there being one solution to the quadratic, i.e. $x_1 = x_2$. When the discriminant is negative, solutions are imaginary or complex.

3 Method of Transform

Now we develop a technique for factoring quadratic expressions that is not in the common teachings.

Starting with a general quadratic expression

$$Ax^2 + Bx + C \ ,$$

split the middle term B into the sum of two unknown variables f_1 and f_2:

$$Ax^2 + Bx + C = Ax^2 + f_1 x + f_2 x + C$$

Equivalently, we may recast the right side in factored form

$$Ax^2 + Bx + C = \frac{1}{C}(f_1 x + C)(f_2 x + C) \ ,$$

or also equivalently:

$$Ax^2 + Bx + C = \frac{1}{A}(Ax + f_1)(Ax + f_2)$$

3.1 Transform Kernel

For the above forms to remain consistent, the restriction on $f_{1,2}$ emerges:

$$f_1 + f_2 = B$$
$$f_1 \cdot f_2 = AC$$

This is the quantification of the unavoidable kernel - the 'hard part' of factoring - that requires us to dream up two numbers whose sum is B and whose product is AC.

3.2 Transformed Quadratic

As a system of two equations and two unknowns, the restriction on $f_{1,2}$ can be conjoined to a single equation which is itself quadratic

$$f^2 - Bf + AC = 0 \ ,$$

and the problem is now transformed into finding solutions for f.

Interestingly, this exposes a deeper quality of factoring problems, in that the nonlinear system of equations, i.e. the kernel, can be transposed into yet another quadratic equation - yet another factoring problem. Students are quietly asked to solve this the 'hard' way without knowing it.

All we've done is transform one quadratic problem into another, but notice now that the leading coefficient is now absorbed elsewhere. For this reason, it's often easy enough to 'eyeball' the solutions for f_1 and f_2. When this isn't the case, the standard gamut of factoring techniques are all the easier to apply to the f-equation.

3.3 Completing the Rectangle

The so-called method of transform can be regarded as 'completing the rectangle'. In terms of $f_{1,2}$, note we can generally write

$$Ax^2 + Bx + C = \left(\frac{f_1}{\sqrt{C}}x + \sqrt{C}\right)\left(\frac{f_2}{\sqrt{C}}x + \sqrt{C}\right) \ ,$$

or equivalently,

$$Ax^2 + Bx + C = \left(\sqrt{A}x + \frac{f_1}{\sqrt{A}}\right)\left(\sqrt{A}x + \frac{f_2}{\sqrt{A}}\right) \ ,$$

and so on.

$f_1 x$	C
Ax^2	$f_2 x$

Completing the rectangle is a way to frame a factored quadratic expression as the product of two unequal 'lengths'. No matter how this is done though, one inevitably confronts the kernel represented by the transformed quadratic.

3.4 Special Cases

Unit Leading Coefficient

In the special case that $A = 1$, the method of transform returns something *almost* tautological in the sense that

$$x^2 + Bx + C$$

implies

$$f^2 - Bf + C = 0\,.$$

The left-hand expressions are the same up to a minus sign on the active variable, so we identify

$$f = -x\,.$$

That is, if we find solutions to f, their negations are the solutions to x.

3.5 Examples: Method of Transform

Example 1

Factor:

$$6x^2 + 11x + 4$$

Identify $B = 11$ and $AC = 24$ to write

$$f^2 - 11f + 24 = 0\,,$$

from which we discern

$$(f - 8)(f - 3) = 0\,.$$

With two solutions for f, proceed using the split-term method:

$$\begin{aligned} 6x^2 + 11x + 4 &= 6x^2 + 8x + 3x + 4 \\ &= 2x(3x + 4) + (3x + 4) \\ &= (2x + 1)(3x + 4) \end{aligned}$$

Example 2

Factor:

$$3x^2 - 2x - 8$$

Identify $B = -2$ and $AC = -24$ to write

$$f^2 + 2f - 24 = 0\,,$$

from which we discern

$$(f + 6)(f - 4) = 0\,.$$

With two solutions for f, proceed using the split-term method:

$$\begin{aligned} 3x^2 - 2x - 8 &= 3x^2 - 6x + 4x - 8 \\ &= 3x(x - 2) + 4(x - 2) \\ &= (3x + 4)(x - 2) \end{aligned}$$

Example 3

Factor:

$$15x^2 + 14x - 8$$

Identify $B = 14$ and $AC = -120$ to write

$$f^2 - 14f - 120 = 0\,,$$

from which we discern

$$(f + 6)(f - 20) = 0\,.$$

With two solutions for f, proceed by completing the rectangle:

$$\begin{aligned} 15x^2 + 14x - 8 &= \frac{-1}{8}(-6x - 8)(20x - 8) \\ &= (5x - 2)(3x + 4) \end{aligned}$$

Example 4

Factor:

$$36x^2 - 121y^2$$

Identify $B = 0$ and $AC = -36 \cdot 121y^2$ to write

$$f^2 = 36 \cdot 121y^2\,,$$

from which we discern

$$f = \pm 66y\,.$$

With two solutions for f, proceed by completing the rectangle:

$$\begin{aligned} 36x^2 - 121y^2 &= \frac{1}{A}(Ax + 66y)(Ax - 66y) \\ &= (6x + 11y)(6x - 11y) \end{aligned}$$

Example 5

Factor:

$$x^2 + 3x - 28$$

Identify $B = 3$ and $AC = -28$ to write

$$f^2 - 3f - 28 = 0\,,$$

reminiscent of the original problem, up the the variable change $f = -x$. Proceed by standard means to find $f_1 = 4$, $f_2 = -7$. The solutions in x are the negatives of these, so we finally have

$$x^2 + 3x - 28 = (x - 4)\,(x + 7)\ .$$

4 Problems: Factoring Quadratics

Use any factoring method to prove the following:

Problem 5

$$4x^2 + 15x + 9 = (4x + 3)\,(x + 3)$$

Problem 6

$$6x^2 + 23x - 4 = (6x - 1)\,(x + 4)$$

Problem 7

$$10x^2 + x - 3 = (5x + 3)\,(2x - 1)$$

Problem 8

$$15x^2 - 7x - 4 = (3x + 1)\,(5x - 4)$$

Problem 9

$$4x^2 + 2x - 12 = 2\,(2x - 3)\,(x - 2)$$

Problem 10

$$8x^2 - 2xy - 3y^2 = (2x + y)\,(4x - 3y)$$

Problem 11

$$(x + 4)^3 - 9x - 36 = (x + 4)\,(x + 1)\,(x + 7)$$

Problem 12

$$x^2 + 2x + 1 - y^2 = (x + 1 + y)\,(x + 1 - y)$$

Problem 13

$$x^2 - y^2 + 12y - 36 = (x + y - 6)\,(x - y + 6)$$

Problem 14

$$3x^3 + 5x^2y + xy^2 - y^3 = (3x - y)\,(x + y)^2$$

5 Cubic Expressions

A *cubic expression* is a *third-order* polynomial:

$$Ax^3 + Bx^2 + Cx + D$$

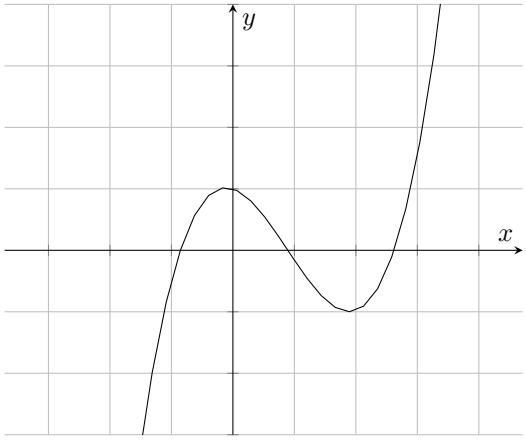
As a generalized quadratic expression, one might wonder how a cubic expression could be factored, and whether there exists an order-three analog to the quadratic formula. As it turns out, this answer isn't so trivial to come by. Unlike quadratic expressions, whose mysteries were shooed away millennia ago, it took until the sixteenth century to crack the problem of cubics.

5.1 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.

5.2 Factoring Cubic Equations

We now develop a method to factor any cubic expression, which amounts to looking for x-intercepts in the cubic equation

$$y = Ax^3 + Bx^2 + Cx + D\,.$$

5.3 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

$$\begin{aligned} b &= -\frac{B^2}{3A^2} + \frac{C}{A} \\ -c &= 2\left(\frac{B}{3A}\right)^3 - \frac{C}{A}\left(\frac{B}{3A}\right) + \frac{D}{A} \end{aligned}$$

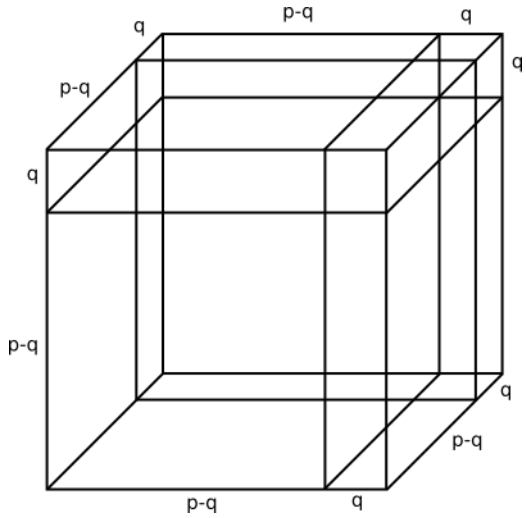
to arrive at the equation of the *depressed cubic*:

$$z^3 + bz = c$$

Henceforth we'll work in the variable z, keeping in mind that we get back to x using the initial subsitution.

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

$$p^3 = (p-q)^3 + 3q^2\,(p-q) + 3\,(p-q)^2\,q + q^3\,,$$

readily simplifying to

$$(p-q)^3 + 3pq\,(p-q) = p^3 - q^3\,.$$

Looking closely, note the above is a depressed cubic equation. Identifying

$$\begin{aligned} p - q &= z \\ 3pq &= b \\ p^3 - q^3 &= c\,, \end{aligned}$$

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).

For shorthand, the Ferro-Tartaglia formula is also written

$$z_0 = \left(\frac{c}{2} + q\right)^{1/3} + \left(\frac{c}{2} - q\right)^{1/3} ,$$

where

$$q = \sqrt{\frac{c^2}{4} + \frac{b^3}{27}} .$$

Example 1

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:

$$z = p - q = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

5.4 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.

Example 2

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:

$$z^3 - \frac{x}{3} - \frac{2}{27} = \left(z - \frac{2}{3}\right)\left(z + \frac{1}{3}\right)^2$$

Depressed Cubic in Disguise

Consider the rather exotic quantity

$$\left(7 + \sqrt{50}\right)^{1/3} + \left(7 - \sqrt{50}\right)^{1/3} ,$$

which might seem impossible to evaluate, namely because $7 - \sqrt{50}$ is surely negative, hinting of complex numbers. Proceeding with caution, let us store the

whole quantity in a variable x, and then calculate x^3:

$$\begin{aligned} x^3 &= \left(\left(7+\sqrt{50}\right)^{1/3} + \left(7-\sqrt{50}\right)^{1/3} \right)^3 \\ &= 14 + 3\left(\left(7+\sqrt{50}\right)^{1/3} \left(7-\sqrt{50}\right)^{1/3} \right) \\ &\quad \cdot \left(\left(7+\sqrt{50}\right)^{1/3} + \left(7-\sqrt{50}\right)^{1/3} \right) \\ &= 14 + 3\left(-1\right)^{1/3} x \end{aligned}$$

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:

$$\left(7+\sqrt{50}\right)^{1/3} + \left(7-\sqrt{50}\right)^{1/3} = 2$$

6 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 creep up on the issue, let us factor the expression:

$$x^3 + 5x^2 - 2x - 24$$

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

$$q = \sqrt{\frac{c^2}{4} + \frac{b^3}{27}} = \sqrt{-\frac{25}{3}}\,.$$

Alarmingly, there is a minus sign inside the square root term.

6.1 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

$$\begin{aligned} 11 \cdot 14 \cdot 8 &= \tilde{U}\left(\tilde{U} + 3\tilde{V}\right)\left(\tilde{U} - 3\tilde{V}\right) \\ 1 \cdot 12 \cdot 10 &= \tilde{V}\left(\tilde{U} + \tilde{V}\right)\left(\tilde{U} - \tilde{V}\right)\,, \end{aligned}$$

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

$$z_1 = -\frac{7}{3}$$
$$z_2 = -\frac{4}{3}\,.$$

Restoring the x variable, these read

$$x_1 = -4$$
$$x_2 = -3\,.$$

Finally then, we have

$$x^3 + 5x^2 - 2x - 24 = (x-2)\,(x+4)\,(x+3)\ ,$$

finishing the example.

Chapter 3

Polynomial Division

1 Introduction

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:

$$x+1\,\Big)\;\overline{\begin{array}{ccc} x^2 & -9x & -10 \end{array}}$$

For the process to yield useful results, the numerator should always contain a higher-degree polynomial than the denominator.

1.1 Long Division Algorithm

Divide the first term in the dividend by the first term in the divisor to get $x^2/x = x$. Place the result (x) in the quotient field (above the line). Then, distribute x into the divisor and subtract the result from the dividend:

$$x+1\,\Big)\;\begin{array}{ccc} & x & \\ \hline x^2 & -9x & -10 \\ x^2 & x & \\ \hline & -10x & -10 \end{array}$$

The 'bottom line' of the above is $-10x-10$, which may now regard as the updated dividend, and the process is ready to repeat. Dividing the respective leading terms, we find $-10x/x = -10$, and update as follows:

$$x+1\,\Big)\;\begin{array}{ccc} & x & -10 \\ \hline x^2 & -9x & -10 \\ x^2 & x & \\ \hline & -10x & -10 \\ & -10x & -10 \\ \hline & & 0 \end{array}$$

With a new dividend of zero, the process halts, and we can read off the answer:

$$\frac{x^2 - 9x - 10}{x + 1} = x - 10$$

1.2 Remainders

Polynomial division doesn't always finish so cleanly as the example chosen above. Taking a more informative case, consider the ratio

$$\frac{\left(x^4 + x + 1\right)^2}{x^2 - 1} = \frac{x^8 + 2x^5 + 2x^4 + x^2 + 2x + 1}{x^2 - 1}.$$

Setting up and doing the hard work, we have:

$$
\begin{array}{rl|lllllll}
 & & x^6 & & +x^4 & +2x^3 & +3x^2 & +2x & +4 \\
\hline
x^2-1 \Big) & x^8 & & +2x^5 & +2x^4 & & +x^2 & +2x & +1 \\
 & x^8 & -x^6 & & & & & & \\
\hline
 & & x^6 & +2x^5 & +2x^4 & & +x^2 & +2x & +1 \\
 & & x^6 & & -x^4 & & & & \\
\hline
 & & & 2x^5 & +3x^4 & & +x^2 & +2x & +1 \\
 & & & 2x^5 & & -2x^3 & & & \\
\hline
 & & & & 3x^4 & +2x^3 & +x^2 & +2x & +1 \\
 & & & & 3x^4 & & -3x^2 & & \\
\hline
 & & & & & 2x^3 & +4x^2 & +2x & +1 \\
 & & & & & 2x^3 & & -2x & \\
\hline
 & & & & & & 4x^2 & +4x & +1 \\
 & & & & & & 4x^2 & & -4 \\
\hline
 & & & & & & & 4x & +5
\end{array}
$$

The next step *would* be to try dividing $4x$ by x^2, however the result (and any following it) will contain factors of x^{-1}. This is a sign to halt the division process and tuck the leftovers into a remainder term, namely $(4x+5)\,/\,(x^2-1)$. In doing so, we write the final result:

$$\frac{x^8+2x^5+2x^4+x^2+2x+1}{x^2-1} = x^6+x^4+2x^3+3x^2+2x+4+\frac{4x+5}{x^2-1}$$

1.3 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 $x=0$.

2 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 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

$$-d = bA+aB$$
$$c = A+B\,,$$

solved by

$$A = \frac{ac+d}{a-b}$$
$$B = \frac{bc+d}{b-a}\,,$$

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.

Example 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:

$$A = -7$$
$$B = 9$$

Step 3: Assemble the result:

$$\frac{2x+1}{(x-3)(x-4)} = \frac{-7}{x-3} + \frac{9}{x-4}$$

Example 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:

$$A = 1/2a$$
$$B = 1/2a$$

Step 4: Assemble the result:

$$\frac{1}{a^2-x^2} = \frac{1}{2a}\left(\frac{1}{a-x} + \frac{1}{a+x}\right)$$

2.1 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:

$$\frac{1}{(x-a)^3(x-b)} = \frac{A_1}{x-a} + \frac{A_2}{(x-a)^2} + \frac{A_3}{(x-a)^3} + \frac{B}{x-b}$$

2.2 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 $\left(x^2+ax+b\right)^2$ occurs, extra terms are needed:

$$\frac{1}{(x^2+ax+b)^2(x-c)} = \frac{A_1x+B_1}{x^2+bx+c} + \frac{A_2x+B_2}{(x^2+bx+c)^2} + \frac{C}{x-c}$$

Example 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:

$$\frac{1}{(x-1)(x+1)(x^2+1)} = \frac{A}{x-1} + \frac{B}{x+1} + \frac{Cx+D}{x^2+1}$$

Step 3: Multiply through by the left-hand denominator:

$$1 = A(x+1)\left(x^2+1\right) + B(x-1)\left(x^2+1\right) + (Cx+D)(x-1)(x+1)$$

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:

$$\frac{1}{x^4-1} = \frac{1}{4}\left(\frac{1}{x-1} - \frac{1}{x+1}\right) - \frac{1}{2}\left(\frac{1}{x^2+1}\right)$$

2.3 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

$$x^2+x\;\Big)\;\overline{\qquad\qquad}\;\;x^3\quad+4\quad\;,$$

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:

$$\frac{x^3+4}{x^2+x} = x - 1 + \frac{4}{x} - \frac{3}{x+1}$$

---

3 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:

$$x-a\;\Big)\;\overline{\qquad\qquad}\;\;x^n\quad -a^n$$

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:

$$\frac{x^n - a^n}{x-a} = x^{n-1} + a^1x^{n-2} + a^2x^{n-3} + a^3x^{n-4} + \frac{a^4x^{n-4} - a^n}{x-a}$$

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:

$$x^n - a^n = (x-a)\left(x^{n-1} + a^1x^{n-2} + a^2x^{n-3} + a^3x^{n-4}\right) + \left(a^4x^{n-4} - a^n\right)$$

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:

$$x^n - a^n = (x-a)\left(x^{n-1} + a^1x^{n-2} + a^2x^{n-3} + \cdots + a^{j-1}x^{n-n}\right) + \left(a^jx^{n-j} - a^n\right)$$

By choosing $j = n$, the remainder term vanishes entirely, leaving

$$x^n - a^n = (x-a)\left(x^{n-1} + a^1 x^{n-2} + a^2 x^{n-3} + \cdots + a^{n-1}\right) . \tag{3.1}$$

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.

3.1 Sigma Notation

In order to avoid always writing Equation (3.1) as a long polynomial basted with exponents, we will often use condensed *sigma notation* as follows:

$$x^n - a^n = (x-a)\left(\sum_{k=1}^{n} a^{k-1} x^{n-k}\right) \tag{3.2}$$

The symbol Σ is the uppercase Greek 'sigma'.

Sometimes it's convenient to work the modified version that replaces a^n with a:

$$x^n - a = \left(x - a^{1/n}\right)\left(\sum_{k=1}^{n} a^{(k-1)/n} x^{n-k}\right) \tag{3.3}$$

3.2 Examples

Example 1

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:

$$x^3 - 8 = (x-2)\left(x^2 + 2x + 4\right)$$

Example 2

Factor:

$$x^4 - 9$$

Step 1: Identify variables:

$$n = 4$$
$$a = 3$$

Step 2: Write the factored expression in summation notation:

$$x^4 - 9 = \left(x - \sqrt{3}\right)\left(\sum_{k=1}^{4} 9^{(k-1)/4} x^{4-k}\right)$$

Step 3: Simplify:

$$\begin{aligned} x^4 - 9 &= \left(x - \sqrt{3}\right)\left(x^3 + \sqrt{3}x^2 + 3x + 3\sqrt{3}\right) \\ &= \left(x - \sqrt{3}\right)\left(x^2\left(x + \sqrt{3}\right) + 3\left(x + \sqrt{3}\right)\right) \\ &= \left(x - \sqrt{3}\right)\left(x + \sqrt{3}\right)\left(x^2 + 3\right) \end{aligned}$$

4 Recursive Sequences

Equation (3.1) 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.

4.1 Applied Polynomial Division

By making the substitution $a = (-1/x)^n$, we use Equation (3.3) 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 nth result as C_n, we equivalently write

$$C_n = \frac{x^n - (-1/x)^n}{x + 1/x} \tag{3.4}$$

$$= \begin{cases} C_1 = 1 \\ C_2 = x^1 - x^{-1} \\ C_3 = -C_1 + x^2 + x^{-2} \\ C_4 = -C_2 + x^3 - x^{-3} \\ C_5 = -C_3 + x^4 + x^{-4} \\ C_6 = -C_4 + x^5 - x^{-5} \end{cases}$$

Recursion Relations

By inspection of the above, the coefficients C_n are subect to *recursion relations*:

$$C_n = -C_{n-2} + x^{n-1} - x^{-(n-1)} \tag{3.5}$$

$$C_{n+1} = -C_{n-1} + x^n + x^{-n} \tag{3.6}$$

4.2 Large-n Recursion

Supposing we choose any even-valued n, the coefficient C_n and its next neighbor relate by the recursion relations (3.5), (3.6). 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:

$$C_{n+1} \approx \frac{-C_n}{x}$$

5 Lucas Numbers

In deriving Equation (3.4), we manged 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

$$\phi \cdot \psi = -1$$
$$\phi + \psi = 1 \, .$$

Then, making the association

$$x = \phi$$
$$1/x = -\psi \, ,$$

the coefficients C_n represented in Equation (3.4) specialize to

$$F_n = \frac{\phi^n - \psi^n}{\phi - \psi} \, . \tag{3.7}$$

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,

$$F_n = -F_{n-2} + L_{n-1} \, . \tag{3.8}$$

5.1 Lucas Generating Formula

The terms

$$L_n = \phi^n + \psi^n \tag{3.9}$$

are called *Lucas numbers*. These are a bit tricky to evaluate, but nonetheless with some grit one can produce:

$$L_1 = \phi^1 + \psi^1 = 1$$
$$L_2 = \phi^2 + \psi^2 = (\phi + \psi)^2 - 2\phi\psi = L_1^2 + 2$$
$$L_3 = \phi^3 + \psi^3 = \phi^2 + 1 + \psi^2 = L_1 + L_2$$
$$L_4 = \phi^4 + \psi^4 = \left(\phi^2 + \psi^2\right)^2 - 2\phi^2\psi^2 = L_2^2 - 2$$
$$L_5 = \phi^5 + \psi^5 = \phi^4 + \phi^2 + 1 + \psi^2 + \psi^4 = L_3 + L_4$$
$$L_6 = \phi^6 + \psi^6 = \left(\phi^3 + \psi^3\right)^2 - 2\phi^3\psi^3 = L_3^2 + 2$$

Evidently, the pattern in L_n splits between odd and even channels

$$L_{n \text{ odd}} = L_{n-1} + L_{n-2}$$
$$L_{n \text{ even}} = L_{n/2}^2 - 2 \cdot (-1)^{n/2} \, ,$$

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 $L_{n\ \mathrm{odd}}$ makes reference to the *two* previous terms, it suffices to write

$$L_n = L_{n-1} + L_{n-2} \tag{3.10}$$

as a single formula applying to both odd and even Lucas numbers.

5.2 Lucas Sequence

Listing the Lucas numbers in order gives the *Lucas sequence*:

$$\{L\} = \{2, 1, 3, 4, 7, 11, 18, 29, \dots\}$$

6 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 (3.10), namely

$$F_n = F_{n-1} + F_{n-2}\,. \tag{3.11}$$

Despite its obvious truth, the proof of the above is reserved for the more general treatment of the Lucas-Fibonacci system (see below).

6.1 Fibonacci Sequence

Listing the Fibonacci numbers in order gives the *Fibonacci sequence*:

$$\{F\} = \{1, 1, 2, 3, 5, 8, 13, 21, \dots\}$$

6.2 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

$$\begin{aligned}
\tilde{F}_1 &= 1 = F_1 \\
\tilde{F}_2 &= -\frac{(\phi+\psi)\,\cancel{(\phi-\psi)}}{\cancel{(\phi-\psi)}} = -1 = -F_2 \\
\tilde{F}_3 &= \frac{\cancel{(\phi-\psi)}\left(\phi^2 + \phi\psi + \psi^2\right)}{\cancel{(\phi-\psi)}} = -1 + L_2 = F_3 \\
\tilde{F}_4 &= \frac{\left((-\psi)^2 + (-\phi)^2\right)\cancel{(\phi-\psi)}\cancel{((-\psi)+(-\phi))^{-1}}}{\cancel{(\phi-\psi)}} \\
&= -\left(L_1^2 + 2\right) = -F_4\,,
\end{aligned}$$

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:

$$\{F\} = \{\dots, -8, 5, -3, 2, -1, 1, 0, 1, 1, 2, 3, 5, 8, \dots\}$$

6.3 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

$$x_1 = \phi = \frac{1+\sqrt{5}}{2} \approx 1.618034\ldots$$
$$x_2 = \psi = -\frac{1}{\phi} = \frac{1-\sqrt{5}}{2} \approx -0.618034\ldots\,.$$

Golden Ratio

The constant

$$\phi = \frac{1+\sqrt{5}}{2} \approx 1.618034\ldots$$

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 nth Lucas or Fibonacci number can be straightforwardly expressed:

$$L_n = \frac{\left(1+\sqrt{5}\right)^n + \left(1-\sqrt{5}\right)^n}{2^n}$$
$$F_n = \frac{\left(1+\sqrt{5}\right)^n - \left(1-\sqrt{5}\right)^n}{2^n\sqrt{5}}$$

7 General L-F Numbers

The Lucas-Fibonnacci rabbithole was entered by setting $C_2 = 1$ as it occurs as Equation (3.4). Of course, this can all be 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

$$x_1 \cdot x_2 = -1$$
$$x_1 + x_2 = p\,.$$

Pursuing this, we stumble upon a new recursion statement analogous to Equation (3.8), namely

$$C_n = -C_{n-2} + \tilde{L}_{n-1}\,, \tag{3.12}$$

where $\tilde{L}$ is a generalized Lucas number

$$\tilde{L}_n = x_1^n + x_2^n\,,$$

obeying the recursion relation

$$\tilde{L}_n = p\tilde{L}_{n-1} + \tilde{L}_{n-2}\,. \tag{3.13}$$

7.1 Recursion Relations

To derive a robust recursion relation, let us write three instances of Equation (3.8) based on respective indices n, $n+1$, and $n+2$:

$$C_n = -C_{n-2} + \tilde{L}_{n-1}$$
$$C_{n+1} = -C_{n-1} + \tilde{L}_n$$
$$C_{n+2} = -C_n + \tilde{L}_{n+1}$$

Multiply the first equation by a factor of -1, and the second by a factor of $-p$

$$-C_n = C_{n-2} - \tilde{L}_{n-1}$$
$$-pC_{n+1} = pC_{n-1} - p\tilde{L}_n$$
$$C_{n+2} = -C_n + \tilde{L}_{n+1}$$

Next, take the sum of all three equations and regroup terms:

$$(C_{n+2} - pC_{n+1} - C_n) + (C_n - pC_{n-1} - C_{n-2}) = \left(\tilde{L}_{n+1} - p\tilde{L}_n - \tilde{L}_{n-1}\right)$$

The right side is identically zero by Equation (3.13). The rest can only be true if

$$C_n = pC_{n-1} + C_{n-2}\,, \tag{3.14}$$

which we take as the generalized recursion relation. In the special case $p = 1$, the above reduces to the Fibonacci case represented by Equation (3.11).

7.2 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 (3.14), we easily find the rest:

$$\{F_{p=1}\} = \{1, 1, 2, 3, 5, 8, 13, 21, \ldots\}$$
$$\{F_{p=2}\} = \{1, 2, 5, 12, 29, 70, 169, 408, \ldots\}$$
$$\{F_{p=3}\} = \{1, 3, 10, 33, 109, 360, 1189, 3927, \ldots\}$$
$$\{F_{p=4}\} = \{1, 4, 17, 72, 305, 1292, 5473, 23184, \ldots\}$$

Chapter 4

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:

$$\frac{1}{1-x} = \sum_{k=1}^{n} x^{k-1} + \frac{x^n}{1-x} \tag{4.1}$$

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*,

$$\frac{1}{1-x} = 1+x+x^2+x^3+\cdots = \sum_{k=0}^{\infty} x^k \tag{4.2}$$

provided $|x| < .1$

Example

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

2.1 Long Division Method

The most brutal way to derive the geometric series is to perform polynomial division on the quantity $1/(1-x)$:

$$
\begin{array}{r|l}
 & 1 \quad x \quad x^2 \quad x^3 \\
\hline
1-x & 1 \\
 & \\
 & 1 \quad -x \\
\hline
 & \quad x \\
 & \quad x \quad -x^2 \\
\hline
 & \qquad x^2 \\
 & \qquad x^2 \quad -x^3 \\
\hline
 & \qquad\quad x^3 \\
 & \qquad\quad x^3 \quad -x^4 \\
\hline
 & \qquad\qquad x^4
\end{array}
$$

After a few terms in, it's clear that such an exercise leads to Equation (4.1) again.

2.2 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

$$G = \frac{1 \cancel{-x + x} \cancel{-x^2 + x^2} + \cdots - x^{n+1}}{1 - x} \,,$$

and simplify to recover Equation (4.1):

$$1 + x + x^2 + x^3 + \cdots + x^n = \frac{1 - x^{x+1}}{1 - x}$$

2.3 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 $q \to \infty$, this result is indistinguishable from Equation (4.2).

3 Manipulations

3.1 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

$$\begin{aligned}\left(\frac{1}{1-x}\right)^2 = 1 &+ x + x^2 + x^3 + \cdots \\ &+ x + x^2 + x^3 + x^4 + \cdots \\ &+ x^2 + x^3 + x^4 + x^5 + \cdots \ ,\end{aligned}$$

simplifying to:

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + 5x^4 + \cdots \quad (4.3)$$

As an exercise in brute force algebra, higher powers can be handled as well:

$$\frac{1}{(1-x)^3} = 1 + 3x + 6x^2 + 10x^3 + 15x^4 + \cdots \quad (4.4)$$

$$\frac{1}{(1-x)^4} = 1 + 4x + 10x^2 + 20x^3 + 35x^4 + \cdots \quad (4.5)$$

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 the (standard left-aligned) 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 nth column and reading down the triangle predicts the expansion of $1/(1-x)^n$.

3.2 Negative Argument

In the geometric series, setting $x \to -x$ 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:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \cdots \quad (4.6)$$

$$\frac{1}{(1+x)^2} = 1 - 2x + 3x^2 - 4x^3 + 5x^4 - \cdots \quad (4.7)$$

$$\frac{1}{(1+x)^3} = 1 - 3x + 6x^2 - 10x^3 + 15x^4 - \cdots \quad (4.8)$$

$$\frac{1}{(1+x)^4} = 1 - 4x + 10x^2 - 20x^3 + 35x^4 - \cdots \quad (4.9)$$

These can be predicted by the compliment Pascal triangle based on subtraction rather than addition:

$$\begin{array}{ccccccc} 1 & & & & & & \\ 1 & -1 & & & & & \\ 1 & -2 & 1 & & & & \\ 1 & -3 & 3 & -1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ 1 & -5 & 10 & -10 & 5 & -1 & \\ 1 & -6 & 15 & -20 & 15 & -6 & 1 \end{array}$$

Choosing the nth column and reading downward predicts the expansion of $1/(1+x)^n$.

3.3 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 (4.3), (4.6) to expand the right side, we find, after simplifying:

$$\frac{1}{1-x^2} = 1 + x^2 + x^4 + x^6 + \cdots \tag{4.10}$$

Evidently, replacing $x \to x^2$ is as straightforward as it looks. All powers are doubled. In a similar way, we can try $x \to -x^2$ to establish

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots . \tag{4.11}$$

Note also that the sum of the above results essentially repeats the problem with all powers doubled again:

$$\frac{1}{1-x^4} = 1 + x^4 + x^8 + x^{12} + \cdots \tag{4.12}$$

$$\frac{1}{1+x^4} = 1 - x^4 + x^8 - x^{12} + \cdots \tag{4.13}$$

3.4 Pascal Transform

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 the Pascal triangle based on the quantity $(1-z)$.

However, Equations (4.6) - (4.9) already claim the columns of the complement Pascal triangle. Using this information to replace the right side entirely gives a curious representation of $1/z$:

$$\frac{1}{z} = \frac{1}{1+z} + \frac{1}{(1+z)^2} + \frac{1}{(1+z)^3} + \cdots \tag{4.14}$$

In light of this shifty move, the above no longer converges for $0 < z < 2$ as before the so-called Pascal transform. 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 (4.9) 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 real numbers, particularly neighboring fractions, for instance:

$$\frac{1}{3} = \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \frac{1}{4^4} + \cdots$$

4 Repeating Decimals

The geometric series helps make sense of decimal numbers whose digits eventually repeat. Consider a number of the format

$$N = 0.abcd\ldots qabcd\ldots q ,$$

where the sequence $abcd\ldots q$ is Q digits in length. As a sum, N can be written

$$N = \left(\frac{a}{10} + \frac{b}{100} + \frac{c}{1000} + \cdots + \frac{q}{10^Q}\right) \times \left(1 + \frac{1}{10^Q} + \frac{1}{10^{2Q}} + \frac{1}{10^{3Q}} + \cdots\right) ,$$

which has been factored into a product of two terms.

For the first term we can define the shorthand

$$N' = \frac{a}{10} + \frac{b}{100} + \frac{c}{1000} + \cdots + \frac{q}{10^Q}$$

as the truncation of N before the sequence repeats. The second term is precisely a geometric series:

$$1 + \frac{1}{10^Q} + \frac{1}{10^{2Q}} + \frac{1}{10^{3Q}} + \cdots = \frac{1}{1-10^{-Q}}$$

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 $0.12312323\ldots$ has $N' = 0.123$ with $Q = 3$:

$$0.123123123\cdots = \frac{0.123}{1-10^{-3}} = \frac{123}{1000-1} = \frac{123}{999}$$

For another example, the special case $N' = 0.9$ with $Q = 1$ tells us

$$0.999\cdots = \frac{0.9}{1-10^{-1}} = \frac{9}{10-1} = \frac{9}{9} = 1\,,$$

which means $0.999\ldots$ repeating forever is indistinguishable from 1:

$$0.999\cdots = 1$$

5 Zeno's Paradox

5.1 The 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.

5.2 Linear Motion

Had the Greeks known about the geometric series, particularly the notion of convergence, then maybe there would have been no paradox, as we can make easy work of the situation.

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, but let us check:

$$\begin{aligned}\text{Spatial sum} &= \frac{X}{2} + \frac{X}{4} + \frac{X}{8} + \frac{X}{16} + \cdots \\ &= \frac{X}{2}\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots\right)\end{aligned}$$

The parenthesized sum is a geometric series equivalent to $1/\left(1-1/2\right) = 2$, leaving

$$\text{Spatial sum} = X\,,$$

simple as that.

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:

$$\begin{aligned}X/2 &= Vt_1 \\ X/4 &= Vt_2 \\ X/8 &= Vt_3 \\ X/2^j &= Vt_j\end{aligned}$$

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 evidence of a paradox.

6 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, to 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

$$B = 6\,.$$

Part II

Pre-Calculus

Chapter 5

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 5.1. 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.

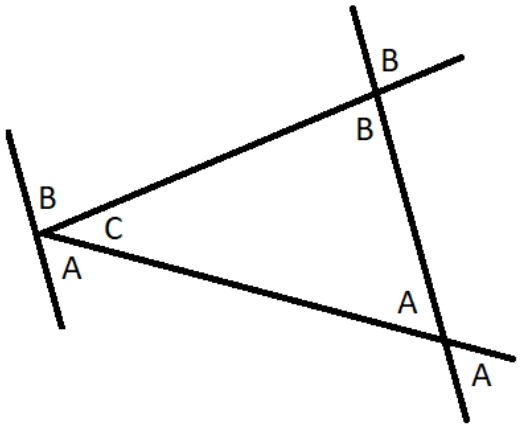


Figure 5.1: Triangle made from three lines.

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

$$1^\circ = \frac{\pi}{180}\, rad \qquad 1\, rad = \frac{180^\circ}{\pi}\ ,$$

which extrapolates to the following:

$$0^\circ = 0\, rad$$
$$45^\circ = \pi/4\, rad$$
$$90^\circ = \pi/2\, rad$$
$$135^\circ = 3\pi/4\, rad$$
$$360^\circ = 2\pi\, rad$$

The primary domain for angles is represented by

$$0 \le \theta < 360^\circ$$
$$0 \le \theta < 2\pi\ ,$$

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

Triangles whose sides and angles obey certain relationships may have standard names.

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.

Isosceles 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 one internal angle greater than 90°.

1.3 Right Triangles

A *right triangle* is any triangle that has two sides meeting at $90° = \pi/2\,rad$. Labeling either of the 'unused' angles as θ, the sides of the right triangle take on unique names as shown in Figures 5.2, 5.3.

- 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*.

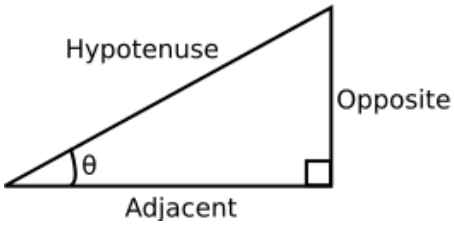


Figure 5.2: Right triangle.

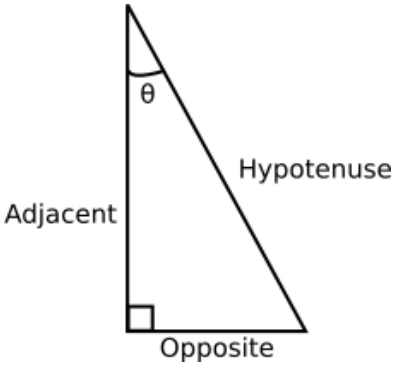


Figure 5.3: Right triangle.

Area of a Triangle

From geometry, we know that the area of any triangle is given by

$$\text{Area} = \frac{\text{Base} \times \text{Height}}{2}\,.$$

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 5.4 shows a typical right triangle with hypotenuse c, opposite a, and adjacent b.

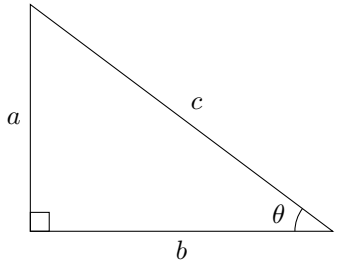


Figure 5.4: Right triangle.

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 5.5.

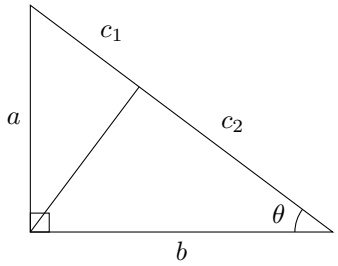


Figure 5.5: Similar right triangles.

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:

$$a^2 + b^2 = c^2 \tag{5.1}$$

Using the triangle side names in place of a, b, c yields the equivalent statement:

$$\text{Opposite}^2 + \text{Adjacent}^2 = \text{Hypotenuse}^2$$

1.5 Sine, Cosine, Tangent

On a right triangle, the opposite, adjacent, and hypotenuse can be stacked into ratios. These ratios have designated names:

$$\text{Sine} = \frac{\text{Opposite}}{\text{Hypotenuse}} \tag{5.2}$$
$$\text{Cosine} = \frac{\text{Adjacent}}{\text{Hypotenuse}} \tag{5.3}$$
$$\text{Tangent} = \frac{\text{Opposite}}{\text{Adjacent}} = \frac{\text{Sine}}{\text{Cosine}} \tag{5.4}$$

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 (5.2)-(5.4) 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 (5.2), (5.3), 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 square 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:

$$\sin^2(\theta) + \cos^2(\theta) = 1 \tag{5.5}$$

2 Circles

In the Cartesian plane, a circle is most generally described by

$$(x-h)^2 + (y-k)^2 = R^2 \,, \tag{5.6}$$

where the center of the circle is located at (h, k) and the radius is R as depicted in Figure 5.6.

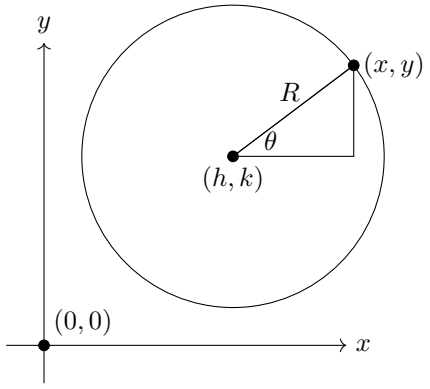


Figure 5.6: Circle.

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

$$A = \pi R^2 \, .$$

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.

Inscription Problem

Let ABC be a triangle with right angle A and hypotenuse $|BC|$ as shown in Figure 5.7. If the inscribed circle of radius R touches the hypotenuse at D, show that:

$$|CD| = \frac{|AC| + |BC| - |AB|}{2}$$

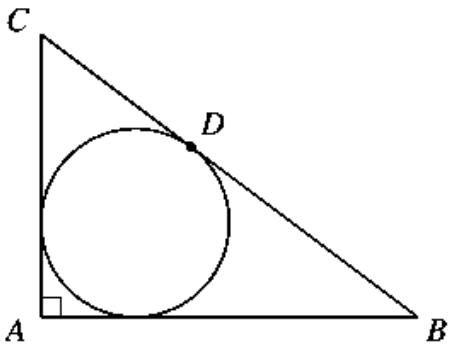


Figure 5.7: Inscribed Circle in Right Triangle

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

$$|AB| - |BD| = |AC| - |CD| \, .$$

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:

$$x(\theta) = h + R\cos(\theta) \tag{5.7}$$
$$y(\theta) = k + R\sin(\theta) \tag{5.8}$$

The above represents the *parameterized* equation of a circle. To quickly recover Equation (5.6), solve for $\cos(\theta)$, $\sin(\theta)$, respectively, and exploit Equation (5.5).

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 5.8.

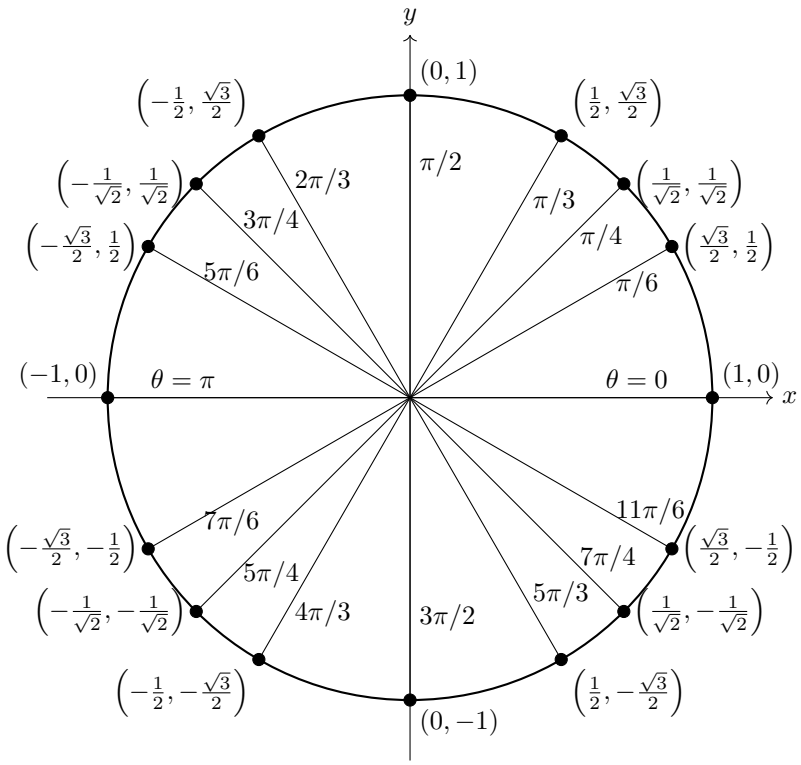


Figure 5.8: Unit circle.

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 5.9. 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

$$xx_0 + yy_0 = 1\,. \tag{5.9}$$

It just happens that the length of the tangent line 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.

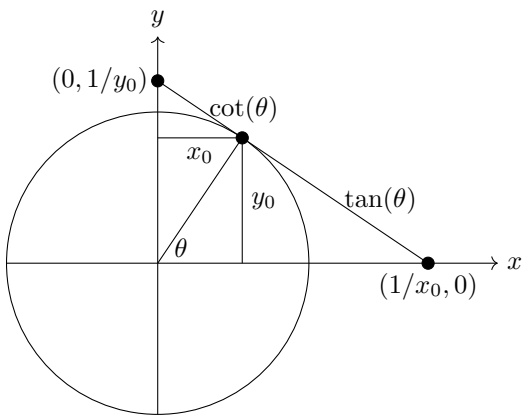


Figure 5.9: Unit circle with tangent line.

2.5 Cotangent

The *cotangent* (of angle θ) is defined as

$$\cot(\theta) = \frac{1}{\tan(\theta)} = \frac{\cos(\theta)}{\sin(\theta)} . \tag{5.10}$$

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 5.9, 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 (5.4) and Equation (5.5), simplifies to

$$\frac{1}{\cos(\theta)} = \frac{1}{x_0} .$$

Since x_0 is already claimed as the cosine of theta, we have:

$$\sec(\theta) = \frac{1}{\cos(\theta)} \tag{5.11}$$

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 x_0 is already claimed as the sine of theta, we have:

$$\csc(\theta) = \frac{1}{\sin(\theta)} \tag{5.12}$$

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

$$\sin(\theta \pm 2n\pi) = \sin(\theta) \tag{5.13}$$
$$\cos(\theta \pm 2n\pi) = \cos(\theta) \tag{5.14}$$
$$\tan(\theta \pm 2n\pi) = \tan(\theta) , \tag{5.15}$$

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 Angles

If we replace θ by $-\theta$, the symmetry of the unit circle demands:

$$\sin(-\theta) = -\sin(\theta) \tag{5.16}$$
$$\cos(-\theta) = \cos(\theta) \tag{5.17}$$
$$\tan(-\theta) = -\tan(\theta) \tag{5.18}$$

Phase Shift Pi

A phase shift of π radians jumps exactly across the unit circle. Accordingly, we have:

$$\sin(\theta \pm \pi) = -\sin(\theta) \tag{5.19}$$
$$\cos(\theta \pm \pi) = -\cos(\theta) \tag{5.20}$$

Phase Shift Pi/2

As they're defined, it turns out that $\sin(\theta)$ and $\cos(\theta)$ are related by the phase $\pi/2$:

$$\sin\left(\theta + \frac{\pi}{2}\right) = \cos(\theta) \tag{5.21}$$
$$\cos\left(\theta + \frac{\pi}{2}\right) = -\sin(\theta) \tag{5.22}$$
$$\sin\left(\theta - \frac{\pi}{2}\right) = -\cos(\theta) \tag{5.23}$$
$$\cos\left(\theta - \frac{\pi}{2}\right) = \sin(\theta) \tag{5.24}$$

Similar equations apply when θ is replaced with $-\theta$:

$$\sin\left(\frac{\pi}{2} - \theta\right) = \cos(\theta) \tag{5.25}$$
$$\cos\left(\frac{\pi}{2} - \theta\right) = \sin(\theta) \tag{5.26}$$

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 (5.5), 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:

$$1 + \cot^2(\theta) = \csc^2(\theta) \tag{5.27}$$
$$\tan^2(\theta) + 1 = \sec^2(\theta) \tag{5.28}$$

For an interesting sanity check, take the sum of the two above equations to come up with

$$(\tan(\theta) + \cot(\theta))^2 = \csc^2(\theta) + \sec^2(\theta)\,, \tag{5.29}$$

which is the summary of Figure 5.9.

3.2 Angle-Sum Formulas

Consider the sum of two angles α, β, as they embed in the unit circle as shown in Figure 5.10. The triangle swept out by β has adjacent side $\cos(\beta)$ and opposite side $\cos(\beta)$. Each of these sides is the hypotenuse of a pair of right triangles whose sides are the products denoted in the Figure.

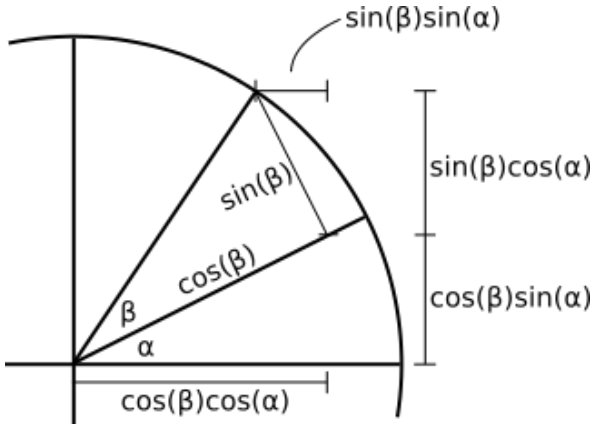


Figure 5.10: Angle-sum analysis.

Expressing $\sin(\alpha+\beta)$, $\cos(\alpha+\beta)$ in terms of the products of individual terms, we find by inspection the *angle-sum formulas*:

$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) \tag{5.30}$$

$$\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \tag{5.31}$$

Using these two results, we can easily calculate the tangent of $\alpha+\beta$:

$$\tan(\alpha \pm \beta) = \frac{\tan(\alpha) \pm \tan(\beta)}{1 \mp \tan(\alpha)\tan(\beta)} \tag{5.32}$$

3.3 Product Formulas

Starting with the angle-sum formulas (5.30), (5.31), it's straightforward to derive the *product formulas*:

$$2\sin(\alpha)\cos(\beta) = \sin(\alpha+\beta) + \sin(\alpha-\beta) \tag{5.33}$$

$$2\cos(\alpha)\sin(\beta) = \sin(\alpha+\beta) - \sin(\alpha-\beta) \tag{5.34}$$

$$2\cos(\alpha)\cos(\beta) = \cos(\alpha+\beta) + \cos(\alpha-\beta) \tag{5.35}$$

$$2\sin(\alpha)\sin(\beta) = \cos(\alpha-\beta) - \cos(\alpha+\beta) \tag{5.36}$$

3.4 Double-Angle Formulas

Starting with the product formulas, let $\alpha = \beta = \theta$ to derive the *double-angle formulas*:

$$\sin(2\theta) = 2\sin(\theta)\cos(\theta) \tag{5.37}$$
$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta) \tag{5.38}$$
$$\tan(2\theta) = \frac{2\tan(\theta)}{1 - \tan^2(\theta)} \tag{5.39}$$

3.5 Half-Angle Formulas

Starting with equation (5.38), replace the $\sin^2(\theta)$ term and also replace $\theta \to \theta/2$ 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*:

$$\sin\left(\frac{\theta}{2}\right) = \pm\sqrt{\frac{1}{2} - \frac{\cos(\theta)}{2}} \tag{5.40}$$

$$\cos\left(\frac{\theta}{2}\right) = \pm\sqrt{\frac{1}{2} + \frac{\cos(\theta)}{2}} \tag{5.41}$$

$$\tan\left(\frac{\theta}{2}\right) = \frac{\sin(\theta)}{1+\cos(\theta)} \tag{5.42}$$

$$\cot\left(\frac{\theta}{2}\right) = \frac{\sin(\theta)}{1-\cos(\theta)} \tag{5.43}$$

$$\sec\left(\frac{\theta}{2}\right) = \frac{2\cos(\theta/2)}{1+\cos(\theta)} \tag{5.44}$$

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:

$$\sin(\alpha)+\sin(\beta) = 2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right) \tag{5.45}$$

Replace β with $-\beta$ to get the second superposition relationship for free:

$$\sin(\alpha)-\sin(\beta) = 2\sin\left(\frac{\alpha-\beta}{2}\right)\cos\left(\frac{\alpha+\beta}{2}\right) \tag{5.46}$$

These are both called *superposition relationships.*

Superposition of Cosines

Starting with the superposition relationships above, introduce the phase shifts:

$$\begin{aligned}\alpha &\to \alpha+\pi/2 \\ \beta &\to \beta-\pi/2\end{aligned}$$

Inserting these and simplifying gives two more superposition relationships for the cosine:

$$\cos(\alpha)-\cos(\beta) = -2\sin\left(\frac{\alpha+\beta}{2}\right)\sin\left(\frac{\alpha-\beta}{2}\right) \tag{5.47}$$

$$\cos(\alpha)+\cos(\beta) = 2\cos\left(\frac{\alpha-\beta}{2}\right)\cos\left(\frac{\alpha+\beta}{2}\right) \tag{5.48}$$

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:

$$\arcsin(\sin(\theta)) = \theta \tag{5.49}$$

$$\arccos(\cos(\theta)) = \theta \tag{5.50}$$

$$\arctan(\tan(\theta)) = \theta \tag{5.51}$$

$$\operatorname{arccsc}(\csc(\theta)) = \theta \tag{5.52}$$

$$\operatorname{arcsec}(\sec(\theta)) = \theta \tag{5.53}$$

$$\operatorname{arccot}(\cot(\theta)) = \theta \tag{5.54}$$

Inverse Trig Nomenclature

Confusingly enough, there is another way to write $\arcsin(\theta)$, $\arccos(\theta)$, etc., using the nomenclature

$$\begin{aligned}\sin^{-1}(\theta) &= \arcsin(\theta) \\ \cos^{-1}(\theta) &= \arccos(\theta)\ ,\end{aligned}$$

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:

$$\arcsin(1/x) = \operatorname{arccsc}(x) \tag{5.55}$$

$$\operatorname{arccsc}(1/x) = \arcsin(x) \tag{5.56}$$

$$\arccos(1/x) = \operatorname{arcsec}(x) \tag{5.57}$$

$$\operatorname{arcsec}(1/x) = \arccos(x) \tag{5.58}$$

$$\arctan(1/x) = \operatorname{arccot}(x) \tag{5.59}$$

$$\operatorname{arccot}(1/x) = \arctan(x) \tag{5.60}$$

To prove any of the above will demonstrate how to handle the rest. Choosing the arctan case,

$$\begin{aligned}A &= \arctan(1/x) \\ B &= \operatorname{arccot}(x)\ ,\end{aligned}$$

and then

$$\begin{aligned}\tan(A) &= 1/x \\ \cot(B) &= x\ .\end{aligned}$$

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

$$\begin{aligned} x &= \cos(\theta) \\ \arccos(x) &= \theta \\ \sin(\arccos(x)) &= \sin(\theta) \ , \end{aligned}$$

resulting in:

$$\sin(\arccos(x)) = \sqrt{1-x^2} \tag{5.61}$$

Divide through by x to get a second result:

$$\tan(\arccos(x)) = \frac{\sqrt{1-x^2}}{x} \tag{5.62}$$

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

$$\begin{aligned} x &= \sin(\theta) \\ \arcsin(x) &= \theta \\ \cos(\arcsin(x)) &= \cos(\theta) \ , \end{aligned}$$

resulting in:

$$\cos(\arcsin(x)) = \sqrt{1-x^2} \tag{5.63}$$

Similarly we can establish:

$$\tan(\arcsin(x)) = \frac{x}{\sqrt{1-x^2}} \tag{5.64}$$

Arctangent

Consider a new right triangle with hypotenuse $\sqrt{x^2+1}$, adjacent side 1, and opposite side x. For this case, we have

$$\begin{aligned} x &= \tan(\theta) \\ \arctan(x) &= \theta \ , \end{aligned}$$

implying

$$\begin{aligned} \cos(\arctan(x)) &= \cos(\theta) \\ \sin(\arctan(x)) &= \sin(\theta) \ . \end{aligned}$$

From these, conclude:

$$\cos(\arctan(x)) = \frac{1}{\sqrt{x^2+1}} \tag{5.65}$$

$$\sin(\arctan(x)) = \frac{x}{\sqrt{x^2+1}} \tag{5.66}$$

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:

$$\sin(\operatorname{arccsc}(x)) = \frac{1}{x} \tag{5.67}$$

$$\cos(\operatorname{arccsc}(x)) = \frac{\sqrt{x^2-1}}{x} \tag{5.68}$$

$$\tan(\operatorname{arccsc}(x)) = \frac{1}{\sqrt{x^2-1}} \tag{5.69}$$

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:

$$\sin(\operatorname{arcsec}(x)) = \frac{\sqrt{x^2-1}}{x} \tag{5.70}$$

$$\cos(\operatorname{arcsec}(x)) = \frac{1}{x} \tag{5.71}$$

$$\tan(\operatorname{arcsec}(x)) = \sqrt{x^2-1} \tag{5.72}$$

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:

$$\tan(\operatorname{arccot}(x)) = \frac{1}{x} \tag{5.73}$$

$$\sin(\operatorname{arccot}(x)) = \frac{1}{\sqrt{x^2+1}} \tag{5.74}$$

$$\cos(\operatorname{arccot}(x)) = \frac{x}{\sqrt{x^2+1}} \tag{5.75}$$

5 Trigonometry Tables

Trigonometry tables are lists of data containing key values of $\sin(\theta)$, $\cos(\theta)$. Contained in the tables that follow are the data generated by a trip around the unit circle.

5.1 Standard Trigonometry Tables

First Quadrant

Angle (rad)	Angle (°)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
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 (°)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
$9\pi/16$	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 (°)	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$	$\csc(\theta)$	$\sec(\theta)$	$\cot(\theta)$
$17\pi/16$	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.

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:

$$\sin(\theta) = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots$$
$$\cos(\theta) = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots$$

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 for very small angles θ, the sine and cosine become:

$$\sin(\theta) \approx \theta - \cancel{\frac{\theta^3}{3!} + \cdots}$$
$$\cos(\theta) \approx 1 - \cancel{\frac{\theta^2}{2!} + \cdots},$$

or move concisely,

$$\theta \text{ small:} \qquad \sin(\theta) \approx \theta$$
$$\theta \text{ small:} \qquad \cos(\theta) \approx 1\,.$$

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 (5.30), (5.31), namely

$$\sin(\alpha+\beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$$
$$\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)\,,$$

and put the following restrictions on α, β:

$$0 \le \alpha < 2\pi$$
$$|\beta| \ll \alpha$$

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:

$$\sin(\alpha+\beta) \approx \sin(\alpha) + \cos(\alpha)\,\beta \tag{5.76}$$
$$\cos(\alpha+\beta) \approx \cos(\alpha) - \sin(\alpha)\,\beta \tag{5.77}$$

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:

$$\beta = 83^\circ - 78.75^\circ = 4.25^\circ = 0.0742\,rad$$

With β in radians, we can plug into Equation (5.77) straightforwardly:

$$\cos(83^\circ) \approx \cos(78.75^\circ) - \sin(78.75^\circ)\,(0.0742)$$
$$\cos(83^\circ) \approx 0.1223$$

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.

Example 1

Given $\sin(20^\circ) = 0.342$ and $\cos(20^\circ) = 0.940$, estimate $\sin(22^\circ)$ and $\cos(22^\circ)$.

Let

$$\alpha = 20^\circ = 20^\circ\left(\frac{\pi}{180^\circ}\right) = 0.349$$
$$\beta = 2^\circ = 2^\circ\left(\frac{\pi}{180^\circ}\right) = 0.0349\,,$$

so then:

$$\sin(22^\circ) \approx 0.342 + 0.0349\,(0.940) \approx 0.375$$
$$\cos(22^\circ) \approx 0.940 + 0.0349\,(0.342) \approx 0.928$$

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:

$$\begin{aligned} S^2 &= S\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} - \cdots\right) \\ &= (1)\,x^2 - \left(\frac{2}{3!}\right)x^4 + \left(\frac{2}{5!} + \frac{1}{3!3!}\right)x^6 \\ &\quad - \left(\frac{2}{7!} + \frac{2}{3!5!}\right)x^8 + \cdots \end{aligned}$$

$$\begin{aligned} C^2 &= C\left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} - \cdots\right) \\ &= 1 - \left(\frac{2}{2!}\right)x^2 + \left(\frac{2}{4!} + \frac{1}{2!2!}\right)x^4 \\ &\quad - \left(\frac{2}{6!} + \frac{2}{2!4!}\right)x^6 \\ &\quad + \left(\frac{2}{8!} + \frac{2}{2!6!} + \frac{1}{4!4!}\right)x^8 - \cdots \end{aligned}$$

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

$$\begin{aligned} \sin^2(\theta) &= \theta^2 - \frac{1}{3}\theta^4 + \frac{2}{45}\theta^6 - \frac{1}{315}\theta^8 + \cdots \\ \cos^2(\theta) &= 1 - \theta^2 + \frac{1}{3}\theta^4 - \frac{2}{45}\theta^6 + \frac{1}{315}\theta^8 - \cdots , \end{aligned}$$

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

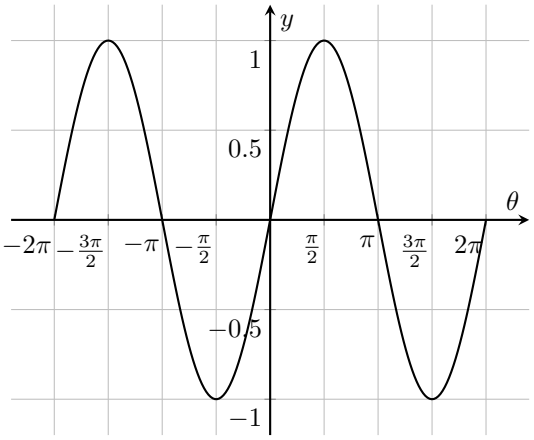


Figure 5.11: $y = \sin(\theta)$.

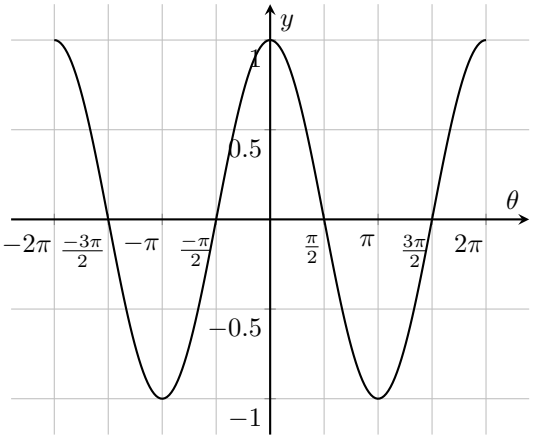


Figure 5.12: $y = \cos(\theta)$.

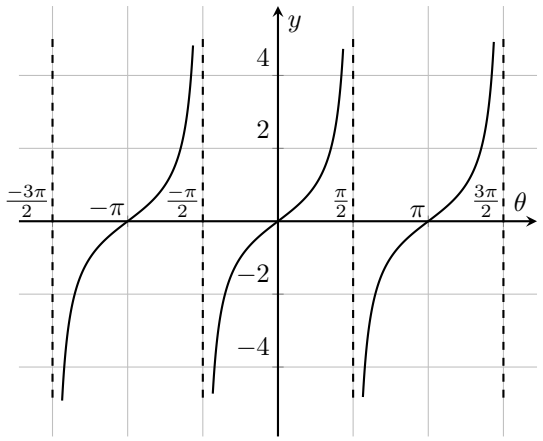


Figure 5.13: $y = \tan(\theta)$.

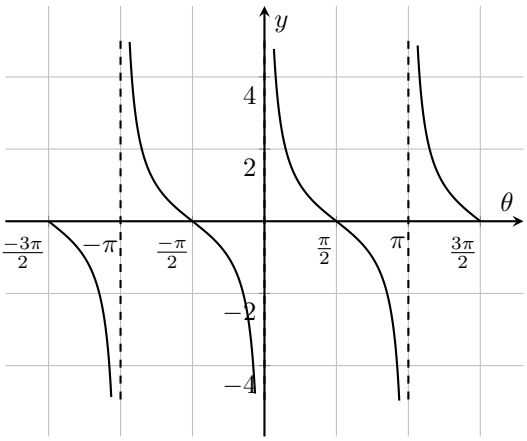


Figure 5.16: $y = \cot(\theta) = 1/\tan(\theta)$.

Cosecant, Secant, Cotangent

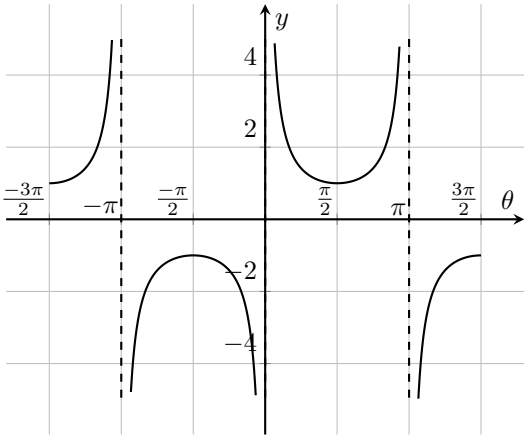


Figure 5.14: $y = \csc(\theta) = 1/\sin(\theta)$.

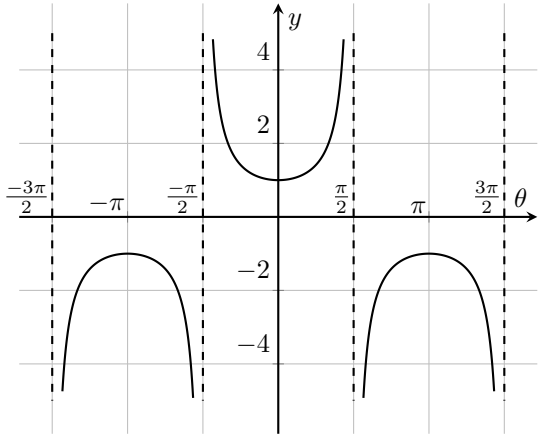


Figure 5.15: $y = \sec(\theta) = 1/\cos(\theta)$.

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 (5.65), (5.66) and invert these to solve for the arcsine and arccosine, respectively:

$$\arcsin(x) = \arctan\left(\frac{x}{\sqrt{1-x^2}}\right),\ x^2 \leq 1$$
$$\arccos(x) = \arctan\left(\frac{\sqrt{1-x^2}}{x}\right),\ 0 < x \leq 1$$

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

$$\arccos(x) = \pi + \arctan\left(\frac{\sqrt{1-x^2}}{x}\right),\ -1 \leq x < 0$$

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 5.17, 5.18.

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

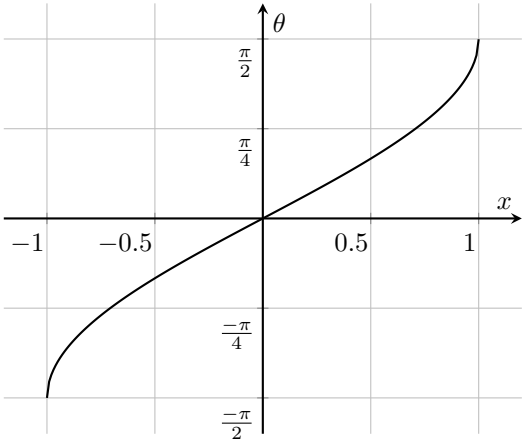


Figure 5.17: $\theta = \arcsin(x)$, $|x| \leq 1$.

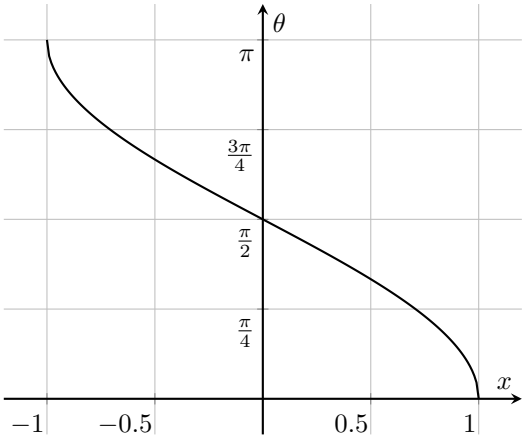


Figure 5.18: $\theta = \arccos(x)$, $|x| \leq 1$.

Arccosecant, Arcsecant

For the arccosecant and arcsecant we repeat a similar analysis to the above starting with Equations (5.69), (5.72). 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 5.17, 5.18.

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	$\pi/2$

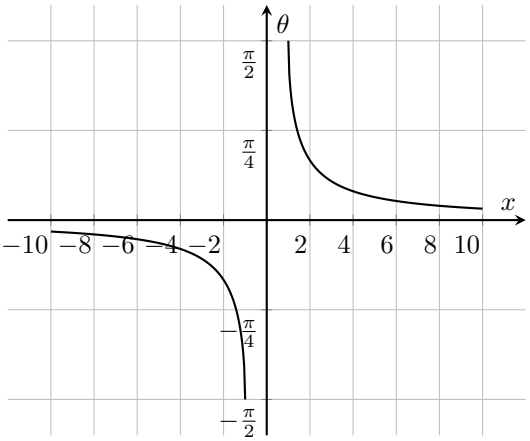


Figure 5.19: $\theta = \operatorname{arccsc}(x)$, $|x| \geq 1$.

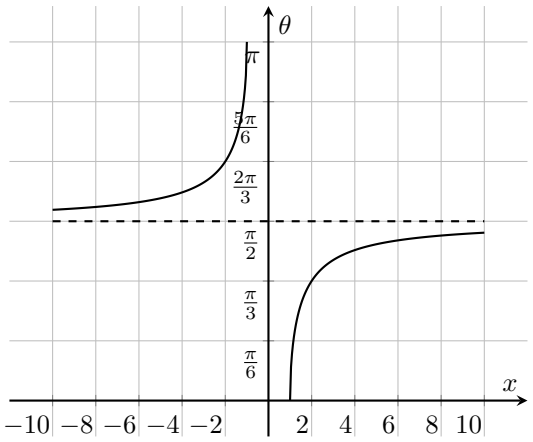


Figure 5.20: $\theta = \text{arcsec}\,(x)$, $|x| \geq 1$.

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 (5.73) to write

$$\text{arccot}\,(x) = \arctan\left(\frac{1}{x}\right)\,,$$

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 5.21, 5.22.

x	$\arctan(x)$	$\text{arccot}(x)$
$-\infty$	$-\pi/2$	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

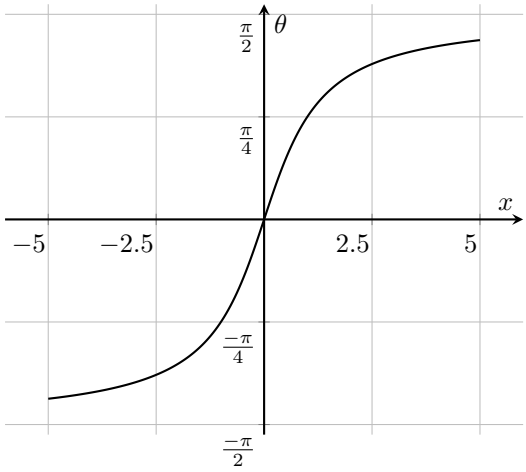


Figure 5.21: $\theta = \arctan\,(x)$

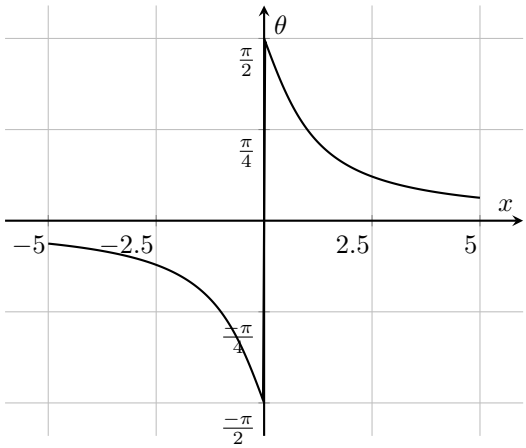


Figure 5.22: $\theta = \text{arccot}\,(x)$

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*.

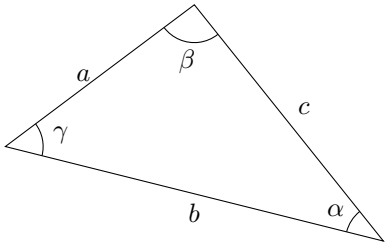


Figure 5.23: Arbitrary triangle.

Consider a triangle with three sides and three angles as labeled in Figure 5.23. 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 5.24. Note too that side b is now broken into the sum $b_1 + b_2 = b$.

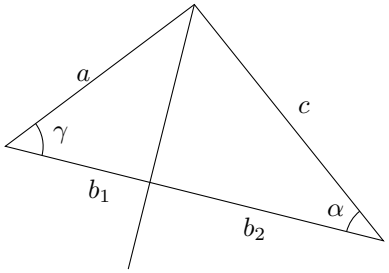


Figure 5.24: Arbitrary triangle with line intersecting side b at a right angle.

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

$$\begin{aligned} \cos^2(\gamma) &= 1 - \sin^2(\gamma) \\ \cos^2(\alpha) &= 1 - \sin^2(\alpha) \end{aligned}$$

and simplify like mad.

At the end, arrive at the all-powerful law of cosines:

$$a^2 + c^2 - 2ac\cos(\beta) = b^2 \tag{5.78}$$

By symmetry, since we could have sliced the triangle two more ways, there two more expressions of the same law:

$$b^2 + c^2 - 2bc\cos(\alpha) = a^2 \tag{5.79}$$

$$a^2 + b^2 - 2ab\cos(\gamma) = c^2 \tag{5.80}$$

6.2 Inscribed Angle

A tricky analysis starts with a circle of any radius R with a diameter $AB = 2R$ as shown in Figure 5.25. 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.)

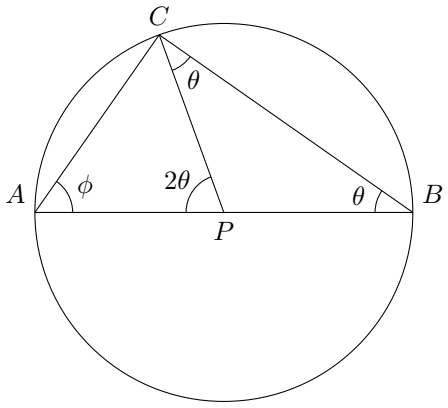


Figure 5.25: Inscribed angle.

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:

$$\begin{aligned} r^2 &= 2xR \\ r &= 2R\cos(\phi) \\ r &= 2R\sin\left(\frac{\pi}{2} - \phi\right) \end{aligned}$$

The idea now is to 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.

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:

$$\angle APC = 2\theta$$

6.3 Law of Sines

There is another relationship called the *law of sines* that is obeyed by all triangles.

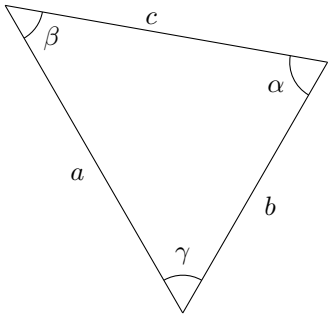


Figure 5.26: Arbitrary triangle.

Quick Derivation

Consider a triangle with three sides and three angles as labeled in Figure 5.26. Using the same process that led to Figure Figure 5.24, 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:

$$\frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \tag{5.81}$$

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 (5.81) but we can also interpret the law of sines as it relates to the area of the triangle:

$$\frac{2T}{abc} = \frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \tag{5.82}$$

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 5.82. Generally, such a circle has a radius R with a center point that may or may not lie within the confines of the triangle.

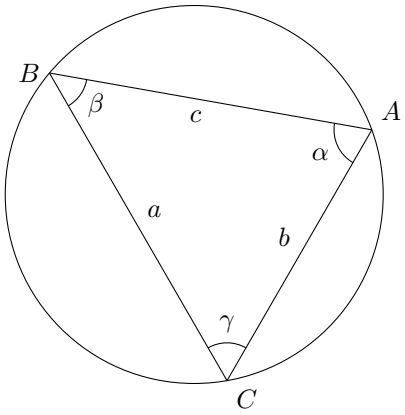


Figure 5.27: Arbitrary triangle with circumcircle.

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 5.28.

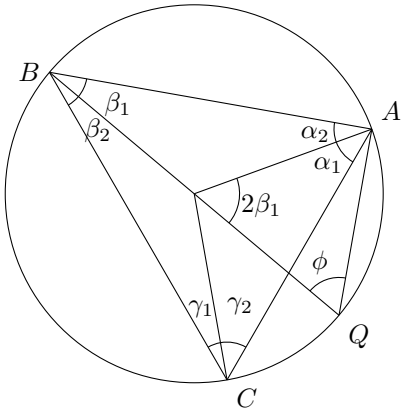


Figure 5.28: Arbitrary triangle with detailed circumcircle.

The angles α, β γ are now partitioned (not bisected) according to

$$\begin{aligned}\alpha &= \alpha_1 + \alpha_2 \\ \beta &= \beta_1 + \beta_2 \\ \gamma &= \gamma_1 + \gamma_2 \,.\end{aligned}$$

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 5.27, 5.28 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:

$$\frac{1}{2R} = \frac{\sin(\alpha)}{a} = \frac{\sin(\beta)}{b} = \frac{\sin(\gamma)}{c} \tag{5.83}$$

Taking Equations (5.82), (5.83) 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

$$\begin{aligned}z &\leq x + y \\ x &\leq y + z \\ y &\leq z + x\end{aligned}$$

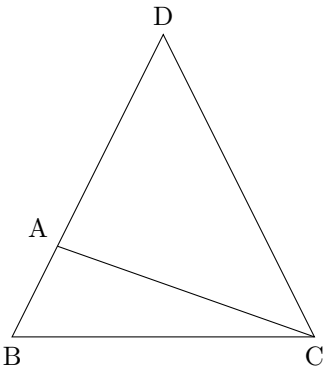


Figure 5.29: Triangle inequality.

To prove this, consider the triangle ABC depicted in Figure 5.29. 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.

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*:

$$x(r, \theta) = r\cos(\theta) \tag{5.84}$$

$$y(r, \theta) = r\sin(\theta) \tag{5.85}$$

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 5.30.

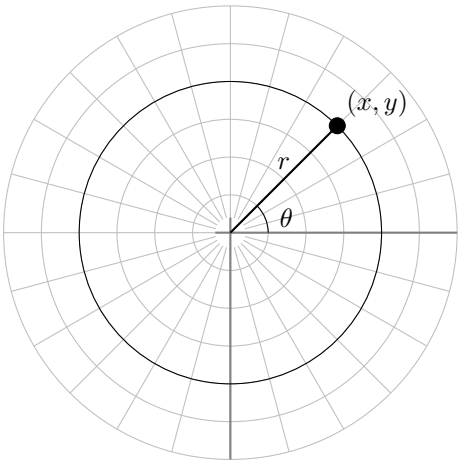


Figure 5.30: Polar coordinate system.

Radial and Angular Coordinates

In terms of x and y, the r- and θ-variables are straightforwardly isolated. Squaring Equations (5.84), polarcoordsy and taking their sum and square root gives the formula for r

$$r = \sqrt{x^2 + y^2}\,. \tag{5.86}$$

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:

$$\theta = \arctan\left(\frac{y}{x}\right) \tag{5.87}$$

As a consistency check, we should be able to apply $\sin()$ or $\cos()$ to both sides of Equation (5.87) to recover the x, y equations. Using the trig identities (5.65), (5.66), 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 Straight 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 (5.84), (5.85) for polar coordinates and the same line becomes

$$r = \frac{b}{\sin(\theta) - m\cos(\theta)}\,. \tag{5.88}$$

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:

$$r = \frac{b}{\sqrt{1+m^2}}\csc(\theta - \arctan(m))\,. \tag{5.89}$$

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 (5.30), (5.31), 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.

7.4 Offset Circles

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 5.31.

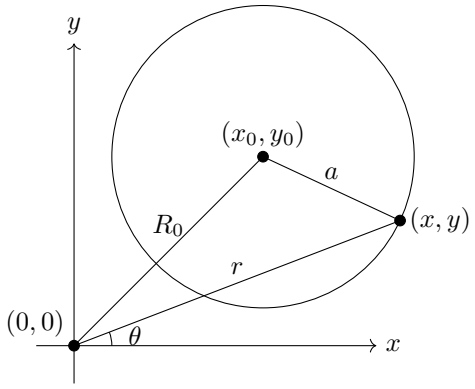


Figure 5.31: Offset circle.

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:

$$r^2 + R_0^2 - 2 r R_0 \cos(\theta - \theta_0) = a^2$$

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:

$$r = R_0 \cos(\theta - \theta_0) \pm \sqrt{a^2 - R_0^2 \sin^2(\theta - \theta_0)} \quad (5.90)$$

7.5 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 5.32.

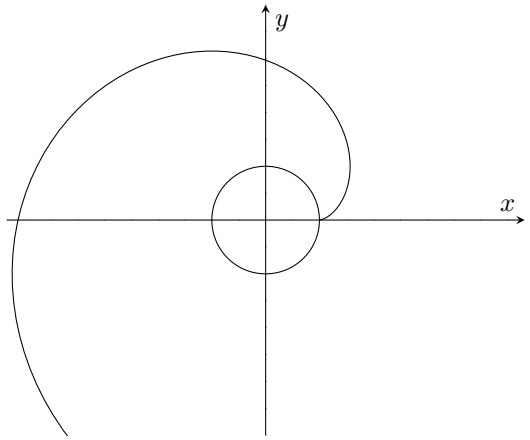


Figure 5.32: The involute.

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 $(\pi/2 : \pi/2)$.

8 Lissasjous Curves

...

Chapter 6

Conic Sections

1 Ellipse

1.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labled *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

$$\frac{R}{Q} = e < 1 \,. \tag{6.1}$$

Sketched in Figure 6.1 are some of the points that obey such a rule.

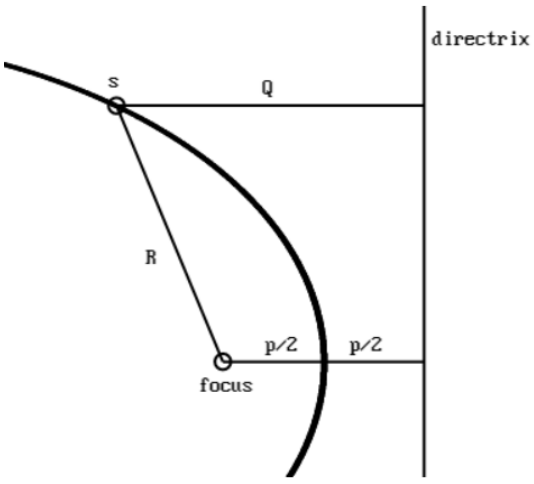


Figure 6.1: Points obeying $R/Q = e < 1$ as measured from a directrix and a focus separated by distance p. The origin is at the focus.

To determine the proper shape defined by the rule, begin with Equation (6.1) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \tag{6.2}$$

$$Q = p - x \tag{6.3}$$

Inserting the above into Equation (6.1) and completing the square in x, one finds

$$\frac{(x + c)^2}{a^2} + \frac{y^2}{b^2} = 1 \,, \tag{6.4}$$

describing an *ellipse* centered at $x = -c$. The *semimajor axis* a, *semiminor axis* b, and offset c relate to e, p by:

$$a = \frac{ep}{1 - e^2} \tag{6.5}$$

$$b = \frac{ep}{\sqrt{1 - e^2}} \tag{6.6}$$

$$c = ae \tag{6.7}$$

The ellipse has two *vertex* points at $(-c \pm a, 0)$, and two *covertex* points at $(-c, \pm b)$.

Problem 1

Derive Equation (6.4) simultaneously with Equations (6.5), (6.6), (6.7).

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

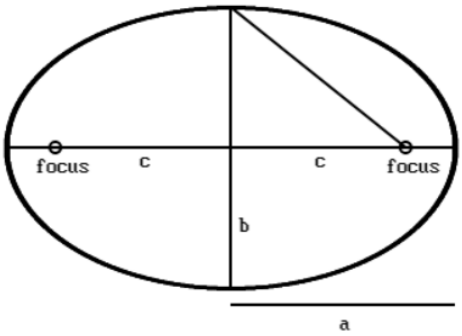


Figure 6.2: Ellipse displaying internal relations beteween a, b, c.

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 6.2. Note that the minor axis b is *never* greater than the major axis a.

Problem 2

Derive the internal relations:

$$a^2 - b^2 = c^2 \tag{6.8}$$

$$e = \sqrt{1 - \frac{b^2}{a^2}} \tag{6.9}$$

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. Prove this is equal to b^2/a.

1.4 Symmetry

Reflected Origin

We decided by writing Equation (6.2) 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 complimentary focus with its own directrix in the mirror image of the ellipse as shown in Figure 6.3. Should we wish to choose to rebuild using the 'left' focus as the origin, the resulting equation is complimentary to Equation (6.4), with the sign on c reversing:

$$\frac{(x-c)^2}{a^2} + \frac{y^2}{b^2} = 1$$

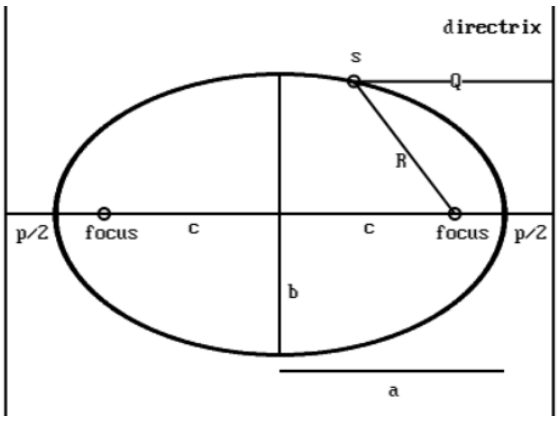


Figure 6.3: Vertical symmetry of the ellipse implies another directrix and focus.

1.5 Translations

Centered Origin

Placing the origin at the geometric center, the most symmetric equation of the ellipse reads

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\,. \tag{6.10}$$

Having no offset term, the focii are located symmetrically at $x = \pm c$.

Shifted Origin

An ellipse centered at the point (x_0, y_0) is represented by

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1\,. \tag{6.11}$$

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 focii. Answer: major $= 2\sqrt{5}$, minor $= 2$, center $= (2, 0)$, $e = 2/\sqrt{5}$, vertices $= \left(2, \pm 2\sqrt{5}\right)$, covertices $= (2 \pm 2, 0)$, focii $= (2, \pm 4)$

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 focii. Answer: major $= 2\sqrt{2}$, minor $= 2$, center $= (0, -1)$, $e = 1/\sqrt{2}$, vertices $= \left(\pm 2\sqrt{2}, -1\right)$, covertices $= (0, -1 \pm 2)$, focii $= (\pm 2, -1)$

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

$$x = r\cos(\theta)$$
$$y = r\sin(\theta)\ ,$$

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:

$$r = \sqrt{x^2 + y^2}$$
$$\theta = \arctan\left(\frac{y}{x}\right)$$

The definition (6.1) combined with Equations (6.2), (6.3) 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

$$r = \frac{pe}{1 + e\cos(\theta)} \,. \tag{6.12}$$

Equation (6.12) 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 (6.10):

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

By comparing the above to the fundamental identity of triginometry, namely

$$\cos(\phi)^2 + \sin(\phi)^2 = 1$$

for any angle ϕ, we cannot help but make the association

$$x = a\cos(\phi) \tag{6.13}$$
$$y = b\sin(\phi) \,. \tag{6.14}$$

Equations (6.13), (6.14) constitute a *parametric representation* of the ellipse.

Problem 9

Check that Equations (6.13), (6.14) combine to recover Equation (6.10).

1.8 Interior Identities

Sum of Radii

Consider a point (x, y) on an ellipse centered on the origin, and let

$$r_1 = \sqrt{(x + c)^2 + y^2}$$
$$r_2 = \sqrt{(x - c)^2 + y^2}$$

be the distance from (x, y) to each respective focus as shown in Figure 6.4. 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

$$A^2 = 2\left(x^2 + y^2 + c^2\right) + 2\sqrt{(x^2e^2 + a^2 + 2cx)\left(x^2e^2 + a^2 - 2cx\right)} \,,$$

simplifying further to

$$A^2 = 2\left(x^2e^2 + a^2\right) + 2\left(a^2 - x^2e^2\right) \,.$$

Performing the final cancelation, we find $A^2 = 4a^2$, or

$$r_1 + r_2 = 2a \,. \tag{6.15}$$

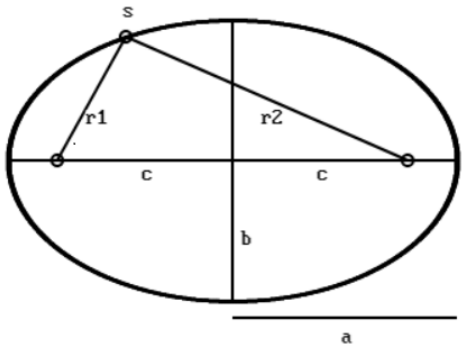


Figure 6.4: Line segments $r_{1,2}$ connect each focus to the point s on the ellipse. The sum $r_1 + r_2$ always equals the constant $2a$.

Drawing an Ellipse

The interior length identity (6.15) 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 (6.15).

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 (6.15) that the sum of the interior radii $r_1 + r_2$ in the ellipse always yields the constant $2a$. Natuarally 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

$$r_1 - r_2 = 2xe\,. \tag{6.16}$$

Problem 12
Derive Equation (6.16).

Decoupled Identities

Having Equations (6.15) and (6.16) in hand, we can isolate each of $r_{1,2}$ to yield a pair of tight formulas representing the ellipse:

$$r_1 = a + xe \tag{6.17}$$
$$r_2 = a - xe \tag{6.18}$$

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 6.6. The value of m_s is straightforwardly attained by implicit differentiation[1] of (6.10), which comes out to

$$m_s = \frac{-b^2}{a^2}\frac{x}{y}\,. \tag{6.19}$$

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 6.5.

[1] A trick from calculus.

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 (6.10) to eliminate y_1, y_2 such that

$$m_c = \frac{b\sqrt{1 - x_2^2/a^2} - b\sqrt{1 - x_1^2/a^2}}{x_2 - x_1}\,.$$

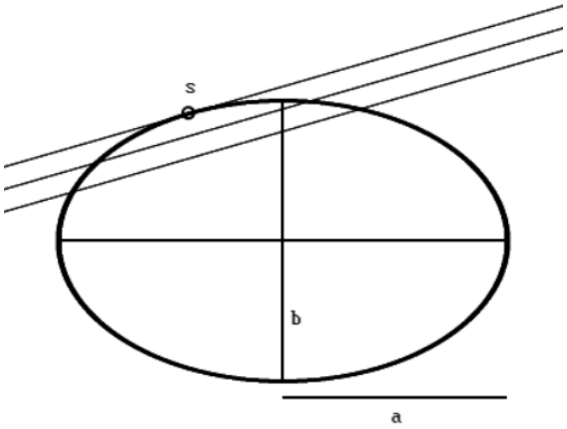


Figure 6.5: A chord cutting through the ellipse is constructed on the ellipse's edge so the endpoints share a common location s. The extension of the chord at s is the tangent line to the ellipse having slope m_s. The origin is at the center.

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

$$(x_1 + h)^2 \approx x_1^2 + 2x_1h + \cancel{h^2}\,.$$

The argument inside the square root shall be handled by the approximation

$$\lim_{z \text{ small}} \sqrt{1 + z} \approx 1 + \frac{z}{2} - \frac{z^2}{8} + \cdots\,, \tag{6.20}$$

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 (6.19) for the slope of the ellipse at point s.

1.10 Reflection Property

Consider an ellipse centered on the origin as shown in Figure 6.6, with respective focii 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:

$$\begin{aligned}\text{Angle } Asf_1 &= \theta\\ \text{Angle } Bsf_2 &= \phi\\ \text{Slope of } AB &= m_s\\ \text{Slope of } r_1 &= m_1\\ \text{Slope of } r_2 &= m_2\end{aligned}$$

The slopes m_1, m_2 are straightforward to write by inspsection of Figure 6.6:

$$\begin{aligned}m_1 &= \frac{y}{x+c}\\ m_2 &= \frac{y}{x-c}\end{aligned}$$

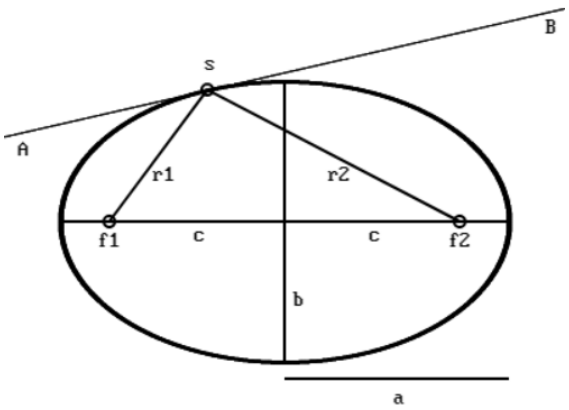


Figure 6.6: A ray emerging from one focus will reflect from the ellipse to the other focus.

Next, we'll need to use the angle-sum identity for tangent, namely

$$\tan(\alpha+\beta) = \frac{\tan(\alpha)\pm\tan(\beta)}{1\mp\tan(\alpha)\tan(\beta)}\,, \tag{6.21}$$

and observe again from the Figure that

$$\begin{aligned}\tan(\theta) &= \frac{m_1-m_s}{1+m_1m_s}\\ \tan(\phi) &= \frac{m_2-m_s}{1+m_2m_s}\,.\end{aligned}$$

Simplifying each expression delivers

$$|\tan(\theta)| = |\tan(\phi)| = \frac{b^2}{cy}\,, \tag{6.22}$$

telling us that $\theta = \phi$ and the proof is done.

Problem 14

Prove Equation (6.22).

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 6.7. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by (6.19). 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

$$y_q = y + (x - x_q)/m_s\,. \tag{6.23}$$

The normal line intersects the x-axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$\begin{aligned}q_0 &= x + m_sy = x + \left(\frac{-b^2}{a^2}\frac{x}{y}\right)y\\ q_0 &= x\left(1-\frac{b^2}{a^2}\right) = xe^2\end{aligned} \tag{6.24}$$

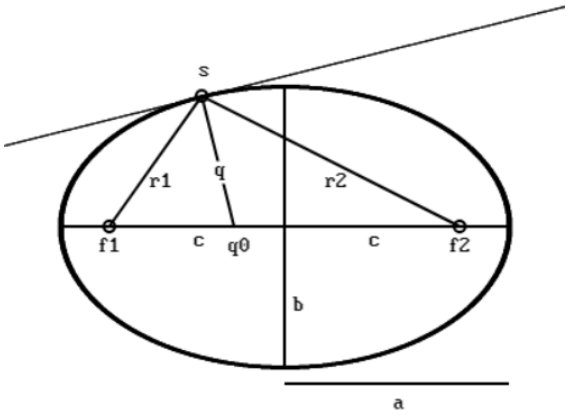


Figure 6.7: A point s on the ellipse implies a normal line q that intersects the x-axis at $x = q_0$. The origin is at the focus.

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

$$\frac{x}{\tilde{x}} - \frac{y}{\tilde{y}}\left(1 - e^2\right) = e^2 \, .$$

1.12 Vector Analysis (Optional)

Three Vectors

Refering again to Figure 6.7, let us construct three vectors from straight lines:

$$\begin{aligned} &\text{Line } f_1 s = \vec{r}_1 \\ &\text{Line } f_2 s = \vec{r}_2 \\ &\text{Line } q_0 s = \vec{q} \end{aligned}$$

Introducing the unit vector $\hat{x}$ that points horizontally to the right, we can jot down three ways to get from the origin to point s:

$$\vec{s} = -c\hat{x} + \vec{r}_1 \tag{6.25}$$
$$\vec{s} = c\hat{x} + \vec{r}_2 \tag{6.26}$$
$$\vec{s} = q_0\hat{x} + \vec{q} \tag{6.27}$$

Solving (6.25), (6.26) for $\vec{r}_1$, $\vec{r}_2$ respectively, we can divide each by its own magnitude to write unit vectors:

$$\hat{r}_1 = \frac{\vec{s} + c\hat{x}}{r_1}$$
$$\hat{r}_2 = \frac{\vec{s} - c\hat{x}}{r_2}$$

Recovering Reflection Property

In terms of the vectors $\vec{r}_1$, $\vec{r}_1$, $\vec{q}$, the reflection property of the ellipse can be proposed by writing

$$\hat{r}_1 \cdot \vec{q} = \hat{r}_2 \cdot \vec{q} \, , \tag{6.28}$$

which is to claim that the angle formed bewtween either $\hat{r}_j$ and $\vec{q}$ is the same. Proceeding carefully, one can simplify the above down to

$$\frac{b^2}{a}\left(\frac{\cancel{a + xe}}{r_1}\right) = \frac{b^2}{a}\left(\frac{\cancel{a - xe}}{r_2}\right) , \tag{6.29}$$

and the claim is proven. The parenthesized terms cancel due to Equations (6.17), (6.18). Evidently, we discover that the dot product between either unit vector $\hat{r}_{1,2}$ and the normal vector $\vec{q}$ equals a constant:

$$\hat{r}_{1,2} \cdot \vec{q} = \frac{b^2}{a} \tag{6.30}$$

Problem 17

Derive Equation (6.29) from (6.28).

Recovering Interior Length Identites

Next, we make use of the reflection property of the ellipse to realize that the sum of $\hat{r}_1$ and $\hat{r}_2$ must be parallel to (i.e. proportional to) the vector $\vec{q}$. Following this lead, we first take the sum

$$\hat{r}_1 + \hat{r}_2 = \vec{s}\left(\frac{1}{r_1} + \frac{1}{r_2}\right) + c\left(\frac{1}{r_1} - \frac{1}{r_2}\right)\hat{x} \, ,$$

simplifying down to

$$\hat{r}_1 + \hat{r}_2 = \left(\frac{r_1 + r_2}{r_1 r_2}\right)\left(\vec{s} + \frac{e}{2}\left(r_2 - r_1\right)\hat{x}\right) \tag{6.31}$$

Note that the interior length identity $r_1 + r_2 = 2a$ was used along the way. Comparing the above to Equation (6.27), we must have

$$q_0 = \frac{e}{2}\left(r_1 - r_2\right) \, . \tag{6.32}$$

Problem 18

Derive Equations (6.31), (6.32) and then recover Equation (6.16).

2 Hyperbola

2.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labled *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

$$\frac{R}{Q} = e > 1 \,. \tag{6.33}$$

Sketched in Figure 6.8 are some of the points that obey such a rule.

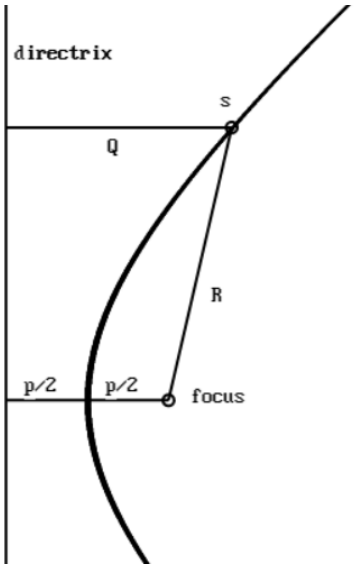


Figure 6.8: Points obeying $R/Q = e > 1$ as measured from a directrix and a focus separated by distance p. The origin is at the focus.

To determine the proper shape defined by the rule, begin with (6.33) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \tag{6.34}$$

$$Q = p + x \tag{6.35}$$

Inserting the above into (6.33) and completing the square in x, one finds

$$\frac{(x+c)^2}{a^2} - \frac{y^2}{b^2} = 1 \,, \tag{6.36}$$

describing a *hyperbola*. The constants a, b, c relate to e, p by:

$$a = \frac{ep}{e^2 - 1} \tag{6.37}$$

$$b = \frac{ep}{\sqrt{e^2 - 1}} \tag{6.38}$$

$$c = ae \tag{6.39}$$

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 (6.36) roughly reads

$$\frac{x^2}{a^2} \approx \frac{y^2}{b^2} \,,$$

implying a pair of straight lines having slope

$$m_\pm = \pm\frac{b}{a} \tag{6.40}$$

that are *asymptopes* to the hyperbola.

To find an exact equation for each asymptote, begin with the equation of the hyperbola (6.36) 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 6.8 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

$$y = \pm\frac{b}{a}\,(x + c) \,. \tag{6.41}$$

Our findings are summarized in Figure 6.9.

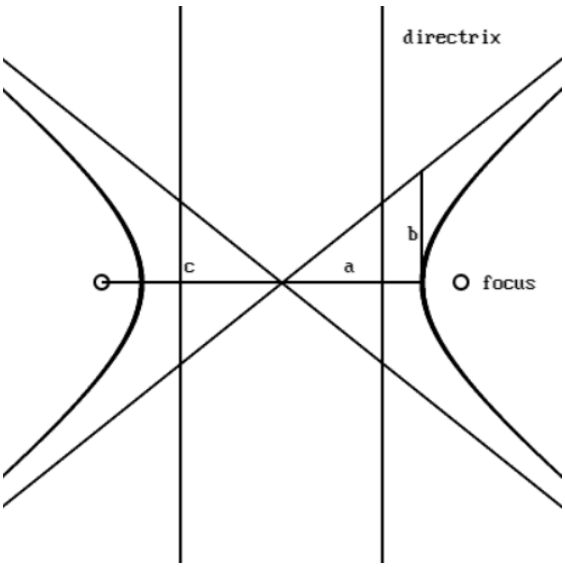


Figure 6.9: Vertical symmetry of the hyperbola implies another directrix and focus.

2.3 Internal Relations

Problem 19

Derive the internal relations:

$$a^2 + b^2 = c^2 \tag{6.42}$$

$$e = \sqrt{\frac{b^2}{a^2} + 1} \tag{6.43}$$

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. Sbow that c is the distance from the focus to the intersection of the asymptotes.

2.4 Symmetry

Reflected Origin

We decided by wtiring Equation (6.34) 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 6.9. Should we wish to choose to rebuild using the 'left' focus as the origin, the resulting equation is complimentary to (6.36), 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

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1\,. \tag{6.44}$$

Having no offset term, the focii are located symmetrically at $x = \pm c$.

Shifted Origin

A hyperbola centered at the point (x_0, y_0) is represented by

$$\frac{(x-x_0)^2}{a^2} - \frac{(y-y_0)^2}{b^2} = 1\,. \tag{6.45}$$

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 focii. Answer: major = 4, minor = 3, center = $(0,0)$, $e = 5/4$, vertices = $(0, \pm 3)$, $y = \pm 3x/4$, focii = $(0, \pm 5)$

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 focii. Answer: major = $\sqrt{3}$, minor = $2\sqrt{2}$, center = $(1/2, 3/2)$, $e = \sqrt{11/3}$, vertices = $\left(1/2, 3/2 \pm \sqrt{3}\right)$, $y = \pm\sqrt{3/8}x \mp \sqrt{3/32} + 3/2$, focii = $\left(1/2, 3/2 \pm \sqrt{11}\right)$

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 $\left(6, \sqrt{15}\right)$. 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

$$x = r\cos(\theta)$$
$$y = r\sin(\theta)\ ,$$

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:

$$r = \sqrt{x^2 + y^2}$$
$$\theta = \arctan\left(\frac{y}{x}\right)$$

The definition (6.33) combined with (6.34), (6.35) 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

$$r = \frac{pe}{1 - e\cos(\theta)}\ . \tag{6.46}$$

Equation (6.46) 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 (6.44):

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

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

$$x = a\cosh(\phi) \tag{6.47}$$
$$y = b\sinh(\phi)\ . \tag{6.48}$$

Equations (6.47), (6.48) constitute a parametric representation of the hyperbola.

Problem 25

Check that Equations (6.47), (6.48) combine to recover Equation (6.44).

2.8 Interior Identities

Difference of Radii

Consider a point (x, y) on a hyperbola centered on the origin, and let

$$r_1 = \sqrt{(x + c)^2 + y^2}$$
$$r_2 = \sqrt{(x - c)^2 + y^2}$$

be the distance from (x, y) to each respective focus as shown in Figure 6.10. 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

$$A^2 = 2\left(x^2 + y^2 + c^2\right) - 2\sqrt{\left(x^2e^2 + a^2 + 2cx\right)\left(x^2e^2 + a^2 - 2cx\right)}\ ,$$

simplifying further to

$$A^2 = 2\left(x^2e^2 + a^2\right) + 2\left(a^2 - x^2e^2\right)\ .$$

Performing the final cancellation, we find $A^2 = 4a^2$, or

$$r_1 - r_2 = 2a\ . \tag{6.49}$$

Problem 26

Derive Equation (6.49).

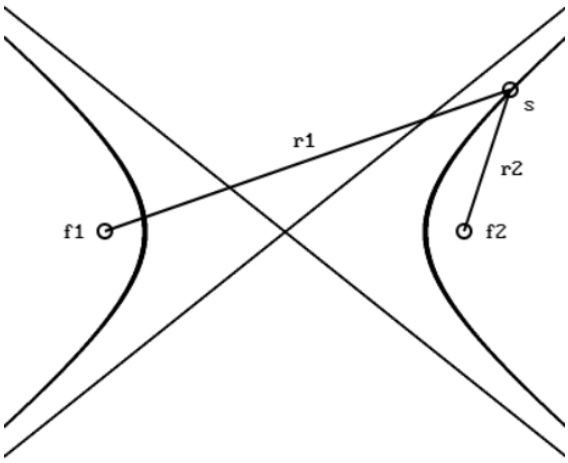


Figure 6.10: Line segments $r_{1,2}$ connect each focus to the point s on the hyperbola. The difference $r_1 - r_2$ always equals a constant $2a$.

Locating Ships

The interior length identity (6.49) 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 (6.49) that the difference of the interior radii $r_1 - r_2$ in the hyperbola always yields the constant $2a$. Natuarally 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

$$r_1 + r_2 = 2xe\,. \tag{6.50}$$

Problem 27

Derive Equation (6.50).

Decoupled Identities

Having Equations (6.49) and (6.50) in hand, we can isolate each of $r_{1,2}$ to yield a pair of tight formulas representing the hyperbola:

$$r_1 = a + xe \tag{6.51}$$
$$r_2 = -a + xe \tag{6.52}$$

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 6.11. The value of m_s is straightforwardly attained by implicit diffentiation of Equation (6.44), which comes out to

$$m_s = \frac{b^2}{a^2}\frac{x}{y}\,. \tag{6.53}$$

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

$$\frac{x\tilde{x}}{a^2} - \frac{y\tilde{y}}{b^2} = 1\,.$$

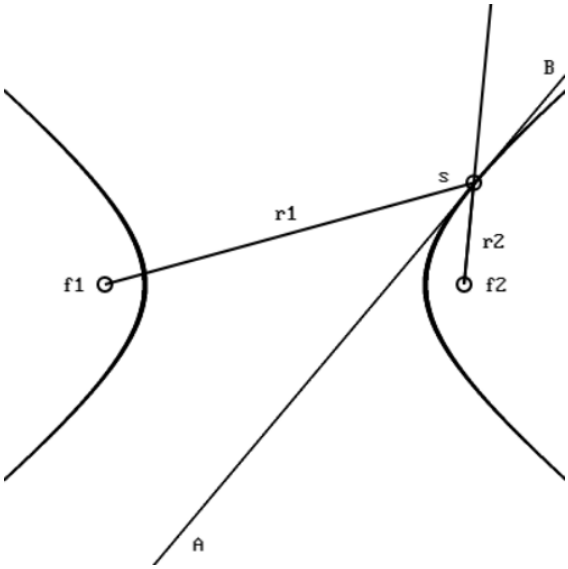


Figure 6.11: An incoming ray aimed at f_2 intersects the hyperbola at $s = (x, y)$. The reflected ray goes toward f_1. The angle formed between the incoming ray and the tangent line AB is is identical to angle Asf_2. The origin is at the focus.

2.10 Reflection Property

Consider a hyperbola centered on the origin as shown in Figure 6.11, with respective focii 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:

$$\begin{aligned}\text{Angle } Asf_1 &= \theta\\ \text{Angle } Asf_2 &= \phi\\ \text{Slope of } AB &= m_s\\ \text{Slope of } r_1 &= m_1\\ \text{Slope of } r_2 &= m_2\end{aligned}$$

The slopes m_1, m_2 are straightforward to write by inspsection of Figure 6.11:

$$m_1 = \frac{y}{x + c}$$
$$m_2 = \frac{y}{x - c}$$

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

$$\tan(\theta) = \frac{m_1 - m_s}{1 + m_1 m_s}$$
$$\tan(\phi) = \frac{m_2 - m_s}{1 + m_2 m_s} .$$

Simplifying each expression delivers

$$|\tan(\theta)| = |\tan(\phi)| = \frac{b^2}{cy} , \qquad (6.54)$$

telling us that $\theta = \phi$ and the proof is done.

Problem 29

Prove Equation (6.54).

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 6.12. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by Equation (6.53). 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

$$y_q = y + (x - x_q)/m_s . \qquad (6.55)$$

The normal line intersects the x-axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$q_0 = x + m_s y = x + \left(\frac{b^2}{a^2}\frac{x}{y}\right) y$$
$$q_0 = x\left(1 + \frac{b^2}{a^2}\right) = xe^2 \qquad (6.56)$$

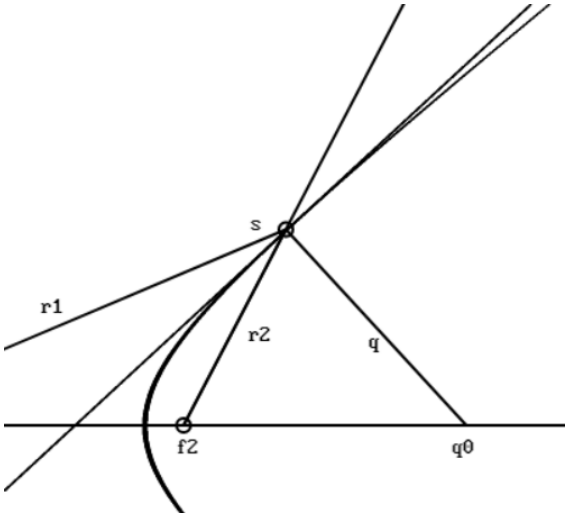


Figure 6.12: A point s on the hyperbola implies a normal line q that intersects the x-axis at $x = q_0$. The origin is at the focus.

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

$$\frac{x}{\tilde{x}} + \frac{y}{\tilde{y}}\left(e^2 - 1\right) = e^2 .$$

3 Parabola

3.1 Definition

In the Cartesian plane, consider a point labeled *focus* that is distance p from a vertical line labled *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

$$\frac{R}{Q} = 1 . \qquad (6.57)$$

Sketched in Figure 6.13 are some of the points that obey such a rule.

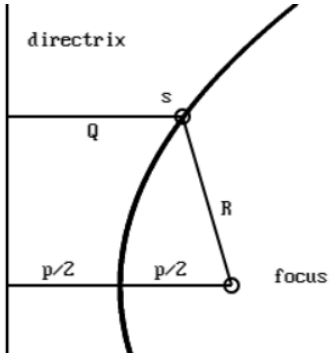


Figure 6.13: Points obeying $R/Q = 1$ as measured from a directrix and a focus separated by distance p. The origin is at the focus.

To determine the proper shape defined by the rule, begin with (6.57) and discern from inspection that R, Q can be written:

$$R = \sqrt{x^2 + y^2} \tag{6.58}$$

$$Q = p + x \tag{6.59}$$

Inserting the above into (6.57), one finds

$$y^2 = p^2 + 2px\,, \tag{6.60}$$

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 6.14:

$$x^2 = p^2 + 2py \tag{6.61}$$

The sign on the p-term determines the opening direction in either case.

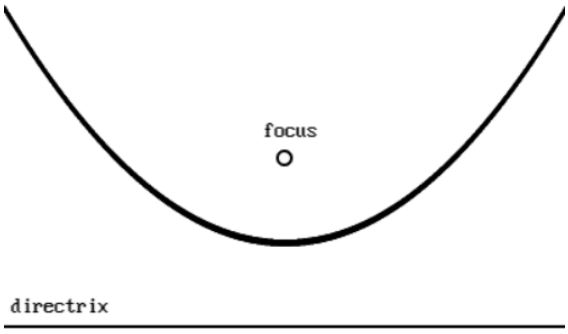


Figure 6.14: The directrix-focus construction rotated by ninety degrees produces an up- (or down-) opening parabola.

3.3 Parabolic Expressions

In Equations (6.60), (6.61), the value p^2 acts to translate the parabola vertically or horzontally. The factor of $2p$ 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 (6.60) is more generally written as

$$x = x_0 + a\,(y - y_0)^2\ . \tag{6.62}$$

By the same token, Equation (6.61) is more generally written as

$$y = y_0 + a\,(x - x_0)^2\ . \tag{6.63}$$

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: $\big(-b/2a, c - b^2/4a\big)$, $1/4a$ above the vertex

3.4 Polar Representation

In polar coordinates, recall that a point (x, y) in the Cartesian plane is represented by

$$x = r\cos(\theta)$$
$$y = r\sin(\theta) \ ,$$

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:

$$r = \sqrt{x^2 + y^2}$$
$$\theta = \arctan\left(\frac{y}{x}\right)$$

The definition (6.57) combined with Equations (6.58), (6.59) 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

$$r = \frac{p}{1 - \cos(\theta)} \ . \tag{6.64}$$

Equation (6.64) traces a parabola in the plane.

3.5 Internal Relations

Right Focal Chord

Problem 35

Prove 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 6.15. The value of m_s is straightforwardly attained by implicit diffentiation of Equation (6.60), which comes out to

$$m_s = \frac{p}{y} \ . \tag{6.65}$$

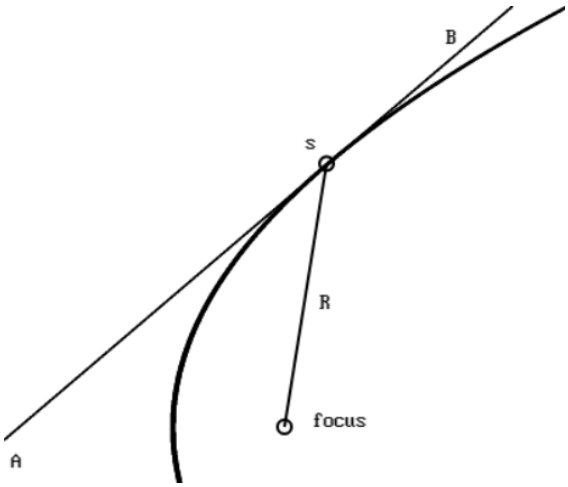


Figure 6.15: A point s on the parabola implies a tangent line AB that represents the instantaneous slope of the parabola. The origin is at the focus.

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 6.16. The slope of the normal line is defined as the negative reciprocal of the tangent's slope, namely $-1/m_s$ given by (6.65). 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

$$y_q = y + (x - x_q)/m_s \ . \tag{6.66}$$

The normal line intersects the x-axis at the point $x_q = q_0$, which we determine by setting $y_q = 0$:

$$q_0 = x + m_s y = x + p \tag{6.67}$$

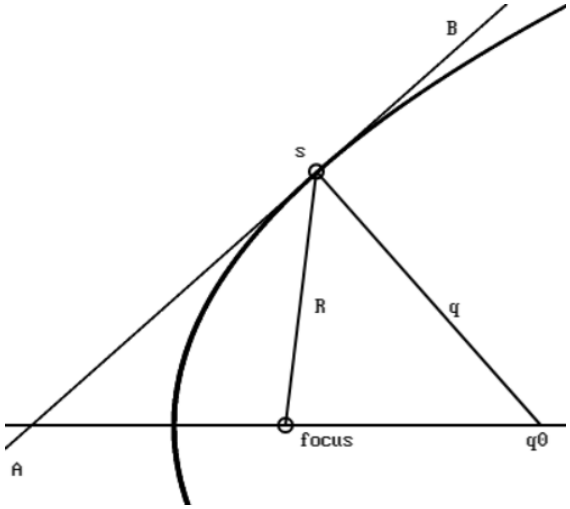


Figure 6.16: A point s on the parabola implies a normal line q that intersects the x-axis at $x = q_0$. The origin is at the focus.

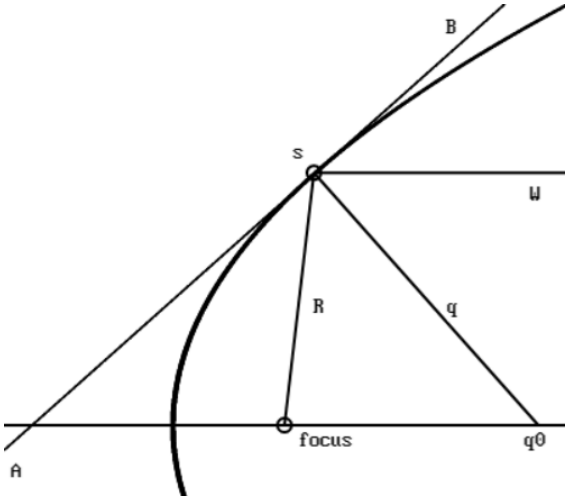


Figure 6.17: A ray emitted from the focus will intersect the parabola at $s = (x, y)$ and reflect parallel to the axis along line sW.

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 6.17, 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.*

To proceed, refer to Figure 6.17 to define:

$$\begin{aligned} \text{Angle } Asf &= \alpha_1 \\ \text{Angle } BsW &= \alpha_2 \\ \text{Angle } fsq_0 &= \beta_1 \\ \text{Angle } Wsq_0 &= \beta_2 \\ \text{Line } sW &\propto \hat{x} \\ \text{Slope of } AB &= m_s \end{aligned}$$

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 $\vec{q}$ that points from q_0 to s:

$$\vec{q} = \vec{s} - q_0 \, \hat{x} \tag{6.68}$$

Note that $\vec{s}$ is equivalent to the position vector $\vec{r}$, 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 $\vec{s}$ projected onto $\vec{q}$ has to equal that of the unit vector $-\hat{x}$ projected onto $\vec{q}$:

$$q \cos(\beta_1) = \hat{r} \cdot \vec{q} = -\hat{x} \cdot \vec{q} = q \cos(\beta_2) \tag{6.69}$$

Substituting $\vec{q}$ from Equation (6.68), we have

$$\hat{r} \cdot (\vec{s} - q_0 \, \hat{x}) = -\hat{x} \cdot (\vec{s} - q_0 \, \hat{x}) \ ,$$

boiling down to

$$r - \cancel{\left(\frac{x + p}{r} \right)} x = -x + q_0 \, .$$

On the right, the quantity $-x+q_0$ is simply the constant p according to Equation (6.67). The parenthesized quantity on the left resolves to *one* according to definition (6.57), bringing the result to

$$r - x = p\,. \tag{6.70}$$

To check that Equation (6.70) is true we employ polar coordinates, $x = r\cos(\theta)$, and the above quickly resolves to the polar representation (6.64) of the parabola, finishing the proof.

Problem 40

Derive Equation (6.70) from Equation (6.69).

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 (6.21) 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:

$$0 = m_s^2 + 2\frac{x}{y}m_s - 1 \tag{6.71}$$

Complete the square in m_s and then solve for m_s. The result boils down to

$$m_s = \frac{p}{y}\,,$$

the formula (6.65) 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 (6.65) from Equation (6.71).

Differential Analysis

Starting from Equation (6.71), multuiply through by y^2 and make the subsitution

$$y^2 = r^2 - x^2\,,$$

where implicit differentiation tells us

$$ym_s = r\frac{dr}{dx} - x\,.$$

With this, Equation (6.71) 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 (6.70), 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

$$x^2 + y^2 = \alpha z^2\,, \tag{6.72}$$

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

$$1 = \frac{x}{x_0} + \frac{z}{z_0}\,. \tag{6.73}$$

Conspiculously absent from Equation (6.73) is any representation of the y-variable. Due to the axial symmety of the cone, one can always choose a coordinate system where the horizontal coodinate on the 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:

$$u\cos(\theta) = x_0 - x \tag{6.74}$$
$$u\sin(\theta) = z \tag{6.75}$$
$$v = y \tag{6.76}$$
$$\tan(\theta) = \frac{z_0}{x_0} \tag{6.77}$$

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

$$1 = \frac{u^2}{z_0^2}\sin^2(\theta)\left(\alpha\tan^2(\theta) - 1\right) + 2\frac{u}{z_0}\sin(\theta) - \tan^2(\theta)\frac{v^2}{z_0^2} \tag{6.78}$$

and furthermore, after a page of algebra:

$$1 = \left(\frac{u}{z_0}\cos(\theta)\frac{\left(1-\alpha\tan^2(\theta)\right)}{\sqrt{\alpha}} - \frac{\cot(\theta)}{\sqrt{\alpha}}\right)^2 + \frac{v^2}{z_0^2}\left(\frac{1-\alpha\tan^2(\theta)}{\alpha}\right) \tag{6.79}$$

Problem 42

Derive Equation (6.78) and Equation (6.79).

Gamma Factor

To help tame the symbolic explosion that has occurred, let us introduce yet another symbol γ such that

$$\gamma = 1 - \alpha\tan^2(\theta)\ . \tag{6.80}$$

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 (6.78) 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 (6.79), it turns out that the γ factor 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:

$$1 = \left(\frac{u}{z_0}\cos(\theta)\frac{\gamma}{\sqrt{\alpha}} - \frac{\cot(\theta)}{\sqrt{\alpha}}\right)^2 + \frac{v^2}{z_0^2}\left(\frac{\gamma}{\alpha}\right) \tag{6.81}$$

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

$$e = \sqrt{1 - \frac{b^2}{a^2}} = \sqrt{1 - \gamma\cos^2(\theta)} = \sin(\theta)\sqrt{1+\alpha}\,. \tag{6.82}$$

Similarly, the eccentricity of the hyperobla is

$$e = \sqrt{1 + \frac{b^2}{a^2}} = \sqrt{1 + \gamma\cos^2(\theta)} = \sqrt{2 - \sin^2(\theta)(1+\alpha)}\,. \tag{6.83}$$

Problem 43

Derive Equation (6.82) and Equation (6.83).

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 distict 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:

$$\begin{array}{cc} & \text{eccentricity} \\ \text{ellipse} & e = \sqrt{1 - b^2/a^2} < 1 \\ \text{hyperbola} & e = \sqrt{1 + b^2/a^2} > 1 \\ \text{parabola} & e = 1 \end{array}$$

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

$$Ax^2 + Bxy + Cy^2 + Dx + Ey = F\,. \tag{6.84}$$

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 (6.84), 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 shame 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 counterclockwise 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

$$\begin{aligned} &Ax^2 + Bxy + Cy^2 \\ &\quad = u^2\left(A\cos^2(\theta) + C\sin^2(\theta) + B\sin(\theta)\cos(\theta)\right) \\ &\quad + v^2\left(A\sin^2(\theta) + C\cos^2(\theta) - B\sin(\theta)\cos(\theta)\right) \\ &\quad + uv\left(-A\sin(2\theta) + C\sin(2\theta) + B\cos(2\theta)\right) \end{aligned}$$

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

$$\tan(2\theta) = \frac{B}{A - C}\,. \tag{6.85}$$

So far then, we can write

$$A'u^2 + C'v^2 + D'u + E'v = F\,, \tag{6.86}$$

where the primed coeffcients 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: $2\theta = \tan^{-1}\left(\sqrt{3}\right)$, $u^2/9 + v^2/4 = 1$)

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

$$v = \left(\frac{a^2 + b^2}{a^2 - b^2}\right)u \pm \frac{ab}{a^2 - b^2}\sqrt{4u^2 - 2\left(a^2 - b^2\right)}\,.$$

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 (6.86) 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 Discrimanant of a Conic

It is possible to determine the type of curve described by the general Equation (6.84) without manually rotating coordinates. To do this, we must calculate the *discrimanant* of the conic, defined by

$$\mathcal{D} = B^2 - 4AC\,. \tag{6.87}$$

It turns out that $\mathcal{D}$ resolves to the same value regardless of the rotation angle of the coordiante 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:

$$\begin{aligned}(B')^2 &= \left(A^2 + C^2 - 2AC\right)\sin^2(2\theta)\\ &\quad + B^2\cos^2(2\theta) + B\,(C - A)\sin(4\theta)\\ -4A'C' &= \left(-A^2 - C^2 + B^2\right)\sin^2(2\theta)\\ &\quad - B\,(C - A)\sin(4\theta)\\ &\quad - 4AC + 2AC\sin^2(2\theta)\end{aligned}$$

Taking the sum of the two results, we see that all of the ugly terms cancel out and the form $B^2 - 4AC$ emerges:

$$(B')^2 - 4A'C' = B^2 - 4AC \tag{6.88}$$

More succintly, we see $\mathcal{D}' = \mathcal{D}$ for any angle θ. Not surprisingly, the coefficients D, D', E, E', F, are not involved in the discrimanant or the classification of the curve.

Problem 49

Derive Equation (6.88).

5.5 Classifying General Conics

Having shown that the discrimanant $B^2 - 4AC$ 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 (6.85) to write

$$\mathcal{D} = B^2 - 4AC = -4A'C'\,. \tag{6.89}$$

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.

	discrimanant
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 (6.89) to express the same hyberbola 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 (6.85), 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 (6.89)

$$B^2 - 4AC = -4A'C'$$

to replace the primed terms with the original variables and the problem is finished:

$$\text{Area} = \frac{\pi}{\sqrt{4AC - B^2}}$$

Chapter 7

Taylor Polynomial

1 Introduction

Students are almost universally exposed, at one point or another, to simple concepts of motion, a study often called *kinematics*. Kinematics is a careful accounting of the *position*, x, of some object (or several objects) as a function of time t. Governing the evolution of the position is the *velocity*, $v(t)$, which is itself governed by the *acceleration*, $a(t)$.

1.1 Constant Acceleration

To keep things simple, a study of kinematics often limits the acceleration to be constant, or *uniform*, which is convenient for describing many systems, including freefall motion near Earth's surface, or the motion of charges in a uniform electric field. In such a case, the student is provided with a hierarchy of kinematic formulas:

$$\begin{aligned} x(t) &= x_0 + v_0 t + \frac{1}{2}at^2 \\ v(t) &= v_0 + at \\ a(t) &= \text{constant} \end{aligned} \tag{7.1}$$

The initial values x_0, v_0 correspond to the position and velocity at time $t = 0$.

Kinematic Identities

The standard kinematic formulas are reinforced by a flurry of kinematic identities

$$\begin{aligned} x(t) &= x_0 + \overline{v}t \\ x(t) &= x_0 + v(t)\,t - \frac{1}{2}at^2 \\ (v(t))^2 &= v_0^2 + 2a\Delta x\,, \end{aligned}$$

where

$$\begin{aligned} \overline{v}(t) &= \frac{v_0 + v(t)}{2} \\ \Delta x &= x(t) - x_0\,. \end{aligned}$$

Problem 1

Use $x(t) = x_0 + \overline{v}t$ and $v(t) = v_0 + at$ to derive Equation (7.1).

Position Plot

The kinematic equation for position $x(t)$ is *quadratic* in the variable t, thus it's of interest to complete the square in t and write the position x as

$$x(t) = \left(x_0 - \frac{v_0^2}{2a}\right) + \frac{a}{2}\left(t + \frac{v_0}{a}\right)^2\,. \tag{7.2}$$

The vertex of the motion, occurring at $(t_{\text{vert}}, x_{\text{vert}})$, is calculated by setting $t = -v_0/a$, giving

$$(t_{\text{vert}}, x_{\text{vert}}) = \left(\frac{-v_0}{a}, x_0 - \frac{v_0^2}{2a}\right)\,.$$

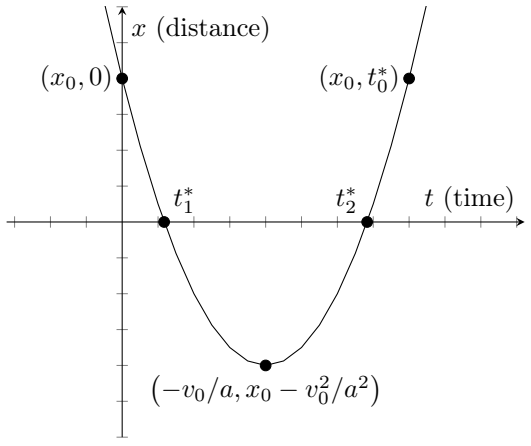
There exists a condition for which the position returns to $x = x_0$, given by

$$t_0^* = \frac{-2v_0}{a}\,.$$

Note that if v_0/a resolves to a positive number, the above condition is not met for positive time values. We may also determine the t-intercepts, corresponding to the point(s) satisfying $x = 0$:

$$t_{1,2}^* = \frac{v_0}{a}\left(1 \pm \sqrt{1 - \frac{2x_0 a}{v_0^2}}\right)$$

The summary of our findings is contained in the following graph, choosing $x_0 > 0$, $v_0 < 0$, $a > 0$ for the sake of demonstration:

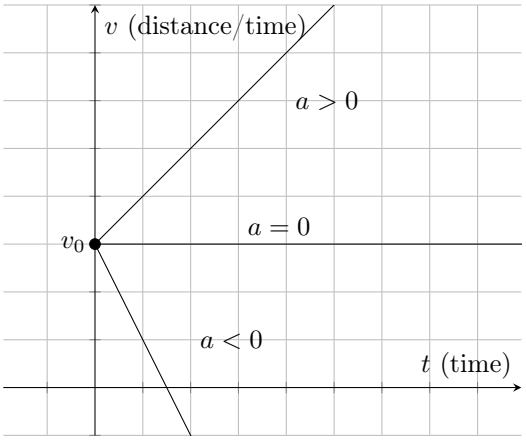


Problem 2

Derive Equation (7.2) from Equation (7.1) and verify the formulas for the vertex and t-intercepts.

Velocity Plot

When the acceleration is uniform, the plot representing $v\,(t)$ is that of a straight line. The slope of the line is defined as the acceleration. Shown below is a single graph with several lines representing various trajectories of a common initial velocity.



2 Uniform Jerk and Beyond

Inevitably during a study of kinematics, one wonders how things change if acceleration is itself allowed to vary with time, a phenomenon called *jerk*.

2.1 Identities

In the case of uniform jerk, represented by j, we may write

$$a\,(t) = a_0 + jt\,,$$

perfectly analogous to $v = v_0 + at$ in the constant-acceleration regime. In the same analogy, certain 'acceleration-jerk' identities can be written, for instance:

$$\begin{aligned} v\,(t) &= v_0 + a_0 t + \frac{1}{2} j t^2 \\ a\,(t)^2 &= a_0^2 + 2j\Delta v \end{aligned}$$

2.2 Time-Shift Analysis

All is well until we try to come up with an equation for position $x\,(t)$. Going by pattern alone, it seems that the new 'jerk' term will depend on t^3, but we can't be sure which coefficient to write. Putting this uncertainty into the unknown coefficient A, we have:

$$x\,(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3$$

To proceed, introduce a shift of time such that

$$t \to t + h\,,$$

where h can be of any value. Inserting this into the above gives

$$\begin{aligned} x\,(t+h) = x_0 &+ v_0\,(t+h) \\ &+ \frac{1}{2} a_0\,(t+h)^2 + \frac{1}{A} j\,(t+h)^3\ , \end{aligned}$$

and now the job is to expand all factors involving $(t+h)$. Doing so, and then combining like terms in powers of h, results in something interesting:

$$\begin{aligned} x\,(t+h) = &\left(x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{A} j t^3 \right) \\ &+ h \left(v_0 + a_0 t + \frac{3}{A} j t^2 \right) \\ &+ \frac{1}{2} h^2 \left(a_0 + \frac{6}{A} j t \right) + \frac{1}{6} h^3\,(j) \end{aligned}$$

From this, we see the *only* way to correctly recover the identities already written is to have

$$A = 6\,,$$

and no other choice suffices.

2.3 Time-Shifted Kinematics

To tighten up the analysis above, define new coefficients of motion x_t, v_t, and a_t such that:

$$\begin{aligned} x_t &= x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{6} j t^3 \\ v_t &= v_0 + a_0 t + \frac{1}{2} j t^2 \\ a_t &= a_0 + jt \end{aligned}$$

Then, the time-shifted position $x\,(t+h)$ can be written in condensed form that buries the explicit t-dependence in favor of h:

$$x\,(t+h) = x_t + v_t h + \frac{1}{2} a_t h^2 + \frac{1}{6} j h^3 \qquad (7.3)$$

This result kills two birds with one stone. Firstly, we arrive at a fully-adjustable equation of kinematics with any t as the 'initial' time value, shifting the burden of evolution to h.

Secondly, we see the additional 'uniform jerk term' in the role of kinematics arrives unambiguously as $(1/6)\,jt^3$:

$$x\,(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{6} j t^3$$

Inverse Relations

It's worthwhile to take the time-shifted kinematic identities for x_t, v_t, etc., and solve instead for x_0, v_0, etc., thereby *inverting* the set of equations. Starting with the a-terms and working back, we find:

$$\begin{aligned} a_0 &= a_t - jt \\ v_0 &= v_t - a_t t + \frac{1}{2} j t^2 \\ x_0 &= x_t - v_t t + \frac{1}{2} a_t t^2 - \frac{1}{6} j t^3 \end{aligned}$$

That is, the inverted version differs by the original up to a minus sign on the effective time variable.

2.4 Uniform Snap

Having cracked the problem of uniform jerk, one wonders next what happens if jerk is allowed to vary in time, a situation describing *snap*. Indeed, if we introduce a uniform snap constant k, we have

$$j(t) = j_0 + kt \,,$$

and the whole argument repeats. Then, there must be some fourth-order correction to the position such that

$$x(t) = x_0 + v_0 t + \frac{1}{2} a_0 t^2 + \frac{1}{6} j_0 t^3 + \frac{1}{B} k t^4 \,.$$

Problem 3

Use time-shift analysis to figure out $B = 24$, and then appropriately append x_t, x_0, v_t, v_0, etc.

2.5 Pattern of Coefficients

Looking at the equation $x(t)$ and the coefficient accompanying each term, we can't help but try to see a pattern to these. Each coefficient originates from expanding powers of $(t+h)^n$, discovered using either brute force or referring to Pascal's triangle.

Delving into polynomial expansion, one inevitably discovers the handiness of the *factorial* operator, (!) which is a way to express the product of descending integers starting with N:

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

with the limit case

$$0! = 1 \,.$$

With this, the coefficients in the kinematic equation for $x(t)$ can be written in tighter fashion:

$$x(t) = x_0 + v_0 t + \frac{1}{2!} a_0 t^2 + \frac{1}{3!} j_0 t^3 + \frac{1}{4!} k t^4$$

3 Change of Base Point

To explore an application of time-shift analysis, consider a trajectory characterized by coefficients

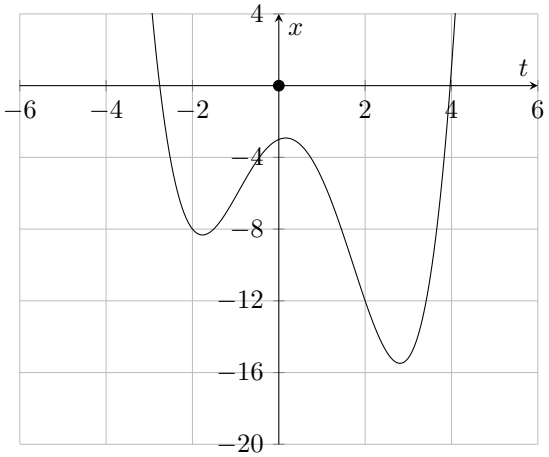
$$x_0 = -3 \qquad v_0 = 1 \qquad \frac{a_0}{2} = -3$$

$$\frac{j_0}{3!} = \frac{-1}{2} \qquad \frac{k}{4!} = \frac{5}{16} \,,$$

such that

$$x(t) = -3 + t - 3t^2 - \frac{1}{2} t^3 + \frac{5}{16} t^4 \,,$$

plotted as follows:



As written, the equation for $x(t)$ is suited such that the 'origin in time' is $t = 0$.

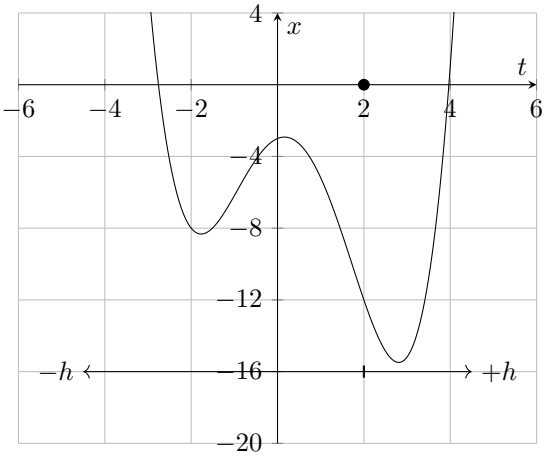
A *change of base point* is analogous to choosing a new origin, i.e. to replace the point that does $t = 0$'s job with something else. In light of time-shift analysis, this amounts to letting, say, $t_p = 2$, and then letting h do the evolution as the effective time variable. The hard work entails calculating the coefficients x_t, v_t, etc.:

$$\begin{aligned} x_t &= \left(x_0 + v_0 t + \frac{a_0}{2} t^2 + \frac{j_0}{6} t^3 + \frac{k}{24} t^4 \right) \Big|_2 = -12 \\ v_t &= \left(v_0 + a_0 t + \frac{j_0}{2} t^2 + \frac{k_0}{6} t^3 \right) \Big|_2 = -7 \\ a_t &= \left(a_0 + j_0 t + \frac{k_0}{2} t^2 \right) \Big|_2 = 3 \\ j_t &= (j_0 + k_0 t) \Big|_2 = 12 \\ k &= \frac{15}{2} \end{aligned}$$

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:



3.1 Reverting

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, use symbolic apparatus to recover the same result using inverse relations

$$\begin{aligned}
j_0 &= (j_t - kt)\Big|_2 \\
a_0 &= \left(a_t - j_t t + \frac{1}{2}kt^2\right)\Big|_2 \\
v_0 &= \left(v_t - a_t t + \frac{1}{2}j_t t^2 - \frac{1}{6}kt^3\right)\Big|_2 \\
x_0 &= \left(x_t - v_t t + \frac{1}{2}a_t t^2 - \frac{1}{6}j_t t^3 + \frac{1}{24}kt^4\right)\Big|_2 ,
\end{aligned}$$

and the original numbers pop back out.

4 Taylor Polynomial

The change of base point procedure can be generalized. Supposing we choose a fixed point in time $t = t_p$ and allow to h change with 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:

$$\begin{aligned}
x(t) = {} & x_{t_p} + v_{t_p}(t - t_p) \\
& + \frac{1}{2}a_{t_p}(t - t_p)^2 + \frac{1}{6}j_{t_p}(t - t_p)^3 + \cdots
\end{aligned}$$

Introducing a generalized notation to represent the velocity, acceleration, jerk, and so on, let us make the associations

$$\begin{aligned}
x_{t_p} &\to x^{(0)}_{t_p} \\
v_{t_p} &\to x^{(1)}_{t_p} \\
a_{t_p} &\to x^{(2)}_{t_p} \\
j_{t_p} &\to x^{(3)}_{t_p} \\
k_{t_p} &\to x^{(4)}_{t_p} ,
\end{aligned}$$

and so on. On the left we've run out of 'named' items after snap, thus the general symbol $x^{(q)}_{t_p}$ is utilized to denote the qth coefficient.

4.1 Generalized Kinematic Equation

In condensed form, $x(t)$ can be written in a most general way using summation notation

$$x(t) = x_{t_p} + \sum_{q=1}^{n} \frac{1}{q!} x^{(q)}_{t_p} (t - t_p)^q , \qquad (7.4)$$

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.

4.2 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

$$\begin{aligned}
\left(x_0 + v_0 t + \frac{a_0}{2}t^2 + \frac{j_0}{6}t^3 + \frac{k_0}{24}t^4 + \cdots\right)\Big|_{t_p} &= x_{t_p} \\
\left(v_0 + a_0 t + \frac{j_0}{2}t^2 + \frac{k_0}{6}t^3 + \cdots\right)\Big|_{t_p} &= v_{t_p} \\
\left(a_0 + j_0 t + \frac{k_0}{2}t^2 + \cdots\right)\Big|_{t_p} &= a_{t_p} \\
(j_0 + k_0 t + \cdots)\Big|_{t_p} &= j_{t_p} ,
\end{aligned}$$

remembering that each of the terms x_{t_p}, v_{t_p}, etc., represent $x^{(0)}_{t_p}$, $x^{(1)}_{t_p}$, etc.

4.3 Inverse Coefficients

The inverse relations to the above, namely the structure that isolates x_0, v_0, etc., can be expressed by:

$$\begin{aligned}
j_0 &= \left(j_t - k_t t + \cdots\right)\Big|_{t_p} \\
a_0 &= \left(a_t - j_t t + \frac{k_t}{2}t^2 - \cdots\right)\Big|_{t_p} \\
v_0 &= \left(v_t - a_t t + \frac{j_t}{2}t^2 - \frac{k_t}{6}t^3 + \cdots\right)\Big|_{t_p} \\
x_0 &= \left(x_t - v_t t + \frac{a_t}{2}t^2 - \frac{j_t}{6}t^3 + \frac{k_t}{24}t^4 - \cdots\right)\Big|_{t_p}
\end{aligned}$$

4.4 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

$$v_{t_p}^{(r)} = x_{t_p}^{(r+1)} \ .$$

4.5 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

$$y(t) = A + Bt + Ct^2 + Dt^3 + \cdots \ ,$$

we're still free to interpret A, B, etc., as kinematic coefficients, i.e.

$$\begin{aligned}
A &= x_0 \\
B &= v_0 \\
C &= a_0/2! \\
D &= j_0/3! \\
E &= k_0/4! \ ,
\end{aligned}$$

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:

$$v(t) = v_0 + a_0 t + \frac{1}{2!} j_0 t^2 + \frac{1}{3!} k_0 t^3 + \cdots$$

Restoring the original coefficients, we find a formula for the slope $y_t^{(1)}$ of the function $y(t)$:

$$y_t^{(1)} = B + 2Ct + 3Dt^2 + 4Et^3 + \cdots$$

5 Area Under a Polynomial

5.1 Displacement as Area

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

$$\begin{aligned}
A_{\text{rectangle}} &= tv_0 \\
A_{\text{triangle}} &= \frac{1}{2} t \Delta v \ ,
\end{aligned}$$

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

$$A_{\text{total}} = v_0 t + \frac{1}{2} a t^2 \ ,$$

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:

$$A_{\text{total}} = x(t) - x_0$$

Riemann Sum (Optional)

Extending the idea of displacement-area-equivalence, it takes little to imagine that the displacement $x(t) - x_0$ is equal to the area under the velocity $v(t)$ plot whether or not the velocity is linear. This is typically justified using a Riemann sum, which approximates a general $v(t)$ as many conjoined straight lines such that

$$x(t) - x_0 = \lim_{N \to \infty} \sum_{q=1}^{N} v\left(t_q^*\right) \Delta t_q \ ,$$

where

$$\Delta t_q = t_q - t_{q-1} \ ,$$

and t_q^* is a value within the interval Δt_q.

In the general case, students of calculus learn to fashion the Riemann sum into a formal integral, and then the whole discussion shifts to techniques of solving integrals.

5.2 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:

$$y(t) = A + Bt + Ct^2 + Dt^3 + \cdots$$

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

$$\begin{aligned} A &= v_0 \\ B &= a_0 \\ C &= j_0/2! \\ D &= k_0/3! \,, \end{aligned}$$

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

$$x(t) - x_0 = v_0 t + \frac{1}{2!}a_0 t^2 + \frac{1}{3!}j_0 t^3 + \frac{1}{4!}k_0 t^4 + \cdots \,.$$

Replacing the kinematic coefficients with the original unknowns, the result

$$x(t) - x_0 = At + \frac{1}{2}Bt^2 + \frac{1}{3}Ct^3 + \frac{1}{4}Dt^4 + \cdots$$

emerges, and like magic, the problem is solved.

Example

For an example, let us exploit the Taylor polynomial to calculate the area under the curve

$$y(t) = (2 + 3t)^2 \,.$$

Begin by expanding $y(t)$ to get

$$y(t) = 4 + 12t + 9t^2 \,,$$

from which we discern

$$\begin{aligned} v_0 &= 4 \\ a_0 &= 12 \\ \frac{1}{2}j &= 9 \,. \end{aligned}$$

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.

6 Euler Exponential

From equation (7.4), it's interesting to conceive of the situation where all 'slope terms' are the same number, i.e.

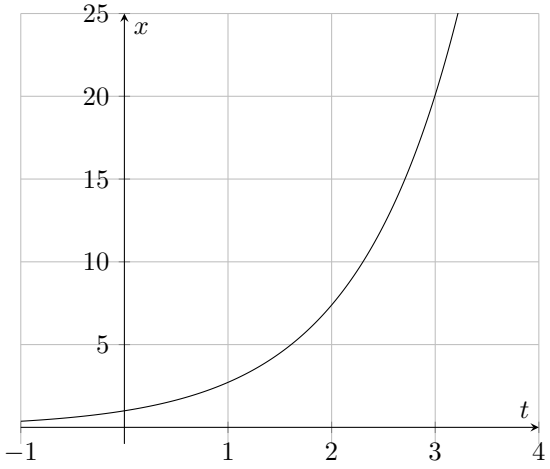
$$x_{t_p}^{(q)} = x_{t_p}^{(0)} = x_{t_p} \,,$$

which assumes (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.

From this, we have

$$x(t) = x_{t_p} \sum_{q=0}^{\infty} \frac{1}{q!} (t - t_p)^q \,,$$

and choosing $x_{t_p} = 1$ and $t_p = 0$ for a moment, we can have a look at this special $x(t)$ in the following plot:



Taking a Limit

To proceed, pluck out the $q = 0$-term and $q = 1$-term from the sum:

$$x(t) = x_{t_p} + x_{t_p} \cdot (t - t_p) + x_{t_p} \sum_{q=2}^{\infty} \frac{1}{q!} (t - t_p)^q$$

Now, to invoke a new restriction on the above, let us insist that the quantity $t - t_p$ become arbitrarily small, approaching but not reaching zero. In the absolute limit $t = t_p$, the above reduces to the tautology $x(t) = x_t$. Just 'before' that though, when $t - t_p$ is a very small number, the entire sum starting from $q = 2$ can be dismissed as negligible, leaving only the low-order terms:

$$x(t) = \lim_{t \to t_p} x_{t_p} + x_{t_p} \cdot (t - t_p) + \cancel{x_{t_p} \sum_{q=2}^{\infty} \frac{1}{q!} (t - t_p)^q}$$

Then, after some quick algebra, we have:

$$\frac{x\left(t\right)}{x_{t_p}} = \lim_{t\to t_p}\left(1+\left(t-t_p\right)\right)$$

6.1 Euler's Constant

From the plot of $x\left(t\right)$, the behavior of the curve seems much less like a polynomial and much more like an exponential. In this spirit, propose such a form for $x\left(t\right)$ namely

$$x\left(t\right) = x_0\, e^t\,,$$

where e is a yet-undetermined constant named to foreshadow the result. In terms of the same constant, it follows that

$$x_{t_p} = x_0\, e^{t_p}\,.$$

Combining this with the above limit analysis, we have

$$\frac{x\left(t\right)}{x_{t_p}} = e^{t-t_p} = \lim_{t\to t_p}\left(1+\left(t-t_p\right)\right)$$

or

$$e = \lim_{u\to 0}\left(1+u\right)^{1/u}\,,$$

where we have set

$$u = t - t_p\,,$$

calling for one more substitution

$$v = \frac{1}{u}\,,$$

so that the limit of u going to zero is replaced with the limit of v going to infinity. Finally, we get

$$e = \lim_{v\to\infty}\left(1+\frac{1}{v}\right)^v\,, \tag{7.5}$$

the 'standard' formula for Euler's constant, evaluating to, approximately,

$$e \approx 2.7182818284590\ldots\,.$$

6.2 Exponential Growth

To summarize so far, we write the Euler exponential as a polynomial such that

$$e^t = \sum_{q=0}^{\infty}\frac{t^q}{q!}\,,$$

which was motivated by setting all slope terms $x_{t_p}^{(q)}$ to be equal. Instead, let us instead insist that, in Equation (7.4), 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 we did before, there is now a factor of α joining the t-variable

$$x\left(t\right) = x_{t_p}\sum_{q=0}^{\infty}\frac{1}{q!}\alpha^q\left(t-t_p\right)^q\,,$$

obeying the limit

$$\frac{x\left(t\right)}{x_{t_p}} = \lim_{t\to t_p}\left(1+\alpha\left(t-t_p\right)\right)\,.$$

Then, using the same substitutions u, v as above leads to another definition of Euler's constant:

$$e^\alpha = \lim_{v\to\infty}\left(1+\frac{\alpha}{v}\right)^v \tag{7.6}$$

As in infinite sum, we take, as a final result:

$$e^{\alpha t} = \sum_{q=0}^{\infty}\frac{\left(\alpha t\right)^q}{q!} \tag{7.7}$$

Setting $\alpha\to-\alpha$, we have another relation to handle backward evolution in time:

$$e^{-\alpha t} = \sum_{q=0}^{\infty}\frac{\left(-\alpha t\right)^q}{q!} \tag{7.8}$$

6.3 Hyperbolic Curves

Two noteworthy combinations of the the Euler exponential equations (7.7), (7.8) can be constructed, namely the *hyperbolic cosine* and the *hyperbolic sine*, given by, respectively:

$$\cosh\left(\alpha t\right) = \frac{e^{\alpha t}+e^{-\alpha t}}{2} \tag{7.9}$$

$$\sinh\left(\alpha t\right) = \frac{e^{\alpha t}-e^{-\alpha t}}{2} \tag{7.10}$$

Simultaneous to these we can get the originals back by taking the sum and difference of each:

$$e^{\alpha t} = \cosh\left(\alpha t\right)+\sinh\left(\alpha t\right)$$
$$e^{-\alpha t} = \cosh\left(\alpha t\right)-\sinh\left(\alpha t\right)$$

Furthermore, given the infinite expansion of e^t, the hyperbolic functions can be written in open form:

$$\cosh\left(t\right) = 1+\frac{t^2}{2!}+\frac{t^4}{4!}+\frac{t^6}{6!}+\cdots \tag{7.11}$$

$$\sinh\left(t\right) = t+\frac{t^3}{3!}+\frac{t^5}{5!}+\frac{t^7}{7!}+\cdots \tag{7.12}$$

It's straightforward to show from the above that the pair of hyperbolic functions obey, for a dimensionless variable t,

$$\left(\cosh\left(t\right)\right)^2-\left(\sinh\left(t\right)\right)^2 = 1\,. \tag{7.13}$$

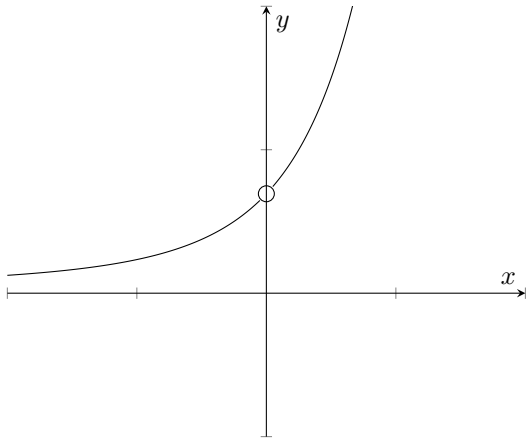
This is somewhat like the standard identity from trigonometry if it weren't for the negative sign. In fact, a whole slew of 'hyperbolic trigonometry' identities can be derived that are analogous to the 'ordinary' trig identities.

6.4 Natural Logarithm

Consider the curious quantity

$$y\left(x\right) = \lim_{x \to 0} \frac{n^x - 1}{x} \,, \tag{7.14}$$

plotted as follows:



From the plot of Equation (7.14), we see the point at $x = 0$ is probably a removeable singularity, which is to say we can come up with an answer for $y\left(0\right)$ given how y behaves in that neighborhood.

Proceed by solving for n to write

$$n = \lim_{x \to 0} \left(1 + xy\right)^{1/x} \,,$$

which is starting to look like the derivation of Euler's constant. By Equation (7.6), the right side evaluates to e^y, telling us

$$y\left(0\right) = \log_e\left(n\right) \,,$$

and for a final answer we take:

$$\ln\left(n\right) = \lim_{x \to 0} \frac{n^x - 1}{x}$$

The missing point in the plot of $y\left(x\right)$ is the natural log of n.

7 Periodic Curves

Consider the generalized (infinite) Taylor polynomial given as Equation (7.4) with $t_p = 0$:

$$x\left(t\right) = \sum_{q=0}^{\infty} \frac{1}{q!} x_0^{(q)} t^q$$

Next, consider a shift of variables $t \to t + w$ such that

$$x\left(t + w\right) = \sum_{q=0}^{\infty} \frac{1}{q!} x_0^{(q)} \left(t + w\right)^q \,,$$

which, up to w replacing h to avoid confusion, resembles the setup for time-shifted kinematics that led to Equation (7.4) originally.

For a familiar but nontrivial exercise, one can expand the right side in powers of w to write a generalization of Equation (7.3), particularly

$$x\left(t + w\right) = \sum_{q=0}^{\infty} \frac{1}{q!} x_t^{(q)} w^q \,. \tag{7.15}$$

A similarly-familiar exercise involves solving for the coefficients $x_t^{(q)}$, which in this case turn out as

$$\begin{aligned}
x_0^{(0)} + x_0^{(1)} t + x_0^{(2)} \frac{t^2}{2} + x_0^{(3)} \frac{t^3}{6} + \cdots &= x_t^{(0)} \\
x_0^{(1)} + x_0^{(2)} t + x_0^{(3)} \frac{t^2}{2} + x_0^{(4)} \frac{t^3}{6} + \cdots &= x_t^{(1)} \\
x_0^{(2)} + x_0^{(3)} t + x_0^{(4)} \frac{t^2}{2} + x_0^{(5)} \frac{t^3}{6} + \cdots &= x_t^{(2)} \,,
\end{aligned}$$

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:

$$\begin{aligned}
x_t^{(0)} + w x_t^{(1)} + w^2 x_t^{(2)} + \cdots = \\
&+ x_0^{(0)} + w x_0^{(1)} + w^2 x_0^{(2)} + \cdots \\
&+ t \left(x_0^{(1)} + w x_0^{(2)} + w^2 x_0^{(3)} + \cdots \right) \\
&+ \frac{t^2}{2} \left(x_0^{(2)} + w x_0^{(3)} + w^2 x_0^{(4)} + \cdots \right) \\
&+ \frac{t^3}{6} \left(x_0^{(3)} + w x_0^{(4)} + w^2 x_0^{(5)} + \cdots \right) + \cdots
\end{aligned}$$

7.1 Periodicity Condition

Suppose, for all times t in the domain of $x\left(t\right)$, the property

$$x\left(t + w\right) = x\left(t\right)$$

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:

$$\begin{aligned} 0 = w &\left(x_0^{(1)} + w x_0^{(2)} + w^2 x_0^{(3)} + \cdots \right) \\ &+ \frac{w^2}{2} \left(x_0^{(2)} + w x_0^{(3)} + w^2 x_0^{(4)} + \cdots \right) \\ &+ \frac{w^3}{6} \left(x_0^{(3)} + w x_0^{(4)} + w^2 x_0^{(5)} + \cdots \right) + \cdots , \end{aligned}$$

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:

$$0 = \left(\sum_{\text{even } k}^{\infty} x_0^{(k_e)} w^{k_e} J_e \right) + \left(\sum_{\text{odd } k}^{\infty} x_0^{(k_o)} w^{k_o} J_o \right)$$

Now, the periodicity condition must also work when w is swapped with $-w$, or any integer multiple of w for that matter. Going with $w \to -w$, 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

$$\begin{aligned} \text{Even} + \text{Odd} &= 0 \\ \text{Even} - \text{Odd} &= 0 , \end{aligned}$$

which is only true if

$$\text{Even} = \text{Odd} = 0 .$$

Each parenthesized sum above, 'odd' and 'even', must separately equal zero.

7.2 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

$$\begin{aligned} 1 &= x_0^{(0)} = x_0^{(4)} = x_0^{(8)} = \cdots \\ -1 &= x_0^{(2)} = x_0^{(6)} = x_0^{(10)} = \cdots \end{aligned}$$

for the even terms, and

$$\begin{aligned} 1 &= x_0^{(1)} = x_0^{(5)} = x_0^{(9)} = \cdots \\ -1 &= x_0^{(3)} = x_0^{(7)} = x_0^{(11)} = \cdots \end{aligned}$$

for the odd terms.

These results let us write two cases for the resulting $x(t)$, namely

$$\cos(t) = 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!} + \cdots \tag{7.16}$$

$$\sin(t) = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!} + \cdots , \tag{7.17}$$

where of course, the variable t can be replaced by the combination αt as done previously.

8 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:

$$U(x) = U_{x_p} + \sum_{q=1}^{n} \frac{1}{q!} U_{x_p}^{(q)} (x - x_p)^q$$

$$T(v) = T_{v_p} + \sum_{q=1}^{n} \frac{1}{q!} T_{v_p}^{(q)} (v - v_p)^q$$

8.1 Conservation of Energy

Now suppose we are interested in the sum of T and U, a quantity labled 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

$$0 = \left(U_{x_p}^{(1)} \Delta x + T_{v_p}^{(1)} \Delta v \right) + \sum_{q=2}^{n} \frac{1}{q!} \left(U_{x_p}^{(q)} \Delta x^q + T_{v_p}^{(q)} \Delta v^q \right) \,.$$

8.2 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:

$$-U_{x_p}^{(1)} = m \frac{\Delta v}{\Delta t} \tag{7.18}$$
$$T_{v_p}^{(1)} = m v_p \tag{7.19}$$

In order for the quantities E, T, U to have units of energy, the constant m can only have units of mass.

8.3 Newton's Second Law

Equation (7.18) 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*:

$$F = m \frac{\Delta v}{\Delta t}$$

8.4 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:

$$U(x) = \text{potential energy}$$

8.5 Kinetic Energy

The second result $T_{v_p}^{(1)} = m v_p$ 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 $m v_p$, it's easy to see that the kinetic energy is generally given by

$$T(v) = \frac{1}{2} m v^2 \,.$$

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:

$$E = \frac{1}{2} m v^2 + U_{x_p} + \sum_{q=1}^{n} \frac{1}{q!} U_{x_p}^{(q)} (x - x_p)^q$$

8.6 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

$$U^{(1)}_{x_p} = 0\,,$$

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:

$$E = \cancel{\frac{1}{2}mv^2} + U_{x_p} + \cancel{\sum_{q=1}^{n} \frac{1}{q!} U^{(q)}_{x_p} (x - x_p)^q}$$

In light of $U^{(1)}_{x_p}$ being zero, the lowest nonzero term in the sum corresponds to $q = 2$, thus we have, for small displacements from x_p:

$$E - U_{x_p} = \frac{1}{2}mv^2 + \frac{1}{2}U^{(2)}_{x_p}(x - x_p)^2$$

The sign on $U^{(2)}_{x_p}$ determines the stability of motion around x_p. For $U^{(2)}_{x_p} < 0$, 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 $U^{(2)}_{x_p} > 0$, the system exhibits small oscillations centered on x_p. The potential energy term

$$U(x) = \frac{1}{2}U^{(2)}_{x_p}(x - x_p)^2$$

can be treated as arising by a spring force centered centered at x_p

$$f(x) = -U^{(2)}_{x_p}(x - x_p)\,,$$

with $U^{(2)}_{x_p}$ being the effective spring constant. Using Newton's second law on this situation gives an equation for the subsequent motion:

$$m x^{(2)}_{t_p} = -U^{(2)}_{x_p}(x - x_p)$$

This particular form is also called *Hooke's law* for springs.

8.7 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

$$U_{grav}(x) = mgx\,,$$

where x is measured vertically from the ground.

Chapter 8

Limits, Functions, Sequences

1 Limits

In mathematics, one speaks of *limit* when a starting value, call it A, is 'becoming' another value B:

$$A \to B$$

The use of the right-arrow ($\to$) 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:

$$dx = \lim_{x \to x_0} x - x_0 \tag{8.1}$$

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:

$$A^2 = A_0^2 + 2A_0 dx + dx^2$$

Now, we know already that A^2 is approximately equal to A_0^2 plus some small correction having to do with dx, namely $2A_0 dx + dx^2$.

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:

$$A^2 \approx A_0^2 + 2A dx$$

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:

$$e^x = \lim_{h\to\infty} \left(1 + \frac{x}{h}\right)^h \tag{8.2}$$

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 $e \approx 2.71828\ldots$.

Problem 2

Show that:

$$\lim_{x\to 1} x^{1/(1-x)} = e$$

Infinite Sum

An infinite limit can take the form of an *infinite sum*, as is the case with geometric series:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots \tag{8.3}$$

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.

Summation Notation

An equivalent statement of the above that uses the 'lim' nomenclature goes like

$$\frac{1}{1-x} = \lim_{n\to\infty} \sum_{j=0}^{n} x^j \,,$$

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 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 $p = \$0.9999\overline{9}$, 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:

$$p = 0.9999\cdots = \frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \frac{9}{10^4} + \cdots$$

Factor 9/10 from the right side

$$p = \frac{9}{10}\left(1 + \frac{1}{10} + \frac{1}{10^2} + \frac{1}{10^3} + \frac{1}{10^4} + \cdots\right) ,$$

and then see that the parenthesized quantity is precisely the geometric series as the right side of Equation (8.3) 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:

$$0.9999\overline{9} = p = 1$$

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 (8.2) 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 (8.3). 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.

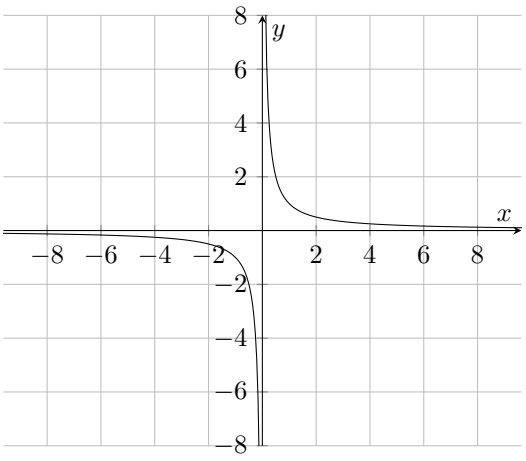


Figure 8.1: Reciprocal curve $y = 1/x$.

To demonstrate, consider the reciprocal curve $y = 1/x$ shown in Figure 8.1, 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

$$\lim_{x\to 0^-} \frac{1}{x} = -\infty \,,$$

whereas on the right, we get a similar statement with all $-$ signs as $+$ signs

$$\lim_{x\to 0^+} \frac{1}{x} = \infty \,,$$

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

$$\begin{aligned} (x < x_0) \text{ or Left-Sided:} &\quad \lim_{x\to x_0^-} \\ (x > x_0) \text{ or Right-Sided:} &\quad \lim_{x\to x_0^+} \,, \end{aligned}$$

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

$$\begin{aligned} \lim_{x\to 0^-} \sqrt{x} &= \text{Undefined} \\ \lim_{x\to 0^+} \sqrt{x} &= 0 \,. \end{aligned}$$

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

$$\lim_{x\to 0^+} x^x = A \,,$$

where A stands for 'Answer'. Next, explicitly check descending x values (starting from a reasonable guess) and inspect for an emerging trend or pattern:

$$\begin{aligned} 0.1^{0.1} &= 0.794328\ldots \\ 0.01^{0.01} &= 0.9549926\ldots \\ 0.001^{0.001} &= 0.99311605\ldots \\ 0.0001^{0.0001} &= 0.999079390\ldots \end{aligned}$$

Evidently, the answer is getting closer to $A = 1$ as x is going to zero. This is motivation enough to write

$$\lim_{x\to 0^+} x^x = 1 \,,$$

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 in only in integers or a subset of them.

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:

$$\lim_{x \to x_0^-} \leftrightarrow \lim_{x \to x_0^+}$$

Two-sidedness apples 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.

In Figure 8.2, 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 $(p, 0)$.

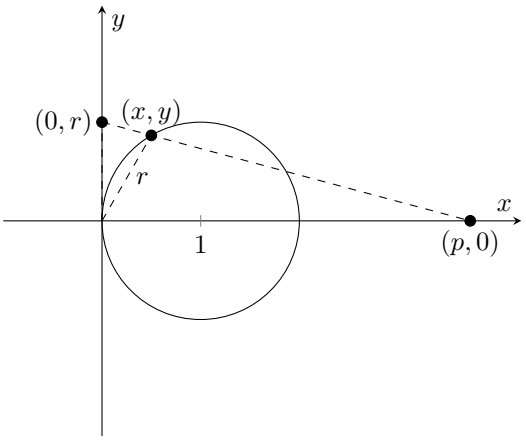


Figure 8.2: Unit circle intersecting straight line.

And now for the problem: Express the distance p in terms of r, and what happens to p as $r \to 0$? 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 $p \to 0$, 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 the 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	$y = a_0 + a_1 x + a_2 x^2 + \cdots$
Square Root	$y = \sqrt{x}$
Exponential	$y = a^x$
Natural Exponential	$y = e^x$
Hyperbolic Cosine	$y = \left(e^x + e^{-x}\right)/2$
Hyperbolic Sine	$y = \left(e^x - e^{-x}\right)/2$
Logarithm	$y = \log(x)$
Natural Log	$y = \ln(x)$
Factorial	$y = x(x-1)(x-2)\cdots 2 \cdot 1$
Cosine	$y = 1 - x^2/2! + x^4/4! - \cdots$
Sine	$y = x - x^3/3! + x^5/5! - \cdots$
Tangent	$y = x + x^3/3 + 2x^5/15 + \cdots$

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

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

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 8.1. 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 $\{x\}$.

Bijective

A curve that is both injective and surjective is called *bijective*. 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 bijective 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

$$f^{-1}(f(x)) = x \tag{8.4}$$

for bijective functions. Applying f to both sides gives a similar statement

$$f\left(f^{-1}(y)\right) = y \ .$$

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:

$$y_1 = f\left(f^{-1}(x_2)\right) = x_2$$

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

$$\lim_{x\to x_0} f(x) = A$$
$$\lim_{x\to x_0} g(x) = B\,.$$

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:

$$\lim_{x\to x_0} (Cf(x)) = C \lim_{x\to x_0} f(x) = CA \tag{8.5}$$

The proof for this begins with a mouthful of mathematical jargon. Starting from the definition

$$\lim_{x\to x_0} f(x) = A\,,$$

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

$$|Cf(x) - CA| < \cancel{|C|}\frac{\epsilon}{\cancel{|C|}}\,,$$

and the proof is done.

Distribution into Sum

The 'limit' construct distributes freely into the sum of two functions:

$$\lim_{x\to x_0} (f(x) \pm g(x)) = \lim_{x\to x_0} f(x) \pm \lim_{x\to x_0} g(x) = A \pm B \tag{8.6}$$

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 (8.6), 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 (8.5) to find

$$\lim_{x\to x_0} (f(x) - g(x)) = A - B\,.$$

Distribution into Product

The 'limit' construct distributes also into the product of two functions:

$$\lim_{x\to x_0} (f(x) \cdot g(x)) = \lim_{x\to x_0} f(x) \cdot \lim_{x\to x_0} g(x) = A \cdot B \tag{8.7}$$

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:

$$0 < |x - x_0| < \delta_1 \implies |f(x) - A| < \sqrt{\epsilon}$$
$$0 < |x - x_0| < \delta_2 \implies |g(x) - B| < \sqrt{\epsilon}$$

Choose δ to be the smaller of $\delta_{1,2}$, and then for $0 < |x - x_0| < \delta$, we find

$$|(f(x) - A) \cdot (g(x) - B)| = |(f(x) - A)| \cdot |(g(x) - B)|\,,$$

and replace the right side to establish

$$|(f(x)-A)\cdot(g(x)-B)| \leq \sqrt{\epsilon}\sqrt{\epsilon}\,.$$

With this, we've proved

$$\lim_{x\to x_0}(f(x)-A)\cdot(g(x)-B)=0\,,$$

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 $x\to x_0$ on every term. Right away, the left side is zero, and we're left with

$$\begin{aligned}\lim_{x\to x_0} f(x)\,g(x) = \\ B\lim_{x\to x_0} f(x)+A\lim_{x\to x_0} g(x)-AB \\ = BA+AB-AB \\ = AB\,,\end{aligned}$$

finishing the proof.

Distribution into Quotient

The 'limit' construct distributes also into the quotient of two functions:

$$\lim_{x\to x_0}\left(\frac{f(x)}{g(x)}\right)=\frac{\lim_{x\to x_0} f(x)}{\lim_{x\to x_0} g(x)}=\frac{A}{B} \tag{8.8}$$

Note of course that B cannot be zero.

For this proof, we need to first spend some effort to establish

$$\lim_{x\to x_0}\frac{1}{g(x)}=\frac{1}{B}\,.$$

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

$$\lim_{x\to x_0}\frac{1}{g(x)}=\frac{1}{B}\,.$$

From here, we have

$$\lim_{x\to x_0}\left(\frac{f(x)}{g(x)}\right)=\lim_{x\to x_0}\left(f(x)\,\frac{1}{g(x)}\right)\,,$$

which by Equation (8.7) is also

$$\lim_{x\to x_0}\left(\frac{f(x)}{g(x)}\right)=\left(\lim_{x\to x_0} f(x)\right)\left(\lim_{x\to x_0}\frac{1}{g(x)}\right)\,,$$

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:

$$\lim_{x\to x_0}(f(x))^n=\left(\lim_{x\to x_0} f(x)\right)^n=A^n \tag{8.9}$$

To begin, begin with

$$\lim_{x\to x_0}(f(x))^n=\lim_{x\to x_0}\left(f(x)^{n-1}\,f(x)\right)\,,$$

which decouples by Equation (8.7) to

$$\lim_{x\to x_0}(f(x))^n=A\cdot\lim_{x\to x_0}\left(f(x)^{n-1}\right)\,.$$

From here, we can use Equation (8.7) again to find

$$\lim_{x\to x_0}\left(f\left(x\right)\right)^n = A^2 \cdot \lim_{x\to x_0}\left(f\left(x\right)^{n-2}\right) ,$$

and repeat this recursively for m steps

$$\lim_{x\to x_0}\left(f\left(x\right)\right)^n = A^m \cdot \lim_{x\to x_0}\left(f\left(x\right)^{n-m}\right) ,$$

stopping when $n = m$.

Note that it's possible to show that Equation (8.9) 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

$$z = \lambda^{1/(\lambda-1)}$$
$$\lambda z = \lambda^{\lambda/(\lambda-1)} ,$$

and similarly:

$$\frac{1}{z} = \lambda^{-1/(\lambda-1)}$$
$$\frac{1}{\lambda z} = \lambda^{-\lambda/(\lambda-1)}$$

While the above is a workable solution, proceed by substituting $q = \lambda/\left(\lambda - 1\right)$ to find

$$\frac{1}{z} = \lambda^{-q/\lambda}$$
$$\frac{1}{\lambda z} = \lambda^{-q} ,$$

also implying

$$z^\lambda = \lambda z .$$

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

$$z_\infty = \lim_{\lambda\to\infty} \lambda^{1/(\lambda-1)} = 1$$

as expected.

As constructed, we know $\lambda = 2$ is an allowed λ-value, but what is the minimum λ? Trying $\lambda = 1$, calculate

$$z_1 = \lim_{\lambda\to 1} \lambda^{1/(\lambda-1)} = e .$$

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

$$f\left(x\right) = \begin{cases} 2x - 1 & x < 1 \\ \sqrt{x} + 1 & x \geq 1 \end{cases} ,$$

which takes different form on either side of the point $x = 1$ as shown in Figure 8.3.

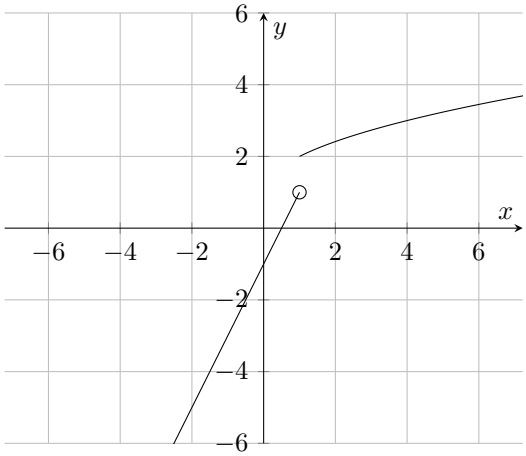


Figure 8.3: Piecewise function.

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*

Smoothness

Modifying the above example, suppose we nixed the $+1$ from the function 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 8.1.

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:

$$\lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^+} f(x) = \lim_{x \to x_0} f(x)$$

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

$$\lim_{x \to x_0^-} f(x) = -1 = \lim_{x \to x_0^+} f(x),$$

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

$$f(x) = \frac{\cancel{(x-1)}(x-2)}{\cancel{x-1}} = x - 2,$$

like there was never a singularity at all.

Natural Logarithm

For a less trivial example, consider the function

$$y = f(x) = \frac{n^x - 1}{x}, \tag{8.10}$$

and let's be interested in the quantity $f(0)$. Immediately we see that $x = 0$ makes Equation (8.10) blow up, but can we remove this point? Consulting a plot of $f(x)$ shown in Figure 8.4, it seems the singularity is removable.

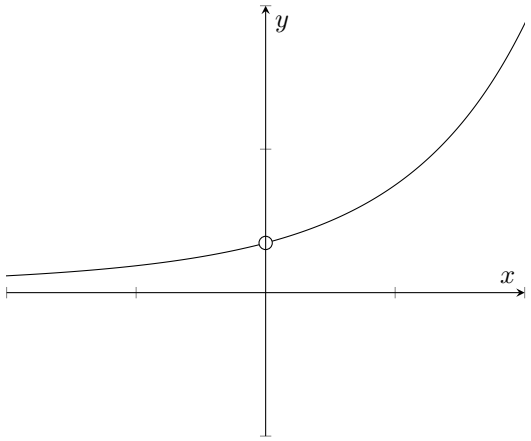


Figure 8.4: Function with removable singularity at $x = 0$.

To proceed, solve Equation (8.10) for n, and set up a limit that pushes x to zero:

$$n = \lim_{x\to 0} (1+xy)^{1/x}$$

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:

$$n = \lim_{h\to\infty} \left(1+\frac{y}{h}\right)^h$$

By Equation (8.2), 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 (8.2):

$$\ln(x) = \lim_{h\to 0} \frac{x^h - 1}{h} \tag{8.11}$$

Sinc Function

An interesting singularity arises in the so-called 'sinc' function, which is defined as $\sin(x)$ divided by x,

$$\operatorname{sinc}(x) = \frac{\sin(x)}{x}\,, \tag{8.12}$$

sketched in Figure 8.5. 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 $y = 1/x$.

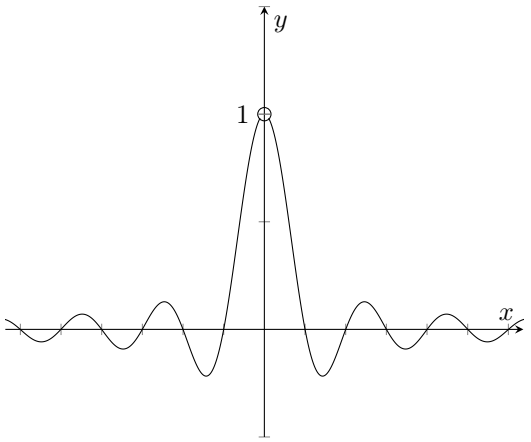


Figure 8.5: Plot of $\sin(x)/x$.

The point $x = 0$ is 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

$$\operatorname{sinc}(0) = \lim_{x\to 0} \frac{\sin(x)}{x}\ .$$

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.

$$\sin(x) = x - x^3/3! + x^5/5! - \cdots\,,$$

and then $\sin(x)/x$ must be

$$\frac{\sin(x)}{x} = 1 - x^2/3! + x^4/5! - \cdots\,,$$

and now the right side has no singularity at all. We can forget about limits and simply set $x = 0$ to find

$$\operatorname{sinc}(0) = \lim_{x\to 0} \frac{\sin(x)}{x} = 1\,, \tag{8.13}$$

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:

$$\lim_{x\to x_0} f(x) = L = \lim_{x\to x_0} h(x)$$

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:

$$\lim_{x\to x_0} f(x) = \lim_{x\to x_0} g(x) = \lim_{x\to x_0} h(x) = L$$

Example 1

Use the squeeze theorem to show that

$$\lim_{x\to 0} x^2 \sin\left(\frac{1}{x}\right) = 0\,.$$

For 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 $x \to 0$, 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

$$\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{\cdots}}}}} = 2\,,$$

where the left side contains an infinite nesting of $\sqrt{x+\cdots}$ 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:

$$\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{\cdots}}}}} = 1$$

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

$$\sqrt{0+\sqrt{0+\sqrt{0+\sqrt{0+\sqrt{\cdots}}}}} = 1\,,$$

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

$$y = \sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{\cdots}}}}}\,,$$

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:

$$y(x) = \begin{cases} 0 & x = 0 \\ 1/2 + \sqrt{1+4x}/2 & x > 0 \end{cases}$$

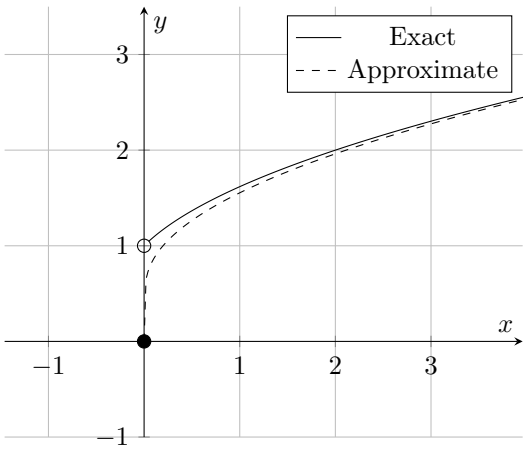


Figure 8.6: A strange beast.

The function $y(x)$ is plotted as a solid curve in Figure 8.6. Note the 'hole' removed 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

$$y\,(1) = \frac{1+\sqrt{5}}{2} = 1.618034\ldots\,,$$

a celebrated number called the *golden ratio*, denoted ϕ. Given the open version for $y\,(x)$, we derive a nifty expression for ϕ:

$$\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{1+\sqrt{\cdots}}}}} = \phi$$

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

$$\{A\} = \{A_1, A_2, A_3, \ldots, A_n\}\ .$$

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 $n \to \infty$ can also be conceived, called an *infinite sequence*. The 'last' term L in an infinite sequence can be expressed as the limit

$$L = \lim_{j\to\infty} A_j\,.$$

Convergent Sequence

For finite L, we may require that that for any positive ϵ, there exists some integer m such that

$$j > m\ \to\ |A_j - L| < \epsilon\,.$$

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):

$$\begin{aligned} \{A\} &= \{10, 20, 1, 2, 0.1, 0.2, 0.01, 0.02, \ldots\} \\ \{B\} &= \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \ldots\right\} \\ \{C\} &= \left\{\frac{9}{10}, \frac{99}{100}, \frac{999}{1000}, \frac{9999}{10000}, \ldots\right\} \end{aligned}$$

Divergent Sequence

A *divergent* sequence is one that has $L \to \pm\infty$. 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):

$$\begin{aligned} \{Q\} &= \{1, 10, 100, 1000, 10000, \ldots\} \\ \{R\} &= \{-1, -2, -3, -4, -5, -6, \ldots\} \end{aligned}$$

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

$$\{A\} = \{1, 2, 3, 1, 2, 3, 1, 2, 3, \ldots\}$$

is periodic.

Recursive Sequence

By studying polynomial division, one can argue into existence a sequence $\{F\}$ obeying the 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:

$$\{F_{p=1}\} = \{1, 1, 2, 3, 5, 8, 13, 21, \ldots\}$$

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 *pi*:

$$\{P\} = \{3, 1, 4, 1, 5, 9, 2, 6, 5, \dots\}$$

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 $f(n) = A_j$. With this, the following is always true:

$$\lim_{x\to\infty} f(x) = L = \lim_{j\to\infty} A_j \tag{8.14}$$

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

$$\lim_{j\to\infty} A_j = L = \lim_{j\to\infty} C_j$$

then

$$L = \lim_{j\to\infty} B_j \, .$$

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):

$$\lim_{j\to\infty} (-|x_j|) = L = -\lim_{j\to\infty} |x_j|$$

In the special case

$$\lim_{j\to\infty} |x_j| \to 0 \, ,$$

thereby making $L \to 0$, the squeeze theorem leads to a stronger conclusion:

$$\lim_{j\to\infty} x_j \to 0$$

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 (8.14):

Addition

$$\lim_{j\to\infty} (A_j \pm B_j) = \lim_{j\to\infty} A_j \pm \lim_{j\to\infty} B_j$$

Scalar Multiplication

$$\lim_{j\to\infty} \lambda A_j = \lambda \lim_{j\to\infty} A_j$$

Product

$$\lim_{j\to\infty} (A_j B_j) = \left(\lim_{j\to\infty} A_j\right)\left(\lim_{j\to\infty} B_j\right)$$

Ratio

As long as the denominator is not zero:

$$\lim_{j\to\infty} \frac{A_j}{B_j} = \frac{\lim_{j\to\infty} A_j}{\lim_{j\to\infty} B_j}$$

Exponent

As long as $A_j > 0$:

$$\lim_{j\to\infty} A_j^q = \left(\lim_{j\to\infty} A_j\right)^q$$

3.5 Geometric Sequence

Consider the geometric sequence

$$\{x_j\} = \left\{x^j\right\}_0^\infty \, .$$

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

$$\lim_{x\to\infty} x^j = L = \lim_{n\to\infty} A_j \, .$$

Handling the easy cases first, we see that $x > 1$ leads to $L \to \infty$, 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

$r^j = \left|x^j\right|$ so we're checking the domain $0 < r < 1$. This reproduces the previous case, and we have

$$\lim_{j\to\infty} r^j = 0 = \lim_{j\to\infty} \left|x^j\right| ,$$

thus the sequence converges.

The case $x = -1$ attempts a limit that does not exist. Writing this case out, one finds the periodic sequence

$$\left\{(-1)^j\right\} = \{1, -1, 1, -1, 1, -1, \dots\} ,$$

exhibiting neither convergence nor divergence.

The ugliest case occurs at $x < -1$. For instance, choosing $x = -2$ leads to

$$\left\{(-2)^j\right\} = \{1, -2, 4, -8, 16, -32, 64 \dots\} ,$$

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:

$$\lim_{j\to\infty} x^j = \begin{cases} 0 & -1 < x < 1 \\ 1 & x = 1 \end{cases} \tag{8.15}$$

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

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 $n \to \infty$, the sum

$$S = \lim_{m\to\infty} s_m = \sum_{j=1}^{\infty} A_j$$

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:

$$\sum_{j=1}^{\infty} (A_j + B_j) = \sum_{j=1}^{\infty} A_j + \sum_{j=1}^{\infty} B_j$$

$$\sum_{j=1}^{\infty} \lambda A_j = \lambda \sum_{j=1}^{\infty} A_j$$

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

$$\lim_{j\to\infty} A_j = 0\,.$$

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 $\{A\}$ *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

$$\frac{1}{2} = \left(\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \cdots\right) - \left(\frac{1}{8} + \frac{1}{32} + \frac{1}{128} + \cdots\right)\,,$$

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 idenfity is worth memorizing:

$$\sum_{j=1}^{n} j = \frac{n\,(n+1)}{2} \tag{8.16}$$

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 $n = 3$,

$$S_3\,(2) = 1 + 2^2 + 3^2 = 27\alpha + 9\beta + 3\gamma\,.$$

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 te hard work is done. The right side can be factored to deliver the final result:

$$\sum_{j=1}^{n} j^2 = \frac{n\,(n+1)\,(2n+1)}{6} \tag{8.17}$$

For completeness, it turns out the result for S_3 is:

$$\sum_{j=1}^{n} j^3 = \left(\frac{n\,(n+1)}{2}\right)^2 \tag{8.18}$$

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

$$\frac{I_n}{2} = \frac{1}{2} + 1 + 2 + 2^2 + \cdots 2^{n-1}$$
$$= \frac{1}{2} + I_{n-1}\,.$$

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:

$$\sum_{j=0}^{n} 2^j = 2^{n+1} - 1 \tag{8.19}$$

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 $k \to m$ 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:

$$C = 1 + \sum_{m=2}^{\infty} \frac{1}{m!} - \sum_{m=2}^{\infty} \frac{1}{m!} = 1$$

Part III

Linear & Complex Algebra

Chapter 9

Vectors and Matrices

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*:

$$\vec{V} = \langle 2, 5, 7 \rangle$$

Vector notation requires a *label* for the vector, $\vec{V}$ in our example, specially marked by an arrow ($\vec{\ }$). On the right, the 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 $\vec{V} = \langle 2, 5, 7 \rangle$ 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 $\vec{V}$ as

$$\vec{V} = \langle V_a, V_b, V_c \rangle$$

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 $\vec{V} = \langle V_a, V_b \rangle$, we plot $\vec{V}$ 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. 9.1 are two *equivalent* representations of the vector $\vec{A} = \langle 2, 6 \rangle$. Note that the vector doesn't 'care' about the choice of base point.

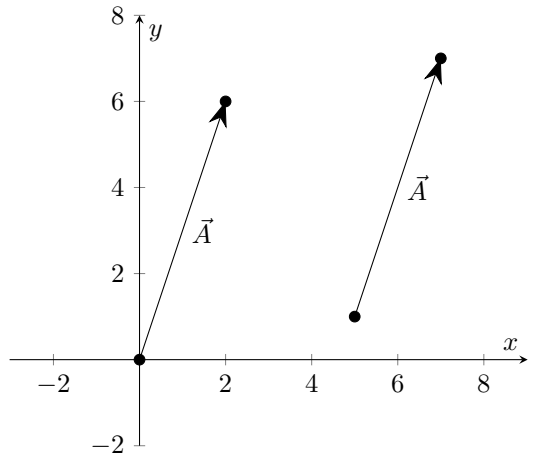


Figure 9.1: Vector $\vec{A} = \langle 2, 6 \rangle$ plotted from two different base points $(0, 0)$ and $(5, 1)$.

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.

1.3 Position Vector

A vector whose base point is the origin $(0, 0)$ is called a *position vector*, often denoted $\vec{R}$ or $\vec{X}$. A position vector $\vec{R} = \langle R_x, R_y \rangle$ is equivalent to the ordered pair (x, y), denoting a unique point in the Cartesian plane.

1.4 Vector 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 $\vec{A}$ of dimension N has a magnitude given by

$$A = \left|\vec{A}\right| = \sqrt{A_1^2 + A_2^2 + \cdots + A_N^2}\,. \tag{9.1}$$

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 $\vec{A}$ 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 $\vec{A} = \langle A_x, A_y\rangle$ is pointing is given by

$$\phi = \arctan\left(\frac{A_y}{A_x}\right) \tag{9.2}$$

To justify (9.2), assume A_x and A_y are two sides of a right triangle such that

$$\begin{aligned} A_x &= A\cos(\phi) \\ A_y &= A\sin(\phi)\ , \end{aligned}$$

and eliminate the magnitude A.

2 Vector Addition

2.1 Definition

Two vectors $\vec{A}$, $\vec{B}$ of equal dimension N can be added by combining like components, resulting in a vector $\vec{C}$ with N components:

$$\vec{C} = \vec{A} + \vec{B} \tag{9.3}$$

The jth component is given by

$$\begin{aligned} C_j &= A_j + B_j \\ j &= 1, 2, 3, \ldots, N\,. \end{aligned} \tag{9.4}$$

Commutativity of Vector Addition

Following immediately from (9.3)-(9.4) is the *commutativity of addition*:

$$\vec{A} + \vec{B} = \vec{B} + \vec{A} \tag{9.5}$$

In particular, (9.5) 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 $\vec{A}$, $\vec{B}$, $\vec{C}$ involves two addition operations. Also following from (9.3)-(9.4) is the *associativity of addition*, telling us that the order of the two addition operations does not effect the result:

$$\vec{A} + \left(\vec{B} + \vec{C}\right) = \left(\vec{A} + \vec{B}\right) + \vec{C} \tag{9.6}$$

2.2 Arrow Trick

The 'arrow' representation of a vector avails a beautiful shortcut for vector addition. Given a pair of two-dimensional vectors $\vec{A}$, $\vec{B}$, 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. 9.2 demonstrates the 'arrow trick' on two example vectors $\vec{A} = \langle 2, 5\rangle$, $\vec{B} = \langle 3, -3\rangle$, whose sum easily comes out to $\vec{C} = \langle 5, 2\rangle$. By plotting $\vec{A}$, $\vec{B}$ as suggested, the sum $\vec{C}$ is visually represented by an arrow beginning at the base of $\vec{A}$ and ending at the tip of $\vec{B}$.

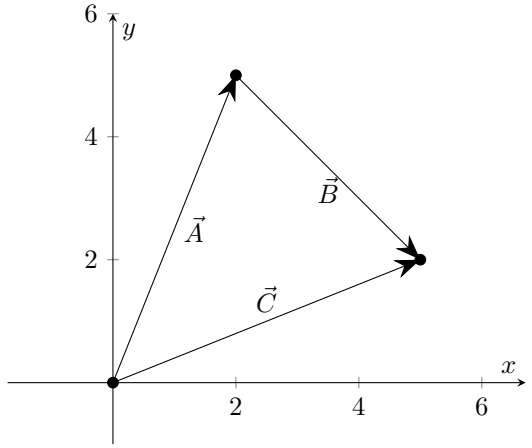


Figure 9.2: Vector addition $\vec{C} = \vec{A} + \vec{B}$.

2.3 Additive Inverse

Given any vector $\vec{A} = \langle A_1, A_2, A_3, \ldots, A_N\rangle$, the *additive inverse* is another vector that reverses the sign

on all components in $\vec{A}$, denoted $-\vec{A}$, where

$$-\vec{A} = \langle -A_1, -A_2, -A_3, \ldots, -A_N \rangle \ . \tag{9.7}$$

2.4 Zero Vector

The so-called *zero vector* is the vector that contains only zeros:

$$\vec{0} = \langle 0, 0, 0, \ldots, 0 \rangle \tag{9.8}$$

For hopefully obvious reasons, turns out that the sum of any vector and its additive inverse always yields the zero vector:

$$\vec{A} + \left(-\vec{A}\right) = \vec{0}$$

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 $\vec{A} + \vec{B} + \vec{C} = 0$.

2.5 Vector Subtraction

With the additive inverse established, the notion of *vector subtraction* can be framed in terms of vector addition. Given two vectors $\vec{A}$, $\vec{B}$, the difference $\vec{D} = \vec{A} - \vec{B}$ can be visualized with the same 'arrow trick', so long as we reverse the direction on $\vec{B}$ as shown in Fig. 9.3.

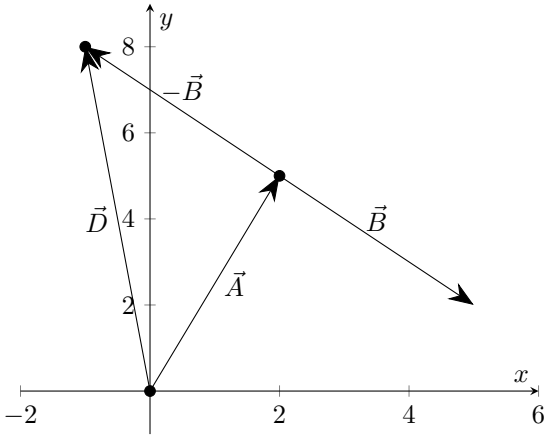


Figure 9.3: Vector subtraction $\vec{D} = \vec{A} - \vec{B}$.

3 Scalar Multiplication

A vector $\vec{A}$ 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 $\vec{B}$:

$$\vec{B} = \alpha\vec{A} = \langle \alpha A_1, \alpha A_2, \alpha A_3, \ldots, \alpha A_N \rangle \tag{9.9}$$

3.1 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 $\vec{A}$, $\vec{B}$ in (9.9) are necessarily parallel.

3.2 Straight 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

$$\vec{b} = \langle 0, b \rangle \ .$$

Required next is a vector $\vec{m}$ that represents the slope of the line, which we capture by writing

$$\begin{aligned} \vec{m} &= \langle m_x, m_y \rangle \\ \frac{m_y}{m_x} &= m \ . \end{aligned}$$

Multiplying $\vec{m}$ 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

$$\vec{r} = \alpha\, \vec{m} + \vec{b}\,, \tag{9.10}$$

where $\vec{r} = \langle x, y \rangle$ is the resulting position vector, as shown in Fig. 9.4. In case (9.10) isn't convincing, one may resolve $\vec{r}$ 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' $\vec{m}_\perp$ is evidently $\vec{m}_\perp = \langle -m_y, m_x \rangle$.

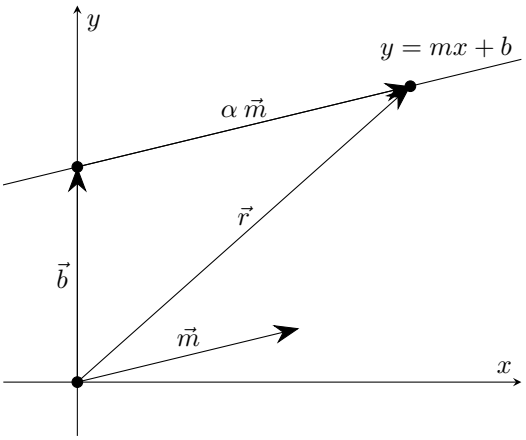


Figure 9.4: Vector construction of a straight line $y = mx + b$.

3.3 Algebraic Properties of Vectors

Associativity with Scalars

If a vector is modified by two scalars, the order in which they're applied does not matter:

$$\alpha\left(\beta\vec{A}\right) = (\alpha\beta)\,\vec{A} = (\beta\alpha)\,\vec{A} = \beta\left(\alpha\vec{A}\right) \tag{9.11}$$

Vector Distributive Properties

Readily provable from the properties of vector addition and scalar multiplication are the distributive properties involving the sum of two scalars:

$$(\alpha+\beta)\,\vec{A} = \alpha\vec{A} + \beta\vec{B} \tag{9.12}$$

$$\alpha\left(\vec{A}+\vec{B}\right) = \alpha\vec{A} + \alpha\vec{B} \tag{9.13}$$

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 $\vec{A}$, $\vec{B}$ of dimension N, the dot product reads

$$\vec{A}\cdot\vec{B} = A_1B_1 + A_2B_2 + \cdots + A_NB_N\ ,$$

or, in summation notation:

$$\vec{A}\cdot\vec{B} = \sum_{j=1}^{N} A_jB_j \tag{9.14}$$

Commutativity Relation

Implicit in the definition (9.14) is the commutativity of the dot product (in any number of dimensions):

$$\vec{A}\cdot\vec{B} = \vec{B}\cdot\vec{A} \tag{9.15}$$

Geometric Interpretation of Vectors

The definition (9.14) becomes more intuitive by studying the $N = 2$ case. Consider two arbitrary vectors given by

$$\vec{A} = \langle A\cos(\phi_A)\,, A\sin(\phi_A)\rangle$$
$$\vec{B} = \langle B\cos(\phi_B)\,, B\sin(\phi_B)\rangle\ ,$$

where A and B are the respecctive magnitudes. Calculating $\vec{A}\cdot\vec{B}$ using the formula provided results in

$$\vec{A}\cdot\vec{B} = AB\left(\cos(\phi_A)\cos(\phi_B) + \sin(\phi_A)\sin(\phi_B)\right)$$
$$\vec{A}\cdot\vec{B} = AB\cos(\phi_B - \phi_A)\ ,$$

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

$$\cos(\theta) = \frac{\vec{A}\cdot\vec{B}}{AB}\ , \tag{9.16}$$

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:

$$\vec{A}\cdot\vec{B} = 0 \tag{9.17}$$

In the Cartesian plane, recall that the slope of a line and another perpendicular line are represented by the vectors

$$\vec{m} = \langle m_x, m_y\rangle$$
$$\vec{m}_\perp = \langle -m_y, m_x\rangle\ ,$$

respectively. We verify these vectors to be orthogonal by calculating

$$\vec{m}\cdot\vec{m}_\perp = -m_xm_y + m_ym_x = 0\,.$$

Vector Magnitude

The dot product is responsible for the formula (9.1) for calculating the magnitude of a vector. Indeed, for an N-dimensional vector $\vec{A}$, we find the dot product with itself to be

$$\vec{A}\cdot\vec{A} = A_1^2 + A_2^2 + A_3^2 + \cdots + A_N^2 \,,$$

which is the square of the magnitude of A. More concisely:

$$A = \left|\vec{A}\right| = \sqrt{\vec{A}\cdot\vec{A}} \tag{9.18}$$

Distributive Property

For three vectors $\vec{A}$, $\vec{B}$, $\vec{C}$ of equal dimension, the dot product obeys the distributive property as one may expect:

$$\vec{A}\cdot\left(\vec{B}+\vec{C}\right) = \vec{A}\cdot\vec{B} + \vec{A}\cdot\vec{C} \tag{9.19}$$

Law of Cosines

An important relation from trigonometry called the *law of cosines* is derived using dot products. Consider the vector sum

$$\vec{A} - \vec{B} = \vec{C} \,,$$

and then square both sides:

$$\left(\vec{A}-\vec{B}\right)\cdot\left(\vec{A}-\vec{B}\right) = \vec{C}\cdot\vec{C}$$
$$\vec{A}\cdot\vec{A} + \vec{B}\cdot\vec{B} - 2\vec{A}\cdot\vec{B} = \vec{C}\cdot\vec{C}$$

Labeling θ as the angle between vectors $\vec{A}$, $\vec{B}$, the above simplifies to the law of cosines:

$$A^2 + B^2 - 2AB\cos(\theta) = C^2 \tag{9.20}$$

Note that all right triangles have $\theta = \pi/2$, in which case (9.20) 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 $N = 3$:

$$\vec{A} = \langle A_x, A_y, A_z\rangle$$
$$\vec{B} = \langle B_x, B_y, B_z\rangle$$

The cross product $\vec{A}\times\vec{B}$ is defined as

$$\vec{A}\times\vec{B} = \langle C_x, C_y, C_z\rangle \,, \tag{9.21}$$

where

$$C_x = A_yB_z - A_zB_y$$
$$C_y = A_zB_x - A_xB_z$$
$$C_z = A_xB_y - A_yB_x \,,$$

and is *orthogonal* to both $\vec{A}$ and $\vec{B}$.

Determinant Notation

The cross product formula (9.21) is tricky to memorize, and can be more transparently represented as a 'block of numbers' (*not* a matrix), sometimes called *determinant notation*:

$$\vec{A}\times\vec{B} = \begin{vmatrix} (x) & (-y) & (z) \\ A_x & A_y & A_z \\ B_z & B_y & B_z \end{vmatrix}$$

Without sweating the details of determinant notation, you can play a matching game between the determinant representation of $\vec{A}\times\vec{B}$ and the formula (9.21) to remember how it goes.

Orthogonality Check

To ensure that $\vec{A}\times\vec{B}$ is mutually orthogonal to $\vec{A}$,

$$\vec{A}\cdot\left(\vec{A}\times\vec{B}\right)$$

to see what comes out. In detail, the former case proceeds as

$$\begin{aligned}
&\vec{A}\cdot\left(\vec{A}\times\vec{B}\right) \\
&\quad = A_xA_yB_z - A_xA_zB_y + A_yA_zB_x \\
&\qquad - A_yA_xB_z + A_zA_xB_y - A_zA_yB_x \\
&\quad = B_z\left(A_xA_y - A_yA_x\right) - B_y\left(A_xA_z - A_zA_x\right) \\
&\qquad + B_z\left(A_xA_y - A_yA_x\right) \\
&\quad = 0 \,.
\end{aligned}$$

This also holds true for the B-case.

Null Case

The cross product of a vector with itself is identically zero:

$$\vec{A}\times\vec{A} = 0 \tag{9.22}$$

Anti-Commutativity Relation

Given the definition (9.21) of the cross product, one sees that swapping $\vec{A}$, $\vec{B}$ puts a minus sign on the result. This is known as the *anti-commutativity* of the cross product:

$$\vec{A}\times\vec{B} = -\vec{B}\times\vec{A} \tag{9.23}$$

Right Hand Rule

There is a trick that allows one to know the direction of $\vec{A} \times \vec{B}$ known as the (oft-dreaded) *right hand rule*. To know the direction of the vector $\vec{A} \times \vec{B}$, 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 $\vec{A}$, let your middle finger be vector $\vec{B}$.
3. Your thumb points along vector $\vec{A} \times \vec{B}$.

Geometric Interpretation

The definition (9.21) becomes more intuitive by studying a special case. Consider the pair of three-dimensional vectors confined to the xy-plane given by

$$\vec{A} = \langle A\cos(\phi_A), A\sin(\phi_A), 0\rangle$$
$$\vec{B} = \langle B\cos(\phi_B), B\sin(\phi_B), 0\rangle \ ,$$

where A and B are the respecctive magnitudes. Calculating $\vec{A} \times \vec{B}$ using the formula provided results in

$$\vec{A} \times \vec{B} = \langle 0, 0, AB(\cos\phi_A \sin\phi_B - \cos\phi_B \sin\phi_A)\rangle$$
$$\vec{A} \times \vec{B} = \langle 0, 0, AB\sin(\phi_B - \phi_A)\rangle \ ,$$

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

$$\sin(\theta) = \frac{\left|\vec{A} \times \vec{B}\right|}{AB} \ , \tag{9.24}$$

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 $\left|\vec{A} \times \vec{B}\right|$ is the area of the parallelogram with sides A, B.

4.3 Vector Identities

Consider three vectors $\vec{A}$, $\vec{B}$, $\vec{C}$, each of three dimensions.

Triple Product

The quantity

$$V = \vec{A} \cdot \left(\vec{B} \times \vec{C}\right) \tag{9.25}$$

is a scalar called the *triple product*. Intuitively, the triple product describes the volume of the parallelpiped with sides A, B, C. One can show by brute force that (9.25) obeys the cyclic relations:

$$\vec{A} \cdot \left(\vec{B} \times \vec{C}\right) = \vec{B} \cdot \left(\vec{C} \times \vec{A}\right) = \vec{C} \cdot \left(\vec{A} \times \vec{B}\right)$$

BAC-CAB Formula

A useful equation known as the *BAC-CAB* identity, reads

$$\vec{A} \times \left(\vec{B} \times \vec{C}\right) = \vec{B}\left(\vec{A} \cdot \vec{C}\right) - \vec{C}\left(\vec{A} \cdot \vec{B}\right) \ . \tag{9.26}$$

The proof of (9.26) is slightly long but straightforward, using (optional) determinant notation to contain the cross product:

$$\vec{A} \times \left(\vec{B} \times \vec{C}\right) = \begin{vmatrix} (x) & (-y) & (z) \\ A_x & A_y & A_z \\ B_yC_z - B_zC_y & B_zC_x - B_xC_z & B_xC_y - B_yC_x \end{vmatrix}$$

$$\begin{aligned}\vec{A} \times \left(\vec{B} \times \vec{C}\right) &= \langle A_yB_xC_y - A_yB_yC_x - A_zB_zC_x + A_zB_xC_z, 0, 0\rangle + \\ &\quad \langle 0, A_zB_yC_z - A_zB_zC_y - A_xB_xC_y + A_xB_yC_x, 0\rangle + \\ &\quad \langle 0, 0, A_xB_zC_x - A_xB_xC_z - A_yB_yC_z + A_yB_zC_y\rangle \\ &= B_x\langle A_yC_y + A_zC_z, 0, 0\rangle - A_xB_x\langle 0, C_y, C_z\rangle + \\ &\quad B_y\langle 0, A_zC_z + A_xC_x, 0\rangle - A_yB_y\langle C_x, 0, C_z\rangle + \\ &\quad B_z\langle 0, 0, A_xC_x + A_yC_y\rangle - A_zB_z\langle C_x, C_y, 0\rangle\end{aligned}$$

$$\vec{A} \times \left(\vec{B} \times \vec{C} \right) = B_x \left\langle \vec{A} \cdot \vec{C}, 0, 0 \right\rangle - A_x B_x \left\langle C_x, C_y, C_z \right\rangle + B_y \left\langle 0, \vec{A} \cdot \vec{C}, 0 \right\rangle - A_y B_y \left\langle C_x, C_y, C_z \right\rangle + B_z \left\langle 0, 0, \vec{A} \cdot \vec{C} \right\rangle - A_z B_z \left\langle C_x, C_y, C_z \right\rangle$$

$$\vec{A} \times \left(\vec{B} \times \vec{C} \right) = \vec{B} \left(\vec{A} \cdot \vec{C} \right) - \vec{C} \left(\vec{A} \cdot \vec{B} \right)$$

5 Polar Representation

In the Cartesian plane, consider a position vector

$$\vec{r} = \langle r_x, r_y \rangle \ .$$

The 'magnitude-and-direction' interpretation of $\vec{r}$ assigns the magnitude r to the hypotenuse of a right triangle, where the adjacent and opposite sides are respectively given by

$$r_x = r \cos(\phi) \tag{9.27}$$
$$r_y = r \sin(\phi) \ , \tag{9.28}$$

congruent with equations (9.1)-(9.2). The angle parameter ϕ is also known as the *phase* of the vector, a dimensionless argument unique on the interval $[0 : 2\pi)$.

5.1 Polar Coordinate System

Equations (9.27)-(9.28) 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). In particular, we take the position vecor in polar coordinates to be

$$\begin{aligned} \vec{r} &= \langle r \cos(\theta), r \sin(\theta) \rangle \\ &= r \langle \cos(\theta), \sin(\theta) \rangle \end{aligned}$$

5.2 Rotated Vectors

Starting with a vector $\vec{r}$ in two dimensions, particularly

$$\vec{r} = \langle r \cos(\phi), r \sin(\phi) \rangle \ ,$$

we may inquire what happens when we modify the phase such that $\phi \to \phi + \theta$, effectively rotating the vector in the plane.

Carrying this out, one writes

$$(\vec{r})' = r \langle \cos(\phi + \theta), \sin(\phi + \theta) \rangle \ ,$$

or, expanding the trigonometry terms,

$$\begin{aligned} r'_x &= r \left(\cos(\phi) \cos(\theta) - \sin(\phi) \sin(\theta) \right) \\ r'_y &= r \left(\sin(\phi) \cos(\theta) + \cos(\phi) \sin(\theta) \right) \ . \end{aligned}$$

Things get interesting when we keep simplifying:

$$r'_x = r_x \cos(\theta) - r_y \sin(\theta) \tag{9.29}$$
$$r'_y = r_x \sin(\theta) + r_y \cos(\theta) \tag{9.30}$$

Written this way, we see that the 'new' components r'_j are a mixture of the 'old' components r_j scaled by trigonometry terms that depend only on θ.

5.3 Rotation Matrix

Equations (9.29)-(9.30) can be packed into a single statement using *matrix notation*:

$$\begin{bmatrix} r'_1 \\ r'_2 \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{9.31}$$

Explicitly, we have made the associations

$$(\vec{r})' = \begin{bmatrix} r'_x \\ r'_y \end{bmatrix} = \begin{bmatrix} r'_1 \\ r'_2 \end{bmatrix}$$
$$\vec{r} = \begin{bmatrix} r_x \\ r_y \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \ ,$$

and the 'block of numbers' containing the trigonometry terms is called the *rotation* matrix, or *rotation operator*, denoted R:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \tag{9.32}$$

Whenever a matrix such as R occurs to the left of a vector such as $\vec{r}$ as in (9.31), there is an implied operation that 'applies' the matrix onto the vector by cross-multiplying certain components. This procedure is captured by writing (9.31) in component form:

$$r'_j = \sum_{k=1}^{2} R_{jk} \, r_k \tag{9.33}$$

6 Basis Vectors

6.1 Unit Vectors

Consider a vector $\vec{V}$ of dimension N, having magnitude V. A special vector $\hat{V}$, called a *unit vector*, is defined as $\vec{V}$ divided by its own magnitude:

$$\hat{V} = \frac{1}{V}\,\vec{V} \tag{9.34}$$

That is, a unit vector always has magnitude one, and points along the original vector. A vector of the form (9.34) is said to be *normalized*.

A more intuitive way to understand unit vectors is to rearrange (9.34) to write

$$\vec{V} = V\,\hat{V}\,,$$

which says that a full vector $\vec{V}$ is the product of the magnitude V and the 'direction' unit vector $\hat{V}$.

Problem 1

What is the vector that bisects the angle between two vectors $\vec{U}$, $\vec{V}$?

6.2 Introduction to Basis Vectors

Consider an arbitrary vector $\vec{V}$ of dimension N. A curious way to express

$$\vec{V} = \langle V_1, V_2, V_3, \ldots, V_N \rangle$$

is to fully pull apart each component so that $\vec{V}$ is the sum of N pure sub-vectors:

$$\begin{aligned}\vec{V} = {} & \langle V_1, 0, 0, \ldots \rangle \\ & + \langle 0, V_2, 0, \cdots \rangle + \langle \cdots, 0, V_3, 0, \cdots \rangle \\ & + \cdots + \langle \cdots, 0, V_N \rangle\end{aligned}$$

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 $\hat{e}_j$.

$$\begin{aligned}\hat{e}_1 &= \langle 1, 0, 0, \cdots, 0 \rangle \\ \hat{e}_2 &= \langle 0, 1, 0, \cdots, 0 \rangle \\ \hat{e}_3 &= \langle 0, 0, 1, \cdots, 0 \rangle \\ &\cdots \\ \hat{e}_N &= \langle 0, 0, 0, \cdots, 1 \rangle\end{aligned} \tag{9.35}$$

There is one basis vector $\hat{e}_j$ for each of the N dimensions in which the vector is situated.

Cartesian Coordinates

In the Cartesian xy-plane, a vector is typically represented as $\vec{V} = \langle V_x, V_y \rangle$, suggesting basis vectors

$$\begin{aligned}\hat{e}_x &= \langle 1, 0 \rangle \\ \hat{e}_y &= \langle 0, 1 \rangle\ .\end{aligned}$$

Note that the same notation extrapolates to three dimensions, in which case

$$\begin{aligned}\hat{e}_x &= \langle 1, 0, 0 \rangle \\ \hat{e}_y &= \langle 0, 1, 0 \rangle \\ \hat{e}_z &= \langle 0, 0, 1 \rangle\end{aligned}$$

are the basis vectors.

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

$$\hat{e}_j \cdot \hat{e}_k = 0\,. \tag{9.36}$$

On the other hand, two of the same basis vector $\hat{e}_k$ obeys

$$\hat{e}_k \cdot \hat{e}_k = 1\,. \tag{9.37}$$

In the general case, any set of basis vectors $\{\hat{e}_j\}$ that obeys (9.36), (9.37) is said to be *orthonormal*.

6.3 Linear Combinations

Having established the notion of basis vectors, we are free to an express arbitrary vector $\vec{V}$ as a *linear combination* of each $\hat{e}_j$, namely

$$\vec{V} = V_1\,\hat{e}_1 + V_2\,\hat{e}_2 + V_3\,\hat{e}_3 + \cdots + V_N\,\hat{e}_N\,, \tag{9.38}$$

or in summation notation:

$$\vec{V} = \sum_{j=1}^{N} V_j\,\hat{e}_j \tag{9.39}$$

In the above, $\vec{V}$ can potentially point to any 'place' in the N-dimensional space in which it lives. Such a place is formally called a *vector space*.

Vector Component Isolation

One may 'solve' for the V_jth component in a vector $\vec{V}$ by exploiting the orthogonality relations (9.36), (9.37). Start with (9.38), and multiply any particular $\hat{e}_k$ into both sides:

$$\hat{e}_k \cdot \vec{V} = V_1\,\hat{e}_k \cdot \hat{e}_1 + V_2\,\hat{e}_k \cdot \hat{e}_2 + \cdots + V_N\,\hat{e}_k \cdot \hat{e}_N$$

Next, observe that *all except one* of the dot products on the right will cancel due to (9.37). The whole sum collapses to the term with $j = k$, namely $V_k\,\hat{e}_k \cdot \hat{e}_k$,

simplifying to V_k. Formally, we have uncovered the obvious yet satisfying statement:

$$V_k = \vec{V} \cdot \hat{e}_k \tag{9.40}$$

With an explicit formula for the V_jth component of a vector, it's curious to see happens by replacing V_j in (9.39). Carrying this out, we can write a component-free way to reference a vector and its contents:

$$\vec{V} = \sum_{j=1}^{N} \left(\vec{V} \cdot \hat{e}_j \right) \hat{e}_j \tag{9.41}$$

Spanning the Vector Space

It's important to notice that a linear combination $\vec{V}$, 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 $\{\hat{e}_j\}$ are said to *span* the vector space.

7 Change of Basis

Consider a two-dimensional vector $\vec{V} = \langle V_x, V_y \rangle$, 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 9.5 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 $\vec{r}$ has a different representation in each basis.

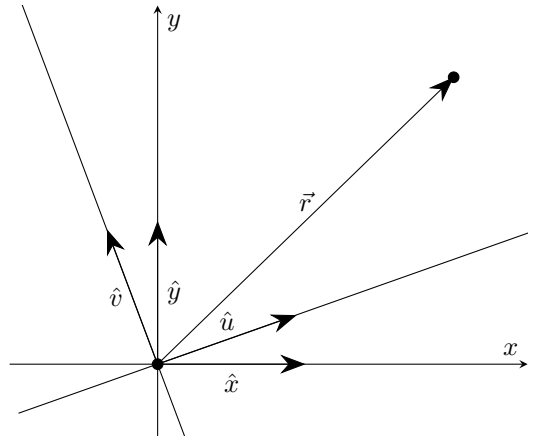


Figure 9.5: Vector $\vec{r}$ as a linear combination in two different bases.

Generalizing this idea to N diensions, we can say that linear combinations of the form (9.39) can be re-expressed in terms of a different set of orthonormal basis vectors $\{\hat{u}_j\}$:

$$(\vec{V})' = \sum_{j=1}^{N} V'_j \, \hat{u}_j \tag{9.42}$$

Analogous to (9.40), the primed components relate to the vector by:

$$V'_k = (\vec{V})' \cdot \hat{u}_k \tag{9.43}$$

7.1 Two Dimensions

Rotated Cartesian Coordinates

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 $\vec{r}$ is expressed in the $\{\hat{e}_j\}$ basis as the linear combination

$$\vec{r} = r_1 \, \hat{e}_1 + r_2 \, \hat{e}_2 \ ,$$

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:

$$\hat{e}_1 = \frac{\sqrt{3}}{2}\,\hat{u}_1 - \frac{1}{2}\,\hat{u}_2$$
$$\hat{e}_2 = \frac{1}{2}\,\hat{u}_1 + \frac{\sqrt{3}}{2}\,\hat{u}_2$$

Then, the vector $\vec{r}$ can be written

$$(\vec{r})' = r_1\left(\frac{\sqrt{3}}{2}\,\hat{u}_1 - \frac{1}{2}\,\hat{u}_2\right) + r_2\,\left(\frac{1}{2}\,\hat{u}_1 + \frac{\sqrt{3}}{2}\,\hat{u}_2\right)$$
$$(\vec{r})' = \left(r_1\frac{\sqrt{3}}{2} + r_2\frac{1}{2}\right)\hat{u}_1 + \left(-r_1\frac{1}{2} + r_2\frac{\sqrt{3}}{2}\right)\hat{u}_2\,,$$

where the components $r_{1,2}$ are finally readable as

$$r_1' = r_1\frac{\sqrt{3}}{2} + r_2\frac{1}{2}$$
$$r_2' = -r_1\frac{1}{2} + r_2\frac{\sqrt{3}}{2}\,,$$

and a form like (9.42) is attained:

$$(\vec{r})' = r_1'\,\hat{u}_1 + r_2'\,\hat{u}_2$$

General Coordinate Rotations

The above example can be easily generalized such that $\hat{u}$ points anywhere in the Cartesian plane, with $\hat{v}$ appropriately perpendicular to $\hat{u}$. To achieve this, we introduce an arbitrary parameter θ such that

$$\hat{u}_1 = \cos(\theta)\,\hat{e}_1 + \sin(\theta)\,\hat{e}_2$$
$$\hat{u}_2 = -\sin(\theta)\,\hat{e}_1 + \cos(\theta)\,\hat{e}_2\,.$$

By straightforward algebra, we find the inverted version to be

$$\hat{e}_1 = \cos(\theta)\,\hat{u}_1 - \sin(\theta)\,\hat{u}_2$$
$$\hat{e}_2 = \sin(\theta)\,\hat{u}_1 + \cos(\theta)\,\hat{u}_2\,.$$

The pairs of equations above are suggestive of a matrix formulation, particularly

$$\begin{bmatrix}\hat{u}_1\\ \hat{u}_2\end{bmatrix} = \begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\hat{e}_1\\ \hat{e}_2\end{bmatrix}\,, \tag{9.44}$$

and

$$\begin{bmatrix}\hat{e}_1\\ \hat{e}_2\end{bmatrix} = \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\hat{u}_1\\ \hat{u}_2\end{bmatrix}\,, \tag{9.45}$$

respectively. Comparing the above to (9.32), we see (9.45) contains the same rotation matrix R that rotates vectors in a fixed basis. Denoting the other matrix in (9.44) as $\tilde{R}$, the component version of the above reads:

$$\hat{u}_j = \sum_{k=1}^{2}\tilde{R}_{jk}\hat{e}_k$$
$$\hat{e}_j = \sum_{k=1}^{2}R_{jk}\hat{u}_k\,.$$

With a convenient representation for the basis vectors $\{\hat{e}_j\}$, an arbitrary linear combination

$$\vec{r} = r_1\,\hat{e}_1 + r_2\,\hat{e}_2$$

becomes

$$(\vec{r})' = r_1\sum_{k=1}^{2}R_{1k}\,\hat{u}_k + r_2\sum_{k=1}^{2}R_{2k}\,\hat{u}_k$$
$$= \sum_{k=1}^{2}\left(\sum_{j=1}^{2}R_{jk}\,r_j\right)\hat{u}_k\,,$$

telling us that the kth component of the vector $(\vec{r})'$ is given by

$$r_k' = \sum_{j=1}^{2}R_{jk}\,r_j\,, \tag{9.46}$$

which is a cousin to equation (9.33). To do a fair comparison, let us swap the j-index and the k-index in (9.46) to write

$$r_j' = \sum_{k=1}^{2}R_{kj}\,r_k\,.$$

Looking carefully, the above differs from (9.33) by the order of the subscripts on the R-term, ultimately equivalent to reversing the sign on θ. Said another way, a 'positive' rotation in the basis vectors with $\vec{r}$ fixed is equivalent to a 'negative' rotation of $\vec{r}$ with the basis fixed.

7.2 N Dimensions

Change of Basis Vectors

At the center of the change-of-basis problem is the issue of relating the two orthonormal bases $\hat{e}_j$, $\hat{u}_j$ to one another. In N dimensions, the basis vectors are related by linear combinations

$$\hat{u}_j = \sum_{k=1}^{N}\tilde{U}_{jk}\,\hat{e}_k \tag{9.47}$$
$$\hat{e}_j = \sum_{k=1}^{N}U_{jk}\,\hat{u}_k\,. \tag{9.48}$$

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 $\hat{e}_m$, $\hat{u}_m$, respectively into (9.47), (9.48):

$$\hat{e}_m \cdot \hat{u}_j = \sum_{k=1}^{N} \tilde{U}_{jk}\, \hat{e}_m \cdot \hat{e}_k$$
$$\hat{u}_m \cdot \hat{e}_j = \sum_{k=1}^{N} U_{jk}\, \hat{u}_m \cdot \hat{u}_k$$

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:

$$\tilde{U}_{jm} = \hat{e}_m \cdot \hat{u}_j = \hat{u}_j \cdot \hat{e}_m \tag{9.49}$$
$$U_{jm} = \hat{u}_m \cdot \hat{e}_j = \hat{e}_j \cdot \hat{u}_m \tag{9.50}$$

From this, deduce also that

$$\tilde{U}_{jm} = U_{mj}\ . \tag{9.51}$$

Linear Combinations

An arbitrary linear combination

$$\vec{V} = \sum_{j=1}^{N} V_j\, \hat{e}_j$$

can be written with all $\hat{e}_j$ replaced according to (9.48):

$$(\vec{V})' = \sum_{j=1}^{N}\sum_{k=1}^{N} U_{jk}\, V_j\, \hat{u}_k = \sum_{k=1}^{N}\left(\sum_{j=1}^{N} U_{jk}\, V_j\right)\hat{u}_k$$

Comparing the above to (9.42) gives a formula for the kth component of the vector $(\vec{V})'$:

$$(V')_k = \sum_{j=1}^{N} U_{jk}\, V_j \tag{9.52}$$

Unity Condition

We can learn a bit more about the matrix components U_{jk}, $\tilde{U}_{jk}$ by eliminating $\hat{u}_k$ between (9.47)-(9.48):

$$\hat{e}_j = \sum_{k=1}^{N} U_{jk}\left(\sum_{m=1}^{N} \tilde{U}_{km}\, \hat{e}_m\right)$$
$$= \sum_{k=1}^{N}\sum_{m=1}^{N}\left(U_{jk}\tilde{U}_{km}\right)\hat{e}_m$$

Next, use the symbol I to represent the quantity

$$I_{jm} = \sum_{k=1}^{N}\left(U_{jk}\tilde{U}_{km}\right)\ ,$$

and the above becomes

$$\hat{e}_j = \sum_{m=1}^{N} I_{jm}\, \hat{e}_m\ .$$

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 (9.36)-(9.37).

8 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.

8.1 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 9.6 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.

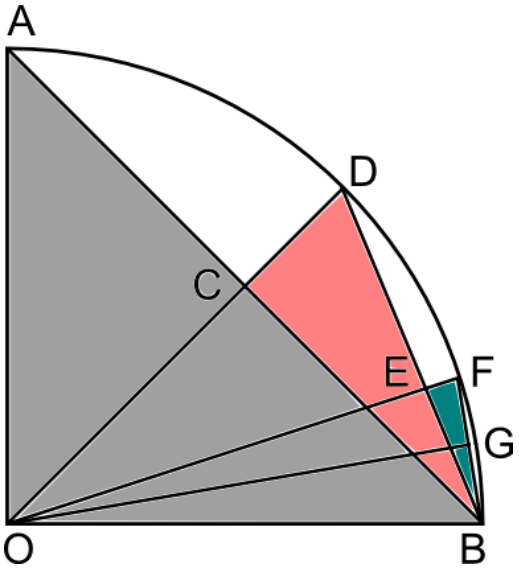


Figure 9.6: Covering a quarter circle with triangles.

To establish some notation, let the lines OA, OB define a pair of unit vectors:

$$\hat{i} = \overline{OB}$$
$$\hat{j} = \overline{OA}$$

Also, let the hypotenuse of AOB be a vector $\vec{h}_0$ such that

$$\vec{h}_0 = \hat{i} - \hat{j}\,,$$

with magnitude

$$\left|\vec{h}_0\right| = \sqrt{2}\,.$$

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

$$A_0 = \frac{1}{2}\left|\hat{i}\right|\left|\hat{j}\right|\,,$$

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 9.6 these are depicted DCA, DCB, respectively.

Analyzing DCB, consider a unit vector $\hat{x}_1$ and a shorter vector $\vec{x}_1$ such that

$$\vec{x}_1 = \overline{OC} = \frac{\hat{i}+\hat{j}}{2}$$
$$\hat{x}_1 = \overline{OD} = \frac{\hat{i}+\hat{j}}{\sqrt{2}}\,,$$

whose difference in length is CD.

Then, the area of DCB is

$$\begin{aligned} A_1 &= \frac{(CD)\,(CB)}{2} \\ &= \frac{1}{2}\left|\hat{x}_1 - \vec{x}_1\right|\frac{1}{2}\left|\vec{h}_0\right|\,. \end{aligned}$$

That is, the base is line CB, whose length is half the magnitude $\vec{h}_0$. The height CB is given by the difference in x-vectors.

Notice that, because $\hat{x}_1$, $\vec{x}_1$ are parallel, the following simplification can be made:

$$\left|\hat{x}_1 - \vec{x}_1\right| = 1 - x_1$$

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 $\vec{h}_1$ and is given by

$$\vec{h}_1 = \hat{i} - \hat{x}_1\,,$$

and has length DB. Note that the vector $\vec{h}_1$ doesn't play into the area of DCB, rather the previous $\vec{h}_0$ is used. Explicitly, the vector $\vec{h}_1$ reads

$$\vec{h}_1 = \left(1 - \frac{1}{\sqrt{2}}\right)\hat{i} + \frac{1}{\sqrt{2}}\,\hat{j}\,,$$

having magnitude

$$\left|\vec{h}_1\right| = \sqrt{2-\sqrt{2}}\,.$$

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

$$\vec{x}_2 = \overline{OE} = \hat{x}_1 + \frac{1}{2}\vec{h}_1$$
$$\hat{x}_2 = \overline{OF} = \vec{x}_2/\left|\vec{x}_2\right|\,,$$

whose difference in length is EF. In detail, it's straightforward to show that

$$\vec{x}_2 = \frac{1}{2}\left(1 + \frac{1}{\sqrt{2}}\right)\hat{i} + \frac{1}{2\sqrt{2}}\,\hat{j}\,,$$

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

$$\begin{aligned} A_2 &= \frac{(EF)\,(EB)}{2} \\ &= \frac{1}{2}\left|\hat{x}_2 - \vec{x}_2\right|\frac{1}{2}\left|\vec{h}_1\right| \\ &= \frac{1}{4}\,(1-x_2)\left|\vec{h}_1\right|\,, \end{aligned}$$

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 $\vec{h}_2$, given by

$$\vec{h}_2 = \hat{i} - \hat{x}_2\,,$$

which is also just like the the formula for $\vec{h}_1$ with the indices bumped by one. Explicitly, the vector $\vec{h}_2$ reads

$$\vec{h}_2 = \left(1 - \frac{(1+1/\sqrt{2})}{\sqrt{2+\sqrt{2}}}\right)\hat{i} + \left(\frac{\sqrt{2-\sqrt{2}}}{2}\right)\hat{j}\,,$$

having magnitude

$$\left|\vec{h}_2\right| = \sqrt{2-\sqrt{2+\sqrt{2}}}\,.$$

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

$$\begin{aligned}\vec{x}_3 &= \hat{x}_2 + \frac{1}{2}\vec{h}_2\\ \hat{x}_3 &= \vec{x}_3/\left|\vec{x}_3\right|\,,\end{aligned}$$

implying the area to be

$$\begin{aligned}A_3 &= \frac{1}{2}\left|\hat{x}_3 - \vec{x}_3\right|\frac{1}{2}\left|\vec{h}_2\right|\\ &= \frac{1}{4}\left(1-x_3\right)\left|\vec{h}_2\right|\,,\end{aligned}$$

and furthermore:

$$\vec{h}_3 = \hat{i} - \hat{x}_3$$

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 nth step, it follows that

$$\begin{aligned}\vec{x}_n &= \hat{x}_{n-1} + \frac{1}{2}\vec{h}_{n-1}\\ \hat{x}_n &= \vec{x}_n/\left|\vec{x}_n\right|\\ \vec{h}_n &= \hat{i} - \hat{x}_n\,,\end{aligned}$$

with

$$A_n = \frac{1}{4}\left(1-x_n\right)\left|\vec{h}_{n-1}\right|\,.$$

While we'll take the above as a workable result, note that

$$\left|\vec{h}_{n-1}\right| = \sqrt{2\left(1-\hat{i}\cdot\hat{x}_{n-1}\right)}\,.$$

With this, the area simplifies to

$$A_n = \frac{\sqrt{2}}{4}\left(1-x_n\right)\sqrt{1-\hat{i}\cdot\hat{x}_{n-1}}\,,$$

which only works for $n \geq 2$.

Also, it's easy to derive

$$\hat{i}\cdot\hat{x}_{n-1} = \cos\left(\frac{\pi}{2^n}\right)$$

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:

$$\begin{aligned}A_4 = &-\frac{2\sqrt{2-\sqrt{2+\sqrt{2}}}}{16}\\ &+\frac{4\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2}}}}}{16}\end{aligned}$$

N-Sided Polygon

Tallying all areas of all triangles up to the Nth step gives one quarter the area of an N-sided polygon with equal angles and equal sides. Doing so, we write

$$P\left(N\right) = \sum_{n=0}^{N} w_n A_n\,,$$

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 $n \geq 2$:

$$P\left(N\right) = \sum_{n=0}^{N}\frac{2^n}{4}\left(1-x_n\right)\left|\vec{h}_{n-1}\right| = \sum_{n=0}^{N} P_n$$

Evaluating P(N)

Now the real work begins. Starting with $N = 0$ and working up, we find

$$\begin{aligned}P_0 &= 2^0 A_0 = \frac{1}{2}\\ P_1 &= 2^1 A_1 = -\frac{1}{2} + \frac{1}{\sqrt{2}}\\ P_2 &= 2^2 A_2 = -\frac{1}{\sqrt{2}} + \sqrt{2-\sqrt{2}}\\ P_3 &= 2^3 A_3 = -\sqrt{2-\sqrt{2}} + 2\sqrt{2-\sqrt{2+\sqrt{2}}}\,,\end{aligned}$$

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 Nth 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

$$P(N) = \frac{2^N}{4}\sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2+\cdots}}}}$$

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:

$$\begin{aligned}
\Pi(0) &= 2 \\
\Pi(1) &= 2.8284271248\ldots \\
\Pi(2) &= 3.0614674589\ldots \\
\Pi(3) &= 3.1214451523\ldots \\
\Pi(4) &= 3.1365484905\ldots \\
\Pi(5) &= 3.1403311570\ldots \\
\Pi(10) &= 3.1415914215\ldots \\
\Pi(15) &= 3.1415926524\ldots \\
\Pi(20) &= 3.1415926536\ldots
\end{aligned}$$

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 $N \to \infty$, the triangles cover the unit circle and the area converges to π:

$$\pi = \lim_{N\to\infty} \Pi(N) \ ,$$

or

$$\pi = \lim_{N\to\infty} 2^N \sqrt{2-\sqrt{2+\sqrt{2+\sqrt{2+\cdots}}}} \, .$$

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.

9 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.

9.1 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

$$A\vec{x} = \vec{y} \, ,$$

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:

$$\sum_{k=1}^{N} A_{jk} \, x_k = y_j \qquad (9.53)$$

$$j = 1, 2, 3, \ldots, M$$

9.2 Matrix Components

Consider two vectors $\vec{x}$, $\vec{y}$, each a linear combination in some N-dimensional basis such that

$$\vec{x} = \sum_{j=1}^{N} x_j \, \hat{e}_j$$
$$\vec{y} = \sum_{j=1}^{M} y_j \, \hat{e}_j \, .$$

While $\vec{y}$ is perfectly happy being expressed as a linear combination in the basis $\{\hat{e}_j\}$, it's instructive to re-express $\vec{y}$ in terms of its brother, $\vec{x}$. To do so, we propose an operator A such that

$$\vec{y} = A\vec{x} \, .$$

To proceed, write the above as

$$\sum_{k=1}^{M} y_k \, \hat{e}_k = \sum_{k=1}^{N} x_k \, A\hat{e}_k \, ,$$

and multiply the basis vector $\hat{e}_j$ (via dot product) into both sides:

$$\sum_{k=1}^{M} y_k \, \hat{e}_j \cdot \hat{e}_k = \sum_{k=1}^{N} x_k \, \hat{e}_j \cdot A\hat{e}_k$$

On the left, every term in the sum vanishes except that with $j = k$, and the above becomes

$$y_j = \sum_{k=1}^{N} (\hat{e}_j \cdot A\hat{e}_k) \, x_k$$
$$j = 1, 2, 3, \ldots, M \, .$$

The parenthesized quantity is what we're after:

$$A_{jk} = \hat{e}_j \cdot A\hat{e}_k \tag{9.54}$$

The term A_{jk} is the component of the matrix A corresponding to the jth row, kth column.

9.3 Projector

Consider the curious quantity

$$P_x = \vec{x} \, \vec{x} \, , \tag{9.55}$$

called the the *projector* of $\vec{x}$. By itself, P_x does nothing - there is no operation between the two copies of $\vec{x}$. What the projector *does* is 'wait' to be multiplied into another vector, resulting in a scaled version of $\vec{x}$. For example, applying the projector to a different vector $\vec{y}$ (of the same dimension as $\vec{x}$) goes like

$$P_x \vec{y} = \vec{x} \, (\vec{x} \cdot \vec{y}) \, .$$

9.4 Identity Operator

Consider a vector $\vec{x}$ as a linear combination in some N-dimensional basis:

$$\vec{x} = \sum_{j=1}^{N} x_j \, \hat{e}_j$$

For any one of the basis vectors $\hat{e}_k$, write the projector

$$P_{e_k} = \hat{e}_k \, \hat{e}_k \, ,$$

and then multiply $\vec{x}$ onto the right side to get

$$P_{e_k} \vec{x} = \hat{e}_k \, \hat{e}_k \cdot \vec{x} = x_k \, \hat{e}_k$$

By summing over the index k, the right side is identically $\vec{x}$:

$$\left(\sum_{k=1}^{N} P_{e_k} \right) \vec{x} = \sum_{k=1}^{N} x_k \, \hat{e}_k = \vec{x}$$

For the left side to also equal $\vec{x}$, the parenthesized quantity must be equivalent to 'multiplying by one', which we call the *identity* operator:

$$I = \sum_{k=1}^{N} P_{e_k} \tag{9.56}$$

The identity operator leaves a vector unchanged:

$$I\vec{x} = \vec{x}$$

The matrix-equivalence of I is square, has no mixed components, and has ones along the diagonal:

$$I = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{9.57}$$

9.5 Unified Matrix Notation

Recall that a matrix A relates to its components A_{jk} in a way given by (9.54), namely

$$A_{jk} = \hat{e}_j \cdot A\hat{e}_k \, .$$

To establish this directly using projectors, start with $A = IAI$ and watch what happens:

$$A = IAI = \sum_{j=1}^{M} \sum_{k=1}^{N} P_{e_j} A P_{e_k}$$
$$= \sum_{j=1}^{M} \sum_{k=1}^{N} \hat{e}_j \, (\hat{e}_j \cdot A\hat{e}_k) \, \hat{e}_k$$

The parenthesized quantity is precisely A_{jk}. Evidently, the symbolic notation unifies with the index notation in the equation

$$A = \sum_{j=1}^{M} \sum_{k=1}^{N} \hat{e}_j \, (A_{jk}) \, \hat{e}_k \tag{9.58}$$

The presence of $\hat{e}_j \, \hat{e}_k$ is like a projector - it couples the component to the operator.

10 Matrix Operations

10.1 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:

$$A_{jk} + B_{jk} = C_{jk} \qquad \begin{cases} j = 1, 2, 3, \ldots, M \\ k = 1, 2, 3, \ldots, N \end{cases} \tag{9.59}$$

10.2 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:

$$\alpha A_{jk} = B_{jk} \qquad \begin{cases} j = 1, 2, 3, \ldots, M \\ k = 1, 2, 3, \ldots, N \end{cases} \tag{9.60}$$

10.3 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:

$$AB \neq BA \tag{9.61}$$

Multiplication Formula

To derive the formula for matrix multiplication, begin with the following 'unified' representation (9.58) 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

$$\begin{aligned} AB &= \sum_{m=1}^{M} \sum_{k=1}^{K} \hat{e}_m \, (A_{mk}) \, \hat{e}_k \sum_{k'=1}^{K} \sum_{n=1}^{N} \hat{e}_{k'} \, (B_{k'n}) \, \hat{e}_n \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sum_{n=1}^{N} \hat{e}_m \, (A_{mk}) \, \hat{e}_k \, \hat{e}_{k'} \, (B_{k'n}) \, \hat{e}_n \end{aligned}$$

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 (9.36)-(9.37), namely

$$\hat{e}_k \cdot \hat{e}_{k'} = \begin{cases} 1 & k = k' \\ 0 & k \neq k' \end{cases} \, ,$$

which has the effect of equating $k = k'$ in the above, eliminating one of the sums. So far then, we have

$$\begin{aligned} AB = C &= \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{n=1}^{N} (A_{mk} B_{kn}) \, \hat{e}_m \, \hat{e}_n \\ C &= \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\sum_{k=1}^{K} A_{mk} B_{kn} \right) \hat{e}_m \, \hat{e}_n \, . \end{aligned}$$

The symbol C has replaced the quantity AB on the left. Comparing the right side to (9.58), we conclude that the component C_{mn} of matrix C is given by the famed *matrix multiplication* formula:

$$C_{mn} = \sum_{k=1}^{K} A_{mk} B_{kn} \qquad \begin{cases} m = 1, 2, 3, \ldots, M \\ n = 1, 2, 3, \ldots, N \end{cases} \tag{9.62}$$

Equation (9.62) 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:

$$(AB)\,C = A\,(BC) \tag{9.63}$$

10.4 Change of Basis

A square matrix A with components A_{jk} in the basis $\{\hat{e}_j\}$ can be represented by (9.58):

$$A = \sum_{j=1}^{N} \sum_{k=1}^{N} \hat{e}_j\,(A_{jk})\,\hat{e}_k\ .$$

Under a change of basis $\{\hat{e}_j\} \to \{\hat{u}_j\}$, we can use (9.48)

$$\hat{e}_j = \sum_{k=1}^{N} \tilde{U}_{jk}\,\hat{u}_k$$

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} due to (9.51). The parenthesized quantity is precisely the formula for the component A'_{mn} of the transformed matrix

$$A'_{mn} = \sum_{j=1}^{N} \sum_{k=1}^{N} U_{mj} A_{jk} \tilde{U}_{kn}\ , \tag{9.64}$$

or in symbolic form,

$$A' = U A \tilde{U}\ .$$

Note that the above verifies the associativity rule (9.63) for matrix multiplication. The order in which the sums are taken directly corresponds to which matrices are multiplied first. As a bonus, (9.64) tells us exactly how to take the product of three square matrices.

Chapter 10

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

$$\sqrt[3]{\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} = U + \sqrt{-1}\,V$$
$$\sqrt[3]{\frac{c}{2} - \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} = U - \sqrt{-1}\,V$$

for two unknown coefficients U and V. Then, when we assemble x_0 again, the V-terms cancel,

$$x_0 = U + \cancel{\sqrt{-1}\,V} + U - \cancel{\sqrt{-1}\,V} = 2U\,,$$

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

$$\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}} = U\left(U^2 - 3V^2\right) + \sqrt{-1}\,V\left(3U^2 - V^2\right)\,,$$

availing the connection:

$$\text{Real part of}\left(\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}\right) = U\left(U^2 - 3V^2\right)$$
$$\text{Imag. part of}\left(\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}\right) = V\left(3U^2 - V^2\right)$$

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

$$x^3 + bx - c = (x - x_0)\left(x^2 + x_0 x + \frac{c}{x_0}\right)\,.$$

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:

$$x^2 + x_0 x + \frac{c}{x_0} = 0$$

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

$$\begin{aligned} 18 &= U\left(U^2 - 3V^2\right) \\ 26 &= V\left(3U^2 - V^2\right)\,. \end{aligned}$$

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

Formally, let the complex number z exist as as an ordered pair of two (real) numbers called *components* a, b such that

$$z = (a, b)\ . \tag{10.1}$$

Complex Conjugate

For every complex number z, there exists the *complex conjugate*, also a complex number, denoted $\overline{z}$ or z^*, by flipping the sign on b:

$$\overline{z} = z^* = (a, -b)\ . \tag{10.2}$$

Relationship to Real Numbers

A subtlety worth highlighting is that the complex number $z = (a, 0)$ is the same as the scalar $z = a$:

$$(a, 0) \leftrightarrow a \tag{10.3}$$

Problem 2

Check that the complex conjugate of the complex conjugate recovers the complex number:

$$\overline{\overline{z}} = z \tag{10.4}$$

2.2 Complex Arithmetic

Scalar Multiplication

Complex numbers can be 'scaled' by real numbers called *scalars*:

$$\lambda z = (\lambda a, \lambda b) \tag{10.5}$$

Complex Addition

For two complex numbers $z_1 = (a_1, b_1)$, $z_2 = (a_1, b_1)$, their sum is

$$z_1 + z_2 = (a_1 + a_2, b_1 + b_2)\ . \tag{10.6}$$

Complex Multiplication

For two complex numbers $z_1 = (a_1, b_1)$, $z_2 = (a_1, b_1)$, their product is simultaneously defined by a commutation relation and a conjugate relation:

$$z_1 \cdot z_2 = z_2 \cdot z_1 \tag{10.7}$$

$$\overline{z_1} \cdot \overline{z_2} = \overline{z_1 \cdot z_2} \tag{10.8}$$

Conspicuously absent from the list of axioms is an explicit formula for complex multiplication. For completeness, the formula for complex multiplication reads

$$z_1 \cdot z_2 = (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1)\ , \tag{10.9}$$

but this is does not need to be axiomatic unless you're in a hurry. Instead, we'll soon derive Equation (10.9) from the equations preceding it.

2.3 Properties of Addition

Isolating Complex Components

Given a complex number $z = (a, b)$ and its complex conjugate $\overline{z} = (a, -b)$, take their sum and difference, respectively, to write a pair of relations that 'solve for' a, b:

$$(a, 0) = \frac{z + \overline{z}}{2} \tag{10.10}$$

$$(0, b) = \frac{z - \overline{z}}{2} \tag{10.11}$$

Note that a and b in isolation are each real numbers. As they appear in $z = (a, b)$, 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:

$$z_1 + z_2 = z_2 + z_1 \tag{10.12}$$
$$\overline{z_1} + \overline{z_2} = \overline{z_1 + z_2} \tag{10.13}$$

Problem 3

Verify Equations (10.12) and (10.13) 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 (10.9) 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 $a_{1,2}$, $b_{1,2}$

$$z_1 \cdot z_2 = (a_1, b_1) \cdot (a_2, b_2) = (q, r) \ ,$$

where

$$q = \alpha\, a_1 a_2 + \beta\, a_1 b_2 + \gamma\, a_2 b_1 + \delta\, b_1 b_2$$
$$r = \tilde{\alpha}\, a_1 a_2 + \tilde{\beta}\, a_1 b_2 + \tilde{\gamma}\, a_2 b_1 + \tilde{\delta}\, b_1 b_2 \ ,$$

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 (10.8), 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:

$$q = \alpha\, a_1 a_2 + \delta\, b_1 b_2$$
$$r = \tilde{\beta}\, a_1 b_2 + \tilde{\gamma}\, a_2 b_1 \ ,$$

Impose the commutation relation (10.7), 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:

$$q = \alpha\, (a_1 a_2 - b_1 b_2)$$
$$r = \tilde{\beta}\, (a_1 b_2 + a_2 b_1)$$

Let $b_2 = 0$. The corresponding product becomes

$$z_1 \cdot z_2 = \left(\alpha\, a_1 a_2, \tilde{\beta}\, a_2 b_1\right) = a_2 \left(\alpha\, a_1, \tilde{\beta}\, b_1\right) \ ,$$

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 $z_j = (a_j, b_j)$ with $j = 1, 2, 3$ obey the *associative property*

$$(z_1 \cdot z_2) \cdot z_3 = z_1 \cdot (z_2 \cdot z_3) \ , \tag{10.14}$$

shown by brute force:

$$\begin{aligned}
(z_1 \cdot z_2) \cdot z_3 &= (a_1 a_2 - b_1 b_2, a_1 b_2 + b_1 a_2) \cdot (a_3, b_3) \\
&= (a_1 a_2 a_3 - b_1 b_2 a_3 - a_1 b_2 b_3 - b_1 a_2 b_3, \\
&\qquad a_1 a_2 b_3 - b_1 b_2 b_3 + a_1 b_2 a_3 + b_1 a_2 a_3) \\
&= (a_1\, (a_2 a_3 - b_2 b_3) - b_1\, (b_2 a_3 + a_2 b_3)\,, \\
&\qquad a_1\, (a_2 b_3 + b_2 a_3) + b_1\, (b_2 b_3 - a_2 a_3)) \\
&= (a_1, b_1) \cdot (a_2 a_3 - b_2 b_3, a_2 b_2 + b_2 a_3) \\
&= z_1 \cdot (z_2 \cdot z_3)
\end{aligned}$$

Distributive Property

Complex numbers $z_j = (a_j, b_j)$ with $j = 1, 2, 3$ also obey the *distributive property*:

$$z_1 \cdot (z_2 + z_3) = (z_1 \cdot z_2) + (z_2 \cdot z_3) \ , \tag{10.15}$$

also shown by brute force:

$$\begin{aligned}
z_1 \cdot (z_2 + z_3) &= (a_1, b_1) \cdot (a_2 + a_3, b_2 + b_3) \\
&= (a_1 a_2 + a_1 a_3 - b_1 b_2 - b_1 b_3, \\
&\qquad a_1 b_2 + a_1 b_3 + a_2 b_1 + a_3 b_1) \\
&= (a_1 a_2 - b_1 b_2, a_1 b_2 + a_2 b_1) + \\
&\qquad (a_1 a_3 - b_1 b_3, a_1 b_3 + a_3 b_1) \\
&= (z_1 \cdot z_2) + (z_1 \cdot z_3)
\end{aligned}$$

2.5 Complex Magnitude

The quantity

$$|z| = \sqrt{z \cdot \overline{z}} \tag{10.16}$$

is called the *magnitude* of z, and is always a non-negative real number:

$$\begin{aligned} |z| &= \sqrt{z \cdot \overline{z}} = \sqrt{(a,b) \cdot (a,-b)} \\ &= \sqrt{(a^2+b^2, 0)} = \sqrt{a^2+b^2} \end{aligned}$$

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:

$$d = \frac{z_1}{z_2} = \frac{(a_1, b_1)}{(a_2, b_2)}$$

To make a useful complex number from this, multiply the top and bottom by $\overline{z}_2$:

$$d = \frac{(a_1, b_1)}{(a_2, b_2)} \frac{(a_2, -b_2)}{(a_2, -b_2)} = \frac{z_1 \cdot \overline{z}_2}{|z_2|^2} \tag{10.17}$$

Explicitly, the above means:

$$d = \frac{1}{a_2^2 + b_2^2} (a_1 a_2 + b_1 b_2, -a_1 b_2 + a_2 b_1) \tag{10.18}$$

Problem 4

Prove that the complex division operation obeys its own conjugate relation:

$$\overline{z}_1 / \overline{z}_2 = \overline{z_1 / z_2} \tag{10.19}$$

2.7 Generalized Complex Arithmetic

If we seek to write a tighter set of axioms, one could start with the general equations

$$\begin{aligned} z_1 \star z_2 &= z_2 \star z_1 \\ \overline{z_1} \star \overline{z_2} &= \overline{z_1 \star z_2} \end{aligned}$$

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 (10.6) using only Equations (10.12) and (10.13).

2.8 Imaginary Unit

Having established the notion of complex multiplication, now consider the product:

$$(0,1) \cdot (0,1) = (0 - 1 \cdot 1, 0) = -1$$

The left side has two instances of $(0,1)$, which ought to mean:

$$\sqrt{(0,1) \cdot (0,1)} = (0,1)$$

Suddenly though, without ever asking, we have an answer for the meaning of $\sqrt{-1}$. Apply the square root operation across the whole equation to find

$$(0,1) = \sqrt{-1}\,,$$

known as the *imaginary unit*. Often, the fundamental unit is denoted as i, meaning

$$i^2 = -1\,.$$

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 Complex Numbers as Operators

For a complex number $z = (a, b)$, consider the 'identity' statement

$$(1,0) \cdot (a,b) = (a,b)\ . \tag{10.20}$$

Equation (10.20) is the 'least invasive' operation in which z_1 can particpate, which is to merely multiply by one.

Seeking a corresponding identity for $(0,1)$, we find, by complex multiplication,

$$(0,1) \cdot (a,b) = (-b,a)\ . \tag{10.21}$$

The 'operator' $(0,1)$ 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 $(1,0)$ and $(0,1)$ as 'basis vectors', and write a linear combination with two undetermined coefficients α, β:

$$\alpha\,(1,0) + \beta\,(0,1) = (\alpha, 0) + (0, \beta) = (\alpha, \beta)$$

Almost obviously, such an operation is nothing more than a complex number (α, β). Trying this 'operator' on a different complex number $z = (a, b)$, we simply have

$$(\alpha, \beta) \cdot (a, b) = (\alpha a - \beta b, \alpha b + \beta a) \ , \tag{10.22}$$

or more concisely,

$$(\alpha, \beta) \cdot z = z' = (a', b') \ .$$

3.2 Rotation Operator

Let us find an operator (just a complex number) $z_\phi = (\alpha_\phi, \beta_\phi)$ that acts on $z = (a, b)$ such that the components change but the magnitude does not. From Equation (10.22), 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

$$|z'| = |z| \sqrt{\alpha_\phi^2 + \beta_\phi^2} \, . \tag{10.23}$$

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

$$\alpha_\phi = \cos(\phi) \tag{10.24}$$
$$\beta_\phi = \sin(\phi) \ , \tag{10.25}$$

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 $(1, 0)$, $(0, 1)$. Let us therefore take the *complex rotation operator* to be the complex number

$$z_\phi = (\cos(\phi), \sin(\phi)) \ . \tag{10.26}$$

3.3 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

$$|z| = r \tag{10.27}$$

is the product

$$z = r z_\phi = (r \cos(\phi), r \sin(\phi)) = (a, b) \ . \tag{10.28}$$

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

$$a = r \cos(\phi) \tag{10.29}$$
$$b = r \sin(\phi) \ , \tag{10.30}$$

easily inverted:

$$r = |z| = \sqrt{a^2 + b^2} \tag{10.31}$$
$$\phi = \arctan\left(\frac{b}{a}\right) \tag{10.32}$$

As a matter of terminology, recall that the a- and b- components of z are the real and imaginary parts, i.e.

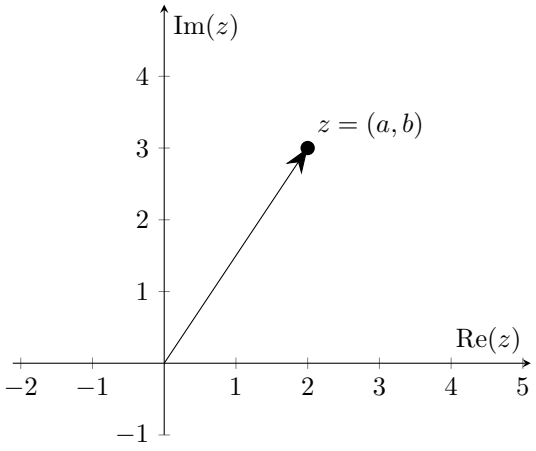
$$\begin{aligned} a &= \mathrm{Re}(z) \\ b &= \mathrm{Im}(z) \ . \end{aligned}$$

The phase angle ϕ is often denoted $\mathrm{Arg}(z)$, or 'argument of z'. All together, an equivalent statement of (10.32) reads

$$\mathrm{Arg}(z) = \arctan\left(\frac{\mathrm{Im}(z)}{\mathrm{Re}(z)}\right) \ . \tag{10.33}$$

3.4 Real and Imaginary Axes

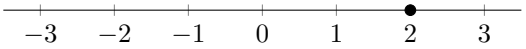
The r, ϕ interpretation of complex numbers leads to an almost-identical mathematical apparatus needed for plane polar coordinates. Complex numbers occupy a space that is analogous to the Cartesian xy-plane, except the x-axis is replaced by the real axis, and the y-axis is replaced by the imaginary axis. In this connection we speak freely of the *complex plane*:



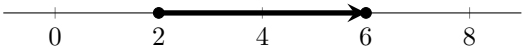
Any 'location' on the $\mathrm{Im}(z) = 0$ line is a purely real number $z = (a, 0) = a$, whereas anywhere on the $\mathrm{Re}(z) = 0$ is a purely imaginary number $z = (0, b)$. Any off-axis location is a complex number $z = (a, b)$.

3.5 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:

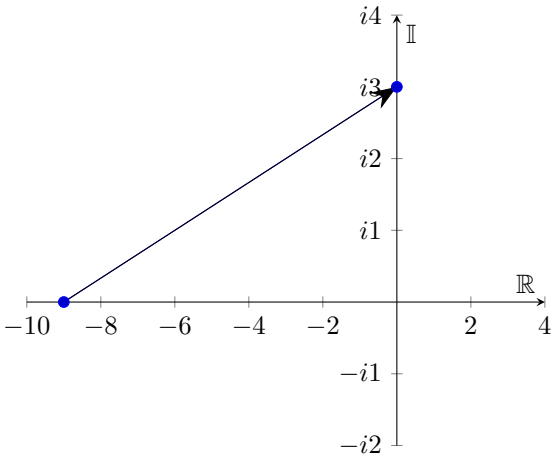


In fact, all of the 'well-behaved' operations that one could possibly perform on x will land *somewhere* on the number line.

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

$$x^2 = -9$$
$$x = \sqrt{-9} = \sqrt{-1} \times \sqrt{9} = 3\sqrt{-1}$$

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:



3.6 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 (10.29)-(10.30) 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:

$$\cos(\phi_1 - \phi_2) = \frac{a_1 a_2 + b_1 b_2}{|z_1|\,|z_2|} \tag{10.34}$$

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

$$\mathbf{z}_1 \cdot \mathbf{z}_2 = |z_1|\,|z_2| \cos(\phi_1 - \phi_2)\ . \tag{10.35}$$

Note that the quantity $a_1 a_2 + b_1 b_2$ can be written yet another way, namely

$$a_1 a_2 + b_1 b_2 = \frac{1}{2}\left(\overline{z}_1 \cdot z_2 + z_1 \cdot \overline{z}_2\right)\ , \tag{10.36}$$

allowing a tight relationship to be written:

$$\mathbf{z}_1 \cdot \mathbf{z}_2 = \frac{1}{2}\left(\overline{z}_1 \cdot z_2 + z_1 \cdot \overline{z}_2\right) \tag{10.37}$$

Problem 8

Write equations analogous to (10.34), (10.35), and (10.36) to derive the cross product analog to (10.37):

$$\mathbf{z}_1 \times \mathbf{z}_2 = \frac{1}{2}\left(\overline{z}_1 \cdot z_2 - z_1 \cdot \overline{z}_2\right) \tag{10.38}$$

Problem 9

Derive:

$$\overline{z}_1 \cdot z_2 = (\mathbf{z}_1 \cdot \mathbf{z}_2, \mathbf{z}_1 \times \mathbf{z}_2) \tag{10.39}$$

Problem 10

Using vectors as an analogy, derive the triangle inequality for complex numbers:

$$||z_1| - |\,z_2| \le |z_1 + z_2| \le |z_1| + |z_2| \tag{10.40}$$

3.7 Embedded Complex Numbers

As a matter of curiosity, one may wonder if it makes sense to work with 'embedded complex numbers' such as

$$((a,b),c) \qquad (a,(b,c)) \ .$$

In the same way that equation (10.3) permits the association $(a,0) \leftrightarrow a$, let us extend this by writing

$$((a,b),0) \leftrightarrow (a,b) \ .$$

Using the above, we find, for the first case:

$$\begin{aligned} ((a,b),c) &= ((a,b),0) + (0,c) \\ &= (a,b) + (0,c) \\ &= (a,b+c) \end{aligned}$$

Evidently, the 'embedded' complex number $((a,b),c)$ flattens down to the 'ordinary' complex number $(a,b+c)$. Let's work out the second case in similar fashion:

$$\begin{aligned} (a,(b,c)) &= (a,0) + (0,(b,c)) \\ &= (a,0) + (0,1)\cdot((b,c),0) \\ &= (a,0) + (0,1)\cdot(b,c) \\ &= (a,0) + (-c,b) \\ &= (a-c,b) \end{aligned}$$

Once again, the embedded complex number flattens down to a (different) ordinary complex number. For this reason, it follows that any embedding can be flattened to an ordinary complex number, and the whole notion is essentially redundant.

4 Euler's Formula

Now we derive the most important equation in all of complex analysis, called *Euler's formula*.

4.1 Repeated Rotations

To kick things off, recall how the complex rotation operator (10.26) acts on any complex number $z = (a,b)$ to produce a new complex number z' such that

$$z' = (\cos(\phi), \sin(\phi)) \cdot z \ ,$$

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

$$z' = (\cos(\theta_1+\theta_2), \sin(\theta_1+\theta_2)) \cdot z \ ,$$

simplifying nicely to

$$z' = (\cos(\theta_1), \sin(\theta_1)) \cdot (\cos(\theta_2), \sin(\theta_2)) \cdot z \ .$$

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

$$z' = \left(\cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right)\right)^n \cdot z \ .$$

(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:

$$(\cos(\phi), \sin(\phi)) = \left(\cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right)\right)^n \quad (10.41)$$

4.2 Infinite Rotations

Given the re-interpred rotation operator (10.41), 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

$$\cos(\phi/n)$$
$$\sin(\phi/n)$$

as the argument ϕ/n goes to zero. Borrowing from 'elementary' trigonometry and precalculus though, the relations

$$\lim_{n\to\infty} \cos\left(\frac{\phi}{n}\right) = 1 \quad (10.42)$$

$$\lim_{n\to\infty} \sin\left(\frac{\phi}{n}\right) = \frac{\phi}{n} \quad (10.43)$$

apply, letting us write

$$\lim_{n\to\infty} \left(\cos\left(\frac{\phi}{n}\right), \sin\left(\frac{\phi}{n}\right)\right)^n = \lim_{n\to\infty} \left(1, \frac{\phi}{n}\right)^n \ .$$

Summarizing again, we now see that the rotation operator is an infinite repetition of small rotations:

$$(\cos(\phi), \sin(\phi)) = \lim_{n\to\infty} \left(1, \frac{\phi}{n}\right)^n \quad (10.44)$$

4.3 Invoking Euler's Constant

Looking again at the complex number $(1, \phi/n)$, split the components via

$$\left(1, \frac{\phi}{n}\right) = (1,0) + \left(0, \frac{\phi}{n}\right) .$$

First, note that $(1,0)$ 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

$$\left(1, \frac{\phi}{n}\right) = 1 + \frac{1}{n}(0, \phi) .$$

Plugging this back into (10.44) gives

$$(\cos(\phi), \sin(\phi)) = \lim_{n\to\infty} \left(1 + \frac{1}{n}(0,\phi)\right)^n . \quad (10.45)$$

The right-side of the above should remind us of another notion from precalculus, namely Euler's constant, defined as:

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n ,$$

or in more useful form,

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$$

for any real number x.

4.4 Euler's Formula

Euler reasoned that the formula for e^x could be modified (it's *his* formula, after all) to receive complex arguments:

$$e^z = \lim_{n\to\infty} \left(1 + \frac{z}{n}\right)^n$$

If this is the case, the right side of (10.45) can be written as en exponential

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}(0,\phi)\right)^n = e^{(0,\phi)} ,$$

which is amazing, because it means the *left* side of (10.45) is an exponential! Putting these results together, we write

$$(\cos(\phi), \sin(\phi)) = e^{(0,\phi)} , \quad (10.46)$$

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

$$(0,1) = \sqrt{-1} = i ,$$

thus the above becomes:

$$\cos(\phi) + i\sin(\phi) = e^{i\phi}$$

Setting $\phi = \pi$ reveals a remarkable connection between key players of mathematics:

$$0 = 1 + e^{i\pi}$$

4.5 Polar Form of Complex Numbers

Let us revisit the radius-and-phase construction of a complex number. For a complex number $z = (a,b)$, recall that equations (10.29)-(10.32)

$$\begin{aligned} a &= r\cos(\phi) \\ b &= r\sin(\phi) \\ r &= |z| = \sqrt{a^2+b^2} \\ \phi &= \arctan(b/a) \end{aligned}$$

allow us to understand z as a rotation of the real number r 'up into' the complex plane by an angle ϕ:

$$z = (a,b) = r(\cos(\phi), \sin(\phi))$$

With Euler's formula in hand, we may replace the trigonometry terms via (10.46), and arrive at the most elegant expression of a complex number in terms of radius and phase, the so-called polar form:

$$z = r\, e^{(0,\phi)} \quad (10.47)$$

Problem 11

Show that:

$$\cos(n\arccos(x)) = \sum_{k=0}^{n/2} \binom{n}{2k} x^{n-2k}\left(x^2-1\right)^k$$

Hint: Let $x = \cos(\theta)$ and arrive at

$$\cos(n\theta) = \mathrm{Re}\left((\cos(\theta) + i\sin(\theta))^n\right) .$$

4.6 Multiplication and Division

Given the polar form (10.47) of complex numbers, the multiplication formula (10.9) and the division formula (10.18) can be revised. For two complex numbers $z_1(r_1,\phi_1)$, $z_2(r_2,\phi_2)$, we have

$$\begin{aligned} z_1 \cdot z_2 &= r_1 r_2\, e^{(0,\phi_1+\phi_2)} \\ &= r_1 r_2 (\cos(\phi_1+\phi_2), \sin(\phi_1+\phi_2)) \end{aligned} \quad (10.48)$$

for multiplication, and for division:

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{r_1}{r_2}\, e^{(0,\phi_1-\phi_2)} \\ &= \frac{r_1}{r_2}(\cos(\phi_1-\phi_2), \sin(\phi_1-\phi_2)) \end{aligned} \quad (10.49)$$

Problem 12

Derive (10.48) and (10.49).

4.7 Trigonometric Functions

Euler's formula lends to a curious representation of the standard trigonometric functions. Start with (10.46) and let $\phi \to -\phi$ to write

$$(\cos(\phi), -\sin(\phi)) = e^{(0,-\phi)},$$

and then add both equations (also divide by two) to isolate the cosine:

$$(\cos(\phi), 0) = \frac{1}{2}\left(e^{(0,\phi)} + e^{(0,-\phi)}\right) \tag{10.50}$$

An analogous procedure isolates the sine:

$$(0, \sin(\phi)) = \frac{1}{2}\left(e^{(0,\phi)} - e^{(0,-\phi)}\right) \tag{10.51}$$

Recalling that the complex rotation operator can be written

$$z_\phi = (\cos(\phi), \sin(\phi)),$$

note that the above can expressed more tightly in accordance with (10.10)-(10.11):

$$(\cos(\phi), 0) = \frac{z_\phi + \overline{z}_\phi}{2}$$
$$(0, \sin(\phi)) = \frac{z_\phi - \overline{z}_\phi}{2}$$

4.8 Hyperbolic Functions

Continuing the discussion of trigonometric functions, isolate the real number ϕ and apply the operator $(0,1)$, resulting in $(0,\phi)$ without question. Next take the complex number $(0,\phi)$ and apply the same operator to find

$$(0,1)\cdot(0,\phi) = -\phi\,.$$

The reason we do this is to start with Euler's formula (10.46), and multiply each instance of ϕ by the complex number $(0,1)$, giving

$$(\cos((0,\phi)), \sin((0,\phi))) = e^{-\phi}\,.$$

A similar set of steps leads to a version with $-\phi \to \phi$, or

$$(\cos((0,\phi)), -\sin((0,\phi))) = e^{\phi}\,.$$

Notice the quantity on the right is a real number, which must mean

$$(0, \sin((0,\phi))) = (0,1)\cdot\sin((0,\phi))$$

is also real, thus $\sin((0,\phi))$ by itself is imaginary.

The above can be written:

$$\cos((0,\phi)) + (0,1)\cdot\sin((0,\phi)) = e^{-\phi}$$
$$\cos((0,\phi)) - (0,1)\cdot\sin((0,\phi)) = e^{\phi}$$

Take the sum, and the sin-term cancels, allowing $\cos((0,\phi))$ to be isolated:

$$\cos((0,\phi)) = \frac{1}{2}\left(e^{\phi} + e^{-\phi}\right)$$

Similarly, take the difference and the cos-term vanishes:

$$-(0,1)\cdot\sin((0,\phi)) = \frac{1}{2}\left(e^{\phi} - e^{-\phi}\right)$$

It turns out that these 'complex trigonometry' terms above have special names, the *hyperbolic cosine* and *hyperbolic sine*, respectively:

$$\cosh(\phi) = \cos((0,\phi)) = \frac{1}{2}\left(e^{\phi} + e^{-\phi}\right) \tag{10.52}$$

$$\sinh(\phi) = -(0,1)\cdot\sin((0,\phi)) = \frac{1}{2}\left(e^{\phi} - e^{-\phi}\right) \tag{10.53}$$

5 Roots and Branches

5.1 Complex Natural Logarithm

Start again with Euler's polar form (10.47)

$$z = r\, e^{(0,\phi)},$$

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

$$\begin{aligned} e^{(0,\phi\pm 2\pi n)} &= (\cos(\phi \pm 2\pi n), \sin(\phi \pm 2\pi n)) \\ &= (\cos(\phi), \sin(\phi)) \\ &= e^{(0,\phi)} \end{aligned}$$

holds for any integer n.

Now things get interesting. The presence of e in (10.47) beckons trying the natural logarithm (ln) on both sides, resulting in:

$$\ln(z) = \ln(r) + (0,\phi) \tag{10.54}$$

Unlike $e^{(0,\phi)}$, the complex natural logarithm does not 'reset' at 2π. The phase term $(0,\phi)$ in (10.54) 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

$$z^{1/2} = \left(r\,e^{(0,\phi)}\right)^{1/2} = r^{1/2}\,e^{(0,\phi/2)}\,. \tag{10.55}$$

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

$$\sqrt{z\,(r,\phi)} = \sqrt{r}\,e^{(0,\phi/2)}$$
$$\sqrt{z\,(r,\phi+2\pi)} = -\sqrt{r}\,e^{(0,\phi/2)}\,,$$

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 nth root problem

$$z^n = a \tag{10.56}$$

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 (10.47), where we recast each variable as

$$z = |z|\,e^{(0,\phi)}$$
$$a = |a|\,e^{(0,\theta+2\pi m)}\,.$$

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 (10.56) gives

$$|z|^n\,e^{(0,n\phi)} = |a|\,e^{(0,\theta+2\pi m)}\,.$$

Then, the radial and angular components on each side are equal:

$$|z|^n = |a|$$
$$n\phi = \theta + 2\pi m$$

Solving the above for $|z|$, ϕ, we have

$$|z| = |a|^{1/n} \tag{10.57}$$

$$\phi = \frac{\theta}{n} + \frac{2\pi m}{n}\,, \tag{10.58}$$

where

$$m = 0, 1, 2, \ldots, (n-1)\,.$$

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 $z = (a,b)$, $w = (c,d)$. Starting with the polar expression $z = |z|\,e^{(0,\phi)}$, the complex exponent z^w calculation is slightly nontrivial:

$$\begin{aligned} z^w &= \left(|z|\,e^{(0,\phi)}\right)^{(c,d)} \\ &= e^{\ln|z|(c,d)}\,e^{(-d\phi,0)}\,e^{(0,c\phi)} \\ z^w &= \exp\left(c\ln|z| - d\phi, d\ln|z| + c\phi\right) \end{aligned} \tag{10.59}$$

6 Complex Functions

The complex exponential (10.47), complex natural logarithm (10.54), 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

$$w\,(z) = \left(u\,(x,y)\,, v\,(x,y)\right)\,, \tag{10.60}$$

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

$$z\,(w) = \left(f\,(u,v)\,, g\,(u,v)\right)\,. \tag{10.61}$$

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 (10.54) and the complex square root (10.55) 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 an abrupt jump in phase. Such 'fault lines' are curves called *branch cuts*.

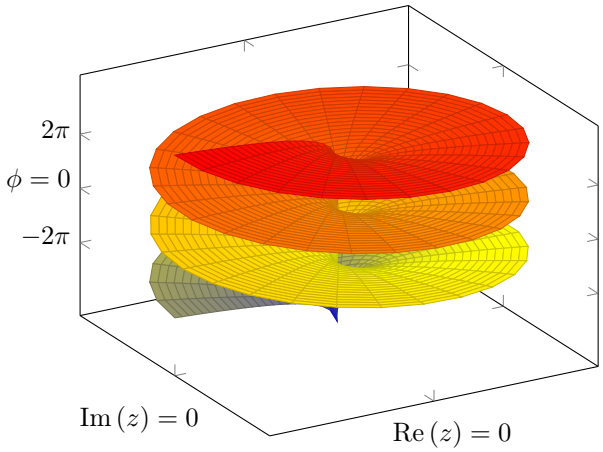


Figure 10.1: Complex natural logarithm $\ln(z) = \ln(r) + (0, \phi)$.

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 (10.55) 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 $w(z) = \ln(z)$, esaily inverted:

$$z(w) = e^w = e^{(u,v)} = (e^u \cos(v), e^u \sin(v))$$

In this case, it makes sense to choose a polar parameterization with $r = e^u$:

$$\begin{aligned} X(r, v) &= r\cos(v) \\ Y(r, v) &= r\sin(v) \\ Z(r, v) &= \ln(r) + v \end{aligned}$$

The 'plot variables' are uppercase symbols X, Y, Z.

Plotting this system with $(r > 0, v \in [-3\pi : 3\pi])$ leads to Fig. 10.1. 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. 10.1. 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. 10.1, and flatting it down onto the complex plane as shown in Figure 10.2. (Color may vary.)

The bold-shaded lines indicate integer r, ϕ, respectively. Specifically, the circle corresponds to $r = 1$ (with $r = 2$ outside of the plot), and the straight spokes correspond to

$$\phi = 0, 1, 2, 3, -3, -2, -1\,,$$

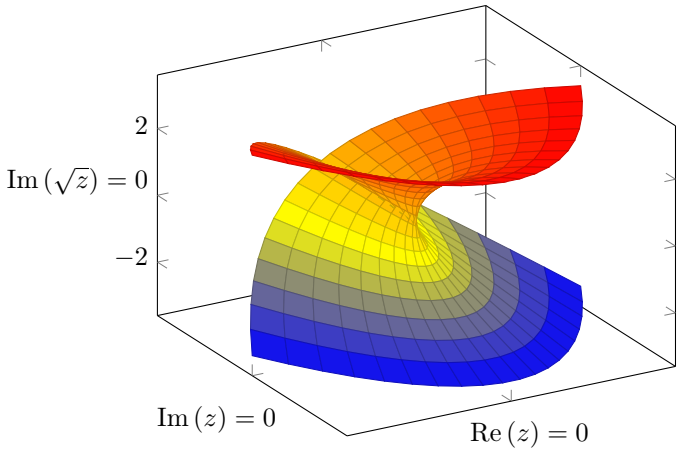


Figure 10.3: Complex square root $z^{1/2}$.

listing counterclockwise from $\phi = 0$.

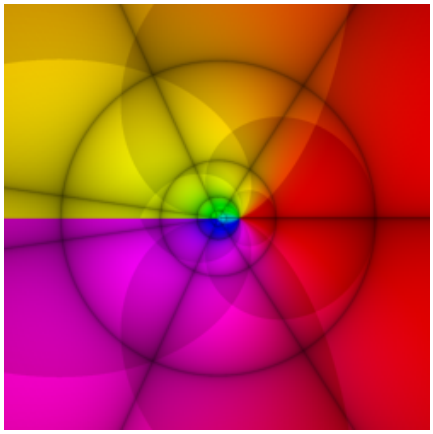


Figure 10.2: Complex logarithm on the branch $(-\pi : \pi]$.

Complex Square Root

The Riemann surface recipe also applies to plotting the square root function $w(z) = z^{1/2}$. Seeking a form like (10.61), we easily write

$$z(w) = w^2 = (u, v) \cdot (u, v) = \left(u^2 - v^2, 2uv\right) .$$

Turning this into a three-dimensional system, we write

$$\begin{aligned} X(u, v) &= u^2 - v^2 \\ Y(u, v) &= 2uv \\ Z(u, v) &= u , \end{aligned}$$

while treating u, v as parameters. Plotting this system near $(u = 0, v = 0)$ leads to Fig. 10.3. 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 10.4. The bold-shaded lines correspond to positive integer outputs of $z^{1/2}$. Below the line $\mathrm{Re}(z) = 0$, the sign on $\mathrm{Im}(z)$ flips, but $\mathrm{Re}(z)$ 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

$$\begin{aligned} z_0 &= \sqrt{|a|}\, e^{(0,\theta/2)} \\ z_1 &= -\sqrt{|a|}\, e^{(0,\theta/2)} . \end{aligned}$$

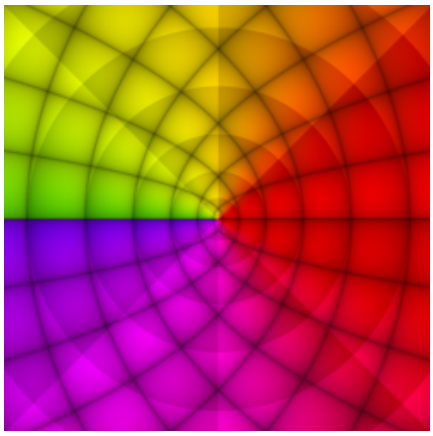


Figure 10.4: Complex square root on the branch $(-\pi : \pi]$.

Chapter 11

Linear Systems

1 Linear Systems

1.1 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

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 &= b_1 \\ A_{21}x_1 + A_{22}x_2 &= b_2\,, \end{aligned} \tag{11.1}$$

admitting a clean matrix representation

$$A\vec{x} = \vec{b}\,,$$

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 $\vec{x}$ depends on the quantity $A_{11}A_{22} - A_{12}A_{21}$, called the *determinant* of the matrix A:

$$\det A = \det \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A_{11}A_{22} - A_{12}A_{21} \tag{11.2}$$

If $\det A$ is nonzero, the vector $\vec{x}$ 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 $\vec{b}$). 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

$$x_1 = \frac{1}{\det A}\det\begin{bmatrix} b_1 & A_{12} \\ b_2 & A_{22}\end{bmatrix} = \frac{\det C_1}{\det A}$$
$$x_2 = \frac{1}{\det A}\det\begin{bmatrix} A_{11} & b_1 \\ A_{21} & b_2\end{bmatrix} = \frac{\det C_2}{\det A}$$

That is, the solution to the two-dimensional linear system (11.1) with nonzero determinant is

$$x_j = \frac{\det C_j}{\det A} \qquad (11.3)$$
$$j = 1, 2\,,$$

with

$$C_1 = \begin{bmatrix} b_1 & A_{12} \\ b_2 & A_{22}\end{bmatrix} \qquad (11.4)$$
$$C_2 = \begin{bmatrix} A_{11} & b_1 \\ A_{21} & b_2\end{bmatrix}\,.$$

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.

1.2 Order-N Formalism

Generalizing the 2×2 linear system to have M equations and N unknowns, we begin with

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1N}x_N &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \cdots + A_{2N}x_N &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \cdots + A_{3N}x_N &= b_3 \\ &\cdots \\ A_{M1}x_1 + A_{M2}x_2 + A_{M3}x_3 + \cdots + A_{MN}x_N &= b_M\,, \end{aligned} \qquad (11.5)$$

admitting the matrix representation $A\vec{x} = \vec{b}$:

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2N} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{M1} & A_{M2} & A_{M3} & \cdots & A_{MN}\end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_N\end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ b_M\end{bmatrix} \qquad (11.6)$$

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 $\vec{x}$, namely

$$x_j = \frac{\det C_j}{\det A} \qquad (11.7)$$
$$j = 1, 2, 3, \ldots, N$$

The matrix C_j is constructed by starting with A and replacing the jth column with $\vec{b}$. That is:

$$C_j = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{1j} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & b_{2j} & \cdots & A_{2N} \\ A_{31} & A_{32} & \cdots & b_{3j} & \cdots & A_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{N1} & A_{N2} & \cdots & b_{Nj} & \cdots & A_{NN}\end{bmatrix} \qquad (11.8)$$

1.3 Row Operations

An important tool set called *row operations* can be applied to matrices. To illustrate these, consider again the N-dimensional linear system (11.5). 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:

$$E^{23}A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44}\end{bmatrix}$$

$$M^2_\alpha A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ \alpha A_{31} & \alpha A_{32} & \alpha A_{33} & \alpha A_{34} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44}\end{bmatrix}$$

$$R^2_3 A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{31}+A_{21} & A_{32}+A_{22} & A_{33}+A_{23} & A_{34}+A_{24} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{41} & A_{42} & A_{43} & A_{44}\end{bmatrix}$$

Note that the subscripts and superscripts on the symbols E, M, R are mere convenience of notation, often-omitted.

2 Determinants

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 (11.2).

2.1 Three Dimensions

Consider the three-dimensional linear system

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 &= b_3 \,, \end{aligned} \tag{11.9}$$

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 $R_2 - R_1$:

$$A \to \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & \frac{A_{11}A_{22}}{A_{21}} - A_{12} & \frac{A_{11}A_{23}}{A_{21}} - A_{13} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

Next, (iii) multiply R_3 by a factor of A_{11}/A_{31}, and (iv) replace R_3 with $R_3 - R_1$:

$$A \to \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & \frac{A_{11}A_{22}}{A_{21}} - A_{12} & \frac{A_{11}A_{23}}{A_{21}} - A_{13} \\ 0 & \frac{A_{11}A_{32}}{A_{31}} - A_{12} & \frac{A_{11}A_{33}}{A_{31}} - A_{13} \end{bmatrix}$$

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 (11.2) to write

$$\begin{aligned} \det A \propto \left(\frac{A_{11}A_{22}}{A_{21}} - A_{12} \right) \left(\frac{A_{11}A_{33}}{A_{31}} - A_{13} \right) \\ - \left(\frac{A_{11}A_{23}}{A_{21}} - A_{13} \right) \left(\frac{A_{11}A_{32}}{A_{31}} - A_{12} \right) , \end{aligned}$$

which after simplifying, becomes

$$\begin{aligned} \det A \left(\frac{A_{21}A_{31}}{A_{11}} \right) \propto A_{11} \left(A_{22}A_{33} - A_{32}A_{23} \right) \\ - A_{12} \left(A_{21}A_{33} - A_{31}A_{23} \right) \\ + A_{13} \left(A_{21}A_{32} - A_{31}A_{22} \right) . \end{aligned}$$

Keepping 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

$$\begin{aligned} \det A = A_{11} \left(A_{22}A_{33} - A_{32}A_{23} \right) \\ - A_{12} \left(A_{21}A_{23} - A_{31}A_{33} \right) \\ + A_{13} \left(A_{21}A_{32} - A_{31}A_{22} \right) . \end{aligned} \tag{11.10}$$

2.2 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

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} ,$$

permits a similar process to reduce the order of the problem. Doing so, the order-four determinant becomes the sum of four terms:

$$\begin{aligned}\det A = A_{11}\det\begin{bmatrix}A_{22} & A_{23} & A_{24}\\ A_{32} & A_{33} & A_{34}\\ A_{42} & A_{43} & A_{44}\end{bmatrix} - A_{12}\det\begin{bmatrix}A_{21} & A_{23} & A_{24}\\ A_{31} & A_{33} & A_{34}\\ A_{41} & A_{43} & A_{44}\end{bmatrix} \\ + A_{13}\det\begin{bmatrix}A_{21} & A_{22} & A_{24}\\ A_{31} & A_{32} & A_{34}\\ A_{41} & A_{42} & A_{44}\end{bmatrix} - A_{14}\det\begin{bmatrix}A_{12} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\\ A_{41} & A_{42} & A_{43}\end{bmatrix}\end{aligned} \tag{11.11}$$

2.3 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 jth row and the kth 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 (11.10), (11.11) can be written:

$$\begin{aligned}\det A_{(3)} &= A_{11}M_{11} - A_{12}M_{12} + A_{13}M_{13}\\ \det A_{(4)} &= A_{11}M_{11} - A_{12}M_{12} + A_{13}M_{13} - A_{14}M_{14}\end{aligned}$$

2.4 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-anddright 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:

$$\begin{aligned}&\text{if } N=1: \quad \det A = \det\begin{bmatrix}A_{11}\end{bmatrix}\\ &\text{if } N>1: \quad \det A = \sum_{\substack{j\le N\\ k=1}}^{N}(-1)^{j+k}A_{jk}M_{jk}\end{aligned} \tag{11.12}$$

Note too that the summation can take place over rows *or* columns, meaning

$$\text{if } N>1: \quad \det A = \sum_{\substack{j\le N\\ k=1}}^{N}(-1)^{j+k}A_{kj}M_{kj}$$

also holds. Note that the variable j is fixed at some integer less than N. Only the k-variable is summed over.

2.5 Properties

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:

$$\det\left(A^T\right) = \det A$$

3 Inverse Matrix

3.1 Definition

Given a square matrix A of dimension N, there *may* exist a special matrix A^{-1} that obeys the property

$$A^{-1}A = AA^{-1} = I\,, \tag{11.13}$$

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 vise 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 $\vec{x}$ to a vector $\vec{b}$ as in $A\vec{x} = \vec{b}$. By multiplying A^{-1} into both sides of $A\vec{x} = \vec{b}$, we end up with

$$A^{-1}A\vec{x} = A^{-1}\vec{b}\,,$$

effectively 'solving' for the vector $\vec{x}$:

$$\vec{x} = A^{-1}\vec{b} \tag{11.14}$$

3.2 Formula

To come up with a formula for the inverse of A, consider another matrix B defined from the components A_{jk} such that

$$B_{jk} = (-1)^{j+k}\, M_{kj}\,, \tag{11.15}$$

where M_{kj} are the matrix minors of A. As innocent as it looks, (11.15) 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 (11.15), 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 (11.12). 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 (11.13), suggests the combination $B/\det A$ is equal to the inverse of A:

$$A^{-1} = \frac{1}{\det A} B \tag{11.16}$$

3.3 Products

The inverse of the product of two matrices is equal to the product of the individual inverses in reversed order:

$$(AB)^{-1} = B^{-1}A^{-1} \tag{11.17}$$

3.4 Cramer's Rule Derived

For a linear system of N dimensions, we start with the dichotomy

$$\begin{aligned} A\vec{x} &= \vec{b} \\ \vec{x} &= A^{-1}\vec{b}\,, \end{aligned}$$

where accounting for (11.16), the 'solution' vector $\vec{x}$ is written

$$\vec{x} = \frac{1}{\det A} B\vec{b}\,,$$

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, in the sense that the calculation $M\vec{b}$ does *not* represent this situation. Instead, the sum constitutes the determinant of a matrix modified from A such that the kth column is replaced by $\vec{b}$. If this situation sounds familiar, it precisely describes the matrix introduced as equation (11.8)

$$C_j = \begin{bmatrix} A_{11} & A_{12} & \cdots & b_{1j} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & b_{2j} & \cdots & A_{2N} \\ A_{31} & A_{32} & \cdots & b_{3j} & \cdots & A_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{N1} & A_{N2} & \cdots & b_{Nj} & \cdots & A_{NN} \end{bmatrix}\,,$$

and the above reduces to Cramer's rule (11.7) for the soltuion of the system:

$$\begin{aligned} x_j &= \frac{\det C_j}{\det A} \\ j &= 1, 2, 3, \ldots, N \end{aligned}$$

3.5 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 (11.2). To calculate the inverse, begin with the B matrix given by (11.15), coming out to

$$B = \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}\,.$$

Then, by the inverse formula (11.16), the inverse of the 2×2 matrix reads:

$$A^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix} \tag{11.18}$$

3.6 Three Dimensions

The three-dimensional matrix with

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}$$

has a total of nine minors M_{jk}, readily readable from A. Constructing the matrix B using

$$B_{jk} = (-1)^{j+k} M_{kj} \,,$$

we find

$$\begin{aligned} B_{11} &= (-1)^2 (A_{22}A_{33} - A_{32}A_{23}) \\ B_{12} &= (-1)^3 (A_{12}A_{33} - A_{32}A_{13}) \\ B_{13} &= (-1)^4 (A_{12}A_{23} - A_{22}A_{13}) \\ B_{21} &= (-1)^3 (A_{21}A_{33} - A_{31}A_{23}) \\ B_{22} &= (-1)^4 (A_{11}A_{33} - A_{31}A_{13}) \\ B_{23} &= (-1)^5 (A_{11}A_{23} - A_{21}A_{13}) \\ B_{31} &= (-1)^4 (A_{21}A_{32} - A_{31}A_{22}) \\ B_{32} &= (-1)^5 (A_{11}A_{32} - A_{31}A_{12}) \\ B_{33} &= (-1)^6 (A_{11}A_{22} - A_{21}A_{12}) \,, \end{aligned}$$

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 (11.16), namely

$$A^{-1} = \frac{1}{\det A} B \,.$$

Note that $\det A$ was already calculated as equation (11.10).

4 Special Matrices

4.1 Transpose and Symmetry

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

$$A^T_{jk} = A_{kj} \,. \tag{11.19}$$

Symmetric Matrix

A square matrix is said to be *symmetric* if the original matrix A is equal to the transposed matrix A^T:

$$\begin{aligned} A &= A^T \\ A_{jk} &= A_{kj} \end{aligned} \tag{11.20}$$

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:

$$\begin{aligned} A &= -A^T \\ A_{jk} &= -A_{kj} \end{aligned} \tag{11.21}$$

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*:

$$A^T = A^{-1} \tag{11.22}$$

4.2 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.

4.3 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*:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \cdots & A_{2n} \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & A_{nn} \end{bmatrix} \tag{11.23}$$

Lower Triangular Form

If a square matrix has $A_{jk} = 0$ when $j < k$, the form is called *lower triangular*:

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 \\ \cdots & \cdots & \cdots & 0 \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \tag{11.24}$$

4.4 Diagonal Form

If a square matrix has $A_{jk} = 0$ when $j \neq k$, the form is called *diagonal*:

$$A = \begin{bmatrix} A_{11} & 0 & 0 & 0 \\ 0 & A_{22} & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & A_{nn} \end{bmatrix} \tag{11.25}$$

For any triangular or diagonal matrix A, the determinant is equal to the product of its diagonal entries:

$$\det A = A_{11}A_{22}\cdots A_{NN} = \prod_{j=1}^{N} A_{jj}$$

4.5 Augmented Matrix

For linear systems characterized by $A\vec{x} = \vec{b}$, where A is an $M \times N$ matrix, we can construct the *augmented* matrix by appending the components of $\vec{b}$ as an extra column:

$$A|b = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} & b_1 \\ A_{21} & A_{22} & \cdots & A_{2N} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ A_{M1} & A_{M2} & \cdots & A_{MN} & b_M \end{bmatrix} \tag{11.26}$$

In the general case, $\vec{b}$ can be replaced with any matrix with M rows.

4.6 Row-Reduced Echelon Form

If (by any means) the a square matrix and a vector $\vec{x}$ can be written as

$$I|x = \begin{bmatrix} 1 & 0 & \cdots & 0 & x_1 \\ 0 & 1 & \cdots & 0 & x_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 & x_N \end{bmatrix}, \tag{11.27}$$

this is called the *row-reduced echelon form.*

5 Elimination

5.1 Linear Systems

Consider an N-dimensional linear system $A\vec{x} = \vec{b}$, represented by the $M = N$-case of (11.5):

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1N}x_N &= b_1 \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \cdots + A_{2N}x_N &= b_2 \\ A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + \cdots + A_{3N}x_N &= b_3 \\ &\cdots \\ A_{N1}x_1 + A_{N2}x_2 + A_{N3}x_3 + \cdots + A_{NN}x_N &= b_N \end{aligned}$$

Equivalently, the above is represented by an augmented matrix $A|b$ of the form (11.26). Next, imagine having done all of the hard work to solve the system

$$\begin{aligned} x_1 + 0 + 0 + \cdots + 0 &= x_1 \\ 0 + x_2 + 0 + \cdots + 0 &= x_2 \\ 0 + 0 + x_3 + \cdots + 0 &= x_3 \\ &\cdots \\ 0 + 0 + 0 + \cdots + x_N &= x_N \,, \end{aligned}$$

which appears like a tautological thing to write, but is in fact the row-reduced echelon form, $I|x$ cataloged as equation (11.27). 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 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

$$\tilde{O}\,(A|b) = I|x\,. \tag{11.28}$$

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

$$A|b = \begin{bmatrix} 1 & 1 & 1 & 5 \\ 2 & 3 & 5 & 8 \\ 4 & 0 & 5 & 2 \end{bmatrix} .$$

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.

$$A|b \xrightarrow{\text{(i)}} \begin{bmatrix} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 4 & 0 & 5 & 2 \end{bmatrix} \xrightarrow{\text{(ii)}} \begin{bmatrix} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 0 & -4 & 1 & -18 \end{bmatrix} \xrightarrow{\text{(iii)}} \begin{bmatrix} 1 & 1 & 1 & 5 \\ 0 & 1 & 3 & -2 \\ 0 & 0 & 13 & -26 \end{bmatrix}$$

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.

$$\xrightarrow{\text{(iv)}} \begin{bmatrix} 1 & 1 & 1 & 5 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{bmatrix} \xrightarrow{\text{(v)}} \begin{bmatrix} 1 & 1 & 0 & 7 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{bmatrix} \xrightarrow{\text{(vi)}} \begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 4 \\ 0 & 0 & 1 & -2 \end{bmatrix} = I|x$$

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 (11.28), we see the operator $\tilde{O}$ is comprised of steps (i)-(vi), each being one particular E, M, R operation.

5.2 Matrix Inverse

Looking again at Equation (11.28), note that the sequence of row operations $\tilde{O}$ applies to A and $\vec{b}$ separately:

$$\begin{aligned} \tilde{O}A &= I \\ \tilde{O}\vec{b} &= \vec{x} \end{aligned}$$

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

$$\begin{aligned} A^{-1}A &= I \\ A^{-1}\vec{b} &= \vec{x} \, . \end{aligned}$$

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:

$$\begin{aligned} \tilde{O}A &= I \\ \left(\tilde{O}A\right) A^{-1} &= IA^{-1} \\ \tilde{O}\left(AA^{-1}\right) &= A^{-1} \\ \tilde{O}I &= A^{-1} \end{aligned}$$

Once again, we see the sequence $\tilde{O}$ is doing the same job as A^{-1}. Rounding up the circumstancial evidence, we see the set of steps $\tilde{O}$ that carries $A \to I$ is the *same* set of steps that carries $I \to A^{-1}$. In the language of augmented matrices, this is summarized by

$$\tilde{O}\left(A|I\right) = I|A^{-1} \, . \tag{11.29}$$

This conspiracy of mathematics is otherwise known as *Gauss-Jordan elimination.*

Two Dimensions

Demonstrating on a 2×2 matrix, begin with $A|I$ as

$$A|I = \begin{bmatrix} A_{11} & A_{12} & 1 & 0 \\ A_{21} & A_{22} & 0 & 1 \end{bmatrix} ,$$

and perform row operations until form $I|A^{-1}$ is attained. In brief detail, the augmented matrix develops as:

$$\begin{aligned} A|I &\to \begin{bmatrix} A_{12}A_{21} - A_{22}A_{11} & 0 & -A_{22} & A_{21} \\ A_{21} & & A_{22} & 0 & 1 \end{bmatrix} \\ A|I &\to \frac{1}{\det A} \begin{bmatrix} 1 & 0 & -A_{22} & A_{21} \\ 0 & 1 & -A_{21} & A_{11} \end{bmatrix} \end{aligned}$$

The final result is none other than (11.18), the formula for the inverse of a 2×2 square matrix:

$$A^{-1} = \frac{1}{A_{11}A_{22} - A_{12}A_{21}} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{bmatrix}$$

6 Eigenvectors and Eigenvalues

An important situation that arises in mathematics and physics is the so-called *eigenvalue problem*

$$A\vec{u} = \lambda\vec{u}\,. \tag{11.30}$$

The matrix A is taken to be square and N-dimensional. The vectors $\vec{u}^{(j)}$ that satisfy (11.30) are called *eigenvectors*, and the corresponding scalar $\lambda^{(j)}$ is called an *eigenvalue*.

6.1 Calculating Eigenvalues

The eivenvalue problem (11.30) can be equivalently framed as

$$(A - \lambda I)\,\vec{u} = 0\,, \tag{11.31}$$

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

$$(A_{11} - \lambda)\,(A_{22} - \lambda) - A_{12}A_{21} = 0\,. \tag{11.32}$$

The only unknown in the equation is λ, which can be isolated using the quadratic formula:

$$\lambda_\pm = \frac{A_{11} + A_{22}}{2} \pm \frac{1}{2}\sqrt{(A_{11} - A_{22})^2 + 4A_{12}A_{21}} \tag{11.33}$$

Note there are two solutions for λ, which we label λ_+, and λ_-, respectively.

6.2 Characteristic Equation

When confronted with the eigenvalue problem (11.31), 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 (11.32). Pausing on this result for a moment, note that a quicker way to get there is to write

$$\det\,(A - \lambda I) = 0\,, \tag{11.34}$$

which is in fact true in any number of dimensions. Equation (11.34) is called the *characteristic equation* of the system.

Characteristic Polynomial

The characteristic equation always 'simplifies' to the *characteristic polynomial*, a single equation embedding λ:

$$P_N\,(\lambda) = C_0 + C_1\lambda + C_2\lambda^2 + \cdots + C_N\lambda^N = 0 \tag{11.35}$$

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.

6.3 Calculating Eigenvectors

Once the eigenvalues λ are known, the components of each eigenvector $\vec{u}$ are readily calculated directly from

$$\begin{aligned} A\vec{u}^{(j)} &= \lambda_j\vec{u}^{(j)} \\ j &= 1, 2, 3, \ldots, N\,. \end{aligned}$$

Two Dimensions

Developing the eigenvalue problem in two dimensions, there are two eigenvalues $\lambda_\pm$ given by (11.33), and let us label the two corresponding eigenvectors $\vec{u}$, $\vec{v}$ such that

$$\begin{aligned} A\vec{u} &= \lambda_+\vec{u} \\ A\vec{v} &= \lambda_-\vec{v}\,. \end{aligned}$$

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:

$$\begin{aligned} \frac{u_1}{u_2} &= \frac{-A_{12}}{A_{11} - \lambda_+} \\ \frac{u_1}{u_2} &= \frac{-\,(A_{22} - \lambda_+)}{A_{21}} \end{aligned} \tag{11.36}$$

As a sanity check, we may eliminate the ratio u_1/u_2 and recover the characteristic equation (11.32). A similar set of steps isolates the ratio v_1/v_2 for the second eigenvalue/eigenvector

$$\frac{v_1}{v_2} = \frac{-A_{12}}{A_{11} - \lambda_-} \tag{11.37}$$

$$\frac{v_1}{v_2} = \frac{-\left(A_{22} - \lambda_-\right)}{A_{21}} \,, \tag{11.38}$$

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 (11.33), and simplify very nicely:

$$\lambda_\pm = a \pm b$$

Denoting the respective eigenvectors $\vec{u}$, $\vec{v}$, we apply (11.36) directly to find

$$\frac{u_1}{u_2} = \frac{-b}{-b} = 1 \,.$$

Meanwhile, (11.37) similarly tells us

$$\frac{v_1}{v_2} = \frac{-b}{b} = -1 \,,$$

and we're done. Evidently, the two eigenvectors are

$$\vec{u} = \langle 1, 1 \rangle$$
$$\vec{v} = \langle 1, -1 \rangle \,,$$

or in normalized form,

$$\hat{u} = \frac{1}{\sqrt{2}} \langle 1, 1 \rangle$$
$$\hat{v} = \frac{1}{\sqrt{2}} \langle 1, -1 \rangle \,.$$

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:

$$\hat{u} = \frac{1}{\sqrt{2}} \langle 1, i \rangle$$
$$\hat{v} = \frac{1}{\sqrt{2}} \langle 1, -i \rangle$$

7 Diagonalization

For the eigenvalue problem (11.30)

$$A\vec{u} = \lambda\vec{u}$$

of dimension N, suppose we already have the list of N eigenvalues λ and N eigenvectors $\vec{u}$.

7.1 Modal Matrix

It's instructive to condense all of the eigenvector information into a new object called the *modal matrix*, denoted C, whose jth column is comprised of the components of the jth eigenvector:

$$\begin{aligned} C &= \begin{bmatrix} u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(N)} \\ u_2^{(1)} & u_2^{(2)} & \cdots & u_2^{(N)} \\ \cdots & \cdots & \cdots & \cdots \\ u_N^{(1)} & u_N^{(2)} & \cdots & u_N^{(N)} \end{bmatrix} \\ &= \begin{bmatrix} \vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(N)} \end{bmatrix} \end{aligned} \tag{11.39}$$

Then, the matrix product AC can be written

$$\begin{aligned} AC &= \begin{bmatrix} \lambda_1\vec{u}^{(1)} & \lambda_2\vec{u}^{(2)} & \cdots & \lambda_N\vec{u}^{(N)} \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 u_1^{(1)} & \lambda_2 u_1^{(2)} & \cdots & \lambda_N u_1^{(N)} \\ \lambda_1 u_2^{(1)} & \lambda_2 u_2^{(2)} & \cdots & \lambda_N u_2^{(N)} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_1 u_N^{(1)} & \lambda_2 u_N^{(2)} & \cdots & \lambda_N u_N^{(N)} \end{bmatrix} \,. \end{aligned}$$

7.2 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:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \tag{11.40}$$

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:

$$\Lambda = C^{-1}AC \tag{11.41}$$

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.

7.3 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.

$$\vec{V} = \sum_{j=1}^{N} V_j \, \hat{e}_j \overset{?}{\rightarrow} (\vec{V})' = \sum_{j=1}^{N} V'_j \, \hat{u}^{(j)}$$

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 $\left\{\hat{u}^{(j)}\right\}$ form an orthogonal basis. That is, it's not always the case that any two eigenvectors are orthogonal.

Hermitian Matrix

Consider two solutions to the eigenvalue problem (11.30),

$$\begin{aligned} A\vec{u}^{(j)} &= \lambda_j \vec{u}^{(j)} \\ A\vec{u}^{(k)} &= \lambda_k \vec{u}^{(k)} \, , \end{aligned}$$

and multiply $\vec{u}^{(k)}$, $\vec{u}^{(j)}$, onto the left and right sides respectively into each:

$$\begin{aligned} \vec{u}^{(k)} \cdot A\vec{u}^{(j)} &= \lambda_j \vec{u}^{(k)} \cdot \vec{u}^{(j)} \\ A\vec{u}^{(k)} \cdot \vec{u}^{(j)} &= \lambda_k \vec{u}^{(k)} \cdot \vec{u}^{(j)} \end{aligned}$$

Looking at the left side of each equation, it appears as if

$$\vec{u}^{(k)} \cdot A\vec{u}^{(j)} = A\vec{u}^{(k)} \cdot \vec{u}^{(j)} \tag{11.42}$$

wants to be true, but simply isn't in the general case. The special that satisfies (11.42) is called a *Hermitian* matrix.

Non-equal Eigenvalues

Pursuing the case where A is Hermitian, the above condenses to:

$$\lambda_j \vec{u}^{(k)} \cdot \vec{u}^{(j)} = \lambda_k \vec{u}^{(k)} \cdot \vec{u}^{(j)}$$

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*:

$$\vec{u}^{(k)} \cdot \vec{u}^{(j)} = 0$$

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.

8 Degenerate Systems

Concerning the eigenvalue problem (11.30)

$$A\vec{u} = \lambda\vec{u} \, ,$$

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 $\vec{u}^{(j)}$, 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.

8.1 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 $A\vec{u} = 1\vec{u}$, resulting in

$$\vec{u}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} .$$

Proceeding to the repeated eigenvalue case, we solve $A\vec{u} = -2\vec{u}$ to get a single eigenvector

$$\vec{u}^{(2,3)} = \frac{1}{\sqrt{10}} \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix} .$$

Note that $\vec{u}^{(1)}$ is linearly independent from $\vec{u}^{(2,3)}$, but not orthogonal. Since there is no obvious way to 'peel apart' the eigenvectors $\vec{u}^{(2,3)}$, the show stops here. The matrix A cannot be diagonalized.

8.2 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 $A\vec{u} = -3\vec{u}$, 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

$$\vec{u}^{(3)} = \frac{1}{\sqrt{11}} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} ,$$

or any multiple.

Proceeding to the repeated eigenvalue case, we solve $A\vec{u} = \vec{u}$ 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:

$$\vec{u}^{(1)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

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:

$$\vec{u}^{(2)} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

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 $\Lambda = C^{-1}AC$.

8.3 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

$$0 = (2-\lambda)\left(-\lambda^2 + \lambda + 2\right) ,$$

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 $A\vec{x} = -\vec{x}$, leading to

$$\vec{u}^{(3)} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 0 \\ -\sqrt{2} \end{bmatrix} .$$

Proceeding to the repeated eigenvalue case, we solve $A\vec{u} = 2\vec{u}$ to get a single eigenvector

$$\vec{u}^{(1,2)} = \frac{1}{\sqrt{3\alpha^2/2 + \beta^2}} \begin{bmatrix} \alpha \\ \beta \\ \alpha/\sqrt{2} \end{bmatrix} ,$$

for two arbitrary constants α, β. The aim here is to tease two mutually orthogonal eigenvectors from the above, which means to require

$$\vec{u}^{(1)} \cdot \vec{u}^{(2)} = 0 .$$

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

$$\vec{u}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$
$$\vec{u}^{(2)} = \frac{1}{\sqrt{3/2}} \begin{bmatrix} 1 \\ 0 \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \sqrt{2/3} \\ 0 \\ 1/\sqrt{3} \end{bmatrix} .$$

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 $\left\{\vec{u}^{(j)}\right\}$ form an orthonormal basis, we may further simplify the above using $C^{-1} = C^T$.

Part IV

Calculus

Chapter 12

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:

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \qquad (12.1)$$

Letting $h = x - x_0$, the definition can be written

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} ,$$

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 of the baggage pertaining to continuity, smoothness, etc. must becomes 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

$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} ,$$

simplifying down to

$$f'(x) = \lim_{h \to 0} 2x - h = 2x .$$

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:

$$\begin{aligned} f'(x_0) &= \lim_{x \to x_0} \frac{x^2 - x_0^2}{x - x_0} \\ &= \lim_{x \to x_0} \frac{(x + x_0)\cancel{(x - x_0)}}{\cancel{x - x_0}} = 2x_0 \end{aligned}$$

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

$$f'(x_0) = \lim_{x \to x_0} \frac{x^n - x_0^n}{x - x_0} .$$

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 $x \to x_0$ 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:

$$\frac{d}{dx}\left(x^n\right) = nx^{n-1} \tag{12.2}$$

Reciprocal

For the reciprocal function $f(x) = 1/x$ we have

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{1/x - 1/x_0}{x - x_0} \\ &= \lim_{x\to x_0} \frac{-\cancel{(x-x_0)}}{xx_0\cancel{(x-x_0)}} = \frac{-1}{x_0^2} . \end{aligned}$$

That is, the slope of the reciprocal function at a point x_0 is $-1/x_0^2$. In summary:

$$\frac{d}{dx}\left(\frac{1}{x}\right) = \frac{-1}{x^2} \tag{12.3}$$

As an exercise, perhaps just a mental one, it's straightforward to show that a horizontally-shifted reciprocal function obeys:

$$\frac{d}{dx}\left(\frac{1}{x+a}\right) = \frac{-1}{(x+a)^2} \tag{12.4}$$

Inverse Square

The inverse square function $f(x)$ can be dealt with using Equation (12.2), but we'll suffer the brute force approach:

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{1/x^2 - 1/x_0^2}{x - x_0} \\ &= \lim_{x\to x_0} \frac{-(x+x_0)\cancel{(x-x_0)}}{x^2x_0^2\cancel{(x-x_0)}} = \frac{-2}{x_0^3} \end{aligned}$$

Like the previous few cases, the denominator is eliminated by algebra and the derivative of the function becomes clear:

$$\frac{d}{dx}\left(\frac{1}{x^2}\right) = \frac{-2}{x^3} \tag{12.5}$$

To go with this is the shifted version

$$\frac{d}{dx}\left(\frac{1}{(x+a)^2}\right) = \frac{-2}{(x+a)^3} , \tag{12.6}$$

where a is a constant.

Square Root

The derivative of the square root $f(x) = \sqrt{x}$ is also straightforward, as:

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{\sqrt{x} - \sqrt{x_0}}{x - x_0} \\ &= \lim_{x\to x_0} \frac{\cancel{\left(\sqrt{x} - \sqrt{x_0}\right)}}{\cancel{\left(\sqrt{x} - \sqrt{x_0}\right)}\left(\sqrt{x} + \sqrt{x_0}\right)} = \frac{1}{2\sqrt{x_0}} \end{aligned}$$

In one line, this is:

$$\frac{d}{dx}\left(\sqrt{x}\right) = \frac{1}{2\sqrt{x}} \tag{12.7}$$

1.3 Exponential Derivatives

Standard Exponential

The exponential function $f(x) = n^x$ is a bit more tricky. Following traditional setup, we have

$$f'(x_0) = \lim_{x\to x_0} \frac{n^x - n^{x_0}}{x - x_0} = n^{x_0} \lim_{x\to x_0} \frac{n^{x-x_0} - 1}{x - x_0} ,$$

where the term n^{x_0} can be factored outside of the limit.

Substituting $h = x - x_0$, the remaining limit becomes

$$f'(x_0) = n^x \lim_{h\to 0} \frac{n^h - 1}{h} ,$$

is decidedly equivalent to the natural log of n. In summary:

$$\frac{d}{dx}\left(n^x\right) = n^x \ln(n) \tag{12.8}$$

Natural Exponential

A special case of Equation (12.8) has $n = e$, as in Euler's e, which gets rid of the ln-term because $\ln(e) = 1$. This tells us

$$\frac{d}{dx}\left(e^x\right) = e^x\,, \tag{12.9}$$

meaning $f(x) = e^x$ is its own derivative.

By making repeated use of Equation (12.2) for handling powers, one can show easily that another way to express e^x is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots\,. \tag{12.10}$$

It's worth mentioning too that the definition for e as an infinite limit, i.e. Equation

$$e = \lim_{h\to 0}\left(1 + \frac{1}{h}\right)^h$$

can be derived from the definition of the derivative. Supposing we nothing 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

$$f'(x_0) = \lim_{x\to x_0}\frac{E^x - E^{x_0}}{x - x_0}\,,$$

and E^{x_0} can be factored out:

$$f'(x_0) = E^{x_0}\left(\lim_{x\to x_0}\frac{E^{x-x_0} - 1}{x - x_0}\right)$$

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

$$\lim_{x\to x_0}\frac{E^{x-x_0} - 1}{x - x_0} = 1\,,$$

and then solve for E:

$$E = \lim_{x\to x_0}\left(1 + (x - x_0)\right)^{1/(x-x_0)}$$

By substituting $h = 1/(x - x_0)$, the right side becomes identical to e. 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

$$f'(x_0) = \lim_{x\to x_0}\frac{e^{x^2} - e^{x_0^2}}{x - x_0}$$

Using Equation (12.10), the numerator expands out to

$$x^2 - x_0^2 + \frac{\left(x^2\right)^2 - \left(x_0^2\right)^2}{2!} + \frac{\left(x^2\right)^3 - \left(x_0^2\right)^3}{3!} + \cdots\,,$$

which is an algebraic mess, because we need to factor $x - x_0$ out of the whole expression. Going in chunks, we find:

$$\begin{aligned}
x^2 - x_0^2 &= (x - x_0)(x + x_0)\\
\left(x^2\right)^2 - \left(x_0^2\right)^2 &= (x - x_0)(x + x_0)\left(x^2 + x_0^2\right)\\
\left(x^2\right)^3 - \left(x_0^2\right)^3 &= (x - x_0)\\
&\qquad \left(x^2 + xx_0 + x_0^2\right)\left(x^3 + x_0^3\right)\\
\left(x^2\right)^4 - \left(x_0^2\right)^4 &= (x - x_0)\\
&\qquad (x + x_0)\left(x^2 + x_0^2\right)\left(x^4 + x_0^4\right)\\
\left(x^2\right)^5 - \left(x_0^2\right)^5 &= (x - x_0)\\
&\qquad \left(x^4 + x^3x_0 + x^2x_0^2 + xx_0^3 + x_0^4\right)\\
&\qquad \left(x^5 + x_0^5\right)
\end{aligned}$$

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 (12.10). Finally, the answer:

$$\frac{d}{dx}\left(e^{x^2}\right) = 2x\,e^{x^2} \tag{12.11}$$

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

$$f'(x) = \lim_{x\to x_0}\frac{b^{x^2} - b^{x_0^2}}{x - x_0} = \lim_{h\to 0}\frac{b^{(x_0+h)^2} - b^{x_0^2}}{h}$$

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

$$f'(x) = b^{x_0^2} \lim_{h\to 0} \frac{b^{2x_0 h} - 1}{h} .$$

The remaining limit almost looks familiar as a natural logarithm. Make the substitution $h = 2x_0 k$ to bring it into form:

$$f'(x) = 2x_0\, b^{x_0^2} \lim_{k\to 0} \frac{b^k - 1}{k} .$$

The remaining limit now is unambiguously equivalent to the natural log of b. Finally, we find:

$$\frac{d}{dx}\left(b^{x^2}\right) = 2x\, b^{x^2} \ln(b) \tag{12.12}$$

The special case $b = e$ recovers Equation (12.11).

X to the X

A notorious derivative to figure out is that of $f(x) = x^x$. Setting up the calculation, we have

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{x^x - x_0^{x_0}}{x - x_0} \\ &= \lim_{x\to x_0} \frac{x^{x_0}}{x - x_0}\left(x^{x-x_0} - \left(\frac{x_0}{x}\right)^{x_0}\right) , \end{aligned}$$

and let $h = x - x_0$:

$$f(x) = \lim_{h\to 0} x^{x_0}\left(\frac{x^h - (1 - h/x)^{x_0}}{h}\right)$$

Recalling the limit-based expression for the natural logarithm, namely

$$\ln(x) = \lim_{h\to 0} \frac{x^h - 1}{h} ,$$

and the above becomes:

$$f(x) = \lim_{h\to 0} x^{x_0}\left(\ln(x) + \frac{1 - (1 - h/x)^{x_0}}{h}\right)$$

The right-hand limit is best handled in isolation. Letting

$$A = \lim_{h\to 0} \frac{1 - (1 - h/x)^{x_0}}{h} ,$$

rearrange to write

$$\lim_{h\to 0}(1 - Ah) = \lim_{h\to 0}\left(1 - \frac{h}{x}\right)^{x_0} ,$$

and raise each side to the $1/h$ power:

$$\lim_{h\to 0}(1 - Ah)^{1/h} = \lim_{h\to 0}\left(1 - \frac{h}{x}\right)^{x_0/h}$$

With one more substitution $q = 1/h$, this is

$$\lim_{q\to\infty}\left(1 - \frac{A}{q}\right)^q = \lim_{q\to\infty}\left(1 - \frac{1}{qx}\right)^{qx_0}$$

Keep in mind that $q \to \infty$ also means $x \to x_0$, 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:

$$\frac{d}{dx}(x^x) = x^x(\ln(x) + 1) \tag{12.13}$$

1.4 Logarithmic Derivatives

Natural Logarithm

A keystone function is the natural logarithm, $f(x) = \ln(x)$. Setting up the derivative calculation, we have

$$\begin{aligned} f'(x_0) &= \lim_{x\to x_0} \frac{\ln(x) - \ln(x_0)}{x - x_0} \\ &= \frac{1}{x_0} \lim_{x\to x_0} \frac{\ln(x/x_0)}{x/x_0 - 1} , \end{aligned}$$

suggesting a substitution $x/x_0 = k$, and the limit becomes a matter of k approaching 1:

$$f'(x_0) = \frac{1}{x_0} \lim_{k\to 1} \frac{\ln(k)}{k - 1}$$

With another substitution $j = k - 1$, this is

$$f'(x_0) = \frac{1}{x_0} \lim_{j\to 0} \frac{\ln(1 + j)}{j} ,$$

equivalent to

$$f'(x_0) = \frac{1}{x_0} \lim_{j\to 0} \ln\left((1 + j)^{1/j}\right) .$$

If it doesn't look familiar yet, let $h = 1/j$ to get

$$f'(x_0) = \frac{1}{x_0} \lim_{h\to\infty} \ln\left(\left(1 + \frac{1}{h}\right)^h\right) .$$

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

$$\frac{d}{dx}(\ln(x)) = \frac{1}{x} \tag{12.14}$$

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

$$f'(x_0) = \lim_{x\to x_0} \frac{\ln(1+x) - \ln(1+x_0)}{x - x_0},$$

suggesting a substitution $z = 1 + x$. The above becomes

$$f'(x_0) = \lim_{z\to z_0} \frac{\ln(z) - \ln(z_0)}{z - z_0},$$

which now looks identical to the the vanilla case in the variable z. Reversing the z-substitution gives the final answer:

$$\frac{d}{dx}(\ln(1+x)) = \frac{1}{1+x} \tag{12.15}$$

Nonlinear Natural Logarithm

For the function $f(x) = x\ln(x)$, the derivative calculation begins as

$$f'(x_0) = \lim_{x\to x_0} \frac{x\ln(x) - x_0\ln(x_0)}{x - x_0}$$

This one is best attacked with polynomial division, which leads to

$$f'(x_0) = \lim_{x\to x_0} \left(\ln(x) + x_0\left(\frac{\ln(x) - \ln(x_0)}{x - x_0}\right)\right),$$

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

$$\frac{d}{dx}(x\ln(x)) = \ln(x) + 1 \tag{12.16}$$

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

$$\frac{d}{dx}\left(x^2\ln(x)\right) = 2x\ln(x) + x\,. \tag{12.17}$$

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

$$f'(x_0) = \lim_{x\to x_0} \frac{\ln(x)/x - \ln(x_0)/x_0}{x - x_0}\,.$$

Using polynomial division and simplifying, we get to the intermediate step:

$$f'(x_0) = \frac{\ln(x_0)}{x_0^2} + \lim_{x\to x_0} \frac{1}{x_0^3}\left(\frac{x_0^2\ln(x) - x^2\ln(x_0)}{x - x_0}\right)$$

The remaining limit almost looks like Equation (12.16), 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

$$\lim_{x\to x_0} \frac{\ln(x) - \ln(x_0)}{x - x_0} = \frac{1}{x_0},$$

and then deduce the following:

$$\lim_{x\to x_0} \frac{x_0^2\ln(x)}{x - x_0} = \lim_{x\to x_0} \frac{x_0^2\ln(x_0)}{x - x_0} + x_0$$
$$\lim_{x\to x_0} \frac{x^2\ln(x_0)}{x - x_0} = \lim_{x\to x_0} \frac{x^2\ln(x)}{x - x_0} - \frac{x^2}{x_0}$$

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:

$$f'(x_0) = \frac{\ln(x_0)}{x_0^2} + \lim_{x\to x_0} \frac{1}{x_0^3}\left(x_0 + \frac{x^2}{x_0}\right) + \lim_{x\to x_0} \frac{1}{x_0^3}\left(\frac{x_0^2\ln(x_0) - x^2\ln(x)}{x - x_0}\right)$$

The latter term in the above contains the (negative) derivative of $x^2\ln(x)$ and can be replaced by Equation (12.16). This gets rid of all singularities, and we can simplify to get the answer:

$$\frac{d}{dx}\left(\frac{\ln(x)}{x}\right) = \frac{1 - \ln(x)}{x^2} \tag{12.18}$$

Modified Natural Logarithm

For the function $f(x) = \ln\left(1 + x^2\right)$, the derivative calculation begins as

$$f'(x_0) = \lim_{x\to x_0} \frac{\ln\left(1+x^2\right) - \ln\left(1+x_0^2\right)}{x - x_0}$$

Letting $h = x - x_0$ and simplifying using the rules for manipulating logarithms, the above reduces way down to

$$f'(x_0) = \lim_{h\to 0} \ln\left(\left(1 + \frac{2x_0h + h^2}{1 + x_0^2}\right)^h\right).$$

In the limit $h \to 0$, the h^2-term is negligible, and the rest, after staring for long enough, contains the definition of the natural exponential:

$$f'(x_0) = \ln\left(e^{(2x_0h)/(1+x_02)}\right)$$

Since the natural log and the natural exponential are mutually-annihilating, we get the result:

$$\frac{d}{dx}\left(\ln\left(1+x^2\right)\right) = \frac{2x}{1+x^2} \tag{12.19}$$

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

$$f'(x) = \lim_{x\to x_0} \frac{\sin(x) - \sin(x_0)}{x - x_0} .$$

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

$$f'(x) = \cos(x_0) \lim_{x\to x_0} \frac{2}{x - x_0}\left(\sin\left(\frac{x-x_0}{2}\right)\right)$$

Let $h = (x - x_0)/2$ to uncover a sinc function lurking about:

$$f'(x) = \cos(x_0) \lim_{h\to 0}\left(\frac{\sin(h)}{h}\right)$$

Recall that we spent some effort deciding that the parenthesized quantity is identically one, and we're left with the answer:

$$\frac{d}{dx}(\sin(x)) = \cos(x) \tag{12.20}$$

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

$$f'(x) = -\sin(x_0) \lim_{h\to 0}\left(\frac{\sin(h)}{h}\right) ,$$

and the same sinc is also present. From this we conclude

$$\frac{d}{dx}(\cos(x)) = -\sin(x) . \tag{12.21}$$

Tangent

For the tangent $f(x) = \tan(x)$, we start with

$$f'(x) = \lim_{x\to x_0} \frac{\tan(x) - \tan(x_0)}{x - x_0} ,$$

and then the trick is to divide out $\cos^2(x)$ from the limit:

$$f'(x) = \frac{1}{(\cos(x_0))^2} \lim_{x\to x_0} \frac{\sin(x)\cos(x_0) - \sin(x_0)\cos(x)}{x - x_0}$$

The remaining limit contains a trig identity for the sum of two angles, namely $x - x_0$. This again resolves to $\operatorname{sinc}(0)$, as was the case with the previous two trig functions. The result is

$$\frac{d}{dx}(\tan(x)) = \frac{1}{(\cos(x))^2} . \tag{12.22}$$

Cotangent

For the cotangent $f(x) = \cot(x)$, we start with

$$f'(x) = \lim_{x\to x_0} \frac{\cot(x) - \cot(x_0)}{x - x_0} ,$$

and then the trick is to divide out $\sin^2(x)$ from the limit:

$$f'(x) = \frac{1}{(\sin(x_0))^2} \lim_{x\to x_0} \frac{\cos(x)\sin(x_0) - \cos(x_0)\sin(x)}{x - x_0}$$

The remaining limit contains a trig identity for the sum of two angles, namely $x_0 - x$. This resolves to $-\operatorname{sinc}(0)$, and the final result is

$$\frac{d}{dx}(\cot(x)) = \frac{-1}{(\sin(x))^2} . \tag{12.23}$$

Secant

For the secant function $f(x) = 1/\cos(x)$, we start with

$$f'(x) = \lim_{x\to x_0} \frac{1/\cos(x) - 1/\cos(x_0)}{x - x_0} ,$$

and then divide out $-\cos^2(x)$ from the limit:

$$f'(x) = \frac{-1}{(\cos(x_0))^2} \lim_{x\to x_0} \frac{\cos(x) - \cos(x_0)}{x - x_0}$$

The limit now looks like the derivative of the cosine function and can be replaced by Equation (12.21), namely $-\sin(x_0)$. Reporting the result in standard form, we find

$$\frac{d}{dx}(\sec(x)) = \tan(x)\sec(x) . \tag{12.24}$$

Cosecant

For the cosecant function $f(x) = 1/\sin(x)$, we start with

$$f'(x) = \lim_{x \to x_0} \frac{1/\sin(x) - 1/\sin(x_0)}{x - x_0} ,$$

and then divide out $-\sin^2(x)$ from the limit:

$$f'(x) = \frac{-1}{(\sin(x_0))^2} \lim_{x \to x_0} \frac{\sin(x) - \sin(x_0)}{x - x_0}$$

The limit now looks like the derivative of the sine function and can be replaced by Equation (12.20), namely $\cos(x_0)$. Reporting the result in standard form, we find

$$\frac{d}{dx}(\csc(x)) = -\cot(x)\csc(x) . \tag{12.25}$$

Squared Argument

For the sine function with a squared argument $f(x) = \sin(x^2)$, we first write

$$f'(x) = \lim_{x \to x_0} \frac{\sin(x^2) - \sin(x_0^2)}{x - x_0} .$$

Using the same trig identity that helped with the regular sine case, the above becomes

$$f'(x) = \lim_{x \to x_0} \left(\frac{2}{x - x_0} \right) \sin\left(\frac{(x - x_0)(x + x_0)}{2} \right) \cos\left(\frac{x^2 + x_0^2}{2} \right) .$$

Lett $h = (x - x_0)/2$ and simplify a little to get

$$f'(x) = \cos(x_0^2) \lim_{h \to 0} \frac{\sin(2x_0 h)}{h} .$$

There seems to be an extra term in the remaining sinc 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

$$f'(x) = \cos(x_0^2)\, 2x_0 \lim_{k \to 0} \left(\frac{\sin(k)}{k} \right) .$$

The final limit is identically one, and in conclusion,

$$\frac{d}{dx}(\sin(x^2)) = \cos(x^2)\, 2x . \tag{12.26}$$

The intermediate steps would be nearly the same had we started with cosine instead of sine, which would result in:

$$\frac{d}{dx}(\cos(x^2)) = -\sin(x^2)\, 2x . \tag{12.27}$$

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

$$f'(x) = \lim_{x \to x_0} \frac{x\sin(x) - x_0\sin(x_0)}{x - x_0}$$

To crack this one, add and subtract the quantity $x_0\sin(x)$ from the numerator, and then repack everything to get

$$f'(x) = \lim_{x \to x_0} \frac{\sin(x_0)(x - x_0)}{x - x_0} + \lim_{x \to x_0} \frac{x_0(\sin(x) - \sin(x_0))}{x - x_0} .$$

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

$$\frac{d}{dx}(x\sin(x)) = \sin(x) + x\cos(x) . \tag{12.28}$$

The same recipe works for the $x\cos(x)$ case. Leaving the details as an exercise, the result is

$$\frac{d}{dx}(x\cos(x)) = \cos(x) - x\sin(x) . \tag{12.29}$$

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*:

$$\lim_{x \to 0} \sin(x) = x - \cancel{\frac{x^3}{3!} + \cdots} \tag{12.30}$$

$$\lim_{x \to 0} \cos(x) = 1 - \cancel{\frac{x^2}{2!} + \cdots} \tag{12.31}$$

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

$$\lim_{\phi \to 0} \sin(\theta + \phi) = \lim_{\phi \to 0} (\sin(\theta) + \phi \cos(\theta))$$
$$\lim_{\phi \to 0} \cos(\theta + \phi) = \lim_{\phi \to 0} (\cos(\theta) - \phi \sin(\theta)) \ .$$

Solve the first equation for $\cos(\theta)$ and the second equation for $-\sin(\theta)$:

$$\cos(\theta) = \lim_{\phi \to 0} \frac{\sin(\theta + \phi) - \sin(\theta)}{\phi}$$
$$-\sin(\theta) = \lim_{\phi \to 0} \frac{\cos(\theta + \phi) - \cos(\theta)}{\phi}$$

This pair of results is none other than the derivative formulas for sine and cosine. By a change of variables these exactly reproduce Equations (12.20), (12.21).

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

$$p'(x) = \lim_{x \to x_0} \frac{f(x)\, g(x) - f(x_0)\, g(x_0)}{x - x_0} \ ,$$

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

$$p'(x_0) = \lim_{x \to x_0} \left(\frac{f(x)\, g(x) - f(x_0)\, g(x)}{x - x_0} \right) - \lim_{x \to x_0} \left(\frac{f(x_0)\, g(x_0) - f(x_0)\, g(x)}{x - x_0} \right) \ ,$$

and simplify:

$$p'(x_0) = g(x_0) \lim_{x \to x_0} \left(\frac{f(x) - f(x_0)}{x - x_0} \right) + f(x_0) \lim_{x \to x_0} \left(\frac{g(x) - g(x_0)}{x - x_0} \right)$$

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 $g(x)$:

$$\frac{d}{dx}(fg) = \frac{df}{dx} g + f \frac{dg}{dx} \tag{12.32}$$

Or, a more economical way to write the same thing:

$$(fg)' = f'g + fg'$$

All of these notations are mixed ant 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.

Example 1

Let $p(x) = x^2 \ln(x)$. Identifying $f(x) = x^2$ and $g(x) = \ln(x)$, we have

$$\begin{aligned} p'(x) &= f'g + fg' \\ &= \ln(x) \frac{d}{dx} \ln\left(x^2\right) + x^2 \frac{d}{dx} (\ln(x)) \\ &= 2x \ln(x) + x \ , \end{aligned}$$

in agreement with Equation (12.16).

Example 2

Let $p(x) = \ln(x)/x$. 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 (12.18).

Example 3

Let $p(x) = x \sin(x)$. 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 (12.28).

Example 4

Let $p(x) = x\cos(x)$. 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 (12.29).

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

$$r'(x) = \lim_{x\to x_0} \frac{f(x)/g(x) - f(x_0)/g(x_0)}{x - x_0} ,$$

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 $g(x)$:

$$\frac{d}{dx}\left(\frac{f}{g}\right) = \frac{1}{g^2}\left(\frac{df}{dx}g - f\frac{dg}{dx}\right) \tag{12.33}$$

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 the right kind of problem.

Example 5

Let $r(x) = \ln(x)/x$. 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 (12.18).

Example 6

Let $r(x) = \tan(x)$. 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 (12.22).

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 $\cos\left(x^2\right)$ 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:

$$f'(x_0) = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

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:

$$x'(t_0) = \lim_{t \to t_0} \frac{x(t) - x(t_0)}{t - t_0}$$

The derivative of f, though, now looks like this:

$$f'(x(t_0)) = \lim_{x(t) \to x(t_0)} \frac{f(x(t)) - f(x(t_0))}{x(t) - x(t_0)}$$

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 $t \to t_0$. Next comes the key move, which is to multiply both sides by $x'(t_0)$ to get:

$$f'(x(t_0)) \cdot x'(t_0) = \lim_{t \to t_0} \frac{f(x(t)) - f(x(t_0))}{\cancel{x(t) - x(t_0)}} \cdot \frac{\cancel{x(t) - x(t_0)}}{t - t_0}$$

The quantity $x - x_0$ conveniently cancels, and we can tidy things up to write:

$$f'(x_0) \cdot x'(t_0) = \lim_{t \to t_0} \frac{f(x(t)) - f(x(t_0))}{t - t_0}$$

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

$$\frac{df}{dt} = \frac{df}{dx} \cdot \frac{dx}{dt}\,, \tag{12.34}$$

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

Example 7

Consider the function $f(x) = e^{x^2}$. 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 (12.11).

Example 8

Consider the function $f(x) = b^{x^2}$. 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 (12.12).

Example 9

Consider the function $f(x) = \ln(x^x)$. 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 (12.13) 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

$$\frac{d}{dx}(\ln(x^x)) = \ln(x) + 1\,. \tag{12.35}$$

Example 10

Consider the function $f(x) = \ln(1 + x)$. 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 (12.15).

Example 11

Consider the function $f(x) = \ln\left(1+x^2\right)$. Letting $u = 1+x^2$, f becomes a composite function $f(x) = \ln(u)$. The derivative of f with respect to x proceeds as:

$$\begin{aligned}\frac{df}{dx} &= \frac{df}{du}\cdot\frac{du}{dx}\\ &= \left(\frac{d}{du}\ln(u)\right)\left(\frac{d}{dx}\left(1+x^2\right)\right)\\ &= \frac{1}{u(x)}2x\\ &= \frac{2x}{1+x^2}\,,\end{aligned}$$

in agreement with Equation (12.19).

Example 12

Consider the function $f(x) = \sin\left(x^2\right)$. Letting $u = x^2$, f becomes a composite function $f(x) = \sin(u)$. The derivative of f with respect to x proceeds as

$$\begin{aligned}\frac{df}{dx} &= \frac{df}{du}\cdot\frac{du}{dx}\\ &= \left(\frac{d}{du}\sin(u)\right)\left(\frac{d}{dx}x^2\right)\\ &= \cos(u(x))\,2x\\ &= \cos\left(x^2\right)2x\,,\end{aligned}$$

in agreement with Equation (12.26).

Logarithm Trick

The chain rule allows for some interesting cheats when calculating derivatives. Suppose we have a function $f(x)$ that seems to 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

$$\frac{d}{dx}\left(\ln\left(f(x)\right)\right) = \frac{1}{f(x)}f'(x)\,,$$

or

$$\frac{df}{dx} = f(x)\,\frac{d}{dx}\left(\ln\left(f(x)\right)\right)\,, \tag{12.36}$$

which we'll call the *logarithm trick*.

Example 13

Consider the function $f(x) = b^{x^2}$. Using the logarithm trick, we find

$$\begin{aligned}\frac{df}{dx} &= b^{x^2}\frac{d}{dx}\left(\ln\left(b^{x^2}\right)\right)\\ &= b^{x^2}\ln(b)\,\frac{d}{dx}x^2\\ &= 2x\,b^{x^2}\ln(b)\,,\end{aligned}$$

in agreement with Equation (12.12).

Example 14

Consider the function $f(x) = x^x$. Using the logarithm trick, we find

$$\begin{aligned}\frac{df}{dx} &= x^x\frac{d}{dx}\left(\ln\left(x^x\right)\right)\\ &= x^x\frac{d}{dx}\left(x\ln(x)\right)\\ &= x^x\left(\ln(x)+1\right)\,,\end{aligned}$$

in agreement with Equation (12.13).

Power Rule Revisited

Recall that the derivative of a function $f(x) = x^n$ is established by Equation (12.2), namely

$$\frac{d}{dx}\left(x^n\right) = nx^{n-1}\,.$$

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 (12.2) holds for any n. To prove this, start with

$$f(x) = x^n = e^{\ln\left(x^n\right)} = e^{n\ln(x)}\,.$$

By the chain rule, we then find

$$\frac{d}{dx}\left(x^n\right) = e^{n\ln(x)}\frac{n}{x} = nx^{n-1}$$

and we're done.

Problem 1

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 (12.12). (ii) Let $f(x) = g(x) = 2^{x^2}$ and use Equation (12.32). (iii) Let $f(g) = 4^g$ and $g(x) = x^2$ and calculate df/dx. (iv) Use Equation (12.36).

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' (×) 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):

$$y'(x) = \pm b\left(\frac{1}{2\sqrt{1 - x^2/a^2}}\right)\left(-2x/a^2\right)$$

$$y'(x) = -\frac{b^2}{a^2}\frac{x}{y}$$

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:

$$\frac{dy}{dx} = \frac{-b^2}{a^2}\frac{x}{y}$$

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

$$y = \left(\frac{-b^2}{a^2}\frac{x_0}{y_0}\right)x + b ,$$

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:

$$\frac{d}{dx}\arccos(x) = \frac{-1}{\sqrt{1-x^2}} \tag{12.37}$$

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:

$$\frac{d}{dx}\arcsin(x) = \frac{1}{\sqrt{1-x^2}} \tag{12.38}$$

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 $\cos^2(\theta) = 1/\left(1+\tan^2(\theta)\right)$ and the answer emerges:

$$\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2} \tag{12.39}$$

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:

$$\frac{d}{dx}\operatorname{arcsec}(x) = \frac{1}{x\sqrt{x^2-1}} \tag{12.40}$$

$$\frac{d}{dx}\operatorname{arccsc}(x) = \frac{-1}{x\sqrt{x^2-1}} \tag{12.41}$$

$$\frac{d}{dx}\operatorname{arccot}(x) = \frac{-1}{1+x^2} \tag{12.42}$$

3.2 Hyperbolic Derivatives

Sinh, Cosh, Tanh

The hyperbolic trigonometric functions

$$\cosh(x) = \frac{e^x + e^{-x}}{2}$$
$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

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

$$f'(x) = \lim_{x\to x_0}\frac{\sinh(x) - \sinh(x_0)}{x - x_0}\,,$$

which, by a similar process the led to Equation (12.20), gives:

$$\frac{d}{dx}(\sinh(x)) = \cosh(x) \tag{12.43}$$

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

$$\frac{d}{dx}(\cosh(x)) = \sinh(x) \tag{12.44}$$

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:

$$\frac{d}{dx}(\tanh(x)) = (\operatorname{sech}(x))^2 \tag{12.45}$$

Coth, Sech, Csch

The hyperbolic cotangent, hyperbolic secant, and hyperbolic cosecant are also straightforwardly handled:

$$\frac{d}{dx}(\coth(x)) = -(\operatorname{csch}(x))^2 \tag{12.46}$$

$$\frac{d}{dx}(\operatorname{sech}(x)) = -\operatorname{sech}(x)\tanh(x) \tag{12.47}$$

$$\frac{d}{dx}(\operatorname{csch}(x)) = -\operatorname{csch}(x)\coth(x) \tag{12.48}$$

Arccosh, Arcsinh, Arctanh

The inverse hyperbolic functions are also straightforward to handle using the ordinary trig case as an analogy:

$$\frac{d}{dx}(\operatorname{arccosh}(x)) = \frac{1}{\sqrt{x^2 - 1}} \tag{12.49}$$

$$\frac{d}{dx}(\operatorname{arcsinh}(x)) = \frac{1}{\sqrt{x^2 + 1}} \tag{12.50}$$

$$\frac{d}{dx}(\operatorname{arctanh}(x)) = \frac{1}{1 - x^2} \tag{12.51}$$

Arcsech, Arcscsch, Arccoth

$$\frac{d}{dx}(\operatorname{arcsech}(x)) = \frac{-1}{x\sqrt{1 - x^2}} \tag{12.52}$$

$$\frac{d}{dx}(\operatorname{arccsch}(x)) = \frac{-1}{x\sqrt{x^2 + 1}} \tag{12.53}$$

$$\frac{d}{dx}(\operatorname{arccoth}(x)) = \frac{1}{1 - x^2} \tag{12.54}$$

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:

$$y - y_0 = f'(x_0)(x - x_0) \tag{12.55}$$

4.2 Mean Value Theorem

The *mean value theorem* is a piece of mathematics that concretizes something that my 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

$$f'(x_m) = \frac{f(b) - f(a)}{b - a}. \tag{12.56}$$

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 point 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

$$L(x_0) = \lim_{x \to x_0} \frac{f(x)}{g(x)} ,$$

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:

$$R = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{g(x) - g(x_0)} \left(\frac{\cancel{x - x_0}}{\cancel{x - x_0}} \right)$$

The $0/0$ Case

Now impose the condition $f(x_0) = 0$ and $g(x_0) = 0$, and the ratio becomes

$$R(x_0) = \lim_{x \to x_0} \frac{f(x)}{g(x)} ,$$

which is in fact identical to $L(x_0)$. This can only mean, for points x_0 that cause L to blow up,

$$L(x_0) = \lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{f'(x_0)}{g'(x_0)} , \qquad (12.57)$$

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 $f(x_0) \to \infty$ and $g(x_0) \to \infty$.

Flipping the problem on its head slightly, one may write

$$L(x_0) = \lim_{x \to x_0} \frac{1/g(x)}{1/f(x)} ,$$

and then attack this using the chain rule. Doing so, we get

$$L(x_0) = \frac{g'(x_0)}{f'(x_0)} \lim_{x \to x_0} \left(\frac{f(x)}{g(x)} \right)^2$$
$$L(x_0) = \frac{g'(x_0)}{f'(x_0)} \left(L(x_0) \right)^2 ,$$

and ultimately,

$$L(x_0) = \frac{f'(x_0)}{g'(x_0)} ,$$

familiar already as Equation (12.57).

The Infinite-x Case

Equation (12.57) is reinforced again by investigating the case there L blows up for $x \to \infty$. To proceed, define a variable $t = 1/x$ such that

$$L(x_0) = \lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{t \to 0} \frac{f(1/t)}{g(1/t)} ,$$

which transforms the problem into a $0/0$-like problem.

Running the chain rule on the right side, we further find:

$$L(x_0) = \lim_{t \to 0} \frac{f'(1/t)}{g'(1/t)} \cancel{\left(\frac{-t^2}{-t^2} \right)}$$
$$L(x_0) = \lim_{x \to \infty} \frac{f'(x)}{g'(x)}$$

Examples

You are encouraged to work through each of the following. For a bonus, pick out the example that helps establish that $0^0 = 1$.

Example 1

$$\lim_{x \to 0} \frac{\tan(x)}{x} = 1$$

Example 2

$$\lim_{x \to 0} \frac{1 - \cos(x)}{x^2} = \lim_{x \to 0} \frac{\sin(x)}{2x} = \frac{1}{2}$$

Example 3

$$\lim_{x\to 0}\frac{e^x-1-x}{\sin^2(x)}=\frac{1}{2}$$

Example 4

$$\lim_{x\to 0^+}x\ln(x)=\lim_{x\to 0^+}\frac{\ln(x)}{1/x}=0$$

Example 5

$$\lim_{x\to 0}\frac{\ln(x)}{x^p}=0$$

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:

$$\lim_{w\to 0}f\left(x_c-\frac{w}{2}\right)=\lim_{w\to 0}f\left(x_c+\frac{w}{2}\right)\qquad(12.58)$$

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

$$f'(x_0)=\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}\,,$$

and then using the shift of variables

$$x\to x_c+w/2$$
$$x_0\to x_c-w/2$$

such that $x-x_0=w$, we write the symmetric form

$$f'(x_c)=\lim_{w\to 0}\frac{f(x_c+w/2)-f(x_c-w/2)}{w}\,.$$

This is perhaps a more 'natural' representation of the derivative compared to what we've been working with. Enforcing Equation (12.58) 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).

Examples

Example 6

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.

From the information given we can write the volume and surface area of the can:

$$V=\pi r^2h$$
$$A=2\pi rh+2\pi r^2$$

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.

$$\frac{dA}{dr}=0=-\frac{2V}{r_c^2}+4\pi r_c\,,$$

implying $h_c=2r_c$. Evidently, the most efficient can has the height equal to the diameter.

Example 7

Prove that

$$e^\pi>\pi^e\,.$$

First use the logarithm operator to get like symbols on their own sides:

$$\pi\ln(e)=e\ln(\pi)$$
$$\frac{\ln(e)}{e}>\frac{\ln(\pi)}{\pi}$$

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.

Example 8

Find the largest rectangle that fits inside a 3-4-5 right triangle where one of the rectangle's edges lies on the hypotenuse.

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 $(0, y_*)$, $(x_*, 0)$, having length

$$b = \sqrt{x_*^2 + y_*^2},$$

and obeying the ratio

$$\frac{y_*}{x_*} = \frac{3}{4}.$$

The height of the inscribed rectangle is

$$h = (4 - x_*)\sin(\theta),$$

where $\sin(\theta) = 3/5$ from geometry.

The area of the rectangle is $A = bh$, or, all in terms of one variable:

$$\begin{aligned} A(x_*) &= \sqrt{x_*^2 + y_*^2}\,(4 - x_*)\,\frac{3}{5} \\ &= x_*(4 - x_*) \end{aligned}$$

The critical point x_c is found by calculating $dA/dx_* = 0$, namely

$$\frac{dA}{dx_*} = 0 = 4 - 2x_c,$$

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 $A = 3$.

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 Sheet

A circular ice sheet of radius $r(t)$ in meters and area $A(t)$ is melting at a rate of $-\alpha\, m^2/s$. How quickly is the radius decreasing?

For this situation, we 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 Rocket

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, $r'(t)$.

Use the Pythagorean theorem to get started

$$r^2 = D^2 + v_0^2 t^2,$$

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:

$$r'(r) = \frac{v_0^2 t}{\sqrt{D^2 + v_0^2 t^2}}$$

5 Second Derivative

5.1 Slope of Slope

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:

$$\text{Slope at a point} = f'(x) = \frac{d}{dx}f(x) = f^{(1)}(x)$$

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:

$$f''(x_0) = \lim_{x \to x_0} \frac{f'(x) - f'(x_0)}{x - x_0}$$

Carefully replacing each f'-term with the definition again, we get

$$f''(x_0) = \lim_{x \to x_0} \frac{1}{x - x_0} \left(\lim_{w \to x} \frac{f(w) - f(x)}{w - x} - \frac{f(x) - f(x_0)}{x - x_0} \right),$$

or, after some algebra,

$$f''(x_0) = \lim_{w \to x} \lim_{x \to x_0} \left(\frac{1}{x - x_0} \right)^2 \left(\lambda f(w) - (\lambda + 1) f(x) + f(x_0) \right),$$

where

$$\lambda = \frac{x - x_0}{w - x}.$$

The above contains two simultaneous limits, namely $w \to x$ and $x \to x_0$. 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

$$f''(x_0) = \lim_{h \to 0} \lim_{x \to x_0} \frac{f(x + h) - 2f(x) + f(x - h)}{h^2},$$

simplifying once more to the standard formula for the second derivative:

$$f''(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) + f(x_0 - h) - 2f(x_0)}{h^2} \tag{12.59}$$

In practice, one does not need to directly deploy Equation (12.59) 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 Stability at Critical Point

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 12.1.

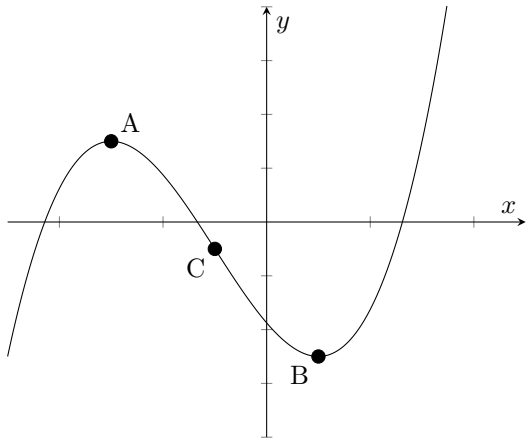


Figure 12.1: Cubic curve having two critical points and one inflection point.

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:

$$x_A = -3/2$$
$$x_B = 1/2$$

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 12.1.

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:

$$f''(x) = \begin{cases} < 0 & \text{Concave down} \\ = 0 & \text{Inflection} \\ > 0 & \text{Concave up} \end{cases}$$

6 Taylor's Theorem

6.1 Kinematic Motivation

In freshman kinematics, one encounters the equations of motion under uniform acceleration

$$x(t) = x_0 + v_0 t + \frac{1}{2}at^2$$
$$v(t) = v_0 + at\,,$$

where $x(t)$ and $v(t)$ are the position and velocity, respectively, with their initial values written as

$$x(0) = x_0$$
$$v(0) = v_0\,,$$

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:

$$v(t) = v_0 + a_0 t + \frac{1}{2}jt^2$$
$$a(t) = a_0 + jt$$

The equation for $x(t)$ is a little more tricky though. Going from the pattern, there should be a new term proportional to jt^3, but the leading coefficient must be left as a variable

$$x(t) = x_0 + v_0 t + \frac{1}{2}a_0 t^2 + \frac{1}{A}jt^3\,,$$

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

$$t \to t + h\,,$$

where h is any constant. Inserting this into the above gives

$$x(t+h) = x_0 + v_0(t+h) + \frac{1}{2}a_0(t+h)^2 + \frac{1}{A}j(t+h)^3\,,$$

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:

$$x(t+h) = \left(x_0 + v_0 t + \frac{1}{2}a_0 t^2 + \frac{1}{A}jt^3\right) + h\left(v_0 + a_0 t + \frac{3}{A}jt^2\right) + \frac{1}{2}h^2\left(a_0 + \frac{6}{A}jt\right) + \frac{1}{6}h^3(j)$$

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

$$x(t+h) = x_t + v_t h + \frac{1}{2}a_t h^2 + \frac{1}{6}jh^3 ,$$

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 *snap*, denoted k. (Beyond this the derivatives aren't conventionally named.) Going through the exercise, one finds

$$x(t+h) = x_t + v_t h + \frac{a_t}{2!}h^2 + \frac{j_t}{3!}h^3 + \frac{k_t}{4!}h^4 + \cdots$$

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

$$x(t) = x_{t_p} + v_{t_p}(t - t_p) + \frac{1}{2!}a_{t_p}(t - t_p)^2 + \frac{1}{3!}j_{t_p}(t - t_p)^3 + \cdots$$

Using a generalized notation to represent the velocity, acceleration, jerk, and so on, let us make the associations

$$\begin{aligned} x_{t_p} &\to x^{(0)}_{t_p} \\ v_{t_p} &\to x^{(1)}_{t_p} \\ a_{t_p} &\to x^{(2)}_{t_p} \\ j_{t_p} &\to x^{(3)}_{t_p} \\ k_{t_p} &\to x^{(4)}_{t_p} , \end{aligned}$$

and so on. On the left we've run out of 'named' items after snap, thus the general symbol $x^{(q)}_{t_p}$ is utilized to denote the qth coefficient.

Taylor Polynomial

In condensed form, $x(t)$ can be written in a most general way using summation notation

$$x(t) = x_{t_p} + \sum_{q=1}^{n} \frac{1}{q!} x^{(q)}_{t_p} (t - t_p)^q + R_n(t) , \quad (12.60)$$

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

$$p(x) = f(x_0) + \sum_{q=1}^{n} \frac{1}{q!} f^{(q)}(x_0)(x - x_0)^q + R_n(x) , \quad (12.61)$$

where $f^{(q)}(x_0)$ is the qth 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:

$$h_n(x) = \begin{cases} (f(x) - p(x)) / ((x - x_0)^n) & x \neq a \\ 0 & x = a \end{cases}$$

It is required that $f(x)$ is an n-times differentiable function and $p(x)$ is the Taylor polynomial appearing in Equation (12.61).

The theorem is considered proven when we show that $h_n(x) = 0$ for all x near x_0. Setting this up, begin with

$$H_n = \lim_{x\to x_0} h_n(x) = \lim_{x\to x_0} \frac{f(x) - p(x)}{(x-x_0)^n},$$

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:

$$\lim_{x\to x_0} \frac{\frac{d}{dx}(f(x) - p(x))}{\frac{d}{dx}((x-x_0)^n)} = \lim_{x\to x_0} \frac{f^{(1)}(x) - p^{(1)}(x)}{n(x-x_0)^{n-1}}$$

The right side is again an indeterminate ratio, calling for another application of L'Hopital's rule:

$$H_n = \lim_{x\to x_0} \frac{f^{(2)}(x) - p^{(2)}(x)}{n(n-1)(x-x_0)^{n-2}}$$

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:

$$H_n = \frac{1}{n!}\left(\lim_{x\to x_0} \frac{f^{(n-1)}(x) - p^{(n-1)}(x)}{x - x_0}\right)$$

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:

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

The derivatives f', f'', etc., are straightforwardly attained from $f(x)$:

$$f'(x_0) = 1/(x - x_0)^2$$
$$f''(x_0) = 2!/(x - x_0)^3$$
$$f'''(x_0) = 3!/(x - x_0)^4$$

Substituting these into $p(x)$ and performing the obvious cancellations gives:

$$p(x) = \frac{1}{1-x_0}\left(1 + \lambda + \lambda^2 + \cdots\right),$$

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 $|x| < 1$. 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,

$$\begin{aligned} f'(x_0) &= \cos(x_0) \\ f''(x_0) &= -\sin(x_0) \\ f'''(x_0) &= -\cos(x_0) \\ f''''(x_0) &= \sin(x_0) \end{aligned}$$

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

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots .$$

Choosing a juicier example, let $x_0 = \pi/2$ and all the signs change:

$$\sin(x) = 1 - \frac{(x-\pi/2)^2}{2!} + \frac{(x-\pi/2)^4}{4!} + \cdots$$

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,

$$\begin{aligned} f(x) &\approx f(x_0) + f'(x_0)(x - x_0) \\ g(x) &\approx g(x_0) + g'(x_0)(x - x_0) \end{aligned}$$

with the understanding that x is near x_0.

Denoting

$$\Delta x = x - x_0 ,$$

the product $P(x)$ reads

$$\begin{aligned} P(x) \approx\ & f(x_0)\, g(x_0) \\ & + \Delta x \left(f'(x_0)\, g(x) + f(x)\, g'(x_0)\right) \\ & + \cancel{(\Delta x)^2\, f'(x_0)\, g'(x_0)} \end{aligned}$$

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 (12.32), i.e. the product rule. After simplifying, we can summarize by writing the first-order approximation to $P(x)$:

$$P_1(x) = f(x_0)\, g(x_0) + \frac{d}{dx}\left(f(x)\, g(x)\right)\Big|_{x_0} \Delta x \tag{12.62}$$

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

$$\frac{1}{g(x)} \approx \lim_{x \to x_0} \frac{1}{g(x_0) + g'(x_0)(x - x_0)} ,$$

and then factor out $1/g(x_0)$:

$$\frac{1}{g(x)} \approx \frac{1}{g(x_0)} \lim_{x \to x_0} \frac{1}{1 + \lambda} ,$$

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:

$$\frac{1}{1+\lambda} = 1 - \lambda + \lambda^2 - \lambda^3 + \cdots$$

Of course, terms λ^2 and above are omitted in the first-order approximation, thus we have

$$\frac{1}{g(x)} \approx \frac{1}{g(x_0)} \lim_{x \to x_0} \left(1 - \frac{g'(x_0)}{g(x_0)}(x - x_0)\right) .$$

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:

$$Q(x) \approx \frac{f(x_0)}{g(x_0)} + \Delta x \left(\frac{f'(x_0)\, g(x) - f(x)\, g'(x_0)}{\left(g(x_0)\right)^2}\right)$$

The latter term in the above is Δx multiplied by the derivative of the quotient $f(x)/g(x)$ per Equation (12.33), i.e. the quotient rule. After simplifying, we can summarize by writing the first-order approximation to $Q(x)$:

$$Q_1(x) = \frac{f(x_0)}{g(x_0)} + \frac{d}{dx}\left(\frac{f(x)}{g(x)}\right)\Big|_{x_0} \Delta x \tag{12.63}$$

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:

$$C_1(x) = f(g(x_0)) + f'(g(x_0))\, g'(x_0)\, \Delta x \quad (12.64)$$

Second Derivative

The formula for the second derivative can also be wriggled from the Taylor polynomial. First write the standard approximation of $f(x)$:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

The same function can be approximated from a different base point, namely $x \to 2x_0 - x$. Writing this out, we have

$$f(2x_0 - x) \approx f(x_0) + f'(x_0)(x_0 - x) + \frac{f''(x_0)}{2!}(x_0 - x)^2 + \frac{f'''(x_0)}{3!}(x_0 - x)^3 + \cdots \ ,$$

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 and reshuffle a little to write

$$\frac{f(x) + f(2x_0 - x) - 2f(x_0)}{(x - x_0)^2} \approx f''(x_0) + \frac{2}{2!}\cancel{f''''(x_0)\,\Delta x^2} \ ,$$

where terms containing powers of Δx^2 and above are negligible in a first-order approximation.

In the infinitesimal limit $x \to x_0$, the above reduces to the exact formula for the second derivative, Equation (12.59). 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:

$$f^{(1)}(x_0) = r(x_0 + a)^{r-1}$$
$$f^{(2)}(x_0) = r(r-1)(x_0 + a)^{r-2}$$
$$f^{(3)}(x_0) = r(r-1)(r-2)(x_0 + a)^{r-3}$$
$$f^{(q)}(x_0) = \frac{r!}{(r-q)!}(x_0 + a)^{r-q}$$

As a sum, the approximation for $f(x)$ then reads

$$f(x) \approx (x_0 + a)^r + \sum_{q=1}^{n} \frac{r!}{q!\,(r-q)!}(x_0 + a)^{r-q}(x - x_0)^q \ .$$

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 has a special name called the *binomial coefficients*, which follow a special notation:

$$\binom{r}{q} = \frac{r!}{q!\,(r-q)!} \quad (12.65)$$

In terms of binomial coefficients, the sum representing $f(x)$ is written

$$(x + a)^r \approx a^r \sum_{q=0}^{n} \binom{r}{q}\left(\frac{x}{a}\right)^q \quad (12.66)$$

valid for 'small' x. This is called the *binomial expansion* formula. The above can be written in open form for more practical use:

$$(x + a)^r \approx a^r + ra^{r-1}x + \frac{r(r-1)}{2!}a^{r-2}x^2 + \frac{r(r-1)(r-2)}{3!}a^{r-3}x^3 + \cdots$$

Examples

Example 1

Expand $\sqrt{1+x}$ for small x.

$$\sqrt{1+x} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \cdots \tag{12.67}$$

Example 2

Expand $1/\sqrt{1+x}$ for small x.

$$\frac{1}{\sqrt{1+x}} \approx 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \cdots \tag{12.68}$$

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:

$$\begin{aligned}
f^{(1)}(x_0) &= 1/(1+x_0) \\
f^{(2)}(x_0) &= -1/(1+x_0)^2 \\
f^{(3)}(x_0) &= 2/(1+x_0)^3 \\
f^{(4)}(x_0) &= -3\cdot 2/(1+x_0)^4 \\
f^{(q)}(x_0) &= (-1)^{q-1}(q-1)!/(1+x_0)^q
\end{aligned}$$

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:

$$\ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \tag{12.69}$$

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:

$$\begin{aligned}
f^{(1)}(x_0) &= \frac{1}{1+x_0^2} \\
f^{(2)}(x_0) &= \frac{-2}{(1+x_0^2)^2} \\
f^{(3)}(x_0) &= \frac{6x_0^2 - 2}{(1+x_0^2)^3} \\
f^{(4)}(x_0) &= \frac{-24x_0\left(x_0^2-1\right)}{(1+x_0^2)^4} \\
f^{(5)}(x_0) &= \frac{24\left(5x_0^4 - 10x_0^2 + 1\right)}{(1+x_0^2)^5}
\end{aligned}$$

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 $x = 0$:

$$\arctan(x) \approx x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \tag{12.70}$$

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:

$$\begin{aligned}
\arctan(x)_{x\approx 1} = &\frac{\pi}{4} + \frac{x-1}{2} \\
&- \frac{(x-1)^2}{4} + \frac{(x-1)^3}{12} \\
&- \frac{(x-1)^5}{40} + \frac{(x-1)^6}{48} - \cdots
\end{aligned} \tag{12.71}$$

Arctangent near Two

For the sake of completeness, the arctangent near $x = 2$ case works out to be:

$$\begin{aligned}
\arctan(x)_{x\approx 2} = &\arctan(2) + \frac{x-2}{5} \\
&- \frac{2(x-2)^2}{25} + \frac{11(x-2)^3}{375} - \cdots
\end{aligned} \tag{12.72}$$

Arctangent of Two

It's worth pausing a moment on the quantity $\arctan(2)$, which is required to evaluate Equation (12.72).

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 (12.71) 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:

$$\arctan(x) = \frac{\pi}{2} - \arctan\left(\frac{1}{x}\right) \tag{12.73}$$

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 (12.70), (12.71) 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 $\pm\pi/2$, $\pm 3\pi/2$, 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:

$$\begin{aligned}
f^{(1)}(0) &= 1 \qquad & f^{(2)}(0) &= 0 \\
f^{(3)}(0) &= 2 \qquad & f^{(4)}(0) &= 0 \\
f^{(5)}(0) &= 16
\end{aligned}$$

Plugging this information into Taylor's theorem yields a useful approximation to the tangent function:

$$\tan(x) \approx x + \frac{x^3}{3} + \frac{2x^5}{15} + O\left(x^7\right) \tag{12.74}$$

The symbol $O\left(x^7\right)$ 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

$$O\left(x^7\right) = \frac{17x^7}{315} \ ,$$

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:

$$\begin{aligned}
f^{(1)}(\pi/4) &= 2 \qquad & f^{(2)}(\pi/4) &= 4 \\
f^{(3)}(\pi/4) &= 16 \qquad & f^{(4)}(\pi/4) &= 80 \\
f^{(5)}(\pi/4) &= 512
\end{aligned}$$

Not forgetting the shift by x_0 units, the approximation for the tangent near $x = \pi/4$ reads

$$\begin{aligned}
\tan(x)_{x\approx\pi/4} = 1 &+ 2\left(x - \frac{\pi}{4}\right) \\
&+ 2\left(x - \frac{\pi}{4}\right)^2 + \frac{8}{3}\left(x - \frac{\pi}{4}\right)^3 \\
&+ \frac{10}{3}\left(x - \frac{\pi}{4}\right)^4 + \frac{64}{15}\left(x - \frac{\pi}{4}\right)^5 \\
&+ O\left(x - \frac{\pi}{4}\right)^6
\end{aligned} \tag{12.75}$$

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

$$\cot(z) \approx -z - \frac{z^3}{3} - \frac{2z^5}{15} - O\left(z^7\right)\,. \tag{12.76}$$

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:

$$\tan(z) \approx -\frac{1}{z} + \frac{z}{3} + \frac{z^3}{45} + O\left(z^5\right) \tag{12.77}$$

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:

$$\cot(x) \approx \frac{1}{x} - \frac{x}{3} - \frac{x^3}{45} - O\left(x^5\right) \tag{12.78}$$

6.10 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

$$a = -g = -9.8 m/s^2\,.$$

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:

$$x(t) = x_0 + v_0 t + \frac{1}{2}at^2$$

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 $v(t)$:

$$\frac{dv}{dt} = -g - bv \tag{12.79}$$

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:

$$v(t) = v_0 + A_1 t + A_2 t^2 + A_3 t^3 + \cdots$$

Without knowing much else about $v(t)$, we can still compute the derivative:

$$\frac{dv}{dt} = A_1 + 2A_2 t + 3A_3 t^2 + 4A_4 t^3 + \cdots$$

Now, plug both of these into $-g = bv + dv/dt$:

$$\begin{aligned} -g = {} & bv_0 + bA_1 t + bA_2 t^2 + bA_3 t^3 + \cdots \\ & + A_1 + 2A_2 t + 3A_3 t^2 + 4A_4 t^3 + \cdots \end{aligned}$$

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:

$$v(t) = v_0 + A_1 t - \frac{bA_1}{2!}t^2 + \frac{b^2A_1}{3!}t^3 - \frac{b^3A_1}{4!}t^4 + \cdots$$

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

$$v(t) = \frac{-g}{b} + \left(v_0 + \frac{g}{b}\right)e^{-bt} . \tag{12.80}$$

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

$$v_{\text{terminal}} = \lim_{t\to\infty} v(t) = \frac{-g}{b} .$$

Position

It just happens that the Frobenius method works for attaining $x(t)$. Postulating

$$x(t) = x_0 + B_1 t + B_2 t^2 + B_3 t^3 + \cdots ,$$

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

$$B_2 = -\frac{1}{2}\left(g + bv_0\right) ,$$

and the position equation takes the form:

$$x(t) = x_0 - \frac{gt}{b} + \frac{1}{b}\left(v_0 + \frac{g}{b}\right)\left(1 - e^{-bt}\right) \tag{12.81}$$

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 (12.80) 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 (12.81) and expanding the exponential to second order gives:

$$x(t) \approx x_0 - \frac{gt}{b} + \frac{1}{b}\left(v_0 + \frac{g}{b}\right)\left(bt - \frac{1}{2}b^2t^2\right) ,$$

which simplifies to

$$x(t) \approx x_0 + v_0 t - \frac{1}{2}(g + v_0 b)\, t^2$$

as expected.

7 Numerical Methods

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.

7.1 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 12.2. 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.

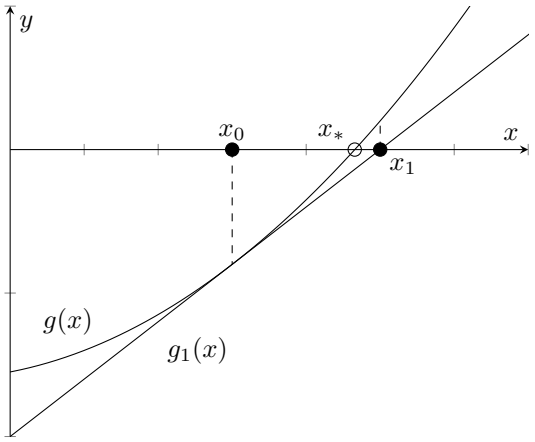


Figure 12.2: Newton's method.

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

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} , \tag{12.82}$$

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

$$x_* \approx 0.7390851332151607\ldots .$$

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

$$x_{n+1} = x_n - \frac{x^3 - 29}{3x^2}$$

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 $x_* \approx 3.072\ldots$, or

$$(29)^{1/3} \approx 3.072316825685847 .$$

Digits of Pi

On a scientific calculator set to radians, type

$$3.14 - \tan(3.14)$$

to get an approximate output

$$\pi \approx 3.14159265\ldots .$$

The reason this works is left as an exercise for the reader.

Second-Order Newton's Method

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:

$$h = \frac{-g'(x_0)}{g''(x_0)} \pm \frac{g'(x_0)}{g''(x_0)}\sqrt{1 - \frac{2g(x_0)\,g''(x_0)}{(g'(x_0))^2}}$$

To deal with the square root term, we turn to another order-two approximation in the form of Equation (12.67). 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:

$$x_{n+1} = x_n - \frac{-g(x_n)}{g'(x_n)}\left(1 + \frac{g(x_n)\,g''(x_n)}{2(g'(x_n))^2}\right) \quad (12.83)$$

7.2 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 that we need to estimate the square root of some number N. Proceed by assuming N to be comprised of some lesser number $Q < N$, along with a smaller contribution $x \ll Q$ such that $Q + x = N$, or also

$$Q^2 + 2Qx + x^2 = N^2 \, .$$

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 $Q + x = N$ is replaced with

$$Q + \frac{N^2 - Q^2}{2Q} \approx N \, .$$

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

$$Q_{n+1} = Q_n + \frac{N^2 - Q_n^2}{2Q_n} \, , \tag{12.84}$$

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

$$x \approx \frac{N^3 - Q^3}{3Q^2} \, ,$$

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 kth 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 $x_* = N$.

7.3 Euler's Method

Numerical methods need not be limited to estimating individual numbers, as estimating entire curves is also fair game.

Revisiting the scenario of kinematics 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. By the Frobenius method we were able to jot down exact solutions to this problem, namely

$$\begin{aligned} v\,(t) &= \frac{-g}{b} + \left(v_0 + \frac{g}{b}\right) e^{-bt} \\ x\,(t) &= x_0 - \frac{gt}{b} + \frac{1}{b}\left(v_0 + \frac{g}{b}\right)\left(1 - e^{-bt}\right) \, , \end{aligned}$$

where after supplying the initial values v_0, x_0, the motion is completely determined.

The haunting question now is, what if we could not easily get hold of solutions for $v\,(t)$, $x\,(t)$? It seems that things fell into place by pure luck in a sense, and if the differential equation had been more complicated, maybe solutions would be be hopelessly tangled up.

A technique called *Euler's method* allows for approximating the path of motion directly from the differential equation. Assuming v_0, x_0 as given, the idea is, much like Newton's method, to calculate the updated information v_1, x_1 using linear approximations.

Forward Euler's Method

Starting with an easy case, consider the frictionless constant-acceleration scenario characterized by

$$dv/dt = -g$$
$$dx/dt = v\,(t)\ .$$

Expanding each of these to first order, we have

$$v\,(t) \approx v\,(t_0) - g\Delta t$$
$$x\,(t) \approx x\,(t_0) + v\,(t_0)\,\Delta t\ .$$

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

$$v_{n+1} = v_n - g\Delta t \tag{12.85}$$
$$x_{n+1} = x_n + v_n\Delta t \tag{12.86}$$

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.

Backward Euler's Method

Recalling the definition of the derivative, namely

$$f'\,(x) = \lim_{h\to 0}\frac{f\,(x+h) - f\,(x)}{h}\ ,$$

note that the definition remains intact by reversing the sign on h:

$$f'\,(x) = \lim_{h\to 0}\frac{f\,(x) - f\,(x-h)}{h}\ .$$

Moreover, no error is made if we simply shift variables $x \to x+h$, so we can also write

$$\lim_{h\to 0} f'\,(x+h) = \lim_{h\to 0}\frac{f\,(x+h) - f\,(x)}{h}\ ,$$

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'd 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 on hand, the backward Euler's method is represented recursively via

$$v_{n+1} = v_n - g\Delta t \tag{12.87}$$
$$x_{n+1} = x_n + v_{n+1}\Delta t\ , \tag{12.88}$$

which is subtlety different from Equations (12.85), (12.86). Note that in a more general case, the g-term would instead be a_{n+1}.

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 n, the energy is

$$E_n = \frac{1}{2}mv_n^2 + mgx_n\ .$$

At the next step $n+1$ the same energy ought to read

$$E_{n+1} = \frac{1}{2}mv_{n+1}^2 + mgx_{n+1}\ .$$

All is fair until we want to substitute v_{n+1} and x_{n+1} into E_{n+1}. That is, nothing says to only use the forward method represented by Equations (12.85), (12.86), or for that matter, nothing forbids the pair of Equations (12.87), (12.88). Which pair is correct? At this point we're obligated to try both, and doing each case carefully, we find:

$$\text{forward: } E_{n+1} = E_n + \frac{1}{2}mg^2\Delta t^2$$
$$\text{backward: } E_{n+1} = E_n - \frac{1}{2}mg^2\Delta t^2$$

Evidently, the energy is conserved to zeroth order and first order by each method. There is, however, a pesky second-order term lingering in each result. This will surely introduce artificiality in the results.

It should be noted that error can be minimized when Δt is very small, but this still leaves the question of, can we do better?

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:

$$v_{n+1} = v_n - g\Delta t \tag{12.89}$$

$$x_{n+1} = x_n + \frac{1}{2}\left(v_n + v_{n+1}\right)\Delta t \tag{12.90}$$

As it turns out, this pair of equations does in fact satisfy $E_{n+1} = E_n$ which you are encouraged to verify.

To see each of Euler's methods in action against a concrete problem, consider the one-dimensional motion of a body that begins at $x_0 = 10\,m$ at $t = 0\,s$ with initial upward speed of $v_0 = 10\,m/s$. For a time step we'll use $\Delta t = 0.1\,s$ over a total of 20 iterations. Writing the appropriate C program and producing graphs with *gnuplot*, we generate the outputs shown in Figure 12.3.

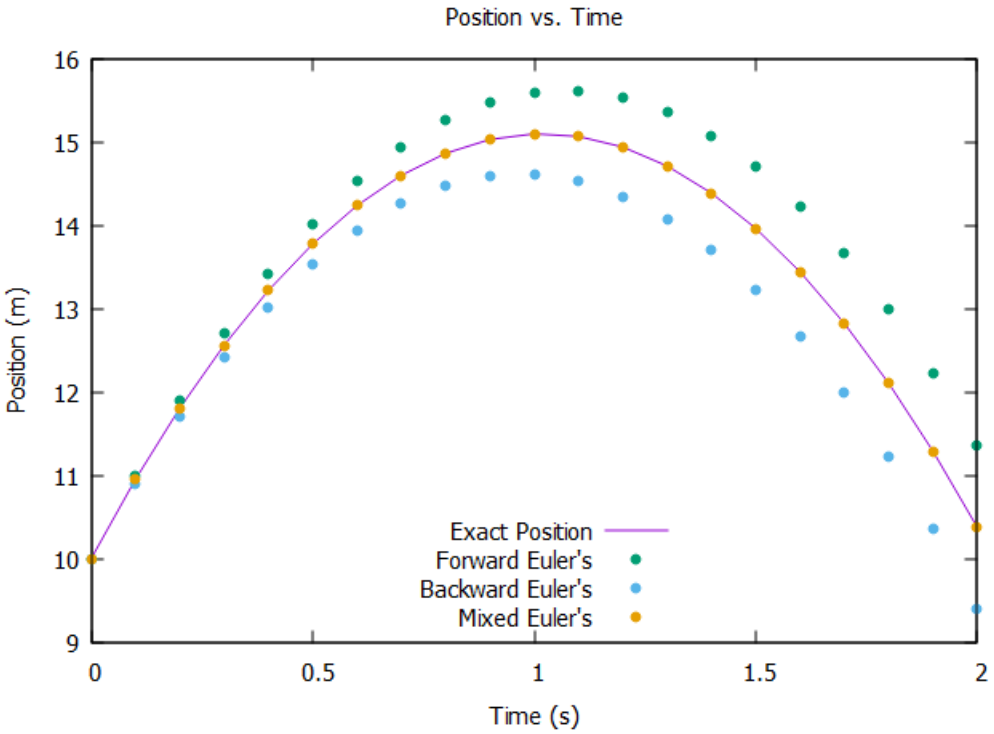


Figure 12.3: Various Euler's method approximations of ideal freefall motion compared to exact solution.

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.

Air Damping Problem

Returning to the problem of kinematics with air damping, governed by

$$\frac{dv}{dt} = -g - bv\,,$$

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 solu-

tions with variations of Euler's method.

For the air damping problem, a set of 'forward' equations are (as always) easy to write explicitly:

$$v_{n+1} = v_n - g\Delta t - bv_n\Delta t$$
$$x_{n+1} = x_n + v_n\Delta t$$

That is, the above is analogous to Equations (12.85), (12.86) and only differ by the presence of the b-term.

As for a 'backward' set of equations, replace downstream n on the right with $n+1$ to get

$$v_{n+1} = v_n - g\Delta t - bv_{n+1}\Delta t$$
$$x_{n+1} = x_n + v_{n+1}\Delta t\,,$$

analogous to Equations (12.87), (12.88). Solving for v_{n+1} and x_{n+1} explicitly, we find these to mean

$$v_{n+1} = \frac{v_n - g\Delta t}{1 + b\Delta t}$$
$$x_{n+1} = x_n + \left(\frac{v_n - g\Delta t}{1 + b\Delta t}\right)\Delta t\,.$$

Finally, we can pursue a set of 'mixed' equations by imposing the average on the downstream terms as

$$v_{n+1} = v_n - g\Delta t - \frac{b}{2}\left(v_n + v_{n+1}\right)\Delta t$$
$$x_{n+1} = x_n + \left(v_n + v_{n+1}\right)\Delta t\,,$$

which are analogous to Equations (12.89), (12.90). After some effort, these can be expressed entirely with $n+1$ on the left and n on the right:

$$v_{n+1} = \frac{v_n\left(1 - b\Delta t/2\right) - g\Delta t}{1 + b\Delta t/2}$$
$$x_{n+1} = x_n + \left(\frac{v_n - g\Delta t/2}{1 + b\Delta t/2}\right)\Delta t$$

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 $b = 3\,s^{-1}$, we generate the output shown in Figure 12.4.

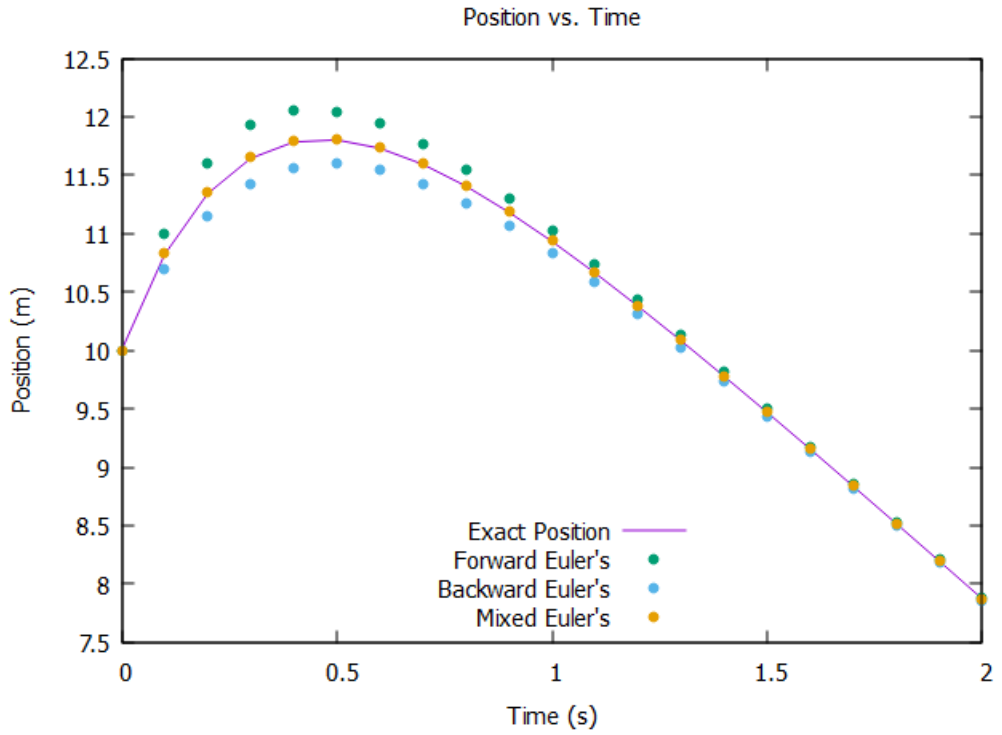


Figure 12.4: Various Euler's method approximations of damped freefall motion compared to exact solution.

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.

8 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*.

8.1 Motivation

A natural way to frame the antiderivative question starts with definition of the derivative

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} ,$$

and then set the left side to a function that is given, call it $g(x)$:

$$g(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

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

$$Q(g(x_0)) = Q\left(\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \right) .$$

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:

$$Q(g(x_0)) = \frac{d}{dx}(r(x))\bigg|_{x_0}$$

Finally, isolate $g(x_0)$ algebraically via

$$g(x_0) = Q^{-1}\left(\frac{d}{dx}(r(x))\bigg|_{x_0} \right) ,$$

where Q^{-1} reverses the manipulations represented by Q.

8.2 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:

$$\ln(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$
$$\ln(x_0) + 1 = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} + 1$$

By adding 1 to each side, the left is suddenly recognizable from Equation (12.35), which reads

$$\frac{d}{dx}(\ln(x^x)) = \ln(x) + 1 .$$

Knowing this, replace the right side of our working equation:

$$\ln(x_0) + 1 = \frac{d}{dx}(\ln(x^x))\bigg|_{x_0}$$
$$\ln(x_0) + 1 = \lim_{x \to x_0} \frac{\ln(x^x) - \ln(x_0^{x_0})}{x - x_0}$$

Now subtract 1 from each side, thereby applying Q^{-1}, and simplify:

$$\ln(x_0) = \lim_{x \to x_0} \frac{(x\ln(x) - x) - (x_0\ln(x_0) - x_0)}{x - x_0}$$

The right side is none other than the derivative of $x\ln(x) - x$, and we're done:

$$\ln(x) = \frac{d}{dx}(x\ln(x) - x) \tag{12.91}$$

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 (12.28), 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:

$$x\cos(x) = \frac{d}{dx}(x\sin(x) + \cos(x)) \tag{12.92}$$

By identical reasoning, one can work out the case $f'(x) = x\sin(x)$ using Equation (12.29). Leaving the details as an exercise, the result is

$$x\sin(x) = \frac{d}{dx}(\sin(x) - x\cos(x)) \ . \tag{12.93}$$

8.3 Powers and Roots

The rule governing powers and roots is covered by Equation (12.2), 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:

$$\frac{\pm x}{\sqrt{1 \pm x^2}} = \frac{d}{dx}\sqrt{1 \pm x^2} \tag{12.94}$$

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 (12.14) going backwards, namely:

$$\frac{1}{x} = \frac{d}{dx}(\ln(x))$$

8.4 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

$$\frac{d}{dx}\left((\ln(x))^2\right) = \frac{2}{x}\ln(x) \ . \tag{12.95}$$

Recognizing the original problem embedded on the right, we have the answer:

$$\frac{\ln(x)}{x} = \frac{1}{2}\frac{d}{dx}\left((\ln(x))^2\right) \ . \tag{12.96}$$

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:

$$\ln(x+1) = \frac{d}{dx}((x+1)\ln(x+1) - x) \tag{12.97}$$

Nonlinear Natural Logarithm

Consider the case $f'(x) = x\ln(x)$. For this, use the product rule to establish

$$\frac{d}{dx}\left(x^2\ln(x)\right) = 2x\ln(x) + x \ .$$

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

$$x = \frac{1}{2}\frac{d}{dx}\left(x^2\right) \ .$$

Substituting this into the above and simplifying gives the result we're after:

$$x\ln(x) = \frac{1}{2}\frac{d}{dx}\left(x^2\left(\ln(x) - \frac{1}{2}\right)\right) \tag{12.98}$$

Modified Natural Logarithm

Finding the antiderivative of the modified natural logarithm $f'(x) = \ln\left(1+x^2\right)$ is a challenge. To begin we'll write something that contains $1+x^2$ and hope for the best, particularly:

$$\frac{d}{dx}\left(x\ln\left(1+x^2\right)\right) = \ln\left(1+x^2\right) + \frac{2x^2}{1+x^2}$$

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:

$$\ln\left(1+x^2\right) = \frac{d}{dx}\left(x\left(\ln\left(1+x^2\right) - 2\right)\right) + 2\frac{d}{dx}\left(\arctan(x)\right) \tag{12.99}$$

8.5 Exponential Antiderivatives

Exponential Times X

To handle the case $f'(x) = xe^x$, use the product rule on the same quantity

$$\frac{d}{dx}\left(xe^x\right) = e^x + xe^x ,$$

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:

$$xe^x = \frac{d}{dx}\left(e^x(x-1)\right) \tag{12.100}$$

Exponential Times X*X

To handle the case $f'(x) = x^2e^x$, use the product rule on the same quantity

$$\frac{d}{dx}\left(x^2e^x\right) = 2xe^x + x^2e^x ,$$

and notice the right-most term contains the answer we want. The middle term would be show-stopper if it weren't for Equation (12.100), which allows the rest to be written as a total derivative:

$$x^2e^x = \frac{d}{dx}\left(e^x\left(x^2 - 2x + 2\right)\right) \tag{12.101}$$

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:

$$\frac{d}{dx}\left(e^x\sin(x)\right) = e^x\sin(x) + e^x\cos(x)$$
$$\frac{d}{dx}\left(e^x\cos(x)\right) = e^x\cos(x) - e^x\sin(x)$$

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:

$$e^x\cos(x) = \frac{1}{2}\frac{d}{dx}\left(e^x\cos(x) + e^x\sin(x)\right) \tag{12.102}$$
$$e^x\sin(x) = \frac{1}{2}\frac{d}{dx}\left(e^x\sin(x) - e^x\cos(x)\right) \tag{12.103}$$

8.6 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 (12.24) comes to mind, namely

$$\frac{d}{dx}\left(\sec(x)\right) = \tan(x)\sec(x) .$$

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 (12.36), the left side is equivalent to the derivative of the natural log of u:

$$\frac{d}{dx}\left(\ln(u)\right) = \tan(x)$$

Reversing the u-substitution, the final answer is

$$\tan(x) = \frac{d}{dx}\left(-\ln\left(\cos(x)\right)\right) . \tag{12.104}$$

By a similar line of reasoning, the cotangent version can also be done, with the details left as an exercise:

$$\cot(x) = \frac{d}{dx}\left(\ln\left(\sin(x)\right)\right) . \tag{12.105}$$

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

$$\begin{aligned}\frac{1}{\cos(x)} = &-\frac{1}{2}\frac{d}{dx}\left(\ln(1-\sin(x))\right) \\ &+ \frac{1}{2}\frac{d}{dx}\left(\ln(1+\sin(x))\right) ,\end{aligned}$$

which can be simplified and the problem is solved:

$$\sec(x) = \frac{1}{2}\frac{d}{dx}\left(\ln\left(\frac{1+\sin(x)}{1-\sin(x)}\right)\right) \tag{12.106}$$

With a little more algebra, the above can be simplified to

$$\sec(x) = \frac{d}{dx}\left(\ln(\sec(x)+\tan(x))\right) .$$

Leaving the details for an exercise, a similar exercise leads to the cosecant version

$$\csc(x) = -\frac{1}{2}\frac{d}{dx}\left(\ln\left(\frac{1+\cos(x)}{1-\cos(x)}\right)\right) , \tag{12.107}$$

or

$$\csc(x) = -\frac{d}{dx}\left(\ln(\csc(x)+\cot(x))\right) .$$

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

$$\frac{d}{dx}\left(\sin(x)\cos(x)\right) = \cos^2(x) - \sin^2(x) ,$$

which is equivalent to both of:

$$\begin{aligned}\frac{d}{dx}\left(\sin(x)\cos(x)\right) &= 1 - 2\sin^2(x) \\ \frac{d}{dx}\left(\sin(x)\cos(x)\right) &= 2\cos^2(x) - 1\end{aligned}$$

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

$$\sin^2(x) = \frac{1}{2}\frac{d}{dx}\left(x - \sin(x)\cos(x)\right) \tag{12.108}$$

$$\cos^2(x) = \frac{1}{2}\frac{d}{dx}\left(x + \sin(x)\cos(x)\right) \tag{12.109}$$

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

$$\frac{d}{dx}\left(\sin^2(x)\right) = -\frac{d}{dx}\left(\cos^2(x)\right) = 2\cos(x)\sin(x) ,$$

and the result is isolated:

$$\cos(x)\sin(x) = \frac{1}{2}\frac{d}{dx}\left(\left(\sin(x)\right)^2\right) \tag{12.110}$$

$$\cos(x)\sin(x) = \frac{-1}{2}\frac{d}{dx}\left(\left(\cos(x)\right)^2\right) \tag{12.111}$$

8.7 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:

$$\frac{d}{dx}\left(x\arccos(x)\right) = \arccos(x) + x\frac{d}{dx}\left(\arccos(x)\right)$$

The derivative of $\arccos(x)$ can be replaced by Equation (12.37), namely

$$\frac{d}{dx}\arccos(x) = \frac{-1}{\sqrt{1-x^2}} ,$$

and the above becomes

$$\frac{d}{dx}\left(x\arccos(x)\right) = \arccos(x) - \frac{x}{\sqrt{1-x^2}} .$$

The square root term is itself the derivative of a function obeying Equation (12.94), 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:

$$\arccos(x) = \frac{d}{dx}\left(x\arccos(x) - \sqrt{1-x^2}\right) \tag{12.112}$$

By similar reasoning, the $\arcsin(x)$ case works out as

$$\arcsin(x) = \frac{d}{dx}\left(x\arcsin(x) + \sqrt{1-x^2}\right) . \tag{12.113}$$

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 (12.39), 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 (12.15) does the job, namely

$$\frac{d}{dx}\left(\ln\left(1+x^2\right)\right) = \frac{2x}{1+x^2}\,.$$

Condensing derivatives on one side with arctan on the other gives the answer:

$$\arctan(x) = \frac{d}{dx}\left(x\arctan(x) - \frac{\ln\left(1+x^2\right)}{2}\right) \tag{12.114}$$

By similar reasoning, the $\operatorname{arccot}(x)$ case works out as

$$\operatorname{arccot}(x) = \frac{d}{dx}\left(x\operatorname{arccot}(x) + \frac{\ln\left(1+x^2\right)}{2}\right)\,. \tag{12.115}$$

8.8 Trigonometric Substitution

Arcsecant and Arccosecant

The last two inverse trig functions, namely the arcsecant 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 (12.40), 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 (12.106) 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

$$q(x) = \ln\left(x + \sqrt{x^2-1}\right)\,.$$

This result alone is worth noting,

$$\frac{1}{\sqrt{x^2-1}} = \frac{d}{dx}\left(\ln\left(x+\sqrt{x^2-1}\right)\right)\,, \tag{12.116}$$

and more importantly, we can write the final answer to the $\operatorname{arcsec}(x)$ antiderivative:

$$\operatorname{arcsec}(x) = \frac{d}{dx}\left(x\operatorname{arcsec}(x) - \ln\left(x+\sqrt{x^2-1}\right)\right) \tag{12.117}$$

A similar exercise gives the $\operatorname{arccsc}(x)$ version:

$$\operatorname{arccsc}(x) = \frac{d}{dx}\left(x\operatorname{arccsc}(x) + \ln\left(x+\sqrt{x^2-1}\right)\right) \tag{12.118}$$

8.9 Hyperbolic Cases

The nontrivial hyperbolic trig antiderivatives are mostly analogous to their ordinary trig counterparts:

$$\tanh(x) = \frac{d}{dx}(\ln(\cosh(x))) \tag{12.119}$$

$$\coth(x) = \frac{d}{dx}(\ln(\sinh(x))) \tag{12.120}$$

The hyperbolic secant and cosecant are a little less obvious, but turn out to be:

$$\operatorname{sech}(x) = \frac{d}{dx}\left(2\arctan\left(\tanh\left(\frac{x}{2}\right)\right)\right) \tag{12.121}$$

$$\operatorname{csch}(x) = \frac{d}{dx}\left(\ln\left(\tanh\left(\frac{x}{2}\right)\right)\right) \tag{12.122}$$

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 arccosh, arcsinh, one straightforwardly finds

$$\operatorname{arccosh}(x) = \frac{d}{dx}\left(x\operatorname{arccosh}(x) - \sqrt{x^2-1}\right) \tag{12.123}$$

$$\operatorname{arcsinh}(x) = \frac{d}{dx}\left(x\operatorname{arcsinh}(x) - \sqrt{x^2+1}\right) . \tag{12.124}$$

Arctanh and Arccoth

The arccosh, arcsinh functions have antiderivatives that are remarkably similar to their trig counterparts:

$$\operatorname{arctanh}(x) = \frac{d}{dx}\left(x\operatorname{arctanh}(x) + \frac{\ln\left(1-x^2\right)}{2}\right) \tag{12.125}$$

$$\operatorname{arccoth}(x) = \frac{d}{dx}\left(x\operatorname{arccoth}(x) + \frac{\ln\left(1-x^2\right)}{2}\right) \tag{12.126}$$

Arcsech and Arccsch

The arcsech, 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 (12.38), (12.50), that

$$f(x) = \arcsin(x)$$
$$g(x) = \operatorname{arcsinh}(x) ,$$

and thus:

$$\operatorname{arcsech}(x) = \frac{d}{dx}(x\operatorname{arcsech}(x) + \arcsin(x)) \tag{12.127}$$

$$\operatorname{arccsch}(x) = \frac{d}{dx}(x\operatorname{arccsch}(x) + \operatorname{arcsinh}(x)) \tag{12.128}$$

9 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

$$U_{spring}(x) = \frac{1}{2}kx^2 .$$

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:

$$\frac{d^2}{dt^2}x(t) = -\frac{k}{m}x(t) \tag{12.129}$$

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:

$$\frac{d^2}{dt^2}\cos\left(\sqrt{\frac{k}{m}}\,t\right) = \frac{-k}{m}\cos\left(\sqrt{\frac{k}{m}}\,t\right)$$
$$\frac{d^2}{dt^2}\sin\left(\sqrt{\frac{k}{m}}\,t\right) = \frac{-k}{m}\sin\left(\sqrt{\frac{k}{m}}\,t\right)$$

Angular Frequency

Both the cosine and the sine seem to satisfy Equation (12.129), 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 (Greek *omega*):

$$\omega = \sqrt{\frac{k}{m}}$$

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

$$x_1(t) = A\cos(\omega t)$$
$$x_2(t) = B\sin(\omega t)$$

fur 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 (12.129). Doing so, we find

$$\begin{aligned}\frac{d^2}{dt^2}x(t) &= -A\omega^2\cos(\omega t) - B\omega^2\sin(\omega t)\\ &= -\omega^2\left(A\cos(\omega t) + B\sin(\omega t)\right)\\ &= -\omega^2 x(t)\\ &= \frac{-k}{m}x(t)\end{aligned}$$

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 introducing 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 13

Integral Calculus

1 Area Under a Curve

1.1 Review

The workhorse equation of differential calculus is undoubtedly the definition of the derivative of a function $f(x)$

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} ,$$

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, all of this derivative-play, i.e. differential calculus, is only half of the total picture. There is, in fact, another important relationship between the function $f(x)$ and its slope $f'(x)$ that is the inverse to the notion of the derivative.

1.2 Motivation

Working in the general case, 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

$$\lim_{x_1 \to x_0} f(x_1) - f(x_0) = f'(x_0) \lim_{x_1 \to x_0} (x_1 - x_0) .$$

Also, consider another point x_2 that is arbitrarily close to x_1, which means

$$\lim_{x_2 \to x_1} f(x_2) - f(x_1) = f'(x_1) \lim_{x_2 \to x_1} (x_2 - x_1) ,$$

and just to start a pattern, consider yet another point x_3 obeying

$$\lim_{x_3 \to x_2} f(x_3) - f(x_2) = f'(x_2) \lim_{x_3 \to x_2} (x_3 - x_2) .$$

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:

$$\Delta x_j = x_{j+1} - x_j$$

Proceed boldly by taking the sum of the equations written above. Doing the right-hand side first, we have

$$\begin{aligned} \text{RHS} = {} & f'(x_0)\,\Delta x_0 + f'(x_1)\,\Delta x_1 \\ & + f'(x_2)\,\Delta x_2 + \cdots , \end{aligned}$$

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

$$\begin{aligned} \text{LHS} = {} & \lim_{x_1 \to x_0} f(x_1) - f(x_0) \\ & + \lim_{x_2 \to x_1} f(x_2) - f(x_1) \\ & + \lim_{x_3 \to x_2} f(x_3) - f(x_2) \\ & + \cdots + \lim_{x_n \to x_{n-1}} f(x_n) - f(x_{n-1}) \\ = {} & f(x_n) - f(x_0) \end{aligned}$$

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

$$f(x_n) - f(x_0) \approx \sum_{j=0}^{n-1} f'(x_j)\,\Delta x_j . \qquad (13.1)$$

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 jth rectangle having height $f'(x_j)$ and width Δx_j.

As a whole, Equation (13.1) 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*.

1.3 Riemann Sums

At face value, Equation (13.1) 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

$$f(x_n) - f(x_0) \approx S = \sum_{j=0}^{n-1} g\left(x_j^*\right) \Delta x \,,$$

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

$$x_j^* = \begin{cases} x_j & \text{Left sum} \\ x_{j+1} & \text{Right sum} \\ \left(x_j + x_{j+1}\right)/2 & \text{Midpoint sum} \end{cases} ,$$

which sample from $g(x)$ differently. Explicitly, these mean:

$$\begin{aligned} \frac{S_{\text{Left}}}{\Delta x} &= g(x_0) + g(x_0 + \Delta x) \\ &\quad + g(x_0 + 2\Delta x) + \cdots + g(x_n - \Delta x) \\ \frac{S_{\text{Right}}}{\Delta x} &= g(x_0 + \Delta x) + g(x_0 + 2\Delta x) \\ &\quad + g(x_0 + 3\Delta x) + \cdots + g(x_n) \\ \frac{S_{\text{Mid}}}{\Delta x} &= g\left(x_0 + \frac{\Delta x}{2}\right) + g\left(x_0 + \frac{3\Delta x}{2}\right) \\ &\quad + g\left(x_0 + \frac{5\Delta x}{2}\right) + \cdots + g\left(x_n - \frac{\Delta x}{2}\right) \end{aligned}$$

Example 1

Using the midpoint sum rule with $n = 10$ bins, approximate the area under the function

$$g(x) = 5x + 2$$

in the domain

$$-2 \le x \le 3 \,.$$

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

$$S_M = \sum_{j=0}^{n-1} g\left(x_j^*\right) \Delta x = \sum_{j=0}^{9} \left(5\left(\frac{-7}{4} + \frac{j}{2}\right) + 2\right) \frac{1}{2} \,,$$

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.

Example 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

$$0 \le x \le 4 \,.$$

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

$$\sum_{k=1}^{n} k = \frac{n\,(n+1)}{2}$$
$$\sum_{k=1}^{n} k^2 = \frac{n\,(n+1)\,(2n+1)}{6} ,$$

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*:

$$\begin{aligned} S_{\text{Trap}} &= \frac{1}{2}\left(S_{\text{Left}} + S_{\text{Right}}\right) \\ &= \frac{1}{2}\sum_{j=0}^{n-1}\left(g\left(x_j\right) + g\left(x_{j+1}\right)\right)\Delta x \\ &= \frac{g\left(x_0\right) + g\left(x_n\right)}{2}\Delta x + \sum_{j=1}^{n-1} g\left(x_j\right)\Delta x \end{aligned}$$

Without summation notation, the above reads

$$\begin{aligned} \frac{S_{\text{Trap}}}{\Delta x} = \frac{g\left(x_0\right)}{2} &+ g\left(x_0 + \Delta x\right) \\ &+ g\left(x_0 + 2\Delta x\right) + \cdots + \frac{g\left(x_n\right)}{2} . \end{aligned}$$

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 $\Delta x \to 0$ and simultaneously $n \to \infty$. 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 $f'\left(x\right)$:

$$f\left(x_n\right) - f\left(x_0\right) = \lim_{n\to\infty}\sum_{j=0}^{n-1} f'\left(x_j\right)\Delta x_j .$$

2.1 Integral Notation

To update the above with cleaner notation, the summation is replaced by the 'integral', literally a giant 'S', via

$$\lim_{\Delta x\to 0}\sum \Delta x \to \int dx ,$$

which also replaces Δx with dx. The limits on the sum turn into *integration limits*, one 'lower' limit and one 'upper' limit:

$$\lim_{\Delta x\to 0}\sum_{j=0}^{n-1} \Delta x \to \int_{x_0}^{x_n} dx$$

All j-subscripts have also been dropped, as x is now understood to be a continuous variable inside the integral.

2.2 Fundamental Theorem

Using integral notation, the above is written

$$f\left(x_n\right) - f\left(x_0\right) = \int_{x_0}^{x_n} f'\left(x\right) dx .$$

While this is workable, it's customary to drop the n-subscript fdrom $f\left(x_n\right)$, and this term becomes $f\left(x\right)$. To prevent a naming conflict on the right, swap the integration variable from x to t:

$$f\left(x\right) - f\left(x_0\right) = \int_{x_0}^{x} f'\left(t\right) dt . \qquad (13.2)$$

This result is called the *fundamental theorem of calculus*, which is the full inversion of the definition of the derivative.

2.3 Role of the Antiderivative

A less tautological way to write Equation (13.2) is

$$f\left(x\right) - f\left(x_0\right) = \int_{x_0}^{x} g\left(t\right) dt ,$$

where $f'\left(t\right)$ is renamed to some given or otherwise evident function $g\left(t\right)$. The left-side function $f\left(x\right)$ is considered unknown.

In order to 'solve' the integral, $g\left(t\right)$ 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 $f(t)$:

$$f(x) - f(x_0) = \int_{x_0}^{x} \frac{d}{dt}\left(f(t)\right) dt$$

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:

$$f(x) - f(x_0) = \int_{x_0}^{x} \frac{d}{\cancel{dt}} f(t) \cancel{dt} = \int_{x_0}^{x} df(t)$$

2.4 Definite Integral

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 $f(f)$ 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:

$$\int_{x_0}^{x} g(t)\, dt = \int_{x_0}^{a} g(t)\, dt + \int_{a}^{x} g(t)\, dt$$

2.5 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:

$$0 = \int_{-x_0}^{x_0} g_{\text{odd}}(t)\, dt \tag{13.3}$$

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

$$f_{\text{odd}}(x_0) = \int_{0}^{x_0} g_{\text{even}}(t)\, dt\,. \tag{13.4}$$

2.6 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.

2.7 Indefinite Integral

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

$$\int f'(x)\,dx = f(x) + C\,, \tag{13.5}$$

which is called the *indefinite integral*.

2.8 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

$$\int g(x)\,dx = \int \frac{d}{dx}\left(f(x)\right)dx = f(x) + C\,.$$

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

$$\frac{d}{dx}\left(\int g(x)\,dx\right) = \frac{d}{dx} f(x) + \frac{\cancel{dC}}{\cancel{dx}}\,,$$

which readily reduces to $g(x) = f'(x)$, provided that:

$$\frac{d}{dx}\left(\int g(x)\,dx\right) = \int \frac{d}{dx}\left(g(x)\right)dx$$

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

$$I = \int_0^{\sqrt{\pi/2}} x\cos\left(x^2\right)dx\,.$$

Right away, note that the function being integrated can be written as a derivative

$$x\cos\left(x^2\right) = \frac{d}{dx}\left(\frac{1}{2}\sin\left(x^2\right)\right)\,,$$

so then

$$I = \int_0^{\sqrt{\pi/2}} \frac{d}{dx}\left(\frac{1}{2}\sin\left(x^2\right)\right)dx\,,$$

and then the integral and derivative operators cancel, leaving only the evaluation:

$$I = \frac{1}{2}\sin\left(x^2\right)\Big|_0^{\sqrt{\pi/2}} = \frac{1}{2}\left(1-0\right) = \frac{1}{2}$$

3.2 Exponents and Roots

Powers

Starting with the power rule for differentiation

$$\frac{d}{dx}\left(x^n\right) = nx^{n-1}\,,$$

replace $n \to n+1$ for convenience and write the same rule:

$$(n+1)\,x^n = \frac{d}{dx}\left(x^{n+1}\right)$$

From this, we can apply the integral operator to derive the rule for integrating powers and roots:

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C \tag{13.6}$$

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:

$$\int dx = x + C \tag{13.7}$$

$$\int x\,dx = \frac{1}{2}x^2 + C \tag{13.8}$$

$$\int x^2\,dx = \frac{1}{3}x^3 + C \tag{13.9}$$

$$\int \sqrt{x}\, dx = \frac{2}{3}x^{3/2} + C \tag{13.10}$$

$$\int x^{3/2}\, dx = \frac{2}{5}x^{5/2} + C \tag{13.11}$$

$$\int x^{-2}\, dx = \frac{-1}{x} + C \tag{13.12}$$

$$\int x^{-3}\, dx = \frac{-1}{2x^2} + C \tag{13.13}$$

$$\int \frac{dx}{\sqrt{x}} = 2\sqrt{x} + C \tag{13.14}$$

$$\int x^{-3/2}\, dx = \frac{-2}{\sqrt{x}} + C \tag{13.15}$$

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.

$$\frac{d}{dx}\left(\ln\left(x\right)\right) = \frac{1}{x}\,,$$

the following must hold:

$$\int \frac{1}{x}\, dx = \ln\left(x\right) + C \tag{13.16}$$

Exponential

Starting with the derivative rule for exponents

$$\frac{d}{dx}\left(n^x\right) = n^x \ln\left(n\right)\,,$$

it must follow that:

$$\int n^x\, dx = \frac{n^x}{\ln\left(n\right)} + C \tag{13.17}$$

Applied Chain Rule

Using the power rule and chain rule for derivatives, it's straightforward to derive

$$\frac{d}{dx}\sqrt{f\left(x\right)} = \frac{f'\left(x\right)}{2\sqrt{f\left(x\right)}}\,.$$

Applying the $\int dx$ operator across both sides and simplifying leads to a useful identity:

$$\int \frac{f'\left(x\right)}{2\sqrt{f\left(x\right)}}\, dx = \sqrt{f\left(x\right)} + C \tag{13.18}$$

It takes some effort to train the eye to make use of identities such as the above. Exploring one case, suppose we have

$$f\left(x\right) = 1 \pm x^2$$

with $f'\left(x\right) = \pm 2x$. Plugging all of this in and simplifying gives a two-channel result:

$$\int \frac{\pm x}{\sqrt{1 \pm x^2}}\, dx = \sqrt{1 \pm x^2} + C \tag{13.19}$$

Problem 1

Prove that the area under a parabolic segment of base b and height h is

$$A = \frac{2}{3}bh\,.$$

Problem 2

Prove that the area of the 'lens' formed between the curves

$$y_1 = x^2$$
$$y_2 = ax + b$$

is

$$A = \frac{1}{6}\left(a^2 + 4b\right)^{3/2}\,.$$

3.3 U-Substitution

The standard integral

$$\int_{x_0}^{x} f'\left(t\right) dt = f\left(x\right) - f\left(x_0\right)$$

can sometimes be made simpler by a technique called *u-substitution*, which entails choosing a function $u\left(x\right)$ 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 $u\left(x\right)$. Once the result is attained as $f\left(u\right)$, reverse-substitute to attain $f\left(x\right)$.

A pragmatic way to choose the *correct* u-substitution can be established. Consider an indefinite integral

$$I = \int f\left(x\right) g\left(x\right) dx$$

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

$$I = \int_0^{\sqrt{\pi/2}} x \cos\left(x^2\right) dx\,.$$

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

$$I = \frac{1}{2} \sin\left(x^2\right) \Big|_0^{\sqrt{\pi/2}} = \frac{1}{2}(1-0) = \frac{1}{2}\,.$$

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 (13.6) 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 3

Use u-substitution to prove Equation (13.19).

Problem 4

Use u-substitution to prove:

$$\int \frac{dx}{1+x} = \ln(1+x) + C \tag{13.20}$$

Problem 5

For a 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

$$x = u^3 + 4$$
$$dx = 3u^2\, du\,.$$

Then, the integral looks much easier in the u-domain:

$$I = 3\int \left(u^4 + 4u\right) du$$

Problem 6

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:

$$\int \frac{d}{dx}H(x)\, dx = \int V(x)\frac{d}{dx}U(x)\, dx + \int U(x)\frac{d}{dx}V(x)\, dx$$

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:

$$\int U\,dV = UV\Big| - \int V\,dU \tag{13.21}$$

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

$$U = x$$
$$dV = \cos(x)\, dx\,,$$

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:

$$\int \ln(x)\,dx = x\ln(x) - x + C \qquad (13.22)$$

Problem 7

Use u-substitution to find the integral of the shifted natural logarithm:

$$\int \ln(1+x)\,dx = (1+x)\ln(1+x) + x + C \quad (13.23)$$

3.7 Label Trick

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:

$$\int \sin(x)\,dx = -\cos(x) + C \tag{13.24}$$

$$\int \cos(x)\,dx = \sin(x) + C \tag{13.25}$$

$$\begin{aligned}\int \tan(x)\,dx &= -\int \frac{d(\cos(x))}{\cos(x)} \\ &= -\ln(\cos(x)) + C \end{aligned}\tag{13.26}$$

$$\begin{aligned}\int \cot(x)\,dx &= \int \frac{d(\sin(x))}{\sin(x)} \\ &= \ln(\sin(x)) + C \end{aligned}\tag{13.27}$$

$$\int \sec(x)\,dx = \ln(\sec(x)+\tan(x)) + C \tag{13.28}$$

$$\int \csc(x)\,dx = -\ln(\csc(x)+\cot(x)) + C \tag{13.29}$$

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:

$$\int \sec^2(x)\,dx = \tan(x) + C \tag{13.30}$$

$$\int \csc^2(x)\,dx = -\cot(x) + C \tag{13.31}$$

$$\int \tan(x)\sec(x)\,dx = \sec(x) + C \tag{13.32}$$

$$\int \cot(x)\csc(x)\,dx = -\csc(x) + C \tag{13.33}$$

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

$$I_1 + I_2 = \int \left(\sin^2(x) + \cos^2(x)\right)dx = \int dx = x\Big|\,,$$

or equivalently

$$I_1 + I_2 = x + C\,.$$

Now integrate I_1 by parts via

$$\begin{aligned} U &= \sin(x) \\ dV &= \sin(x)\,dx \end{aligned}$$

such that

$$\begin{aligned} dU &= \cos(x)\,dx \\ V &= -\cos(x)\ , \end{aligned}$$

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

$$\int \sin^2(x)\,dx = \frac{-\sin(x)\cos(x)}{2} + \frac{x}{2} + C \tag{13.34}$$

$$\int \cos^2(x)\,dx = \frac{\sin(x)\cos(x)}{2} + \frac{x}{2} + C \tag{13.35}$$

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:

$$\int \tan^2(x)\, dx = \tan(x) - x + C \qquad (13.36)$$
$$\int \sec^2(x)\, dx = \tan(x) + C$$

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

$$I_6 = \int \frac{d}{dx}\left(-\cot(x)\right) dx = -\cot(x) + C\,.$$

From the above we get the pair of answers:

$$\int \cot^2(x)\, dx = -\cot(x) - x + C \qquad (13.37)$$
$$\int \csc^2(x)\, dx = -\cot(x) + C$$

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 $du = 2xdx$. After simplifying, we get the answer:

$$\int \arctan(x)\, dx = x\arctan(x) - \frac{1}{2}\ln\left(1+x^2\right) + C \qquad (13.38)$$

The same recipe works for several other inverse trigonometric functions, namely the arccosine, arcsine, and arccotangent:

$$\int \arccos(x)\, dx = x\cos(x) - \frac{1}{2}\ln\left(1-x^2\right) + C \qquad (13.39)$$

$$\int \arcsin(x)\, dx = x\sin(x) + \frac{1}{2}\ln\left(1-x^2\right) + C \qquad (13.40)$$

$$\int \operatorname{arccot}(x)\, dx = x\operatorname{arccot}(x) + \frac{1}{2}\ln\left(1+x^2\right) + C \qquad (13.41)$$

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

$$U = \sin^{m-1}(x)$$
$$dV = \sin(x)\, dx$$

and also

$$dU = (m-1)\sin^{m-2}(x)\cos(x)\, dx$$
$$V = -\cos(x)\,.$$

From this, we have

$$I = -\sin^{m-1}(x)\cos(x)\Big| + (m-1)\int \sin^{m-2}(x)\cos^2(x)\,dx\,.$$

Next, replace $\cos^2(x)$ with $1-\sin^2(x)$ and use the label trick, giving

$$I = -\sin^{m-1}(x)\cos(x)\Big| + (m-1)\int \sin^{m-2}(x)\,dx - (m-1)\,I\,,$$

and solving for I we arrive at a *trigonometric reduction formula*:

$$\int \sin^m(x)\,dx = \frac{-1}{m}\sin^{m-1}(x)\cos(x)\Big| + \frac{m-1}{m}\int \sin^{m-2}(x)\,dx \quad (13.42)$$

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:

$$\int \cos^m(x)\,dx = \frac{1}{m}\cos^{m-1}(x)\sin(x)\Big| + \frac{m-1}{m}\int \cos^{m-2}(x)\,dx \quad (13.43)$$

$$\int \tan^m(x)\,dx = \frac{1}{m-1}\tan^{m-1}(x)\Big| - \int \tan^{m-2}(x)\,dx \quad (13.44)$$

$$\int \csc^m(x)\,dx = \frac{-1}{m-1}\csc^{m-2}(x)\cot(x)\Big| + \frac{m-2}{m-1}\int \csc^{m-2}(x)\,dx \quad (13.45)$$

$$\int \sec^m(x)\,dx = \frac{1}{m-1}\sec^{m-2}(x)\tan(x)\Big| + \frac{m-2}{m-1}\int \sec^{m-2}(x)\,dx \quad (13.46)$$

$$\int \cot^m(x)\,dx = \frac{-1}{m-1}\cot^{m-1}(x)\Big| - \int \cot^{m-2}(x)\,dx \quad (13.47)$$

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

$$\int \sin^m(x)\cos^n(x)\,dx = -\frac{1}{m+n}\sin^{m-1}(x)\cos^{n+1}(x)\Big| + \frac{m-1}{m+n}\int \sin^{m-2}(x)\cos^n(x)\,dx\,. \quad (13.48)$$

Note that this result reproduces Equation (13.42) for $n = 0$.

A different result is attained by letting $u = \cos^{n-1}(x)$:

$$\int \sin^m(x)\cos^n(x)\,dx = \frac{1}{m+n}\sin^{m+1}(x)\cos^{n-1}(x)\Big| + \frac{n-1}{m+n}\int \sin^m(x)\cos^{n-2}(x)\,dx\,. \quad (13.49)$$

Note that this result reproduces Equation (13.43) 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

$$\alpha = m\theta$$
$$\beta = n\theta$$

for non-equal integers m, n.

Next apply the integral operator $\int d\theta$ across the whole equation

$$2\int \sin(m\theta)\cos(n\theta)\,d\theta = \int \sin(m\theta+n\theta)\,d\theta + \int \sin(m\theta-n\theta)\,d\theta\,,$$

and simplify:

$$\int \sin(m\theta)\cos(n\theta)\,d\theta = \frac{-\cos((m+n)\theta)}{2(m+n)} - \frac{\cos((m-n)\theta)}{2(m-n)} + C \quad (13.50)$$

More product formula exploits lead to additional mixed-wavelength integral identities:

$$\int \cos(m\theta)\cos(n\theta)\,d\theta = \frac{\sin((m+n)\theta)}{2(m+n)} + \frac{\sin((m-n)\theta)}{2(m-n)} + C \tag{13.51}$$

$$\int \sin(m\theta)\sin(n\theta)\,d\theta = \frac{-\sin((m+n)\theta)}{2(m+n)} + \frac{\sin((m-n)\theta)}{2(m-n)} + C \tag{13.52}$$

Orthogonality

Evaluating the mixed-wavelength integral identities (13.50)-(13.52), 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

$$\int_{-\pi}^{\pi} \sin(m\theta)\cos(n\theta)\,d\theta = 0$$
$$\int_{-\pi}^{\pi} \cos(m\theta)\cos(n\theta)\,d\theta = 0$$
$$\int_{-\pi}^{\pi} \sin(m\theta)\sin(n\theta)\,d\theta = 0$$

The same results hold when the domain is changed to $[0 : 2\pi]$:

$$\int_{0}^{2\pi} \sin(m\theta)\cos(n\theta)\,d\theta = 0$$
$$\int_{0}^{2\pi} \cos(m\theta)\cos(n\theta)\,d\theta = 0$$
$$\int_{0}^{2\pi} \sin(m\theta)\sin(n\theta)\,d\theta = 0$$

When the wavelengths m, n are one and the same integer m, two results switch to nonzero

$$\int_{-\pi}^{\pi} \cos^2(m\theta)\,d\theta = \int_{0}^{2\pi} \cos^2(m\theta)\,d\theta = \pi$$
$$\int_{-\pi}^{\pi} \sin^2(m\theta)\,d\theta = \int_{0}^{2\pi} \sin^2(m\theta)\,d\theta = \pi\,,$$

and the case that mixes sine and cosine remains zero:

$$\int_{-\pi}^{\pi} \sin(m\theta)\cos(m\theta)\,d\theta = \int_{0}^{2\pi} \sin(m\theta)\cos(m\theta)\,d\theta = 0$$

3.9 Trigonometric Substitution

Each of the following integrals

$$I_1 = \int \frac{dx}{x^2\sqrt{x^2+a^2}}$$
$$I_2 = \int \frac{\sqrt{a^2-x^2}}{x^2}\,dx$$
$$I_3 = \int \frac{dx}{(x^2-a^2)^{3/2}}$$

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 8

Use the above as a starting point to prove:

$$\int \frac{dx}{x^2\sqrt{x^2+a^2}} = -\frac{\sqrt{x^2+a^2}}{a^2x} + C$$

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 9

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:

$$I_3 = \frac{1}{a^2} \int \frac{d(\sin(\theta))}{\sin^2(\theta)} d\theta$$

Problem 10

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

$$\cos(\theta) = \frac{1 - u^2}{1 + u^2}$$
$$\sin(\theta) = \frac{2u}{1 + u^2} \ ,$$

and

$$u = \frac{\sin(\theta)}{1 + \cos(\theta)}$$
$$du = \frac{1}{2}\left(1 + u^2\right) d\theta \ .$$

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 11

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

$$U = \operatorname{arcsec}(x)$$
$$dV = dx$$

such that

$$dU = \frac{dx}{x\sqrt{x^2 - 1}}$$
$$V = x \ .$$

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,

$$\int \sec(\theta)\, d\theta = \ln\left(x + \sqrt{x^2-1}\right) + C\,.$$

Finally, we have the answer:

$$\int \operatorname{arcsec}(x)\, dx = x \operatorname{arcsec}(x) \Big| - \ln\left(x + \sqrt{x^2-1}\right) + C \quad (13.53)$$

Problem 12

Do a similar exercise for the arccosecant:

$$\int \operatorname{arccsc}(x)\, dx = x \operatorname{arccsc}(x) \Big| + \ln\left(x + \sqrt{x^2-1}\right) + C \quad (13.54)$$

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 (13.34). 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 13

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.

3.10 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:

$$\int_a^b \frac{g\left(x\right)}{g\left(b+a-x\right)+g\left(x\right)} \, dx = \frac{b-a}{2} \qquad (13.55)$$

Returning to the problem on hand, the integral J can be written

$$J = \int_0^{\pi/2} \frac{\sqrt{\sin\left(\theta\right)}}{\sqrt{\sin\left(\theta\right)}+\sqrt{\sin\left(\pi/2-\theta\right)}} \, d\theta \, .$$

Comparing this to Equation (13.55), let $a = 0$ and $b = \pi/2$ and the result is half their difference:

$$J = \frac{\pi}{4}$$

3.11 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

$$\frac{d\theta}{dt} = \sqrt{\frac{2g}{L}}\sqrt{\cos\left(\theta\right)-\cos\left(\theta_0\right)} \, ,$$

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:

$$\int \frac{d\theta}{\sqrt{\cos\left(\theta\right)-\cos\left(\theta_0\right)}} = \sqrt{\frac{2g}{L}} \int dt$$

The left side needs some preparation before proceeding. The cosine terms are replaced using the half-angle formula

$$1-\cos\left(\theta\right) = 2\sin^2\left(\frac{\theta}{2}\right) \, .$$

Also define

$$\sin\left(\phi\right) = \frac{1}{a}\sin\left(\frac{\theta}{2}\right) \, ,$$

where

$$a = \sin\left(\frac{\theta_0}{2}\right) \, ,$$

implying

$$d\theta = 2a\frac{\sqrt{1-\sin^2\left(\phi\right)}}{\sqrt{1-a^2\sin^2\left(\phi\right)}} \, d\phi \, .$$

With these substitutions, the integral on hand becomes

$$\int \frac{d\phi}{\sqrt{1-a^2\sin^2\left(\phi\right)}} = \sqrt{\frac{g}{L}} \int dt \, .$$

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:

$$\sqrt{\frac{g}{L}} \int_0^T dt = \int_0^{2\pi} \frac{d\phi}{\sqrt{1-a^2\sin^2\left(\phi\right)}} \, .$$

On the right, use the Taylor expansion of the radical to write

$$\frac{1}{\sqrt{1-a^2\sin^2\left(\phi\right)}} \approx 1 + \frac{1}{2}a^2\sin^2\left(\phi\right) + \frac{3}{8}a^4\sin^4\left(\phi\right) + \cdots \, ,$$

which only works when $a\sin\left(\phi\right)$ 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

$$\int_0^{2\pi} \sin^2\left(\phi\right) d\phi = \pi$$
$$\int_0^{2\pi} \sin^4\left(\phi\right) d\phi = \frac{3\pi}{4} \, ,$$

attainable by elementary means or using a trigonometric reduction formula.

The integral for the period reduces to

$$T \approx 2\pi\sqrt{\frac{L}{g}}\left(1+\frac{a^2}{4}+\frac{9a^4}{64}+\cdots\right) \, .$$

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.

Shifted Natural Logarithm

Starting with Equation (13.20), namely

$$\ln(1+x) + C = \int \frac{dx}{1+x} ,$$

consider the scenario $|x| < 1$.

In this case, the fraction $1/(1+x)$ can be replaced via the geometric series:

$$\ln(1+x) + C = \int \left(1 - x + x^2 - x^3 + \cdots\right) dx$$

The whole right side can be integrated quite easily:

$$\ln(1+x) + C = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

The integration constant is zero by construction, and we end up with an infinite series for the shifted natural logarithm:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

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 $\ln(2)$:

$$\ln(2) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \tag{13.56}$$

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:

$$\frac{d}{dx}\arctan(x) = 1 - x^2 + x^4 - x^6 + \cdots$$

Apply the $\int dx$ operator to both sides and simplify, to get, for $|x| < 1$,

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots .$$

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 $x = 1$,

$$\frac{\pi}{4} \approx 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots , \tag{13.57}$$

called the *Leibniz formula*.

It's important to note that Equations (13.56), (13.57) 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 $x \to 1$ in a formal limit, and making sure divergence does not occur.

Sine of X Squared

Innocent as it appears, the indefinite integral

$$I = \int \sin\left(x^2\right) dx$$

has no elementary solution. To make headway, replace the sine function with its exact polynomial representation, namely

$$\sin\left(x^2\right) = x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \cdots .$$

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:

$$\int \sin\left(x^2\right) dx = \frac{x^3}{3} - \frac{x^7}{42} + \frac{x^{11}}{1320} - \cdots \tag{13.58}$$

3.12 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

$$n! \approx \left(\frac{n}{e}\right)^n . \tag{13.59}$$

Derivation

To begin, write $n!$ in open form, namely

$$n! = n\,(n-1)\,(n-2)\,(n-3)\cdots(2)\,(1)\ ,$$

then take the natural log of both sides to write

$$\ln(n!) = \ln(n) + \ln(n-1) + \ln(n-2) + \cdots\ ,$$

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 (13.59) emerges.

Strange Product

Let us simplify the quantity

$$A = \lim_{n\to\infty} \left(\frac{(n+1)\,(n+2)\cdots(3n)}{n^{2n}} \right)^{1/n}$$

as far as possible.

One way to proceed is to take the natural log of both sides and simplify:

$$\ln(A) = \lim_{n\to\infty} \frac{\ln(n+1) + \cdots + \ln(3n) - 2n\ln(n)}{n}$$

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:

$$\ln(A) = \lim_{n\to\infty} \frac{1}{n}\cdot \left(\ln\left(1+\frac{1}{n}\right) + \ln\left(1+\frac{2}{n}\right) + \cdots + \ln\left(1+\frac{2n}{n}\right) \right)$$

Simplifying, this is

$$\ln(A) = \lim_{n\to\infty} \sum_{j=1}^{2n} \ln\left(1+\frac{j}{n}\right)\frac{1}{n}\ .$$

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

$$A = \lim_{n\to\infty} \left(\frac{1}{n^{2n}} \frac{(3n)!}{n!} \right)^{1/n}\ ,$$

and then Equation (13.59) tells us

$$A \approx \lim_{n\to\infty} \left(\frac{1}{n^{2n}} \frac{(3n)^{3n}}{e^{3n}} \frac{e^n}{n^n} \right)^{1/n}\ ,$$

reducing to $A \approx 3^3/e^2$, as expected. All n-dependence cancels out.

Strange Function

Consider the function

$$A(x) = \lim_{n\to\infty} \left(\frac{1}{n^{xn}} \frac{((x+1)\,n)!}{n!} \right)^{1/n} ,$$

where $x = 2$ reproduces the previous product.

Following similar steps, it's straightforward to show that

$$\ln(A(x)) = \lim_{n\to\infty} \sum_{j=1}^{xn} \ln\left(1 + \frac{j}{n}\right) \frac{1}{n} ,$$

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:

$$S = \int dS = \int_{x_0}^{x_1} \sqrt{1 + (f'(x))^2}\, dx \qquad (13.60)$$

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 14

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 15

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 16

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 $x_0 \le x \le x_1$.

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:

$$V = \int_{x_0}^{x_1} \pi (f(x))^2\, dx \qquad (13.61)$$

Problem 17

Show that a cone of height H and base radius R has volume

$$V = \frac{1}{3} \pi R^2 H .$$

Problem 18

Use elementary methods to show that a *cone frustum* of height H with end radii R_1, R_2 has volume

$$V = \frac{1}{3}\pi \left(R_1^2 + R_1 R_2 + R_2^2\right) H \, .$$

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 19

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

$$\begin{aligned} x(z) &= z\, l/h \\ y(z) &= y\, w/h \, . \end{aligned}$$

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:

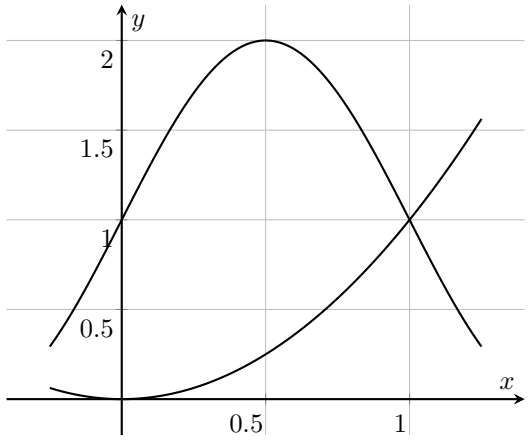
$$V = \int_{x_0}^{x_1} \pi \left((f(x))^2 - (g(x))^2 \right) dx \qquad (13.62)$$

Problem 20

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

$$\begin{aligned} dV_{\text{shell}} = \pi (x + dx)^2 f(x + dx) \\ - \pi (x)^2 f(x) \, , \end{aligned}$$

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:

$$V = \int dV_{\text{shell}} = \int_{x_0}^{x_1} 2\pi x f(x)\, dx \tag{13.63}$$

Problem 21

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 22

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 $y = f(x)$:

$$A = \int_{x_0}^{x_1} 2\pi y\sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \tag{13.64}$$

Gabriel's Horn

Consider the hyperbola

$$y = \frac{1}{x}$$

in the domain

$$1 \le x < \infty .$$

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' was known to seventeenth-century mathematicians, not excluding Hobbes, Wallis, and Galileo, originally brought to public attention by Torricelli.

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

$$\frac{dV}{dx} = \pi\left(\frac{1}{x}\right)^2$$

$$\frac{dA}{dx} = 2\pi\left(\frac{1}{x}\right)\sqrt{1 + \frac{1}{x^4}} ,$$

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 $x \to \infty$, which means the area outpaces the volume in the long run.

4.4 Centroid

...

Problem 23

Show that the centroid of a parabolic segment of height h is $\bar{y} = 2h/5$.

Problem 24

Show that the centroid of a half-ellipse of base $2a$ and vertex height b is $\bar{y} = 4b/3\pi$.

4.5 The Cycloid

Definition

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 13.1.

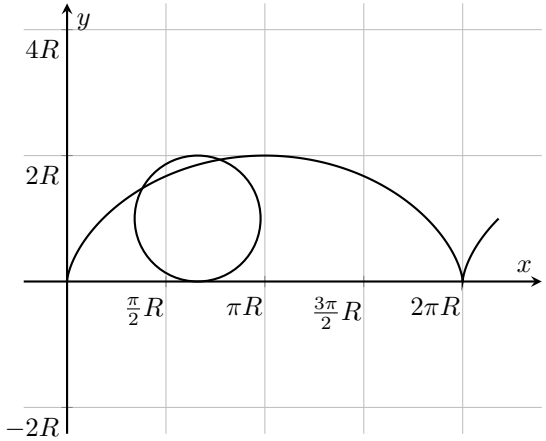


Figure 13.1: The cycloid with generating circle.

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

$$x(\theta) = R\theta - R\sin(\theta) \tag{13.65}$$
$$y(\theta) = R - R\cos(\theta) \,. \tag{13.66}$$

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 a equation for the tangent line to the cycloid:

$$y_{\tan} = y_0 + \cot\left(\frac{\theta}{2}\right)(x - x_0)$$

Replacing x_0, y_0 with their representations in θ gives a neater formula, after some simplifying:

$$y_{\tan} = 2R + \cot\left(\frac{\theta}{2}\right)(x - R\theta)$$

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 25

A stone lodged on the rim of a bicycle tire of radius R dislodges at the height of its cycloidic 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, wee see that the normal line always hits the point of contact between the generating the circle and the line on which it rolls.

Arc Length

The arc length of the cycloid is straightforwardly calculated from Equation (13.60). 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 (13.61). 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 (13.64). 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

$$\begin{aligned} y_0 - y &= R\,(\cos\,(\theta_0) - \cos\,(\theta)) \\ dx &= R\,(1 - \cos\,(\theta))\,d\theta \\ dy &= -R\sin\,(\theta)\,d\theta \,, \end{aligned}$$

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*.

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 $f^{(q)}\,(x_0)$:

$$p\,(x) = f\,(x_0) + \sum_{q=1}^{n} \frac{1}{q!} f^{(q)}\,(x_0)\,(x - x_0)^j + R_n\,(x)$$

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 (13.2), and isolate $f\,(x)$:

$$f\,(x) = f\,(x_0) + \int_{x_0}^{x} f^{(1)}\,(t)\,dt$$

Of course, the function $f^{(1)}\,(t)$ could itself be approximated to first order using the fundamental theorem

$$f^{(1)}\,(t) = f^{(1)}\,(x_0) + \int_{x_0}^{t} f^{(2)}\,(u)\,du \,,$$

which begs substitution into the above, giving:

$$f\,(x) = f\,(x_0) + \int_{x_0}^{x} \left(f^{(1)}\,(x_0) + \int_{x_0}^{t} f^{(2)}\,(u)\,du \right) dt$$

After simplifying, we see the familiar first-order Taylor series term trailed by a messy integral:

$$\begin{aligned} f\,(x) = f\,(x_0) &+ f^{(1)}\,(x_0)\,(x - x_0) \\ &+ \int_{x_0}^{x} \left(\int_{x_0}^{t} f^{(2)}\,(u)\,du \right) dt \end{aligned}$$

Trudging forward, take $f^{(2)}\,(u)$ to first order

$$f^{(2)}\,(u) = f^{(2)}\,(x_0) + \int_{x_0}^{w} f^{(3)}\,(w)\,dw \,,$$

and substitute into the preceding integral. This first means having to solve

$$I = \int_{x_0}^{x} \left(\int_{x_0}^{t} f^{(2)}(x_0)\, du \right) dt\,.$$

Knowing $f^{(2)}(x_0)$ is constant, proceed using brute force to find

$$\begin{aligned} I &= f^{(2)}(x_0) \int_{x_0}^{x} \left(\int_{x_0}^{t} du \right) dt \\ &= f^{(2)}(x_0) \int_{x_0}^{x} (t - x_0)\, dt \\ &= f^{(2)}(x_0) \left(\frac{t^2}{2} - x_0 t \right) \Bigg|_{x_0}^{x} \\ &= \frac{1}{2} f^{(2)}(x_0)\, (x - x_0)^2\ . \end{aligned}$$

Interestingly, this is the second-order term in the Taylor series of $f(x)$. To summarize:

$$\begin{aligned} f(x) = f(x_0) &+ f^{(1)}(x_0)\,(x - x_0) \\ &+ \frac{1}{2} f^{(2)}(x_0)\,(x - x_0)^2 \\ &+ \int_{x_0}^{x} \left(\int_{x_0}^{t} \left(\int_{x_0}^{u} f^{(3)}(w)\, dw \right) du \right) dt \end{aligned}$$

Repeating the steps that got us this far, use the first-order approximation of $f^{(3)}(w)$. The obligatory integral to solve is

$$J = f^{(3)}(x_0) \int_{x_0}^{x} \left(\int_{x_0}^{t} \left(\int_{x_0}^{u} dw \right) du \right) dt\,,$$

which after a bit of grinding, comes out to

$$J = \frac{1}{3!} f^{(3)}(x_0)\, (x - x_0)^3\ .$$

By now we're seeing a pattern, particularly:

$$\begin{aligned} f(x) = f(x_0) &+ f^{(1)}(x_0)\,(x - x_0) \\ &+ \frac{1}{2} f^{(2)}(x_0)\,(x - x_0)^2 \\ &+ \frac{1}{3!} f^{(3)}(x_0)\,(x - x_0)^3 \\ &+ R_3(x)\ , \end{aligned}$$

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 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 $L_0 < L/2$.

6.1 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:

$$\begin{aligned} m\frac{d^2y}{dt^2} &= F^y_{\text{net}} = 0 \\ m\frac{d^2x}{dt^2} &= F^x_{\text{net}} = F_a + F_b\,, \end{aligned}$$

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.)

6.2 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

$$x_a = -L_0 + L/2 - q$$
$$x_b = -L_0 + L/2 + q\,,$$

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

$$m\frac{d^2}{dt^2}x\left(t\right) = -k_a\left(x_a + x\left(t\right)\right) + k_b\left(x_b - x\left(t\right)\right)\,,$$

readily simplifying to

$$m\frac{d^2}{dt^2}x\left(t\right) = -x\left(t\right)\left(k_A + k_b\right)$$

This is a simple harmonic oscillator with effective angular frequency:

$$\omega = \sqrt{\frac{k_a + k_b}{m}}$$

6.3 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 $F^x_{\text{net}} = 0$ for the x-direction, and for the y-direction,

$$F^y_{\text{net}} = 2F_{\text{spring}}\sin\left(\theta\right)\,,$$

where θ is the angle formed between a spring and the horizontal, and from geometry we pick out

$$\sin\left(\theta\right) = \frac{y}{\sqrt{\left(L/2\right)^2 + y^2}}\,.$$

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

$$F^y_{\text{net}} = m\frac{d^2}{dt^2}y\left(t\right) = -2ky\left(1 - \frac{L_0}{\sqrt{\left(L/2\right)^2 + y^2}}\right)\,.$$

Small Vibrations

In the special case that the displacement $|y|$ is always much less than $L/2$, the above becomes

$$F^y_{\text{net}} \approx -2ky\left(1 - \frac{2L_0}{L}\left(1 - \frac{1}{2}\frac{4y^2}{L^2}\right)\right)$$
$$\approx -2ky\left(1 - \frac{2L_0}{L}\right)\,,$$

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

$$m\frac{d^2}{dt^2}y\left(t\right) \approx -\left(\frac{2k}{1 + L_0/p}\right)y\left(t\right)\,.$$

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.

6.4 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:

$$F^y_{\text{net}} \approx -2ky\left(1 - \frac{2l_0}{L}\left(1 - \frac{1}{2}\frac{4y^2}{L^2}\right)\right)$$

Setting $2L_0 = L$, the above simplifies to

$$m\frac{d^2}{dt^2}y\left(t\right) \approx -k\left(\frac{2}{L}\right)^2\left(y\left(t\right)\right)^3\,,$$

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:

$$\frac{1}{2}\frac{d}{dt}\left(\dot{y}^2\right) = \ddot{y}\dot{y} = -\lambda\,\frac{dy}{dt}y^3$$

Multiply dt onto each side to attain a so-called 'differential form'

$$\frac{1}{2}\frac{d}{dt}\left(\dot{y}^2\right)dt = -\lambda\, y^3\, dy \,,$$

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

$$\frac{1}{2}\dot{y}^2 = \frac{\lambda}{4}\left(A^4 - y^4\right) = \frac{\lambda A^4}{4}\left(1 - \left(\frac{y}{A}\right)^4\right) \,,$$

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

$$\begin{aligned} y &= A\cos(\phi) \\ dy &= -A\sin(\phi)\, d\phi \,, \end{aligned}$$

and the above becomes

$$\frac{-A\cancel{\sin(\phi)}\, d\phi}{\cancel{\sqrt{\left(1 - \cos^2(\phi)\right)}}\left(1 + \cos^2(\phi)\right)} = \pm\left(\sqrt{\lambda}\frac{A^2}{2}\right)dt \,,$$

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, and the general form of the above doesn't change when different limits are chosen. Condensing constants and separating out period's relation to the amplitude gives a nifty result:

$$AT = \text{constant}$$

Chapter 14

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 $\vec{m}$ that represents the slope of a straight line, along with a vector $\vec{b} = \langle 0, b \rangle$ as the y-intercept, the whole line can be represented by the vector

$$\vec{r}(\alpha) = \alpha\,\vec{m} + \vec{b}\,,$$

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:

$$x = x(r, \theta) = r\cos(\theta)$$
$$y = y(r, \theta) = r\sin(\theta)$$

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:

$$x = x(t) = r(t)\cos(\theta(t))$$
$$y = y(t) = r(t)\sin(\theta(t))$$

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

$$\cot(\psi) = \frac{r'(\theta)}{r(\theta)} = \frac{d}{d\theta}\left(\ln(r(\theta))\right)\,,$$

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

$$\begin{aligned} x &= g(t) \\ y &= h(t)\ , \end{aligned}$$

wher $y(x)$ 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

$$\Delta A = \frac{1}{2}(r+\Delta r)\,r\sin(\Delta\theta)\ .$$

In the differential limit, we may approximate

$$\begin{aligned} r + \Delta r &\approx r \\ \sin(\Delta\theta) &\approx d\theta\,, \end{aligned}$$

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

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta)\ . \end{aligned}$$

Also consider a neighboring point $(x + dx, y + dy)$, with

$$\begin{aligned} x + dx &= (r + dr)\cos(\theta + d\theta) \\ y + dy &= (r + dr)\sin(\theta + d\theta)\ . \end{aligned}$$

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:

$$\begin{aligned} x + dx &= (r + dr)\left(\cos(\theta)\cos(d\theta) - \sin(\theta)\sin(d\theta)\right) \\ y + dy &= (r + dr)\left(\sin(\theta)\cos(d\theta) + \cos(\theta)\sin(d\theta)\right) \end{aligned}$$

Of course, we can take

$$\begin{aligned} \cos(d\theta) &\approx 1 \\ \sin(d\theta) &\approx d\theta \\ dr \cdot d\theta &\approx 0 \end{aligned}$$

because differential quantities are vanishing, and the above simplifies to

$$\begin{aligned} dx &= -y\, d\theta + \cos(\theta)\, dr \\ dy &= x\, d\theta + \sin(\theta)\, dr\ . \end{aligned}$$

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,

$$dS = \sqrt{r^2 + \left(r'(\theta)\right)^2}\, d\theta\ .$$

This is enough to write the formula for the arc length of a polar function $r(\theta)$ by integrating in the θ-variable:

$$S = \int_{\theta_0}^{\theta_1} \sqrt{r^2 + \left(r'(\theta)\right)^2}\, d\theta$$

4 Position and Basis Vectors

4.1 Cartesian Position Vector

In the Cartesian plane, a point (x, y) can be represented as a *position vector*

$$\vec{R} = \langle x, y\rangle\ .$$

The above is written in explicit 'bracket' notation, which reminds us that the vector $\vec{R}$ is made by going 'over by x', and then 'up by y'.

4.2 Cartesian Basis Vectors

There is an equivalent representation using basis vectors, which goes

$$\vec{R} = x\,\hat{x} + y\,\hat{y}\ ,$$

where $\hat{x}$, $\hat{y}$ are mutually-perpendicular vectors of magnitude 1. The Cartesian basis vectors have a few equivalent representations:

$$\begin{aligned} \hat{x} = \hat{\imath} = \hat{e}_x = \hat{e}_i &= \langle 1, 0\rangle \\ \hat{y} = \hat{\jmath} = \hat{e}_y = \hat{e}_j &= \langle 0, 1\rangle \end{aligned}$$

It's worth noting that the Cartesian basis vectors are fixed in their coordinate system. The vectors $\hat{x}$, $\hat{y}$ never change, and all derivatives of these are patently zero.

4.3 Polar Position Vector

Starting from the Cartesian position vector

$$\vec{R} = x\,\hat{x} + y\,\hat{y}\ ,$$

replace x, y with their equivalent representations in polar coordinates:

$$\vec{R} = r\cos(\theta)\,\hat{x} + r\sin(\theta)\,\hat{y}$$

From the above we can factor the radial term r, leaving a tangle of basis vectors and trig terms:

$$\vec{R} = r\left(\cos(\theta)\,\hat{x} + \sin(\theta)\,\hat{y}\right)$$

Meanwhile, notice that the magnitude of

$$R = \sqrt{\vec{R} \cdot \vec{R}}$$

reduces to $R = r$. This simultaneously means that $\vec{R}$'s unit vector, i.e.

$$\hat{R} = \frac{\vec{R}}{R}$$

is equal to

$$\hat{R} = \cos(\theta)\,\hat{x} + \sin(\theta)\,\hat{y}\,.$$

As a matter of custom, the polar position vector is expressed in lowercase, i.e. $\vec{r}$. This is also true for the polar basis vector $\hat{r}$. The tightest way to write the position vector in polar coordinates is:

$$\vec{r} = r\,\hat{r}$$

4.4 Polar Basis Vectors

Starkly different from the fixed vectors $\hat{x}$, $\hat{y}$, the polar basis vector $\hat{r}$ is not fixed in the plane, and instead varies with θ via

$$\hat{r} = \cos(\theta)\,\hat{x} + \sin(\theta)\,\hat{y}\,.$$

In the same way that $\hat{r}$ points in the direction of increasing r, there ought to exist a second basis vector $\hat{\theta}$, always perpendicular to $\hat{r}$, that 'points' in the direction of increasing θ. This can be attained by rotating $\hat{r}$ by ninety degrees

$$\hat{\theta} = \cos\left(\theta + \frac{\pi}{2}\right)\hat{x} + \sin\left(\theta + \frac{\pi}{2}\right)\hat{y}\,,$$

simplifying to

$$\hat{\theta} = -\sin(\theta)\,\hat{x} + \cos(\theta)\,\hat{y}\,.$$

Notice that the θ-basis vector is not needed to write the position $\vec{r}$.

Matrix Representation

A tight way to express the polar basis vectors $\hat{r}$, $\hat{\theta}$ uses matrix notation:

$$\begin{bmatrix}\hat{r}\\ \hat{\theta}\end{bmatrix} = \begin{bmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}$$

Resolving Cartesian Basis

Given how $\hat{r}$, $\hat{\theta}$ depend on $\hat{x}$, $\hat{y}$, θ, it's useful to solve for $\hat{x}$, $\hat{y}$ in terms of $\hat{r}$, $\hat{\theta}$, θ instead. As a standard 2×2 matrix, the inverse is easy to write down:

$$\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix} = \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}\begin{bmatrix}\hat{r}\\ \hat{\theta}\end{bmatrix}$$

Explicitly, this means:

$$\hat{x} = \cos(\theta)\,\hat{r} - \sin(\theta)\,\hat{\theta}$$
$$\hat{y} = \sin(\theta)\,\hat{r} + \cos(\theta)\,\hat{\theta}$$

5 Intersections

5.1 Line Segments Intersecting

In the Cartesian plane, consider a line segment with endpoints located at $\vec{p}_A$, $\vec{p}_B$. Also consider a second line segment with endpoints located at $\vec{q}_A$, $\vec{q}_B$. 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 parameteized by a vector equation

$$\vec{y}_j = \vec{b}_j + \alpha_j \hat{t}_j\,,$$

where $j = p, q$ toggles between segments. The vector $\vec{b}_j$ is any fixed point on the jth line segment (not necessarily the y-intercept). The unit vector $\hat{t}_j$ registers the slope of the line segment. The parameter α_j is a real number confined to the domain

$$\alpha_j^{\min} \le \alpha \le \alpha_j^{\max}$$

to assure the finite extent of each segment.

Intersection

The condition for intersection is given by $\vec{y}_p = \vec{y}_q$, or

$$\vec{b}_p + \alpha_p \hat{t}_p = \vec{b}_q + \alpha_q \hat{t}_q\,.$$

Letting

$$\Delta\vec{b} = \vec{b}_p - \vec{b}_q$$
$$f = \hat{t}_p \cdot \hat{t}_q$$
$$w_p = \Delta\vec{b} \cdot \hat{t}_p$$
$$w_q = \Delta\vec{b} \cdot \hat{t}_q\,,$$

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 $f = \hat{t} \cdot \hat{q} = \pm 1$, 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.

5.2 Line Intersecting Ellipse

Consider an ellipse located somewhere in the plane parameterized by ϕ such that

$$\vec{r}_e = \vec{r}_0 + \vec{a}\cos(\phi) + \vec{b}\sin(\phi) \ ,$$

where $\vec{r}_0$ is the center, and the mutually-perpendicular vectors $\vec{a}$, $\vec{b}$ orient the major and minor axes.

Also consider a straight line in the plane parameterized by α:

$$\vec{y} = \vec{y}_0 + \alpha\, \hat{v} \ ,$$

where $\vec{y}_0$ is a constant vector, and $\hat{v}$ is the unit tangent vector that tells us the slope of the line.

The case for intersections is given by:

$$\vec{r}_0 + \vec{a}\cos(\phi) + \vec{b}\sin(\phi) = \vec{y}_0 + \alpha\, \hat{v}$$

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 $\vec{a}$, and again by $\vec{b}$, and then isolate the trig terms:

$$\cos(\phi) = \frac{\vec{a} \cdot (\vec{y}_0 - \vec{r}_0) + \alpha\, \vec{a} \cdot \hat{v}}{a^2}$$
$$\sin(\phi) = \frac{\vec{b} \cdot (\vec{y}_0 - \vec{r}_0) + \alpha\, \vec{b} \cdot \hat{v}}{b^2}$$

Let

$$\vec{q}_0 = \vec{y}_0 - \vec{r}_0 \ ,$$

and simplify further:

$$\cos(\phi) = \frac{\hat{a} \cdot \vec{q}_0 + \alpha\, \hat{a} \cdot \hat{v}}{a}$$
$$\sin(\phi) = \frac{\hat{b} \cdot \vec{q}_0 + \alpha\, \hat{b} \cdot \hat{v}}{b}$$

To proceed, use the fundamental trig identity to write

$$1 = \left(\frac{\hat{a} \cdot \vec{q}_0 + \alpha\, \hat{a} \cdot \hat{v}}{a}\right)^2 + \left(\frac{\hat{b} \cdot \vec{q}_0 + \alpha\, \hat{b} \cdot \hat{v}}{b}\right)^2 \ .$$

Blooming out the algebra, the above takes the form

$$0 = A\alpha^2 + B\alpha + C \ ,$$

where:

$$A = \left(\frac{\hat{a} \cdot \hat{v}}{a}\right)^2 + \left(\frac{\hat{b} \cdot \hat{v}}{b}\right)^2$$
$$B = \frac{2\,(\hat{a} \cdot \vec{q}_0)\,(\hat{a} \cdot \hat{v})}{a^2} + \frac{2\left(\hat{b} \cdot \vec{q}_0\right)\left(\hat{b} \cdot \hat{v}\right)}{b^2}$$
$$C = \left(\frac{\hat{a} \cdot \vec{q}_0}{a}\right)^2 + \left(\frac{\hat{b} \cdot \vec{q}_0}{b}\right)^2 - 1$$

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.

5.3 Circle Intersecting Circle

Let the position $\vec{r}_j$ on a circle of radius R_j centered at $\vec{C}_j$ be parameterized by ϕ_j:

$$\vec{r}_j = \vec{C}_j + R_j \left\langle \cos(\phi_j), \sin(\phi_j) \right\rangle$$

For the intersection of two such circles of different radii, we can write

$$\vec{D} = R_2 \left\langle \cos(\phi_2), \sin(\phi_2) \right\rangle - R_1 \left\langle \cos(\phi_1), \sin(\phi_1) \right\rangle \ ,$$

where

$$\vec{D} = \vec{C}_1 - \vec{C}_2 \ .$$

In component form, vector $\vec{D}$ reads

$$D_x = R_2 \cos(\phi_2) - R_1 \cos(\phi_1)$$
$$D_y = R_2 \sin(\phi_2) - R_1 \sin(\phi_1) \ .$$

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:

$$\frac{D_x}{E}\cos(\phi_2) + \frac{D_y}{E}\sin(\phi_2) = 1$$
$$\frac{D_x}{F}\cos(\phi_1) + \frac{D_y}{F}\sin(\phi_1) = 1$$

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:

$$\left(\frac{D}{E}\right)^2 \cos^2(\phi_2) - \frac{2D_x}{E}\cos(\phi_2) + 1 - \left(\frac{D_y}{E}\right)^2 = 0$$

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

$$1 = \frac{D_x^2 \pm D_y D_x\sqrt{D^2/E^2 - 1}}{D^2} + \frac{D_y^2 \pm D_x D_y\sqrt{D^2/E^2 - 1}}{D^2}$$

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

$$\cos\left(\phi_2^+\right) = \frac{D_x E + D_y E\sqrt{D^2/E^2 - 1}}{D^2}$$
$$\sin\left(\phi_2^-\right) = \frac{D_y E - D_x E\sqrt{D^2/E^2 - 1}}{D^2}\,,$$

and the other pair of solutions has the signs outside the roots swapped. The intersection points of the two circles are finally:

$$\vec{X}^1_{\rm int} = \vec{C}_2 + R_2\left\langle \cos\left(\phi_2^+\right), \sin\left(\phi_2^-\right)\right\rangle$$
$$\vec{X}^2_{\rm int} = \vec{C}_2 + R_2\left\langle \cos\left(\phi_2^-\right), \sin\left(\phi_2^+\right)\right\rangle$$

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.

5.4 Circle Intersecting Points

Find the circle

$$(x-h)^2 + (y-k)^2 = R^2$$

that passes through three points in the plane $\vec{q}_j = \langle x_j, y_j\rangle$ with $j = 1, 2, 3$.

Proceed by blooming out the equation for the circle and substitute each data point:

$$q_1^2 - 2x_1 h - 2y_1 k = R^2 - h^2 - k^2$$
$$q_2^2 - 2x_2 h - 2y_2 k = R^2 - h^2 - k^2$$
$$q_3^2 - 2x_3 h - 2y_3 k = R^2 - h^2 - k^2$$

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

$$a_1 + b_1 h + c_1 k = 0$$
$$a_2 + b_2 h + c_2 k = 0 \, .$$

Solving for h, k is a matter of elementary (linear) algebra. For results, we finally have:

$$h = \frac{a_1/b_1 - a_2/b_2}{c_2/b_2 - c_1/b_1}$$
$$k = \frac{a_1/c_1 - a_2/c_2}{b_2/c_2 - b_1/c_1}$$

To solve for the radius R, consider any vector $\vec{P}_j$ that extends from the center of the circle to any given point $\vec{q}_j$:

$$\vec{P}_j = \vec{q}_j - \langle h, k \rangle$$

The radius is the magnitude P.

5.5 Ellipse Intersecting Ellipse

Consider two ellipses in the same plane given by:

$$\vec{r}_j = \vec{k}_j + \vec{a}_j \cos(\phi_j) + \vec{b}_j \sin(\phi_j) \ ,$$

where $j = 1, 2$, and, as usual for an ellipse $\vec{a}_j \cdot \vec{b}_j = 0$. The condition for intersection of the two ellipses is

$$\vec{k}_1 + \vec{a}_1 \cos(\phi_1) + \vec{b}_1 \sin(\phi_1) = \vec{k}_2 + \vec{a}_2 \cos(\phi_2) + \vec{b}_2 \sin(\phi_1) \ .$$

Then, multiply $\vec{a}_1$, $\vec{b}_1$, separately into the above to generate two results:

$$\vec{a}_1 \cdot \vec{k}_1 + a_1^2 \cos(\phi_1) = \vec{a}_1 \cdot \vec{k}_2 + \vec{a}_1 \cdot \vec{a}_2 \cos(\phi_2) + \vec{a}_1 \cdot \vec{b}_2 \sin(\phi_1)$$
$$\vec{b}_1 \cdot \vec{k}_1 + b_1^2 \sin(\phi_2) = \vec{b}_1 \cdot \vec{k}_2 + \vec{b}_1 \cdot \vec{a}_2 \cos(\phi_2) + \vec{b}_1 \cdot \vec{b}_2 \sin(\phi_2)$$

Denoting

$$\Delta\vec{k} = \vec{k}_2 - \vec{k}_1$$
$$A_{jk} = \vec{a}_j \cdot \vec{a}_k$$
$$B_{jk} = \vec{b}_j \cdot \vec{b}_k$$
$$C_{jk} = \vec{a}_j \cdot \vec{b}_k$$

to keep the algebra tame, the above rearrange to

$$\cos(\phi_1) = \frac{\vec{a}_1 \cdot \Delta\vec{k} + A_{12} \cos(\phi_2) + C_{12} \sin(\phi_2)}{a_1^2}$$
$$\sin(\phi_1) = \frac{\vec{b}_1 \cdot \Delta\vec{k} + C_{21} \cos(\phi_2) + B_{12} \sin(\phi_2)}{b_1^2} \, .$$

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.

For a special case, consider the two ellipses

$$\vec{r}_1 = a \cos(\phi_1) \, \hat{x} + b \sin(\phi_1) \, \hat{y}$$
$$\vec{r}_2 = b \cos(\phi_2) \, \hat{x} + a \sin(\phi_2) \, \hat{y} \, .$$

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

$$\vec{r}_1 = \vec{r}_2 = \frac{ab}{\sqrt{a^2 + b^2}} \left(\pm\hat{x} \pm \hat{y} \right) \ .$$

5.6 Ellipse Intersecting Parabola

A parabolic curve

$$y(x) = Ax^2 + Bx + C$$

can be represented in vector notation via

$$\vec{y} = x \, \hat{x} + y(x) \, \hat{y} \, .$$

Let us find the intersection between such a parabola and the ellipse:

$$\vec{r} = \vec{r}_0 + \vec{a} \cos(\phi) + \vec{b} \sin(\phi)$$

The condition for intersection is $\vec{y} = \vec{r}$, or

$$x \, \hat{x} + y(x) \, \hat{y} = \vec{r}_0 + \vec{a} \cos(\phi) + \vec{b} \sin(\phi) \ .$$

Next, multiply $\vec{a}$, $\vec{b}$, separately into the above to generate two results,

$$x \, \vec{a} \cdot \hat{x} + y(x) \, \vec{a} \cdot \hat{y} = \vec{a} \cdot \vec{r}_0 + a^2 \cos(\phi)$$
$$x \, \vec{b} \cdot \hat{x} + y(x) \, \vec{b} \cdot \hat{y} = \vec{b} \cdot \vec{r}_0 + b^2 \sin(\phi) \ ,$$

or:

$$\cos(\phi) = \frac{x a_x + y(x) \, a_y - \vec{a} \cdot \vec{r}_0}{a^2}$$
$$\sin(\phi) = \frac{x b_x + y(x) \, b_y - \vec{b} \cdot \vec{r}_0}{b^2}$$

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 $\vec{r}_0 = 0$, and let $\vec{a} = a\,\hat{x}$, $\vec{b} = b\,\hat{y}$. Then, we have

$$1 = \left(\frac{x}{a}\right)^2 + \left(\frac{y\,(x)}{b}\right)^2 ,$$

or all in terms of one variable,

$$1 = \left(\frac{x}{a}\right)^2 + \left(\frac{Ax^2 + Bx + C}{b}\right)^2 .$$

5.7 Collision of Spheres

Consider two spheres of radius $R_{1,2}$ and mass $m_{1,2}$, each moving with uniform velocity $\vec{v}_{1,2}$. With this setup, the system has a total energy E (scalar) and linear momentum $\vec{P}$ (vector):

$$E = \frac{1}{2}m_1 v_1^2 + \frac{1}{2}m_2 v_2^2$$
$$\vec{P} = m_1\,\vec{v}_1 + m_2\,\vec{v}_2$$

If the spheres are to make contact via an elastic collision without exchanging mass, each sphere emerges with a new velocity vector $\vec{u}_{1,2}$ obeying

$$E = \frac{1}{2}m_1 u_1^2 + \frac{1}{2}m_2 u_2^2$$
$$\vec{P} = m_1\,\vec{u}_1 + m_2\,\vec{u}_2 ,$$

which is to say energy and momentum are conserved throughout the collision. The task is to solve for $\vec{u}_1$, $\vec{u}_2$.

To make the problem easier, we can define a momentum exchange vector $\vec{q}$ such that

$$m_1\,\vec{u}_1 = m_1\,\vec{v}_1 - \vec{q}$$
$$m_2\,\vec{u}_2 = m_2\,\vec{v}_2 + \vec{q}\,.$$

This pair of equations can recover the answers $\vec{u}_{1,2}$ from $\vec{q}$, so the whole problem becomes finding $\vec{q}$.

The two spheres exchange momentum at the point of contact, thus $\vec{q}$ is normal to each sphere's surface at that point. By the same token, $\vec{q}$ is parallel to the vector connecting the center of each sphere. Thus each vector on hand relates by:

$$\vec{q} = |q|\,\hat{n} = |q|\left(\frac{\vec{r}_1 - \vec{r}_2}{|\vec{r}_1 - \vec{r}_2|}\right)$$

Since the positions $\vec{r}_{1,2}$ are given, the direction of $\vec{q}$ is already clear, and the task reduces to finding $|q|$.

To proceed, square each momentum exchange equation to write

$$\frac{1}{2}m_1 u_1^2 = \frac{1}{2}m_1 v_1^2 - \vec{v}_1\cdot\vec{q} + \frac{q^2}{2m_1}$$
$$\frac{1}{2}m_2 u_2^2 = \frac{1}{2}m_2 v_2^2 + \vec{v}_2\cdot\vec{q} + \frac{q^2}{2m_2}$$

Take the sum of these, and notice all kinetic energy terms cancel, leaving

$$0 = \vec{q}\cdot(\vec{v}_1 - \vec{v}_2) - q^2\left(\frac{1}{m_1} + \frac{1}{m_2}\right) .$$

Using $\vec{q} = |q|\,\hat{n}$, finally solve for $|q|$:

$$|q| = \left(\frac{2m_1 m_2}{m_1 + m_2}\right)\hat{n}\cdot(\vec{v}_1 - \vec{v}_2)$$

6 Rotations

6.1 Rotated Vectors

Consider a vector

$$\vec{V} = V_x\,\hat{x} + V_y\,\hat{y}$$

whose magnitude is

$$V = \sqrt{V_x^2 + V_y^2}\,.$$

In terms of a parameter ϕ, the components of $\vec{V}$ can be written

$$V_x = V\cos(\phi)$$
$$V_y = V\sin(\phi)\ .$$

Now, suppose that the angle parameter is increased by another angle θ such that

$$\phi \to \phi + \theta\,,$$

which has the effect of modifying the vector components to new values

$$V_x' = V\cos(\phi + \theta)$$
$$V_y' = V\sin(\phi + \theta)\ ,$$

which have been denoted V_x', V_y'. Right away, one sees that the magnitude

$$V = \sqrt{(V_x')^2 + (V_y')^2}$$

still holds, meaning we aren't changing the length of the vector.

Expanding out the trig terms in the components V_x', V_y' leads us to

$$V_x' = V\left(\cos(\phi)\cos(\theta) - \sin(\phi)\sin(\theta)\right)$$
$$V_y' = V\left(\sin(\phi)\cos(\theta) + \cos(\phi)\sin(\theta)\right)\ ,$$

which simplifies further:

$$V_x' = V_x\cos(\theta) - V_y\sin(\theta)$$
$$V_y' = V_x\sin(\theta) + V_y\cos(\theta)$$

Rotation Matrix

In matrix notation, the above reads

$$\begin{bmatrix} V_x' \\ V_y' \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} V_x \\ V_y \end{bmatrix},$$

which is the same matrix needed to recover the Cartesian basis vectors from the polar ones. This is indeed the standard rotation matrix, denoted R:

$$R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} = \begin{bmatrix} \tilde{R}_{11} & \tilde{R}_{12} \\ \tilde{R}_{21} & \tilde{R}_{22} \end{bmatrix}$$

6.2 Rotated Coordinates

Consider a vector $\vec{A}$ living in the 'standard' Cartesian basis:

$$\vec{A} = A_x\,\hat{x} + A_y\,\hat{y}$$

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 $\hat{u}$, $\hat{v}$ replacing the respective $\hat{x}$, $\hat{y}$ unit vectors.

In this construction, the unit vectors $\hat{u}$, $\hat{v}$ are totally analogous to $\hat{r}$, $\hat{\theta}$, and we can borrow the matrix that brings us from $\hat{x}$, $\hat{y}$ to the rotated system:

$$\begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}$$

In terms of the rotated basis vectors, the original basis vectors are the inversion of the above:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \hat{u} \\ \hat{v} \end{bmatrix}$$

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 $\vec{A}$ in the rotated coordinate system. We frame the rotated vector as

$$\vec{A} = A_u\,\hat{u} + A_v\,\hat{v}\,,$$

and the job is to solve for A_u, A_v. One way to proceed is to substitute $\hat{u}$, $\hat{v}$ and simplify:

$$\begin{aligned} \vec{A} &= A_u\,(\cos(\theta)\,\hat{x} + \sin(\theta)\,\hat{y}) \\ &\quad + A_v\,(-\sin(\theta)\,\hat{x} + \cos(\theta)\,\hat{y}) \\ \vec{A} &= (A_u\cos(\theta) - A_v\sin(\theta))\,\hat{x} \\ &\quad + (A_u\sin(\theta) + A_v\cos(\theta))\,\hat{y} \end{aligned}$$

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.

7 Vector Derivatives

7.1 Rules of Differentiation

The rules for differentiating vector quantities are exactly analogous to those for scalars. For instance, if we have a position vector $\vec{r}(t)$, then the velocity vector is the time derivative by definition:

$$\vec{v}(t) = \frac{d}{dt}\left(\vec{r}(t)\right) = \lim_{h\to 0}\frac{\vec{r}(t+h) - \vec{r}(t)}{h}$$

In the following, suppose c is a constant, $\lambda(t)$ is a function of time, and so too are the vectors $\vec{r}$, $\vec{s}$. Then, we always have:

$$\begin{aligned} \frac{d}{dt}(c\,\vec{r}) &= c\,\frac{d\vec{r}}{dt} \\ \frac{d}{dt}(\lambda\,\vec{r}) &= \frac{d\lambda}{dt}\vec{r} + c\,\frac{d\vec{r}}{dt} \\ \frac{d}{dt}(\vec{r} + \vec{s}) &= \frac{d\vec{r}}{dt} + \frac{d\vec{s}}{dt} \\ \frac{d}{dt}(\vec{r}\cdot\vec{s}) &= \frac{d\vec{r}}{dt}\cdot\vec{s} + \vec{r}\cdot\frac{d\vec{s}}{dt} \\ \frac{d}{dt}(\vec{r}\times\vec{s}) &= \frac{d\vec{r}}{dt}\times\vec{s} + \vec{r}\times\frac{d\vec{s}}{dt} \end{aligned}$$

7.2 Basis Vector Derivatives

Derivatives of the Cartesian basis vectors are all zero. As for polar coordinates, since $\hat{r}$, $\hat{\theta}$ are allowed to swivel about as θ changes, it makes sense to ask about the derivatives of these. Proceeding carefully, we find

$$\frac{d}{d\theta}\hat{r} = -\sin(\theta)\,\hat{x} + \cos(\theta)\,\hat{y} = \hat{\theta}$$

and

$$\frac{d}{d\theta}\hat{\theta} = -\cos(\theta)\,\hat{x} - \sin(\theta)\,\hat{y} = -\hat{r}\,.$$

7.3 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

$$\vec{V} = \frac{d}{dt}\vec{R} = \left(\frac{d}{dt}x\left(t\right)\right)\hat{x} + \left(\frac{d}{dt}y\left(t\right)\right)\hat{y}\,.$$

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'):

$$\omega = \omega\left(t\right) = \frac{d}{dt}\theta\left(t\right)$$

Next, we'll need the time derivative of the *entire* position vector

$$\vec{v} = \frac{d\vec{r}}{dt} = \frac{d}{dt}\left(r\left(t\right)\ \hat{r}\left(t\right)\right)\ ,$$

which calls for the product rule:

$$\vec{v} = \left(\frac{d}{dt}r\left(t\right)\right)\hat{r}\left(t\right) + r\left(t\right)\frac{d}{dt}\hat{r}\left(t\right)$$

Expand the right-most term with the with the chain rule

$$\vec{v} = \frac{dr}{dt}\,\hat{r} + r\frac{d\theta}{dt}\frac{d\hat{r}}{d\theta}\,,$$

and simplify to get the velocity in polar coordinates:

$$\vec{v} = \frac{dr}{dt}\,\hat{r} + r\omega\,\hat{\theta}$$

Problem 1

If the magnitude $|\vec{r}| = r$ is constant in time, show that $\vec{r}$ and $\vec{v}$ are perpendicular. Hint: differentiate $\vec{r}\cdot\vec{r}$

7.4 Acceleration

The acceleration vector is the time derivative of the velocity vector. This is straightforward in Cartesian coordinates:

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{d^2}{dt^2}\vec{R} = \left(\frac{d^2}{dt^2}x\left(t\right)\right)\hat{x} + \left(\frac{d^2}{dt^2}y\left(t\right)\right)\hat{y}$$

Acceleration in Polar Coordinates

As expected, the acceleration in polar coordinates is messy. Starting the calculation, we have

$$\vec{a} = \frac{d\vec{v}}{dt} = \frac{d}{dt}\left(\frac{dr}{dt}\,\hat{r}\right) + \frac{d}{dt}\left(r\omega\,\hat{\theta}\right)\ .$$

Leaving the details for an exercise, the result comes out to

$$\vec{a} = \left(\frac{d^2r}{dt^2} - r\omega^2\right)\hat{r} + \left(r\frac{d\omega}{dt} + 2\frac{dr}{dt}\omega\right)\hat{\theta}\,.$$

Problem 2

Derive the acceleration vector in polar coordinates by taking two derivatives of $\vec{r} = r\,\hat{r}$.

7.5 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:

$$\frac{d}{dt}z = \frac{dr}{dt}\,e^{i\theta} + r\omega\left(i\,e^{i\theta}\right)$$

Comparing this result to the velocity in polar coordinates, a direct analogy emerges:

$$e^{i\theta} \leftrightarrow \hat{r}$$
$$i\,e^{i\theta} \leftrightarrow \hat{\theta}$$

Evidently, the terms $e^{i\theta}$, $\hat{r}$ play similar roles in complex numbers and polar coordinates, which is no surprise since each is of the format $\langle\cos\left(\theta\right),\sin\left(\theta\right)\rangle$. Similar comments apply to the pair $i\,e^{i\theta}$, $\hat{\theta}$.

7.6 Differential Line Element

Starting from the velocity vector, whether it be the Cartesian representation or polar representation

$$\vec{V} = \frac{dx}{dt}\,\hat{x} + \frac{dy}{dt}\,\hat{y}$$
$$\vec{v} = \frac{dr}{dt}\,\hat{r} + r\frac{d\theta}{dt}\,\hat{\theta}\,,$$

and multiply through by dt:

$$d\vec{R} = \vec{V}\,dt = dx\,\hat{x} + dy\,\hat{y}$$
$$d\vec{r} = \vec{v}\,dt = dr\,\hat{r} + r\,d\theta\,\hat{\theta}$$

The terms $d\vec{R}$ (Cartesian), $d\vec{r}$ (polar) are each called the *differential line element*, also called $d\vec{S}$. 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.

7.7 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

$$dS^2 = (dx\,\hat{x} + dy\,\hat{y}) \cdot (dx\,\hat{x} + dy\,\hat{y}) = dx^2 + dy^2 \,.$$

Then, for polar coordinates:

$$dS^2 = d\vec{r} \cdot d\vec{r} = dr^2 + r^2 d\theta^2$$

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

$$S = \int dS = \int \sqrt{d\vec{S} \cdot d\vec{S}} \,.$$

Misguided Missile

Source: TBD

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.

Using polar coordinates, the velocity is

$$\vec{v} = -v\cos(\alpha)\,\hat{r} + v\sin(\alpha)\,\hat{\theta} \,,$$

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.

7.8 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

$$d\vec{R} = dx\,\hat{x} + dy\,\hat{y} \,,$$

and then write two versions of this - one with $dy = 0$, the other with $dx = 0$:

$$\begin{aligned} d\vec{R}_1 &= dx\,\hat{x} + 0\,\hat{y} \\ d\vec{R}_2 &= 0\,\hat{x} + dy\,\hat{y} \end{aligned}$$

This defines two vectors that form the sides of a parallelogram (just a rectangle in this case). The magnitude of the cross product $d\vec{R}_1 \times d\vec{R}_2$ yields the area of said rectangle:

$$dA = \left| d\vec{R}_1 \times d\vec{R}_2 \right| = (dx\,\hat{x}) \times (dy\,\hat{y}) = dx\,dy$$

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

$$d\vec{r} = dr\,\hat{r} + r\,d\theta\,\hat{\theta}\,,$$

and then

$$d\vec{r}_1 = dr\,\hat{r} + 0\,\hat{\theta}$$
$$d\vec{r}_2 = 0\,\hat{r} + r\,d\theta\,\hat{\theta}\,.$$

In this case, $d\vec{r}_1$, $d\vec{r}_2$ form the sides of a parallelogram (not rectangular), whose area is the the magnitude of their cross product:

$$dA = |d\vec{r}_1 \times d\vec{r}_2| = r\,dr\,d\theta$$

8 Plane Curve Analysis

8.1 Tangent Vector

Consider a curve described by a position vector $\vec{r}\,(t)$ in Cartesian coordinates parameterized by a real number t:

$$\vec{r}\,(t) = x\,(t)\,\hat{x} + y\,(t)\,\hat{y}$$

The derivative of $\vec{r}$ 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:

$$\vec{v}\,(t) = \frac{d}{dt}\vec{r}\,(t) = \frac{dx}{dt}\,\hat{x} + \frac{dy}{dt}\,\hat{y} = v_x\hat{x} + v_y\,\hat{y}$$

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:

$$\hat{T} = \frac{\vec{v}}{v} = \frac{v_x\hat{x} + v_y\hat{y}}{\sqrt{v_x^2 + v_y^2}}$$

Going further, we can divide out a factor of dt from the numerator and denominator to write

$$\hat{T} = \frac{dx\,\hat{x} + dy\,\hat{y}}{\sqrt{dx^2 + dy^2}}\,.$$

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

$$\hat{T} = \frac{d\vec{S}}{dS}\,,$$

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

$$\hat{T} = \cos(\phi)\,\hat{x} + \sin(\phi)\,\hat{y}\,,$$

in which case the slope of the curve dy/dx is

$$\frac{dy}{dx} = \frac{\sin(\phi)}{\cos(\phi)} = \tan(\phi)\ .$$

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 $\hat{T}$ can be configured as:

$$\hat{T} = \frac{\hat{x} + (dy/dx)\,\hat{y}}{\sqrt{1 + (dy/dx)^2}} = \frac{\hat{x} + y'\,\hat{y}}{\sqrt{1 + (y')^2}}$$

8.2 Normal Vector

Given the tangent vector $\hat{T}$ to a curve, one can imagine that differential changes in $\hat{T}$ are always perpendicular to the tangent, and are thus perpendicular to the curve:

$$\frac{d}{d\phi}\hat{T} = -\sin(\phi)\,\hat{x} + \cos(\phi)\,\hat{y}$$

One can explicitly check that

$$\hat{T} \cdot \frac{d}{d\phi}\hat{T} = 0\,.$$

This implies the existence of the *normal vector*. In the general case, the normalized vector is crudely defined as

$$\vec{N} = \frac{d}{dt}\vec{T}\,,$$

where the parameter t is considered generic. The unit normal vector divides out its own magnitude

$$\hat{N} = \frac{1}{\left|d\vec{T}/dt\right|}\frac{d\vec{T}}{dt}\,,$$

and by the chain rule, we can essentially swap the t-parameter for the arc length:

$$\hat{N} = \frac{1}{\left|d\vec{T}/dS\right|}\frac{d\vec{T}}{dS}$$

Problem 3

Show that the tangent vector and normal vector always obey the orthogonality relation

$$\vec{T} \cdot \vec{N} = 0 \, .$$

Hint: differentiate $\hat{T} \cdot \hat{T}$ with respect to arc length.

8.3 Acceleration Vector

The position vector

$$\vec{r}\,(t) = x\,(t)\,\hat{x} + y\,(t)\,\hat{y}$$

and the velocity vector

$$\vec{v}\,(t) = \frac{d}{dt}\vec{r}\,(t) = \frac{dx}{dt}\,\hat{x} + \frac{dy}{dt}\,\hat{y}$$

imply the existence of the acceleration vector

$$\vec{a}\,(t) = \frac{d}{dt}\vec{v}\,(t) = \frac{d^2}{dt^2}\vec{r}\,(t) = \frac{d^2x}{dt^2}\,\hat{x} + \frac{d^2y}{dt^2}\,\hat{y} \, .$$

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

$$\vec{v}\,(t) = v\,(t)\,\hat{T}\,(t) \ .$$

Now recalculate the acceleration vector using the above definition, which gives

$$\vec{a}\,(t) = \frac{d}{dt}\vec{v}\,(t) = \frac{dv}{dt}\,\hat{T} + v\,\frac{d\hat{T}}{dt} \, ,$$

or

$$\vec{a}\,(t) = \left(\frac{d^2S}{dt^2}\right)\hat{T} + \left|\frac{d\hat{T}}{dS}\right|\left(\frac{dS}{dt}\right)^2\hat{N} \, ,$$

where $v = dS/dt$ and the definition of the normal vector have been used.

Problem 4

Derive the following:

$$\hat{N} = \frac{1}{v\left|d\hat{T}/dt\right|}\vec{a} - \frac{dv/dt}{v^2\left|d\hat{T}/dt\right|}\vec{v}$$

Traveling Basis Vectors

Since the tangent vector and normal vector are always perpendicular, these form a local set of of basis vectors to the curve:

$$\vec{v} = v\,\hat{T}$$

$$\vec{a} = a_T\,\hat{T} + a_N\,\hat{N}$$

Problem 5

A particle has a known trajectory

$$r\,(t) = \frac{r_0}{\cos\,(\omega t)} \, ,$$

where ω is constant. Find the velocity and acceleration vectors.

8.4 Curvature

The magnitude of the derivative of the normalized tangent vector with respect to arc length, i.e. $\left|d\hat{T}/dS\right|$, is called the curvature, denoted κ (Greek 'kappa'):

$$\kappa = \left|\frac{d\hat{T}}{dS}\right|$$

In terms of the curvature, the normal vector is

$$\hat{N} = \frac{1}{\kappa}\frac{d\hat{T}}{dS} \, ,$$

and the acceleration vector is

$$\vec{a}\,(t) = \left(\frac{d^2S}{dt^2}\right)\hat{T} + \kappa\left(\frac{dS}{dt}\right)^2\hat{N} \, .$$

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

$$\hat{T} = \cos\,(\phi)\,\hat{x} + \sin\,(\phi)\,\hat{y}$$

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 $\hat{T}$

$$\frac{d\hat{T}}{dt} = \frac{d\phi}{dt}\left(-\sin\,(\phi)\,\hat{x} + \cos\,(\phi)\right) \, ,$$

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 $\rho\,(t)$:

$$\rho\,(t) = \frac{dS/dt}{\left|d\hat{T}/dt\right|} = \frac{1}{\left|d\hat{T}/dS\right|} = \frac{1}{\kappa\,(t)}$$

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

$$\vec{v} \times \vec{a} = \left(v\,\hat{T}\right) \times \left(\left(\frac{d^2S}{dt^2}\right)\hat{T} + \kappa\left(\frac{dS}{dt}\right)^2 \hat{N}\right) .$$

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:

$$\left|\hat{T} \times \hat{N}\right| = 1$$

This is enough to isolate the curvature κ in terms of the vectors of motion:

$$\kappa = \frac{|\vec{v} \times \vec{a}|}{v^3}$$

Curvature in Polar Coordinates

For a polar curve

$$\vec{r}\,(t) = r\,(t)\,\hat{r}\,,$$

the velocity and acceleration are

$$\vec{v} = \frac{dr}{dt}\,\hat{r} + r\frac{d\theta}{dt}\,\hat{\theta}$$
$$\vec{a} = \left(\frac{d^2r}{dt^2} - r\left(\frac{d\theta}{dt}\right)^2\right)\hat{r} + \left(r\frac{d^2\theta}{dt^2} + 2\frac{dr}{dt}\frac{d\theta}{dt}\right)\hat{\theta}\,.$$

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 $\vec{v} \times \vec{a}$, 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

$$\vec{v} \times \vec{a} = \left(v_r\hat{r} + v_\theta\hat{\theta}\right) \times \left(a_r\hat{r} + a_\theta\hat{\theta}\right)$$
$$= \left(v_r a_\theta - v_\theta a_r\right)\left(\hat{r} \times \hat{\theta}\right)\,,$$

or, noting $\omega = d\theta/dt$,

$$|\vec{v} \times \vec{a}| = \left|r\frac{dr}{dt}\frac{d\omega}{dt} + 2\omega\left(\frac{dr}{dt}\right)^2 - r\omega\frac{d^2r}{dt^2} + r^2\omega^3\right|\,.$$

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:

$$\kappa = \frac{\left|r^2 + 2\,(dr/d\theta)^2 - r\left(d^2r/d\theta^2\right)\right|}{\left(r^2 + (dr/d\theta)^2\right)^{3/2}}$$

9 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.

9.1 Quadratic Case

In the Cartesian plane, consider three given points $\vec{p}_j = (x_j, y_j)$, where $j = 0, 1, 2$. Such given points are called *control points*.

Next, suppose there is a quadratic curve $\vec{r}\,(t)$ that passes through the first point, and passes through the last point, skipping the middle $j = 1$-point.

Letting t be a dimensionless parameter

$$0 \le t \le 1\,,$$

we know

$$\vec{r}\,(0) = \vec{p}_0$$
$$\vec{r}\,(1) = \vec{p}_2\,,$$

and that the slope $\vec{r}\,(t)$ is

$$\frac{d}{dt}\vec{r}\,(t) = \vec{v}\,(t)\ .$$

This describes an infinite family of curves so far, but now impose the restriction that $\vec{v}\,(0)$ is parallel to the difference $\vec{p}_1 - \vec{p}_0$. Also, let $\vec{v}\,(1)$ be parallel to the difference $\vec{p}_2 - \vec{p}_1$. For shorthand, let us write this as:

$$\vec{v}\,(0) \propto \vec{p}_1 - \vec{p}_0 = \vec{p}_{01}$$
$$\vec{v}\,(1) \propto \vec{p}_2 - \vec{p}_1 = \vec{p}_{12}$$

If we wish for $\vec{r}\,(t)$ 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 $\vec{v}(t)$ already, lets us start a step back and define

$$\vec{a} = \frac{d}{dt} v(t) \ ,$$

which is constant for a quadratic curve.

In terms of two unknown parameters α, β, the constant vector $\vec{a}$ that characterizes the quadratic curve can be written

$$\vec{a} \propto \alpha \, \vec{p}_{01} + \beta \, \vec{p}_{12} \ .$$

Integrate $\vec{a}$ in the t-variable to get the velocity

$$\vec{v}(t) \propto \alpha t \, \vec{p}_{01} + \beta t \, \vec{p}_{12} + \vec{\gamma} \ ,$$

where $\vec{\gamma}$ 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

$$\begin{aligned} \alpha &= -A \\ \beta &= A \\ \vec{\gamma} &= A \, \vec{p}_{01} \ , \end{aligned}$$

where A is an overall proportionality constant. The velocity is given by:

$$\vec{v}(t) = A \left((1-t) \, \vec{p}_{01} + t \, \vec{p}_{12} \right)$$

Now we find the curve $\vec{r}(t)$ by integrating the velocity to write

$$\vec{r}(t) = A \left(\frac{-(1-t)^2}{2} \, \vec{p}_{01} + \frac{t^2}{2} \, \vec{p}_{12} \right) + \vec{\delta} \ ,$$

where $\vec{\delta}$ is a constant.

The constants A, $\vec{\delta}$ are determined by the conditions at $t = 0$, $t = 1$. Writing each of these, we find

$$\begin{aligned} \vec{p}_0 &= \frac{A}{2} \left(-\vec{p}_1 + \vec{p}_0 \right) + \vec{\delta} \\ \vec{p}_2 &= \frac{A}{2} \left(\vec{p}_2 - \vec{p}_1 \right) + \vec{\delta} \ , \end{aligned}$$

implying

$$\begin{aligned} A &= 2 \\ \vec{\delta} &= \vec{p}_1 \ . \end{aligned}$$

Knowing the integration constants, we can go back and rewrite the velocity and its derivative:

$$\begin{aligned} \vec{v}(t) &= 2(1-t) \, \vec{p}_{01} + 2t \, \vec{p}_{12} \\ \vec{a} &= -2 \, \vec{p}_{01} + 2 \vec{p}_{12} \end{aligned}$$

In terms of the given $\vec{p}_j$, the position vector simplifies to

$$\vec{r}(t) = (1-t)^2 \, \vec{p}_0 + 2t(1-t) \, \vec{p}_1 + t^2 \, \vec{p}_2 \ ,$$

and correspondingly:

$$\begin{aligned} \vec{v}(t) &= -2(1-t) \, \vec{p}_0 + 2(1-2t) \, \vec{p}_1 + 2t \, \vec{p}_2 \\ \vec{a} &= 2 \left(\vec{p}_0 - 2\vec{p}_1 + \vec{p}_2 \right) \end{aligned}$$

9.2 Cubic Case

Extending the quadratic case by one, consider four given points $\vec{p}_j = (x_j, y_j)$, where $j = 0, 1, 2, 3$. A curve $\vec{r}(t)$ passes through $\vec{p}_0$ at $t = 0$, and through $\vec{p}_3$ at $t = 1$. The velocity obeys

$$\begin{aligned} \vec{v}(0) &\propto \vec{p}_1 - \vec{p}_0 = \vec{p}_{01} \\ \vec{v}(1) &\propto \vec{p}_3 - \vec{p}_2 = \vec{p}_{23} \ . \end{aligned}$$

Unlike the quadratic case in where the derivative of $\vec{v}(t)$ is a constant $\vec{a}$, we now need an $\vec{a}(t)$ that is (at most) linearly dependent on t. Work toward this by defining two constants

$$\begin{aligned} \vec{p}_{012} &= 2 \left(\vec{p}_0 - 2\vec{p}_1 + \vec{p}_2 \right) \\ \vec{p}_{123} &= 2 \left(\vec{p}_1 - 2\vec{p}_2 + \vec{p}_3 \right) \ . \end{aligned}$$

Notice that $\vec{p}_{012}$ is same as $\vec{a}$ from the quadratic case.

Next, propose a vector $\vec{j}$ as the derivative of $\vec{a}(t)$

$$\frac{d}{dt} \vec{a}(t) = \vec{j} \ ,$$

which is a constant. Similar to the quadratic case, write $\vec{j}$ as a linear combination of vectors

$$\vec{j} \propto \alpha \, \vec{p}_{012} + \beta \, \vec{p}_{123} \ ,$$

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 $\vec{j}$. While $\vec{j}$ 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 $\vec{j}$ playing $\vec{a}$'s role, etc. Transcribing the solution for α, β that works for the quadratic case, one writes

$$\begin{aligned} \vec{a}(t) &= B \left(2(1-t) \, \vec{p}_{012} + 2t \, \vec{p}_{123} \right) \\ \vec{j} &= B \left(-2 \, \vec{p}_{012} + 2 \vec{p}_{123} \right) \ , \end{aligned}$$

where B is a new proportionality constant.

Integrate $\vec{a}(t)$ to get the velocity

$$\vec{v}(t) = B \left(-(1-t)^2 \, \vec{p}_{012} + t^2 \, \vec{p}_{123} \right) + \vec{\gamma} \ ,$$

where $\vec{\gamma}$ is a constant.

Integrate once more to get the position up to another constant $\vec{\delta}$:

$$\vec{r}(t) = B\left(\frac{-(1-t)^3}{3}\,\vec{p}_{012} + \frac{t^3}{3}\,\vec{p}_{123}\right) + t\,\vec{\gamma} + \vec{\delta}$$

To determine B and $\vec{\gamma}$, use $\vec{r}(0) = \vec{p}_0$ and $\vec{r}(1) = \vec{p}_3$, and then subtract one equation from the other to eliminate $\vec{\delta}$. This should result in

$$\vec{p}_3 - \vec{p}_0 = \vec{\gamma} + \frac{B}{3}\left(3\vec{p}_1 - 3\vec{p}_2 + \vec{p}_3 - \vec{p}_0\right) ,$$

which begs the solution

$$\begin{aligned} B &= 3 \\ \vec{\gamma} &= -3\vec{p}_1 + 3\vec{p}_2 \,. \end{aligned}$$

Staying with the velocity equation, substitute $\vec{\gamma}$ and simplify like mad to get

$$\vec{v}(t) = 3(1-t)^2\,\vec{p}_{01} + 6t(1-t)\,\vec{p}_{12} + 3t^2\,\vec{p}_{23} \,.$$

This form is much nicer than the $\vec{r}(t)$ derived a few lines above, so let's integrate the velocity again to get the position

$$\vec{r}(t) = \int \vec{v}(t)\,dt + \vec{\delta}\,,$$

where $\vec{\delta}$ is the same integration constant. Eliminate $\vec{\delta}$ delta using what's known about $t = 0$, $t = 1$, and get the final position:

$$\begin{aligned} \vec{r}(t) = {} & (1-t)^3\,\vec{p}_0 \\ & + 3t(1-t)^2\,\vec{p}_1 + 3t^2(1-t)\,\vec{p}_2 + t^3\,\vec{p}_3 \end{aligned}$$

9.3 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

$$1 - t = w \,.$$

Quadratic

For the quadratic case:

$$\begin{aligned} \vec{v}_2(t) &= 2w\,\vec{p}_{01} + 2t\,\vec{p}_{12} \\ \vec{r}_2(t) &= w^2\,\vec{p}_0 + 2tw\,\vec{p}_1 + t^2\,\vec{p}_2 \end{aligned}$$

Cubic

For the cubic case:

$$\begin{aligned} \vec{v}_3(t) &= 3w^2\,\vec{p}_{01} + 6tw\,\vec{p}_{12} + 3t^2\,\vec{p}_{23} \\ \vec{r}_3(t) &= w^3\,\vec{p}_0 + 3tw^2\,\vec{p}_1 + 3t^2w\,\vec{p}_2 + t^3\,\vec{p}_3 \end{aligned}$$

Binomial Coefficients

Looking at the pattern in the numeric coefficients in each position vector $\vec{r}(t)$, 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, kth 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:

$$\begin{aligned} \vec{r}_2(t) &= \binom{2}{0}w^2\,\vec{p}_0 + \binom{2}{1}tw\,\vec{p}_1 + \binom{2}{2}t^2\,\vec{p}_2 \\ \vec{r}_3(t) &= \binom{3}{0}w^3\,\vec{p}_0 + \binom{3}{1}tw^2\,\vec{p}_1 \\ &\quad + \binom{3}{2}t^2w\,\vec{p}_2 + \binom{3}{3}t^3\,\vec{p}_3 \end{aligned}$$

General Equations

Each of the above lends to summation notation. Converting to this, we find, after restoring $w = 1 - t$:

$$\vec{r}_n(t) = \sum_{j=0}^{n}\binom{n}{j}(1-t)^{n-j}\,t^j\,\vec{p}_j$$

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 it a constant that involves all of the provided control points. For the quadratic case, we discovered

$$\vec{a} = \frac{d^2}{dt^2}\vec{r}_2(t) = 2\left(-\vec{p}_{01} + \vec{p}_{12}\right) ,$$

and for the cubic case,

$$\vec{j} = \frac{d^3}{dt^3}\vec{r}_3(t) = 6\left(-\vec{p}_{012} + \vec{p}_{123}\right) .$$

Going from the pattern, one should suspect that the fourth-order case obeys

$$\vec{k} = \frac{d^4}{dt^4}\vec{r}_4(t) = 24\left(-\vec{p}_{0123} + \vec{p}_{1234}\right) ,$$

and so on for higher orders.

Problem 6

Write down the fourth-order curve $\vec{r}_4(t)$ and determine exactly what is meant by $\vec{p}_{0123}$, $\vec{p}_{1234}$.

10 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

$$\vec{F} = F(r)\,\hat{r} ,$$

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:

$$m\frac{d^2\vec{r}}{dt^2} = m\frac{d\vec{v}}{dt} = m\vec{a} = \vec{F}$$

I avoided saying exactly how $-dU/dx$ becomes $\vec{F}$. 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 an 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:

$$\vec{F}_{12} = -\vec{F}_{21}$$

10.1 Two-Body Problem

Consider two bodies in space, one of mass m_1 at position $\vec{r}_1(t)$, and the other of mass m_2 at position $\vec{r}_2(t)$. The force imposed onto body 1 by body 2 is given by

$$m_1\frac{d^2}{dt^2}\vec{r}_1(t) = m_1\frac{d}{dt}\vec{v}_1(t) = \vec{F}_{12} ,$$

and the force imposed onto particle 2 by particle 1 is given by

$$m_2\frac{d^2}{dt^2}\vec{r}_2(t) = m_2\frac{d}{dt}\vec{v}_2(t) = \vec{F}_{21} .$$

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 $\vec{R}(t)$ such that

$$\vec{R}(t) = \frac{m_1\vec{r}_1(t) + m_2\vec{r}_2(t)}{m_1 + m_2}\,.$$

The time derivative of the center of mass gives a quantity called the *center of velocity*:

$$\vec{V}(t) = \frac{d}{dt}\vec{R}(t) = \frac{m_1\vec{v}_1(t) + m_2\vec{v}_2(t)}{m_1 + m_2}\,.$$

Taking the time derivative of the center of velocity gives something interesting:

$$\begin{aligned}\frac{d^2}{dt^2}\vec{R}(t) &= \frac{m_1\left(d\vec{v}_1(t)/dt\right) + m_2\left(d\vec{v}_2(t)/dt\right)}{m_1 + m_2}\\ &= \frac{\vec{F}_{12} + \vec{F}_{21}}{m_1 + m_2} = \frac{\vec{F}_{12} - \vec{F}_{12}}{m_1 + m_2} = 0\end{aligned}$$

Evidently, the second derivative of the center of mass is precisely zero because $\vec{F}_{12} = -\vec{F}_{21}$, 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 $\vec{V}$ is a constant $\vec{V}_0$.

Relative Displacement

If the distance separating the two bodies is r, define a vector

$$\vec{r}(t) = \vec{r}_1(t) - \vec{r}_2(t)$$

with $|\vec{r}| = r$, capturing the relative displacement between the two.

Listing this with the center of mass $\vec{R}(t)$, we have a system of two equations that can be solved for $\vec{r}_1(t)$, $\vec{r}_2(t)$ separately: (We know everything is a function of t by now, so drop the extra notation.)

$$\begin{aligned}\vec{r}_1 &= \vec{R} + \frac{m_2}{m_1 + m_2}\,\vec{r}\\ \vec{r}_2 &= \vec{R} - \frac{m_1}{m_1 + m_2}\,\vec{r}\end{aligned}$$

Reduced Mass

From the equations above, multiply through by m_1, m_2, respectively, and take two time derivatives:

$$\begin{aligned}m_1\frac{d^2\vec{r}_1}{dt^2} &= m_1\frac{\cancel{d^2\vec{R}}}{\cancel{dt^2}} + \frac{m_1m_2}{m_1 + m_2}\frac{d^2\vec{r}}{dt^2}\\ m_2\frac{d^2\vec{r}_2}{dt^2} &= m_2\frac{\cancel{d^2\vec{R}}}{\cancel{dt^2}} - \frac{m_1m_2}{m_1 + m_2}\frac{d^2\vec{r}}{dt^2}\end{aligned}$$

These results say the same thing, as the left sides are $\vec{F}_{12}$, $\vec{F}_{21}$, respectively, and the right sides differ by the proper negative sign.

Evidently, we have

$$\vec{F}_{12} = \frac{m_1m_2}{m_1 + m_2}\frac{d^2\vec{r}}{dt^2}\,.$$

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 $\vec{r}$ rather than two explicit position vectors $\vec{r}_{1,2}$.

The price we pay is the mass term became a mess. This group of symbols is called the *reduced mass*:

$$m_* = \frac{m_1m_2}{m_1 + m_2}$$

Representing the effective mass of the total system as m_*, the two-body problem is summarized in one equation:

$$\vec{F}_{12} = m_*\frac{d^2\vec{r}}{dt^2} = m_*\vec{a}$$

A handy identity involving the reduced mass, somewhat reminiscent of resistors in parallel, goes as:

$$\frac{1}{m_*} = \frac{1}{m_1} + \frac{1}{m_2}$$

10.2 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 $\vec{L}$ of the two-body system reads

$$\vec{L} = m_*\vec{r}\times\vec{v}\,,$$

where $\vec{r}$ is the relative displacement vector, and $\vec{v}$ is its time derivative. Now calculate the time derivative of $\vec{L}$:

$$\begin{aligned}\frac{d}{dt}\vec{L} &= m_*\frac{d}{dt}\left(\vec{r}\times\vec{v}\right)\\ &= m_*\left(\vec{v}\times\vec{v} + \vec{r}\times\frac{d\vec{v}}{dt}\right)\\ &= \vec{r}\times\vec{F}\end{aligned}$$

For the remaining cross product to vanish, we go back to Newton's original assumption that

$$\vec{F} = F(r)\,\hat{r}\,,$$

which means the force vector and the displacement vector are parallel. Using this, we see that the derivative of $\vec{L}$ 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:

$$\vec{r} = r\,\hat{r}$$
$$\vec{v} = \frac{dr}{dt}\hat{r} + r\frac{d\theta}{dt}\hat{\theta}$$

Remembering $\hat{r} \times \hat{r}$ is zero, we then have

$$\vec{L} = m_* r^2 \frac{d\theta}{dt}\left(\hat{r} \times \hat{\theta}\right) .$$

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 $\vec{L}$, project $\vec{r}$ into both sides:

$$\vec{r} \cdot \vec{L} = m_* \vec{r} \cdot (\vec{r} \times \vec{v}) ,$$

and then make use of the triple product:

$$\vec{r} \cdot \vec{L} = m_* \vec{v} \cdot (\vec{r} \times \vec{r}) = 0$$

10.3 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 $\hat{\theta}$-component of the acceleration vector in polar coordinates. In terms of L, the acceleration vector is

$$\vec{a} = \left(\frac{d^2r}{dt^2} - \frac{L^2}{m_*^2 r^3}\right)\hat{r} + \frac{1}{m_* r}\cancel{\left(\frac{dL}{dt}\right)}\hat{\theta} .$$

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:

$$\frac{d^2r}{dt^2} = \frac{L^2}{m_*^2 r^2 r_0} e\cos(\theta) = \frac{L^2}{m_*^2 r^2}\left(\frac{1}{r} - \frac{1}{r_0}\right)$$

The full acceleration vector then reads

$$\vec{a} = \frac{L^2}{m_*^2 r^2}\left(\cancel{\frac{1}{r}} - \frac{1}{r_0} - \cancel{\frac{1}{r}}\right)\hat{r} ,$$

which simplifies nicely:

$$\vec{a} = \frac{-L^2}{m_*^2 r_0}\frac{\hat{r}}{r^2}$$

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.

$$m_1\frac{d^2\vec{r}_1}{dt^2} = m_*\frac{d^2\vec{r}}{dt^2}$$
$$m_2\frac{d^2\vec{r}_2}{dt^2} = -m_*\frac{d^2\vec{r}}{dt^2} ,$$

we can solve for the absolute acceleration of each body:

$$\vec{a}_1 = \frac{m_*}{m_1}\,\vec{a}$$
$$\vec{a}_2 = \frac{-m_*}{m_2}\,\vec{a}$$

Eliminate $\vec{a}$ between the two equations to recover Newton's third law:

$$m_1\,\vec{a}_1 + m_2\,\vec{a}_2 = 0$$

10.4 Universal Gravitation

Enough ground work has been done to finally write Newton's universal law of gravitation.

Recall the absolute acceleration of each body $\vec{a}_1$, $\vec{a}_2$, and replace the reduced mass m_* and acceleration $\vec{a}$ with expanded forms:

$$\vec{a}_1 = \left(\frac{m_2}{m_1 + m_2}\right)\frac{-L^2}{m_*^2 r_0}\frac{\hat{r}}{r^2}$$
$$\vec{a}_2 = \left(\frac{-m_1}{m_1 + m_2}\right)\frac{-L^2}{m_*^2 r_0}\frac{\hat{r}}{r^2}$$

To get rid of some clutter, let us group the coefficients that occur identically in both equations into an auxiliary constant γ such that

$$\gamma = \frac{1}{m_1 + m_2}\left(\frac{L^2}{m_*^2 r_0}\right) = \frac{L^2}{m_* m_1 m_2 r_0}\,,$$

and then we can forget about γ (momentarily) by trading the equality signs for proportionality symbols:

$$\vec{a}_1 \propto m_2\,\frac{(-\hat{r})}{r^2}$$
$$\vec{a}_2 \propto -m_1\,\frac{(-\hat{r})}{r^2}$$

What we see is the acceleration of body 1 being proportional to the mass of body 2, and vice versa.

Multiply each equation through by m_1, m_2, respectively to turn accelerations into forces:

$$\vec{F}_{12} = m_1\,\vec{a}_1 \propto m_1 m_2\,\frac{(-\hat{r})}{r^2}$$
$$\vec{F}_{21} = m_2\,\vec{a}_2 \propto -m_1 m_2\,\frac{(-\hat{r})}{r^2}$$

Of course, these are saying the same thing due to Newton's third law, so in summary:

$$\vec{F}_{12} \propto -m_1 m_2\,\frac{\hat{r}}{r^2}$$

This comprises all of the ingredients for building the gravitational force. We have a force vector acting along the line connecting two bodies whose strength is proportional the product of the masses and inversely proportional to the square of the separation.

Newton decided to introduce a new proportionality constant G, named after 'gravity', to turn the above back into an equation. We take the gravitational force, finally, to be:

$$\vec{F}_{12} = -G\frac{m_1 m_2}{r^2}\,\hat{r}$$

Note that the force vector bears the 12-subscript and not the other way around. The subscript is often omitted because the unit vector $\hat{r}$ has an implied 12-subscript that goes back to the definition of $\vec{r}$.

To reconcile the constants γ and G, it's easy to work out that

$$G = \gamma m_1 m_2\,.$$

Eliminating γ, we can also write

$$G = \frac{L^2}{m_* r_0}\,.$$

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.

10.5 Equations of Motion

With the law of universal gravitation on hand, we should be able to run the analysis in reverse by starting with $\vec{F}_{12}$ 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:

$$\vec{F}_{12} = -G m_1 m_2 \frac{m_*}{L}\frac{d\theta}{dt}\,\hat{r}$$

Also replace $\vec{F}_{12}$ to keep simplifying

$$m_1\,\vec{a}_1 = m_*\,\vec{a} = -G m_1 m_2 \frac{m_*}{L}\frac{d\theta}{dt}\,\hat{r}\,,$$

and solve for the relative acceleration:

$$\vec{a} = -G\frac{m_1 m_2}{L}\frac{d\theta}{dt}\,\hat{r}$$

Velocity

To proceed, replace the acceleration vector as the derivative of the relative velocity by $\vec{a} = d\vec{v}/dt$. Also replace $\hat{r}$ via $-\hat{r} = d\hat{\theta}/d\theta$ to get

$$\frac{d\vec{v}}{dt} = G\frac{m_1 m_2}{L}\frac{d\theta}{dt}\frac{d\hat{\theta}}{d\theta}\,,$$

simplifying with the chain rule to:

$$d\vec{v} = G\frac{m_1 m_2}{L}\,d\hat{\theta}$$

Integrate both sides of the above to get a vector equation for the velocity

$$\vec{v}\,(t) = G\frac{m_1 m_2}{L}\,\hat{\theta}\,(t) + \vec{v}_0\,,$$

where $\vec{v}_0$ is the integration constant. Letting $\theta = 0$ correspond with the positive x-axis, it must be that $\vec{v}_0 = v_0\,\hat{y}$.

Position

To goal is get hold of a position equation $r(\theta)$. To get closer, calculate the full angular momentum vector:

$$\begin{aligned}\vec{L} &= m_*\vec{r}\times\vec{v}\\ &= m_*\vec{r}\times\left(G\frac{m_1m_2}{L}\,\hat{\theta}+v_0\,\hat{y}\right)\\ &= m_*G\frac{m_1m_2}{L}r\left(\hat{r}\times\hat{\theta}\right)+m_*v_0r\left(\hat{r}\times\hat{y}\right)\end{aligned}$$

To handle the cross products, note that

$$\begin{aligned}\left|\hat{r}\times\hat{\theta}\right| &= 1\\ \left|\hat{r}\times\hat{y}\right| &= \left|\cos(\theta)\right|\ ,\end{aligned}$$

and we can work with just magnitudes:

$$L = m_*G\frac{m_1m_2}{L}r + m_*v_0r\cos(\theta)$$

To help simplify this, recall the proportionality factor γ that preceded G

$$\gamma = \frac{L^2}{m_*m_1m_2r_0}\ ,$$

and work to isolate r:

$$\frac{\gamma r_0}{G} = r\left(1+\frac{\gamma r_0}{G}\frac{m_*v_0}{L}\cos(\theta)\right)$$

The combination $\gamma r_0/G$ is another characteristic length which we'll call R_0:

$$R_0 = \frac{r_0\gamma}{G}$$

Solving for r finally gives the result

$$r(\theta) = \frac{R_0}{1+(m_*R_0v_0/L)\cos(\theta)}\ .$$

With $r(\theta)$ known, the position vector is straightforwardly written:

$$\vec{r} = r(\theta)\,\hat{r}$$

Eccentricity

Comparing the above to the general form of a conic section in polar coordinates, we pick out the eccentricity to be

$$e = \frac{m_*R_0v_0}{L}\ .$$

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:

$$e = \frac{v_0L}{Gm_1m_2}$$

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

$$\frac{Gm_1m_2}{L} = \frac{v_0}{e}\ ,$$

which is only defined for non-circular orbits. With this, we have:

$$\begin{aligned}\vec{v} &= v_0\left(\frac{\hat{\theta}}{e}+\hat{y}\right)\\ \vec{a} &= \frac{-v_0}{e}\frac{d\theta}{dt}\,\hat{r}\end{aligned}$$

10.6 Runge-Lorenz Vector

The two-body problem exhibits conservation of angular momentum via the constant vector $\vec{L}$. There is, in fact, another constant vector of motion lurking about called the *Runge-Lorenz* vector

$$\vec{Z} = \vec{v}\times\vec{L} - Gm_1m_2\,\hat{r}\ .$$

Constant of Motion

Take a time derivative to prove $\vec{Z}$ is constant:

$$\begin{aligned}\frac{d}{dt}\vec{Z} &= \frac{d}{dt}\left(\vec{v}\times\vec{L}\right) - Gm_1m_2\frac{d\hat{r}}{dt}\\ &= \frac{d\vec{v}}{dt}\times\vec{L} + \vec{v}\times\cancel{\frac{d\vec{L}}{dt}} - Gm_1m_2\frac{d\hat{r}}{dt}\end{aligned}$$

Keep simplifying with

$$\frac{d\vec{v}}{dt} = \frac{1}{m_*}\vec{F} = -G\frac{m_1m_2}{m_*r^2}\,\hat{r}\ ,$$

and also with $\vec{L} = m_*\vec{r}\times\vec{v}$, so we have

$$\frac{d}{dt}\vec{Z} = Gm_1m_2\left(-\frac{\hat{r}\times(\vec{r}\times\vec{v})}{r^2} - \frac{d\hat{r}}{dt}\right)\ .$$

Replace $\vec{v}$ with its polar expression and note that

$$\vec{r} \times \vec{v} = \vec{r} \times \left(\frac{dr}{dt}\, \hat{r} + r \frac{d\theta}{dt}\, \hat{\theta} \right) = r^2 \frac{d\theta}{dt} \left(\hat{r} \times \hat{\theta} \right) ,$$

and furthermore, using the BAC-CAB formula:

$$\hat{r} \times (\vec{r} \times \vec{v}) = r^2 \frac{d\theta}{dt}\, \hat{r} \times \left(\hat{r} \times \hat{\theta} \right) = -r^2 \frac{d\theta}{dt}\, \hat{\theta}$$

Summarizing, we find

$$\frac{d}{dt} \vec{Z} = G m_1 m_2 \left(\frac{d\theta}{dt}\, \hat{\theta} - \frac{d\hat{r}}{dt} \right) = 0$$

as proposed.

Perigee

With $\vec{Z}$ known to be constant, we're free to evaluate it at any point along the trajectory. Choose a point $\vec{r}_p = r_p\, \hat{x}$ that has $\vec{v}_p \cdot \vec{r}_p = 0$, called a *perigee*:

$$\begin{aligned} \vec{Z} &= \vec{v}_p \times \vec{L} - G m_1 m_2\, \hat{x} \\ &= \vec{v}_p \times (m_* r_p\, \hat{x} \times \vec{v}_p) - G m_1 m_2\, \hat{x} \\ &= \left(m_* r_p v_p^2 - G m_1 m_2 \right) \hat{x} \end{aligned}$$

At the perigee, the velocity 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 $\vec{Z}$ becomes

$$\vec{Z} = \left(\frac{L^2}{m_* r_p} - G m_1 m_2 \right) \hat{x} \, .$$

We can keep simplifying. Replace L^2 with the expression involving γ:

$$\vec{Z} = G m_1 m_2 \left(\frac{\gamma r_0}{G r_p} - 1 \right) \hat{x} \, .$$

Note that $\gamma r_0 / G$ is the same combination we identified as the characteristic length of a conic trajectory, namely R_0, thus

$$\vec{Z} = G m_1 m_2 \left(\frac{R_0}{r_p} - 1 \right) \hat{x} \, .$$

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 $\vec{Z}$ is:

$$\vec{Z} = G m_1 m_2\, e\, \hat{x}$$

What $\vec{Z}$ 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 7

Take derivatives of

$$r\,(\theta) = \frac{R_0}{1 + e \cos(\theta)}$$

to verify the locations of the perigee and apogee.

Problem 8

Show that:

$$e = \left| \frac{r_p - r_a}{r_p + r_a} \right|$$

Conic Trajectory

The Runge-Lorenz vector

$$\vec{Z} = \vec{v} \times \vec{L} - G m_1 m_2\, \hat{r} \, ,$$

together with its particular expression

$$\vec{Z} = G m_1 m_2\, e\, \hat{x}$$

can be used together to quickly recover the polar equation for conic sections by projecting the position vector across the equation and simplifying:

$$\begin{aligned} \vec{r} \cdot \vec{Z} &= \vec{r} \cdot \left(\vec{v} \times \vec{L} \right) - G m_1 m_2\, \vec{r} \cdot \hat{r} \\ r Z \cos(\theta) &= \vec{L} \cdot (\vec{r} \times \vec{v}) - G m_1 m_2 r \\ G m_1 m_2 r e \cos(\theta) &= \frac{L^2}{m_*} - G m_1 m_2 r \end{aligned}$$

Now solve for $r(\theta)$ and simplify more:

$$\begin{aligned} r(\theta) &= \left(\frac{L^2}{Gm_1m_2m_*}\right)\frac{1}{1+e\cos(\theta)} \\ &= \left(\frac{\gamma r_0}{G}\right)\frac{1}{1+e\cos(\theta)} \\ &= \frac{R_0}{1+e\cos(\theta)} \end{aligned}$$

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

$$a = \frac{R_0}{1-e^2}\,,$$

and also

$$b = \frac{R_0}{\sqrt{1-e^2}}\,.$$

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

$$\begin{aligned} b &= r_*\sin(\theta_*) \\ r_* &= \sqrt{e^2a^2+b^2} \end{aligned}$$

is enough to finish the job. Note that similar relationships can be drawn for hyperbolic orbits.

Problem 9

Show that $\vec{r}\cdot\vec{v} = 0$ is true only at the apogee and perigee.

Dimensionless Runge-Lorenz

The Runge-Lorenz vector can be made into a dimensionless vector $\vec{\epsilon}$ by dividing Gm_1m_2 across the whole equation

$$\vec{\epsilon} = \frac{\vec{v}\times\vec{L}}{Gm_1m_2} - \hat{r}\,,$$

where by the properties of $\vec{Z}$, we also know

$$\vec{\epsilon} = e\,\hat{x}\,.$$

With this setup, write

$$\hat{r} + e\,\hat{x} = \frac{\vec{v}\times\vec{L}}{Gm_1m_2}\,,$$

and then project $\vec{r}$ into each side to recover the equation of a conic section:

$$r(1+e\cos(\theta)) = \frac{\vec{r}\cdot\left(\vec{v}\times\vec{L}\right)}{Gm_1m_2} = R_0$$

10.7 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 10

For a body moving on a path $r = f(\theta)$ obeying Kepler's second law, show that the acceleration is:

$$\vec{a} = \frac{L^2}{m_* r^3}\left(\frac{f''(\theta)}{f(\theta)} - 2\left(\frac{f'(\theta)}{f(\theta)}\right)^2 - 1\right)\hat{r}$$

Problem 11

Show that Kepler's second law works for straight-line motion.

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 $L^2 = Gm_* r_0$:

$$T = 2\pi a^2\sqrt{1-e^2}\,\frac{\sqrt{m_*}}{\sqrt{Gr_0}}$$

To deal with the r_0 term, recall two identities previously used

$$R_0 = \gamma r_0/G$$
$$a = R_0/\left(1-e^2\right)\,,$$

and reason that

$$\sqrt{r_0} = \sqrt{a}\sqrt{1-e^2}\sqrt{m_1 m_2}\,.$$

The period is, after simplifying,

$$T = \frac{2\pi a^{3/2}}{\sqrt{G\left(m_1+m_2\right)}}\,.$$

10.8 Energy Considerations

With the 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:

$$\vec{F} = -\frac{Gm_1 m_2}{r^2}\,\hat{r}$$

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:

$$\vec{F}(\vec{r}) = -\frac{d}{dr}\left(U(r)\right)\hat{r}$$

This is just like the one-dimensional Newton's law $F = -dU/dx$, except the force is a vector, balanced by $\hat{r}$ on the right.

To solve for $U(r)$, project $\hat{r}$ into both sides of the above to get

$$\frac{Gm_1m_2}{r^2} = \frac{d}{dr}\left(U(r)\right) ,$$

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 $d\vec{r}$ to get

$$\int_{r_0}^{r_1} \vec{F}(r) \cdot d\vec{r} = -\int_{r_0}^{r_1} \frac{d}{dr} U(r)\, \hat{r} \cdot d\vec{r} ,$$

where the integral on the right is redundant to the derivative, leaving $U(r)$ evaluated at the endpoints:

$$\int_{r_0}^{r_1} \vec{F}(r) \cdot d\vec{r} = -\left(U(r_1) - U(r_0)\right)$$

Set r_0 to infinity to recover the previous form.

Kinetic Energy

Containing two objects in total, the kinetic energy 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

$$\vec{v} = \vec{v}_1 - \vec{v}_2 .$$

Working out the algebra for this is left as an exercise, but the effort results in

$$T = \frac{1}{2}m_*v^2 + \frac{1}{2}\left(m_1 + m_2\right)V_0^2 ,$$

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

$$\vec{F} = -\frac{Gm_1m_2}{r^2}\,\hat{r} ,$$

and project the velocity vector into each side:

$$\vec{v} \cdot \vec{F} = -\frac{Gm_1m_2}{r^2}\left(\vec{v} \cdot \hat{r}\right)$$

Replace $\vec{F}$ on the left and $\vec{v}$ on the right

$$m_*\left(\vec{v} \cdot \frac{d\vec{v}}{dt}\right) = -\frac{Gm_1m_2}{r^2}\left(\frac{dr}{dt}\,\hat{r} + r\frac{d\theta}{dt}\,\hat{\theta}\right) \cdot \hat{r} ,$$

which simplifies to

$$\frac{1}{2}m_*\frac{d}{dt}\left(\vec{v} \cdot \vec{v}\right) = -\frac{Gm_1m_2}{r^2}\frac{dr}{dt} .$$

(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

$$\frac{d}{dt}\left(T + U\right) = 0 ,$$

as expected.

The Apocalypse Problem

If a planet were suddenly stopped in its orbit, supposed circular, it would fall into the sun in a time which is $\sqrt{2}/8$ times the period of the planet's revolution.

To prove this, begin with the total energy of the system

$$-\frac{Gm_1m_2}{a} = \frac{1}{2}m_*\left(\frac{dr}{dt}\right)^2 - \frac{Gm_1m_2}{r\left(t\right)} ,$$

where a is the radius of the orbit. Solve for dr/dt to get

$$\frac{dr}{dt} = \sqrt{\frac{2G\left(m_1 + m_2\right)}{a}}\sqrt{\frac{a}{r} - 1} ,$$

which can be separated into two equal integrals:

$$\int_a^0 \frac{dr}{\sqrt{a/r - 1}} = \sqrt{\frac{2G\left(m_1 + m_2\right)}{a}}\int_0^{t_*} dt ,$$

where t_* is the answer we're after.

To solve the r-integral, choose the peculiar substitution

$$r = a\cos^2\left(\theta\right)$$
$$dr = -2a\cos\left(\theta\right)\sin\left(\theta\right)d\theta ,$$

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.

10.9 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

$$\vec{F}(r) = -\frac{Gm_0 M}{R^3}\,\vec{r}\,.$$

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

$$\lim_{r\to\infty} U(r) = 0$$

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:

$$U(r_1) = -\int_{\infty}^{R}\vec{F}_{\rm out}\cdot d\vec{r} - \int_{R}^{r_1}\vec{F}_{\rm in}\cdot d\vec{r}\,,$$

where $\vec{F}_{\rm out}$, $\vec{F}_{\rm int}$ 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

$$U(R) = -\frac{Gm_0 M}{R}\,.$$

10.10 Energy and Orbit

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

$$\frac{1}{2}m_* v_0^2 - \frac{Gm_1 m_2}{d} = \frac{1}{2}m_* v^2 - \frac{Gm_1 m_2}{r} = 0\,,$$

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:

$$v_e = \sqrt{\frac{2Gm_1 m_2}{m_* d}} = \sqrt{\frac{2G(m_1+m_2)}{d}}$$

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

$$\vec{v} = v_0 \left(\hat{\theta} + \hat{j}\right)\,,$$

which means

$$v^2 = \vec{v}\cdot\vec{v} = 2v_0^2\frac{R_0}{r}\,.$$

Using the escape velocity in place of v allows us to write

$$\frac{2G\left(m_1+m_2\right)}{r} = 2v_0^2\frac{R_0}{r}\,,$$

or

$$v_0^2 = \frac{G\left(m_1+m_2\right)}{R_0}\,.$$

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

$$\vec{v} = \frac{v_0}{e}\left(\hat{\theta} + e\,\hat{y}\right)\,,$$

which excludes the case of circles. Proceeding carefully, find

$$\begin{aligned} v^2 &= \frac{v_0^2}{e^2}\left(\frac{2R_0}{r} - 1 + e^2\right) \\ &= \frac{2Gm_1m_2}{m_*r} - \frac{Gm_1m_2}{m_*R_0}\left(1-e^2\right)\,, \end{aligned}$$

so the kinetic term is

$$E_{kin} = \frac{Gm_1m_2}{r} - \frac{Gm_1m_2}{2R_0}\left(1-e^2\right)\,.$$

The total energy sums the potential plus the kinetic, which happens to contain equal and opposite $1/r$-like terms, leaving just the constant:

$$E = \frac{-Gm_1m_2}{2R_0}\left(1-e^2\right) = \frac{-Gm_1m_2}{2a}\,,$$

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

$$\vec{v} = \frac{Gm_1m_2}{L}\,\hat{\theta}\,,$$

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

$$L = m_*\sqrt{G\left(m_1+m_2\right)R}\,,$$

and then the square of the velocity is:

$$v^2 = \frac{G\left(m_1+m_2\right)}{R}$$

The time derivative of $\vec{v}$ gives the a familiar equation for the acceleration

$$\vec{a} = -\frac{Gm_1m_2}{L}\frac{d\theta}{dt}\,\hat{r}\,,$$

which for circular orbits simplifies to

$$\vec{a} = \frac{-v^2}{R}\,\hat{r}\,,$$

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 $\vec{L}$ using its definition:

$$\vec{Z} = \vec{v} \times (m_*\vec{r} \times \vec{v}) - Gm_1m_2\,\hat{r}\,,$$

and square the whole equation:

$$\begin{aligned}\vec{Z}\cdot\vec{Z} = {} & |\vec{v} \times (m_*\vec{r} \times \hat{v})|^2 \\ & - 2Gm_1m_2\,\vec{v} \times (m_*\vec{r} \times \hat{v})\cdot\hat{r} + G^2m_1^2m_2^2\end{aligned}$$

For the first term on the right, notice $\vec{v}$ is perpendicular to $\vec{r} \times \vec{v}$, so

$$|\vec{v} \times (m_*\vec{r} \times \hat{v})| = m_*rv^2\,|\sin(\phi)|\,,$$

where ϕ is the angle between $\vec{r}$ and $\vec{v}$.

For the second term, the scalar triple product can be rewritten

$$\vec{v} \times (m_*\vec{r} \times \hat{v})\cdot\hat{r} = m_*\,(\vec{r} \times \vec{v})\cdot(\hat{r} \times \vec{v})\,.$$

The remaining vectors are parallel and the whole quantity simplifies to

$$\vec{v} \times (m_*\vec{r} \times \hat{v})\cdot\hat{r} = m_*rv^2\sin^2(\phi)\,.$$

Rewriting $\vec{Z}\cdot\vec{Z}$ with this in mind, we have

$$\begin{aligned}Z^2 = {} & \left(m_*rv^2\right)^2\sin^2(\phi) \\ & - 2Gm_1m_2m_*rv^2\sin^2(\phi) + G^2m_1^2m_2^2\,,\end{aligned}$$

or

$$\frac{Z^2}{G^2m_1^2m_2^2} = 1 + \sin^2(\phi)\left(q^2 - 2q\right)\,,$$

where

$$q = \frac{m_*rv^2}{Gm_1m_2}\,.$$

simplifying this further gives

$$\frac{Z^2}{G^2m_1^2m_2^2} = \cos^2(\phi) + \sin^2(\phi)\,(1-q)^2$$

Finally, note that the left size is actually the square of the eccentricity, giving, after restoring q:

$$e^2 = \cos^2(\phi) + \sin^2(\phi)\left(1 - \frac{rv^2}{G\,(m_1+m_2)}\right)^2$$

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 $\phi = \pi/2$.

10.11 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

$$\vec{F}_g = -mg\,\hat{y}\,,$$

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:

$$g = \frac{9.8\,m}{s^2}$$

On the other hand, we just went through all the pains of showing that the gravitational force is

$$\vec{F}(\vec{r}) = -\frac{Gm_1m_2}{r^2}\,\hat{r}$$

with potential energy

$$U(r) = -\frac{Gm_1m_2}{r}\,.$$

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 $\hat{r} = \hat{y}$. 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:

$$\vec{F}(y) = -\frac{Gm_1m_2}{(R+y)^2}\,\hat{y}$$
$$U(y) = -\frac{Gm_1m_2}{(R+y)}$$

Next apply binomial expansion to each denominator, particularly:

$$(R+y)^{-2} \approx \frac{1}{R^2} - \frac{2y}{R^3} + \frac{3y^2}{R^4} - \cdots$$
$$(R+y)^{-1} \approx \frac{1}{R} - \frac{y}{R^2} + \frac{y^2}{R^3} - \cdots$$

To first order, the above equations become

$$\vec{F}(y) \approx -\frac{Gm_1m_2}{R^2}\left(1 - \frac{2y}{R}\right)\hat{y}$$
$$U(y) \approx -\frac{Gm_1m_2}{R}\left(1 - \frac{y}{R}\right).$$

We want the force equation to be constant, and it just happens that the quantity $2y/R$ is negligible, so the effective force at the surface is

$$\vec{F}_g = -\frac{Gm_1m_2}{R^2}\,\hat{y}$$

This tells where g comes from:

$$g = \frac{Gm_{\text{Earth}}}{R^2_{\text{Earth}}}.$$

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 just to recover the form mgy, but also the first derivative must equal a constant, which is what we asked of the force.

11 Three Dimensions

11.1 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 14.1.

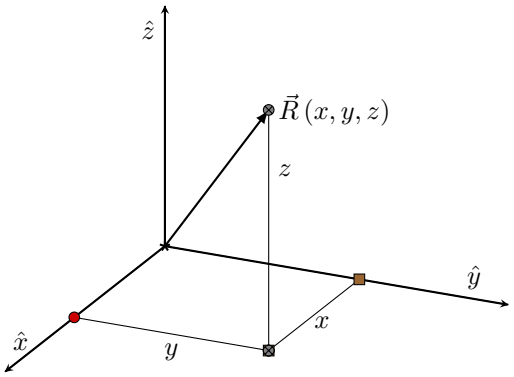


Figure 14.1: Cartesian coordinate system.

The position vector becomes

$$\vec{R} = x\,\hat{x} + y\,\hat{y} + z\,\hat{z},$$

which requires three mutually-perpendicular basis vectors:

$$\hat{x} = \hat{\imath} = \hat{e}_x = \hat{e}_i = \langle 1,0,0\rangle$$
$$\hat{y} = \hat{\jmath} = \hat{e}_y = \hat{e}_j = \langle 0,1,0\rangle$$
$$\hat{z} = \hat{k} = \hat{e}_z = \hat{e}_k = \langle 0,0,1\rangle$$

Quantities like velocity and acceleration simply take on a new z-component:

$$\vec{v} = \frac{dx}{dt}\hat{x} + \frac{dy}{dt}\hat{y} + \frac{dz}{dt}\hat{z}$$
$$\vec{a} = \frac{d^2x}{dt^2}\hat{x} + \frac{d^2y}{dt^2}\hat{y} + \frac{d^2z}{dt^2}\hat{z}$$

The same goes for the differential line element

$$d\vec{S} = dx\,\hat{x} + dy\,\hat{y} + dz\,\hat{z}$$

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*:

$$dV = dx\, dy\, dz$$

11.2 Rotations

...

11.3 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 $\vec{P}$ to hold the coefficients

$$\vec{P} = a\,\hat{x} + b\,\hat{y} + c\,\hat{z}\,,$$

and project this into the position vector

$$\vec{R} = x\,\hat{x} + y\,\hat{y} + z\,\hat{z}$$

to find

$$\vec{R}\cdot\vec{P} = -d\,.$$

11.4 Normal Vector to a Plane

Consider any given point $\vec{R}_0$ that is known to be in the plane. For any point $\vec{R}$ that is also in the plane, the difference

$$\Delta\vec{R} = \vec{R} - \vec{R}_0$$

is a tangent vector that stays embedded in the plane.

This is enough to define the notion of the normal vector to the plane $\vec{N}$ such that

$$\Delta\vec{R}\cdot\vec{N} = 0\,.$$

Peel apart the $\Delta\vec{R}$ term to write:

$$\vec{R}\cdot\vec{N} = \vec{R}_0\cdot\vec{N}$$

The above looks very much like the equation of the plane when written

$$\vec{R}\cdot\vec{P} = -d\,.$$

Subtract the two equations to get

$$\vec{R}\cdot\left(\vec{N} - \vec{P}\right) = \vec{R}_0\cdot\vec{N} + d\,,$$

and then notice d is also equivalent to $-\vec{R}_0\cdot\vec{P}$. The above becomes:

$$\vec{R}\cdot\left(\vec{N} - \vec{P}\right) = \vec{R}_0\cdot\left(\vec{N} - \vec{P}\right)$$

This result needs to be true for all $\vec{R}$, which can only mean the parenthesized quantities are zero, which means $\vec{N} = \vec{P}$. The normal vector is the list of coefficients on x, y, and z:

$$\vec{N} = a\,\hat{x} + b\,\hat{y} + c\,\hat{z}$$

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:

$$\vec{L} = \vec{N}_1\times\vec{N}_2 = \begin{vmatrix}\hat{x} & \hat{y} & \hat{z}\\ a & b & c\\ A & B & C\end{vmatrix} = \langle L_x, L_y, L_z\rangle$$

The vector $\vec{L}$ indicates the intersection of each plane. To establish this, note that the following pair of equations must be true:

$$\vec{N}_1\cdot\vec{L} = 0$$
$$\vec{N}_2\cdot\vec{L} = 0$$

Substitute the proposed form for $\vec{L}$,

$$\vec{N}_1 \cdot \left(\vec{N}_1 \times \vec{N}_2 \right) = 0$$
$$\vec{N}_2 \cdot \left(\vec{N}_1 \times \vec{N}_2 \right) = 0 \,,$$

and notice each as a triple product to finish the job:

$$\vec{N}_2 \cdot \left(\vec{N}_1 \times \vec{N}_1 \right) = 0$$
$$\vec{N}_1 \cdot \left(\vec{N}_2 \times \vec{N}_2 \right) = 0$$

11.5 Point Intersecting Triangle

Consider three points in Cartesian space that mark the vertices of a triangle

$$\vec{P}_j = x_j \, \hat{x} + y_j \, \hat{y} + z_j \, \hat{z} \,,$$

where $j = 1, 2, 3$. Also consider a fourth point in space

$$\vec{Q} = Q_x \, \hat{x} + Q_y \, \hat{y} + Q_j \, \hat{z}$$

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 $\vec{U}$, $\vec{V}$ such that

$$\vec{U} = \vec{P}_2 - \vec{P}_1$$
$$\vec{V} = \vec{P}_3 - \vec{P}_1 \,.$$

Since both vectors are in the same plane, their cross product will by definition point perpendicular to the plane, which is a normal vector:

$$\vec{N} = \vec{U} \times \vec{V}$$

The details of the cross product are straightforward and aren't necessary to spell out here.

Improvising a Coordinate System

The pair of vectors $\vec{U}$, $\vec{V}$, 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 $\vec{p}$ in the plane cab ne expressed as a linear combination of the basis vectors

$$\vec{p} = \alpha \, \vec{U} + \beta \, \vec{V} \,,$$

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

$$\vec{R} = \alpha \, \vec{U} + \beta \, \vec{V} + \gamma \, \vec{N}$$

using a third parameter γ.

In this improvised coordinate system, the point $\vec{P}_1$ is considered to be the origin. That is, if all parameters are zero, points $\vec{p}$ and $\vec{R}$ land back at $\vec{P}_1$.

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 $\vec{Q}$, despite being handed to us in Cartesian coordinates, can also be represented by the vector $\vec{R}$ for some choice of α, β, γ. If $\vec{Q}$ is in the same plane as the three points $\vec{P}_j$ provided, then the γ-parameter should be zero.

Let us propose that point $\vec{Q}$ is located in the *normalized* $\hat{U}$, $\hat{V}$, $\hat{N}$ system with new coefficients α, β, γ (that are different that those that occur in the non-normalized version). Tracing from the origin, we have

$$\vec{Q} = \vec{P}_1 + \alpha \, \hat{U} + \beta \, \hat{V} + \gamma \, \hat{N} \,,$$

and project $\hat{N}$ into both sides to get

$$\vec{Q} \cdot \hat{N} = \vec{P}_1 \cdot \vec{N} + \alpha \cancel{\hat{U} \cdot \hat{N}} + \beta \cancel{\hat{V} \cdot \hat{N}} + \gamma \, \hat{N} \cdot \hat{N} \,,$$

and this lets us solve for gamma, which ought to be zero if $\vec{Q}$ is within the triangle:

$$\gamma = \hat{N} \cdot \left(\vec{Q} - \vec{P}_1 \right)$$

Containment Condition

Supposing point $\vec{Q}$ satisfies $\gamma = 0$, the job now is to figure out whether $\vec{Q}$ is inside or outside the boundaries of the triangle defined by $\vec{P}_j$. Define a third tangent vector $\vec{W}$ that satisfies

$$\vec{U} + \vec{V} + \vec{W} = 0 \,,$$

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 $\vec{Q}$ to be inside the triangle, $\vec{Q}$ must occur to the 'left' of the line formed by each vector $\vec{U}$, $\vec{V}$, $\vec{W}$ simultaneously. (We say 'left' and not 'right' due to the counterclockwise flow of the tangent vectors.)

To see this, write out

$$\vec{U} \times \left(\vec{Q} - \vec{P}_1 \right) = \delta_U \hat{N}$$

for some parameter δ_U. Since $\vec{Q} - \vec{P}_1$ also lies in the plane, the above simplifies

$$U \left| \vec{Q} - \vec{P}_1 \right| \sin(\phi) = \delta_U \, ,$$

where ϕ is the angle formed between $\vec{U}$ and $\vec{Q} - \vec{P}_1$.

If $\vec{Q}$ is to the 'left' of the line made by $\vec{U}$ as described, then $\sin(\phi)$ is a positive quantity, and δ_U is also positive. Repeat for the V- and W-cases to decide if the point is inside the triangle. All δ-parameters must be positive.

11.6 Plane from Three Points

Three points not on the same line define an infinite plane. From the points

$$\vec{P}_1 = 1\,\hat{x} + 0\,\hat{y} + 1\,\hat{z}$$
$$\vec{P}_2 = 0\,\hat{x} + 1\,\hat{y} + 1\,\hat{z}$$
$$\vec{P}_3 = 1\,\hat{x} + 1\,\hat{y} + 0\,\hat{z} \, ,$$

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 $\vec{Q}$ coplanar to the triangle. Proceed by defining a pair of vectors $\vec{U}$, $\vec{V}$:

$$\vec{U} = \vec{P}_2 - \vec{P}_1 = -1\,\hat{x} + 1\,\hat{y} + 0\,\hat{z}$$
$$\vec{V} = \vec{P}_3 - \vec{P}_1 = 0\,\hat{x} + 1\,\hat{y} - 1\,\hat{z}$$

For a change of taste, calculate the normal vector using determinant/bracket notation:

$$\vec{N} = \vec{U} \times \vec{V} = \begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ -1 & 1 & 0 \\ 0 & 1 & -1 \end{vmatrix} = \langle -1, -1, -1 \rangle$$

In other words, the normal vector is

$$\vec{N} = -1\,\hat{x} - 1\,\hat{y} - 1\,\hat{z} \, .$$

Next take any point known to be on the plane, such as $\vec{P}_1$, and propose the vector $\vec{Q}$ is always in the plane if

$$\left(\vec{Q} - \vec{P}_1 \right) \cdot \vec{N} = 0$$

is satisfied. In particular

$$((x-1)\,\hat{x} + (y-0)\,\hat{y} + (z-1)) \cdot \hat{N} = 0 \, ,$$

or, after simplifying:

$$x + y + z - 2 = 0$$

12 Non-Cartesian Coordinates

12.1 Cylindrical Coordinates

Another way of mapping the three-dimensional Cartesian space is with *cylindrical coordinates*. In this system, illustrated in Figure 14.2, 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.

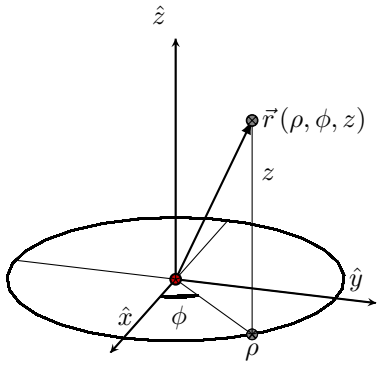


Figure 14.2: Cylindrical coordinate system.

Cylindrical Position Vector

In cylindrical coordinates, the position vector is:

$$\vec{r}\,(\rho, \phi, z) = \rho\,\hat{\rho} + z\,\hat{z} \, ,$$

or in terms of Cartesian directions,

$$\vec{r} = \rho\cos(\phi)\,\hat{x} + \rho\sin(\phi)\,\hat{y} + z\,\hat{z} \, .$$

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:

$$\hat{\rho} = \cos(\phi)\,\hat{x} + \sin(\phi)\,\hat{y}$$
$$\hat{\phi} = -\sin(\phi)\,\hat{x} + \cos(\phi)\,\hat{y}$$

The derivatives follow the familiar pattern:

$$\frac{d\hat{\rho}}{d\phi} = \hat{\phi}$$
$$\frac{d\hat{\phi}}{d\phi} = -\hat{\rho}$$

Cylindrical Line Element

The differential version of the cylindrical position vector $d\vec{S} = d\vec{r}$ gives the *cylindrical line element*:

$$d\vec{S} = d\rho\,\hat{\rho} + \rho\,d\phi\,\hat{\phi} + dz\,\hat{z}$$

Differential Interval

The differential interval in cylindrical coordinates is straightforwardly calculated from the line interval:

$$dS^2 = d\vec{S}\cdot d\vec{S} = d\rho^2 + \rho^2 d\phi^2 + dz^2$$

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:

$$dA_\rho = \left|\rho\,d\phi\,\hat{\phi}\times dz\,\hat{z}\right|$$
$$dA_\phi = \left|d\rho\,\hat{\phi}\times dz\,\hat{z}\right|$$
$$dA_z = \left|d\rho\,\hat{\rho}\times\rho\,d\phi\,\hat{\phi}\right|$$

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

$$dV = d\rho\,\hat{\rho}\cdot\left(\rho\,d\phi\,\hat{\phi}\times dz\,\hat{z}\right)$$

or any of its permutations.

12.2 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.

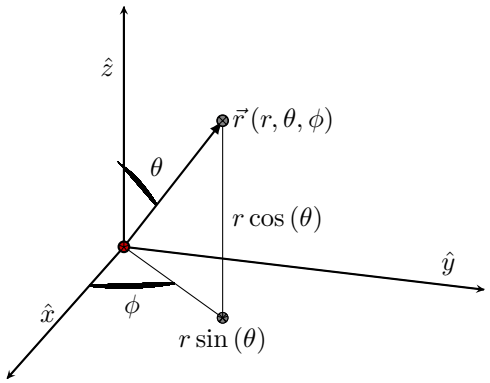


Figure 14.3: Spherical coordinate system.

Spherical Position Vector

In spherical coordinates, the position vector is

$$\begin{aligned}\vec{r}\,(r,\theta,\phi) = {} & r\sin(\theta)\cos(\phi)\,\hat{x}\\ & + r\sin(\theta)\sin(\phi)\,\hat{y}\\ & + r\cos(\theta)\,\hat{z}\ ,\end{aligned}$$

which can be discerned by analyzing Figure 14.3. 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

$$\sqrt{\vec{r}\cdot\vec{r}} = \sqrt{r^2} = r$$

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

$$\vec{r} = r\,\hat{r}\,(r,\theta,\phi)\ ,$$

which tells us the radial basis vector:

$$\hat{r} = \frac{\vec{r}}{r} = \langle\sin(\theta)\cos(\phi)\,,\sin(\theta)\sin(\phi)\,,\cos(\theta)\rangle$$

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 $\hat{r}$, we can quickly write

$$\begin{aligned}\hat{\theta}(\theta,\phi) = &\sin\left(\theta+\frac{\pi}{2}\right)\cos(\phi)\,\hat{x}\\ &+\sin\left(\theta+\frac{\pi}{2}\right)\sin(\phi)\,\hat{y}\\ &+\cos\left(\theta+\frac{\pi}{2}\right)\hat{z}\,,\end{aligned}$$

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 $\hat{r}$, $\hat{\theta}$, $\hat{\phi}$ are easily shown to have unit length

$$\hat{r}\cdot\hat{r} = \hat{\phi}\cdot\hat{\phi} = \hat{\theta}\cdot\hat{\theta} = 1\,,$$

and to be mutually perpendicular

$$\begin{aligned}\hat{r}\cdot\hat{\theta} = \hat{r}\cdot\hat{\phi} &= 0\\ \hat{\theta}\cdot\hat{\phi} = \hat{\theta}\cdot\hat{r} &= 0\\ \hat{\phi}\cdot\hat{r} = \hat{\phi}\cdot\hat{\theta} &= 0\,,\end{aligned}$$

and subordinate to the cross product:

$$\begin{aligned}\hat{r}\times\hat{\theta} &= \hat{\phi}\\ \hat{\theta}\times\hat{\phi} &= \hat{r}\\ \hat{\phi}\times\hat{r} &= \hat{\theta}\end{aligned}$$

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 $R_{r\theta\phi}$:

$$\begin{bmatrix}\hat{x}\\ \hat{y}\\ \hat{z}\end{bmatrix} = R^{-1}_{r\theta\phi}\begin{bmatrix}\hat{r}\\ \hat{\theta}\\ \hat{\phi}\end{bmatrix}$$

Leaving the details to the reader, the required matrix is:

$$R^{-1}_{r\theta\phi} = \begin{bmatrix}\sin(\theta)\cos(\phi) & \cos(\theta)\cos(\phi) & -\sin(\phi)\\ \sin(\theta)\sin(\phi) & \cos(\theta)\sin(\phi) & \cos(\phi)\\ \cos(\theta) & -\sin(\theta) & 0\end{bmatrix}$$

Spherical Velocity Vector

The time derivative of $\vec{r} = r\,\hat{r}\,(r,\theta,\phi)$ gives the velocity vector in spherical coordinates. Doing this carefully, one should find:

$$\vec{v}(t) = \frac{dr}{dt}\,\hat{r} + r\frac{d\theta}{dt}\,\hat{\theta} + r\sin(\theta)\,\frac{d\phi}{dt}\,\hat{\phi}$$

Eliminating the dt-term by the chain rule gives the differential line element

$$d\vec{S} = dr\,\hat{r} + r\,d\theta\,\hat{\theta} + r\sin(\theta)\,d\phi\,\hat{\phi}\,,$$

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

$$dA_r = \left| r\, d\theta\, \hat{\theta} \times r \sin(\theta)\, d\phi\, \hat{\phi} \right| = r^2 \sin(\theta)\, d\theta\, d\phi\,.$$

The other two area elements come from fixing θ, ϕ respectively:

$$dA_\theta = \left| dr\, \hat{r} \times r \sin(\theta)\, d\phi\, \hat{\phi} \right| = r \sin(\theta)\, dr\, d\phi$$
$$dA_\phi = \left| dr\, \hat{r} \times r d\theta\, \hat{\theta} \right| = r\, dr\, d\theta$$

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

$$dV = dr\, \hat{r} \cdot \left(r\, d\theta\, \hat{\theta} \times r \sin(\theta)\, d\phi\, \hat{\phi} \right)$$

or any of its permutations.

13 Curves in Three Dimensions

13.1 Generalizing Plane Curves

In three dimensions, the position vector representing a parametric curve can be written

$$\vec{r}(t) = x(t)\,\hat{x} + y(t)\,\hat{y} + z(t)\,\hat{z}\,.$$

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:

$$\vec{v}(t) = \left(\frac{d}{dt}x(t)\right)\hat{x} + \left(\frac{d}{dt}y(t)\right)\hat{y} + \left(\frac{d}{dt}z(t)\right)\hat{z}$$
$$\vec{a}(t) = \left(\frac{d^2}{dt^2}x(t)\right)\hat{x} + \left(\frac{d^2}{dt^2}y(t)\right)\hat{y} + \left(\frac{d^2}{dt^2}z(t)\right)\hat{z}$$

Tangent Vector

The (unit) tangent vector $\hat{T}$ is given by the velocity vector divided by the speed

$$\hat{T} = \frac{\vec{v}}{v}\,,$$

and the velocity vector can also be written in terms of the speed and the tangent vector:

$$\vec{v}(t) = v\,\hat{T}$$

Another equation for the tangent vector reads

$$\hat{T} = \frac{d\vec{S}}{dS}\,,$$

where $d\vec{S}$ 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

$$\kappa = \left| \frac{d\hat{T}}{dS} \right|\,.$$

The tight formula for curvature also survives migration to three dimensions:

$$\kappa = \frac{|\vec{v} \times \vec{a}|}{v^3}$$

Normal Vector

The (unit) normal vector $\hat{N}$ also needs no modification from the two-dimensional case:

$$\hat{N} = \frac{1}{\kappa}\frac{d\hat{T}}{dS}$$

Acceleration Vector

In terms of the tangent vector, the formula for the acceleration vector remains in tact as well:

$$\vec{a}(t) = \left(\frac{d^2S}{dt^2}\right)\hat{T} + \kappa\left(\frac{dS}{dt}\right)^2\hat{N}$$

13.2 The Binormal Vector

For a curve in three dimensions, the tangent vector $\hat{T}$ and normal vector $\hat{N}$ form the basis vectors for a plane, which itself has a normal vector called the *binormal*. The binormal vector is a unit vector $\hat{B}$, and is defined by

$$\hat{B} = \hat{T} \times \hat{N}\,.$$

There are two cyclic permutations of the definition that contain the same information:

$$\hat{T} = \hat{N} \times \hat{B}$$
$$\hat{N} = \hat{B} \times \hat{T}$$

Torsion

We can learn more about $\hat{B}$ by writing two straightforward consequences of its definition, namely

$$\hat{B} \cdot \hat{T} = 0$$
$$\hat{B} \cdot \hat{B} = 1\,,$$

and apply an arc length derivative to each:

$$\frac{d\hat{B}}{dS} \cdot \hat{T} + \hat{B} \cdot \frac{d\hat{T}}{dS} = 0$$
$$2\hat{B} \cdot \frac{d\hat{B}}{dS} = 0$$

Evidently, the derivative $d\hat{B}/dS$ is perpendicular to both $\hat{T}$ and $\hat{B}$. This can only mean $d\hat{B}/dS$ is parallel to the normal vector:

$$\frac{d\hat{B}}{dS} = -\tau\,\hat{N}$$

The proportionality constant τ (Greek 'tau'), defined with a minus sign, is called the *torsion* along the curve.

13.3 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:

$$\begin{aligned}\frac{d\hat{N}}{dS} &= \frac{d}{dS}\left(\hat{B} \times \hat{T}\right) \\ &= -\tau\,\hat{N} \times \hat{T} + \kappa\,\hat{B} \times \hat{N} \\ &= \tau\hat{B} - \kappa\hat{T}\end{aligned}$$

The 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.

Chapter 15

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 are 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

2.1 Single Integral

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

$$A = \int_{x_0}^{x_1} y(x)\, dx\,.$$

2.2 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 apples 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:

$$A = \int\int_{\mathcal{D}} dx\, dy\,.$$

By obscuring the product $dx\, dy$ into an area element dA, the above can be written free of the Cartesian coordinate system:

$$A = \int\int_{\mathcal{D}} dA$$

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

$$V = \int\int_{\mathcal{D}} f(x, y)\, dx\, dy\,.$$

Integrating Functions

Supposing we need to calculate something more abstract than an area, the double integral apparatus takes any reasonable function in its integrand

$$B = \int\int_{\mathcal{D}} f(x, y)\, dx\, dy\,,$$

where $f(x, y)$ is a generalized surface.

2.3 Polar Integral

To work a specific case, the differential area element in polar coordinates reads

$$dA = r\, dr\, d\theta\,,$$

which means

$$A = \int\int_{\mathcal{D}} r\, dr\, d\theta$$

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

$$A = \int_{\theta_0}^{\theta_1} \int_0^R r \, dr \, d\theta = \frac{1}{2} \left(\theta_1 - \theta_0\right) R^2 \, ,$$

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:

$$\begin{aligned} R\cos\left(\theta_0\right) &= x_0 \\ R\sin\left(\theta_0\right) &= y_0 \\ R\cos\left(\theta_1\right) &= x_1 \\ R\sin\left(\theta_1\right) &= y_1 \end{aligned}$$

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 handled 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 chipped away at. Doing a round of simplifying, reach the intermediate step

$$\begin{aligned} A = \frac{x_1^2}{2} \left(m_1 - \cancel{m_0}\right) & \\ + \int_{x_0}^{x_1} \sqrt{R^2 - x^2} \, dx & \\ - \frac{m_0}{2} \left(x_0^2 - \cancel{y_1^2}\right) \, . & \end{aligned}$$

The remaining integral can be solved with a sine substitution, which has the general solution:

$$\begin{aligned} \int \sqrt{R^2 - x^2} \, dx = \frac{x\sqrt{R^2 - x^2}}{2} & \\ + \frac{R^2}{2} \arctan\left(\frac{x}{\sqrt{R^2 - x^2}}\right) + C & \end{aligned}$$

For the problem on hand, this means

$$\begin{aligned} \int \sqrt{R^2 - x^2} \, dx = \frac{m_0 x_0^2 - m_1 x_1^2}{2} & \\ + \frac{R^2}{2} \left(\arctan\left(\frac{x_0}{y_0}\right) - \arctan\left(\frac{x_1}{y_1}\right)\right) \, . & \end{aligned}$$

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

$$\arctan\left(\cot\left(u\right)\right) = \frac{\pi}{2} - u \, ,$$

and the area becomes

$$A = \frac{R^2}{2} \left(\frac{\pi}{2} - \theta_0 - \frac{\pi}{2} + \theta_1\right) \, ,$$

and finally,

$$A = \frac{1}{2} \left(\theta_1 - \theta_0\right) R^2 \, ,$$

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:

$$I_1^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-a\left(x^2 + y^2\right)} \, dx \, dy$$

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 is 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:

$$\frac{d}{da}\left(-I_1\right) = I_3 = -\frac{d}{da}\left(\sqrt{\frac{\pi}{a}}\right) = \sqrt{\frac{\pi}{4a^3}}$$

2.4 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:

$$V = \int\int\int_{\mathcal{D}} dx\, dy\, dz$$

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

$$B = \int\int\int_{\mathcal{D}} f(x, y, z)\, dx\, dy\, dz\,.$$

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 (a fancy word for 'obscured'):

$$V = \int\int\int_{\mathcal{D}} dV$$

Hurricane Problem

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.

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:

$$T = \frac{d_0\Omega^2}{2} \left(\frac{h}{3} \right) (2\pi) \left(\frac{3a^4}{8} \right) = \frac{\pi}{8} \Omega^2 d_0 h a^4$$

2.5 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 $\vec{D} = D\,\hat{z}$. 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 there entire sphere must be calculated. Choose any element of volume dV inside the sphere at location $\vec{r}$, which is located distance r from the center, at an angle θ from the z-axis.

Let vector $\vec{q}$ denote the line connecting $\vec{D}$ to $\vec{r}$ such that

$$\vec{r} + \vec{q} = D\,\hat{z}\ ,$$

and also let α be the angle between $\hat{z}$ and $\hat{q}$. 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 $\vec{F}$. 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:

$$F = \int_{\mathcal{D}} d\vec{F} \cdot \hat{z} = \int\int\int_{\text{volume}} dF \cos(\alpha)$$

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,

$$q\, dq = rD\sin(\theta)\, d\theta\ ,$$

and rewrite the integral with the intent of integrating over r last. Make you 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.

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

$$\frac{d}{dx} f(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \, .$$

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:

$$\frac{\partial}{\partial x} f(x_0, y) = \lim_{x \to x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} \qquad (15.1)$$

Meanwhile, the slope at the point (x, y_0) allows y to vary while x is constant:

$$\frac{\partial}{\partial y} f(x, y_0) = \lim_{y \to y_0} \frac{f(x, y) - f(x, y_0)}{y - y_0} \qquad (15.2)$$

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

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \frac{\partial}{\partial x} \left(\frac{\partial z}{\partial y} \right) \qquad (15.3)$$

is true. Using brute force, start with

$$\frac{\partial}{\partial y} \left(\frac{\partial z}{\partial x} \right) = \frac{\partial}{\partial y} \left(\lim_{x \to x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} \right) \, ,$$

which becomes

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right) = \lim_{y\to y_0}\lim_{x\to x_0} \frac{f(x,y) - f(x_0,y) - f(x,y_0) + f(x_0,y_0)}{(y-y_0)(x-x_0)} .$$

Now, it takes little to imagine doing a similar calculation with the y-partial derivative first to have

$$\frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial}{\partial x}\left(\lim_{y\to y_0}\frac{f(x,y) - f(x,y_0)}{y-y_0}\right) ,$$

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

$$\lim_{x\to x_0}\lim_{y\to y_0} \leftrightarrow \lim_{y\to y_0}\lim_{x\to x_0}$$

can be assumed.

To prove this, define two new functions

$$X(x,y) = f(x,y) - f(x_0,y)$$
$$Y(x,y) = f(x,y) - f(x,y_0) ,$$

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

$$x_0 < a < x$$
$$y_0 < b < y .$$

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 (15.3) 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 (15.3), 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:

$$\begin{aligned}\partial_{xxx} &= \partial_{x^3} \\ \partial_{xxy} &= \partial_{x^2y} \\ \partial_{yyx} &= \partial_{xy^2} \\ \partial_{yyy} &= \partial_{y^3}\end{aligned}$$

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 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 $F(x)$:

$$\begin{aligned}\frac{d}{dx}F(x) = &\frac{\partial}{\partial x}G(x,b(x)) - \frac{\partial}{\partial x}G(x,a(x)) \\ &+ \frac{\partial}{\partial y}G(x,b(x))\,\frac{db}{dx} \\ &- \frac{\partial}{\partial y}G(x,a(x))\,\frac{da}{dx}\end{aligned}$$

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 $\Delta y \to 0$, the left side is $\partial G/\partial x$. 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

$$\begin{aligned}\frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)\,dt = &\\ &f(x,b(x))\,\frac{db}{dx} - f(x,a(x))\,\frac{da}{dx} \\ &+ \frac{\partial}{\partial x}\int_{a(x)}^{b(x)} f(x,t)\,dt \qquad (15.4)\end{aligned}$$

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 1

Prove the following:

$$\frac{d}{dx}\int_0^x e^{xt^2}\,dt = e^{x^3} + \int_0^x \frac{\partial}{\partial x}e^{xt^2}\,dt$$

3.6 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

$$\Delta x = x - x_0$$
$$\Delta y = y - y_0 \, .$$

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:

$$\begin{aligned} f(x,y) \approx{}& f(x_0,y_0) \\ &+ C_{10}\Delta x + C_{01}\Delta y \\ &+ C_{20}\Delta x^2 + C_{11}\Delta x\Delta y + C_{02}\Delta y^2 \\ &+ C_{30}\Delta x^3 + C_{21}\Delta x^2\Delta y + C_{12}\Delta x\Delta y^2 + C_{03}\Delta y^3 \\ &+ \cdots \end{aligned}$$

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

$$\begin{aligned} C_{10} &= z_x \\ C_{01} &= z_y \end{aligned}$$

$$\begin{aligned} C_{20} &= z_{x^2}/2 \\ C_{11} &= z_{xy} \\ C_{02} &= z_{y^2}/2 \end{aligned}$$

$$\begin{aligned} C_{30} &= z_{x^3}/3! \\ C_{21} &= z_{x^2y}/2 \\ C_{12} &= z_{xy^2}/2 \\ C_{03} &= z_{y^3}/3! \end{aligned}$$

The two-dimensional Taylor's theorem now looks like:

$$\begin{aligned} f(x,y) \approx{}& f(x_0,y_0) \\ &+ z_x\Delta x + z_y\Delta y \\ &+ \frac{z_{x^2}\Delta x^2 + 2z_{xy}\Delta x\Delta y + z_{y^2}\Delta y^2}{2} \\ &+ \frac{z_{x^3}\Delta x^3 + 3z_{x^2y}\Delta x^2\Delta y + 3z_{xy^2}\Delta x\Delta y^2 + z_{y^3}\Delta y^3}{3!} \\ &+ \cdots \end{aligned}$$

Look for a moment at the pattern in the numerical coefficients in the numerator of each term written so far. Jotting these down:

$$\begin{array}{ccccccc} & & & 1 & & & \\ & & 1 & & 1 & & \\ & 1 & & 2 & & 1 & \\ 1 & & 3 & & 3 & & 1 \end{array}$$

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

$$\begin{aligned} z_{x^2}\Delta x^2 + 2z_{xy}\Delta x\Delta y + z_{y^2}\Delta y^2 \\ = (\Delta x\, \partial_x + \Delta y\, \partial y)^2 f(x_0, y_0) \, , \end{aligned}$$

and similarly for all orders.

Switching to summation notation, we finally have the two-dimensional Taylor's theorem

$$f(x,y) \approx f(x_0,y_0) + \sum_{q=1}^{\infty} \frac{1}{q!} (\Delta x\, \partial_x + \Delta y\, \partial y)^q f(x_0,y_0) \tag{15.5}$$

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

4.1 Basis Vectors as Derivatives

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. In other 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:

$$\vec{r} = x\,\hat{x} + y\,\hat{y} + z\,\hat{z}$$
$$\vec{r} = \rho\cos(\phi)\,\hat{x} + \rho\sin(\phi)\,\hat{y} + z\,\hat{z}$$
$$\vec{r} = r\sin(\theta)(\cos(\phi)\,\hat{x} + \sin(\phi)\,\hat{y}) + r\cos(\theta)\,\hat{z}$$

It's customary using geometry to work out 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:

$$\hat{q} = \frac{1}{|\partial\vec{r}/\partial q|}\frac{\partial\vec{r}}{\partial q} \tag{15.6}$$

For example, if we want $\hat{\theta}$ from spherical coordinates, write

$$\frac{\partial\vec{r}}{\partial\theta} = r\cos(\theta)(\cos(\phi)\,\hat{x} + \sin(\phi)\,\hat{y}) - r\sin(\theta)\,\hat{z}\,,$$

whose magnitude is r. Dividing this out delivers the result promised:

$$\frac{1}{r}\frac{\partial\vec{r}}{\partial\theta} = \hat{\theta}$$

4.2 Surface Tangent Vectors

Parametric Surface Tangents

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

$$\vec{r}(u,v) = x(u,v)\,\hat{x} + y(u,v)\,\hat{y} + z(u,v)\,\hat{z}\,,$$

which doesn't necessarily need to framed in the Cartesian system.

Choosing any fixed point (u_0, v_0) on a parameterized surface, there exist a pair of embedded tangent vectors we'll call $\vec{u}$, $\vec{v}$ straightforwardly calculated directly from $\vec{r}(u,v)$:

$$\vec{u}(u_0,v_0) = \left(\frac{\partial}{\partial u}\vec{r}(u,v_0)\right)\Big|_{u_0}$$
$$\vec{v}(u_0,v_0) = \left(\frac{\partial}{\partial v}\vec{r}(u_0,v)\right)\Big|_{v_0}$$

Like all vectors, the tangents $\vec{u}$, $\vec{v}$ can be converted to normal vectors by dividing out the magnitude:

$$\hat{u} = \vec{u}/u$$
$$\hat{v} = \vec{v}/v$$

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

$$\vec{u}(x_0,y_0) = u_x\,\hat{x} + u_z\,\hat{z}$$
$$\vec{v}(x_0,y_0) = v_y\,\hat{y} + v_z\,\hat{z}\,,$$

where without loss of generality, $\vec{u}$ lacks a y-component and $\vec{v}$ lacks an x-component.

The ratios u_z/u_x, v_z/v_y, receptively, are the partial derivatives in disguise, as

$$\frac{u_z}{u_x} = \left(\frac{\partial}{\partial x}f(x,y_0)\right)\Big|_{x_0}$$
$$\frac{v_z}{v_y} = \left(\frac{\partial}{\partial y}f(x_0,y)\right)\Big|_{y_0}\,,$$

which allows the vectors $\vec{u}$, $\vec{v}$ to be written in terms of partial derivatives:

$$\vec{u}(x_0,y_0) = u_x\left(\hat{x} + \left(\frac{\partial}{\partial x}f(x,y_0)\right)\Big|_{x_0}\hat{z}\right)$$
$$\vec{v}(x_0,y_0) = v_y\left(\hat{y} + \left(\frac{\partial}{\partial y}f(x_0,y)\right)\Big|_{y_0}\hat{z}\right)$$

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:

$$\hat{u} = \frac{\hat{x} + f_x\,\hat{z}}{\sqrt{1 + f_x^2}}$$
$$\hat{v} = \frac{\hat{y} + f_y\,\hat{z}}{\sqrt{1 + f_y^2}}$$

4.3 Surface Normal Vector

With a pair of surface tangent vectors $\vec{u}$, $\vec{v}$ in hand for a given point, the cross product of the two yields the vector $\vec{n}$ that is normal to the surface:

$$\vec{n} = \vec{u} \times \vec{v}$$

Parametric Surface Normal

For the parametric surface $\vec{r}(u, v)$, the surface normal is straightforwardly calculated from

$$\vec{n}(u_0, v_0) = \vec{u}(u_0, v_0) \times \vec{v}(u_0, v_0)\ ,$$

which suggests a normalized version

$$\hat{n} = \frac{\vec{u}(u_0, v_0) \times \vec{v}(u_0, v_0)}{|\vec{u}(u_0, v_0) \times \vec{v}(u_0, v_0)|}\ .$$

Of course, there is no need to normalize if we use unit vectors only:

$$\hat{n} = \hat{u} \times \hat{v}$$

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 $\vec{u}(x_0, y_0)$, $\vec{v}(x_0, y_0)$. Explicitly, this is:

$$\vec{n}(x_0, y_0) = \vec{u} \times \vec{v} = \begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ u_x & 0 & u_x f_x \\ 0 & v_y & v_y f_y \end{vmatrix}\ ,$$

or

$$\vec{n} = u_x v_y\,(-f_x\,\hat{x} - f_y\,\hat{y} + \hat{z})\ .$$

Eliminate the stray coefficients by normalizing:

$$\hat{n} = \frac{-f_x\,\hat{x} - f_y\,\hat{y} + \hat{z}}{\sqrt{1 + f_x^2 + f_y^2}}$$

4.4 Tangent Plane

In either picture, whether it be parametric or Cartesian, the tangent vectors $\vec{u}$, $\vec{v}$ 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 $\vec{n}$ 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:

$$\vec{n} \cdot \Delta\vec{x} = 0\ ,$$

where

$$\Delta\vec{x} = \langle x, y, z\rangle - \langle x_0, y_0, z_0\rangle\ .$$

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

$$-f_x\,(x - x_0) - f_y\,(y - y_0) + (z - z_0) = 0\ .$$

Chapter 16

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

$$S = \int \left(T\left(v\left(t\right)\right) - U\left(x\left(t\right)\right)\right) dt$$

is always minimized if $x\left(t\right)$ and $v\left(t\right)$ 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 $x\left(t\right)$ or $v\left(t\right)$ 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

$$\left(x_0, y_0\right), \left(x_1, y_1\right)$$

that represent the initial and final position of any well-behaved curve $y\left(x\right)$. The curve $y\left(x\right)$ shall be considered the 'true' path that connects the initial and final fixed points:

$$y\left(x\right) = \text{true path from } x_0 \text{ to } x_1$$

2.1 Varied Path

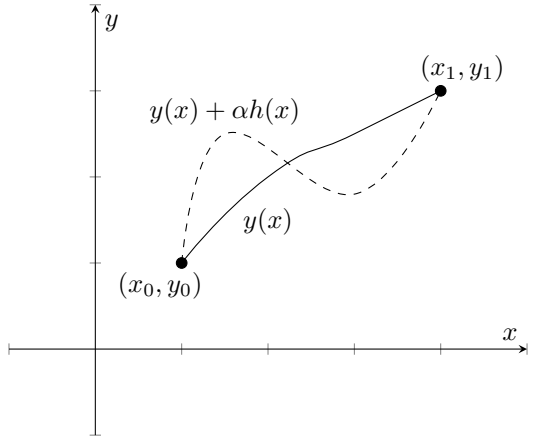
To accompany $y\left(x\right)$ in the plane, introduce a second curve that starts and finishes at the same endpoints, but is allowed to meander in the plane due to an additional term $\alpha h\left(x\right)$ such that

$$f\left(x\right) = y\left(x\right) + \alpha h\left(x\right) \ .$$

The variable α is a dimensionless *variational parameter*. The curve $h\left(x\right)$ is an 'incorrectness' that gives structure to the varied path, and this is decidedly zero at each endpoint:

$$h\left(x_0\right) = h\left(x_1\right) = 0$$

Summarizing this construction, we sketch the following figure in the Cartesian plane:



2.2 Writing the Action

Now, conceive of a new function Λ that depends of 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:

$$S = \int_{x_0}^{x_1} \Lambda(x, f, f')\, dx$$

2.3 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

$$\frac{dA}{d\alpha} = \frac{\partial \Lambda}{\partial x}\frac{\cancel{dx}}{\cancel{d\alpha}} + \frac{\partial \Lambda}{\partial f}\frac{df}{d\alpha} + \frac{\partial \Lambda}{\partial f'}\frac{df'}{d\alpha} \,,$$

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:

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} \left(\frac{\partial \Lambda}{\partial f} h + \frac{\partial \Lambda}{\partial f'} h' \right) dx \,.$$

2.4 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

$$\frac{dS}{d\alpha} = \int_{x_0}^{x_1} \left(\frac{\partial \Lambda}{\partial f} h - h \frac{d}{dx}\left(\frac{\partial \Lambda}{\partial f'} \right) \right) dx + \cancel{\frac{\partial \Lambda}{\partial f'} h \Big|_{x_0}^{x_1}} \,,$$

where the boundary term wholly vanishes because $h(x)$ is zero at each endpoint.

2.5 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:

$$\begin{aligned} dS/d\alpha &\to 0 \\ \alpha &\to 0 \end{aligned}$$

By making such a change, the varied path flattens down to the true path, which means $f(x)$ flattens down to $y(x)$:

$$f(x) \to y(x)$$

2.6 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:

$$0 = \frac{\partial \Lambda}{\partial y} - \frac{d}{dx}\left(\frac{\partial \Lambda}{\partial y_x} \right) \tag{16.1}$$

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.

2.7 Change of Domain

The whole derivation of Equation (16.1) can be repeated in the time domain, in which case t takes the place of x. Following this through, let us substitute

$$\begin{aligned} x &\to t \\ y &\to x\left(t\right) \\ y_x &\to v\left(t\right) \ , \end{aligned}$$

where $v\left(t\right)$ is the velocity $x'\left(t\right)$. While we're at it, relabel the arbitrary function Λ with the the letter L:

$$\Lambda \to L\left(t, x\left(t\right), v\left(t\right)\right)$$

With all this, the Euler-Lagrange equation takes a form in the time domain:

$$0 = \frac{\partial L}{\partial x} - \frac{d}{dt}\left(\frac{\partial L}{\partial v}\right) \tag{16.2}$$

2.8 The Lagrangian

By choosing the proper function the proper function for L, something curious happens with the Euler-Lagrange equation (16.2). After some fiddling, one readily stumbles on the combination

$$L = T\left(v\right) - U\left(x\right) \ , \tag{16.3}$$

called the *Lagrangian*. The functions $T\left(v\right)$, $U\left(x\right)$ are the respective kinetic and potential energies of the body being considered.

By inserting Equation (16.3) into Equation (16.2), we write

$$0 = \frac{\partial\left(T-U\right)}{\partial x} - \frac{d}{dt}\left(\frac{\partial\left(T-U\right)}{\partial v}\right) \ ,$$

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.

2.9 Recovering Newton's Law

For the Newtonian regime, we have that the kinetic energy $T\left(v\right)$ is given by

$$T\left(v\right) = \frac{1}{2}mv^2 \ ,$$

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:

$$m\frac{d}{dt}v\left(t\right) = -\frac{\partial}{\partial x}U\left(x\right)$$

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 apples 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:

$$0 = L_x - d_t L_v$$

3.1 Formal Derivation

With the main ideas established, let us re-state the action calculation

$$S = \int_{x_0}^{x_1} \Lambda\left(x, f, f'\right) dx$$

in slightly different terms. The function $f\left(x\right)$ is still the varied path, except variations are represented by

$$\begin{aligned} f &\to f + \Delta f \\ f_x &\to f_x + \left(\Delta f\right)' \ , \end{aligned}$$

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

$$S[f] = \int_{x_0}^{x_1} \Lambda(x, f, x_x)\, dx\,.$$

Then a variation in S is written

$$\Delta S = \int_{x_0}^{x_1} \Lambda\left(x, f + \Delta f, f_x + (\Delta f)'\right) dx - S[f]\ .$$

By Taylor-expanding the inner quantity to first order, we further have

$$\Lambda\left(x, f + \Delta f, f_x + (\Delta f)'\right) = \Lambda(x, f, f_x) + \frac{\partial \Lambda}{\partial f}\Delta f + \frac{\partial \Lambda}{\partial f_x}(\Delta f)'\ ,$$

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

$$0 = \frac{\partial \Lambda}{\partial f} - \frac{d}{dx}\left(\frac{\partial \Lambda}{\partial f_x}\right)\ .$$

3.2 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:

$$\frac{dA}{dx} = \frac{\partial \Lambda}{\partial x} + \frac{d}{dx}\left(f_x \frac{\partial \Lambda}{\partial f_x}\right)$$

Putting everything on one side gives special form of the Euler-Lagrange equation:

$$0 = \frac{\partial \Lambda}{\partial x} - \frac{d}{dx}\left(\Lambda - f_x \frac{\partial \Lambda}{\partial f_x}\right) \tag{16.4}$$

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.

3.3 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

$$\begin{aligned} x &\to t \\ f &\to x(t) \\ L &\to T(v) - U(x)\ , \end{aligned}$$

and so on.

Since the Lagrangian has no explicit time dependence, may 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.

3.4 More Dimensions

The formal derivation of the Euler-Lagrange equation that generalizes to N simultaneous functions

$$f \to \left\{ f^{(1)}, f^{(2)}, \ldots, f^{(N)} \right\} ,$$

each with a different variation

$$f^{(j)} \to f^{(j)} + \Delta f^{(j)} ,$$

with vanishing variation at the endpoints

$$\Delta f^{(j)} (x_0) = \Delta f^{(j)} (x_1) = 0$$

is straightforwardly written.

The function Λ becomes

$$\Lambda \left(x, f^{(1)}, f^{(2)}, \ldots, f^{(N)}, f_x^{(1)}, f_x^{(2)}, \ldots, f_x^{(N)} \right) ,$$

and gives rise to m Euler-Lagrange equations that all look the same up to index number:

$$0 = \frac{\partial \Lambda}{\partial f^{(j)}} - \frac{d}{dx} \left(\frac{\partial \Lambda}{\partial f_x^{(j)}} \right) \tag{16.5}$$

3.5 Conservation of Energy

When it comes to having more than one dimension, one wonders what the generalization of Equation (16.4) 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:

$$E = E \left(x, f^{(1)}, f^{(2)}, \ldots, f^{(N)}, f_x^{(1)}, f_x^{(2)}, \ldots, f_x^{(N)} \right)$$

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:

$$\begin{aligned} x &\to t \\ \Lambda &\to L = T(v) - U(x) \end{aligned}$$

3.6 Generalized Coordinates

As it pertains to physical systems, The Euler-Lagrange equation (16.5) occurs in the form

$$0 = \frac{\partial L}{\partial q^{(j)}} - \frac{d}{dt} \left(\frac{\partial L}{\partial q_t^{(j)}} \right) , \tag{16.6}$$

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

$$\vec{q} = \langle r, \theta \rangle .$$

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

$$x^{(j)} = x^{(j)} \left(t, q^{(1)}, q^{(2)}, \ldots, q^{(N)} \right)$$

if there are N members in $\{q(t)\}$. To first order, the two sets of coordinates further relate by

$$v^{(j)} = \frac{\partial x^{(j)}}{\partial t} + \sum_{k=1}^{N} \frac{\partial x^{(j)}}{\partial q^{(k)}} q_t^{(k)} .$$

By the chain rule, we further find

$$\frac{\partial x^{(j)}}{\partial q^{(k)}} = \frac{\partial x^{(j)}}{\partial q^{(k)}} \frac{dt}{dt} = \frac{\partial v^{(j)}}{\partial q_t^{(k)}} \, .$$

With the second set of coordinates, the proposed Euler-Lagrange equation reads

$$0 \stackrel{?}{=} \frac{\partial \Lambda}{\partial q^{(j)}} - \frac{d}{dt}\left(\frac{\partial \Lambda}{\partial q_t^{(j)}}\right) \, ,$$

which motivates picking on the inner term:

$$\frac{\partial \Lambda}{\partial q_t^{(j)}} = \sum_{k=1}^{N} \frac{\partial \Lambda}{\partial v^{(k)}} \frac{\partial v^{(k)}}{\partial q_t^{(j)}} = \sum_{k=1}^{N} \frac{\partial \Lambda}{\partial v^{(k)}} \frac{\partial x^{(k)}}{\partial q^{(j)}}$$

Take the time derivative of both sides to get

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial \Lambda}{\partial q_t^{(j)}}\right) = &\sum_{k=1}^{N} \frac{d}{dt}\left(\frac{\partial \Lambda}{\partial v^{(k)}}\right) \frac{\partial x^{(k)}}{\partial q^{(j)}} \\ &+ \sum_{k=1}^{N} \frac{\partial \Lambda}{\partial v^{(k)}} \frac{d}{dt}\left(\frac{\partial x^{(k)}}{\partial q^{(j)}}\right) \, , \end{aligned}$$

and simplify carefully to finish the proof:

$$\begin{aligned} \frac{d}{dt}\left(\frac{\partial \Lambda}{\partial q_t^{(j)}}\right) = & \\ &\sum_{k=1}^{N} \left(\frac{\partial \Lambda}{\partial x^{(k)}} \frac{\partial x^{(k)}}{\partial q^{(j)}} + \frac{\partial \Lambda}{\partial v^{(k)}} \frac{\partial v^{(k)}}{\partial q^{(j)}} \right) \\ &= \frac{\partial \Lambda}{\partial q^{(j)}} \end{aligned}$$

4 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 $\vec{v} = \langle \dot{x}, \dot{y} \rangle$, 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

$$\begin{aligned} T &= \frac{1}{2} m \dot{x}^2 \left(1 + k^2\,(x-a)^{2\alpha-2}\right) \\ U &= \frac{mgk}{\alpha}\,(x-a)^{\alpha} \, , \end{aligned}$$

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:

$$\begin{aligned} 0 = &\ddot{x}\left(1 + k^2\,(x-a)^{2\alpha-3}\right) \\ &+ \dot{x}\frac{k^2}{2}\,(2\alpha-2)\,(x-a)^{2\alpha-3} + gk\,(x-a)^{\alpha-1} \end{aligned}$$

4.1 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:

$$\ddot{w} = -gkw$$

4.2 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

$$k = \tan(\theta)$$

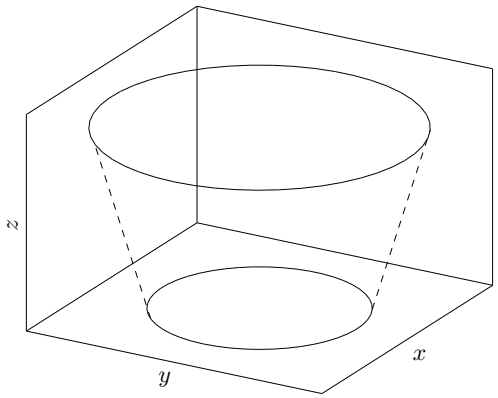
and

$$|a| = g\sin(\theta) \, .$$

5 Minimal Surface

5.1 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 (16.4), 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.

5.2 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.

6 Motion on a Cycloid

6.1 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 (16.4), 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'.

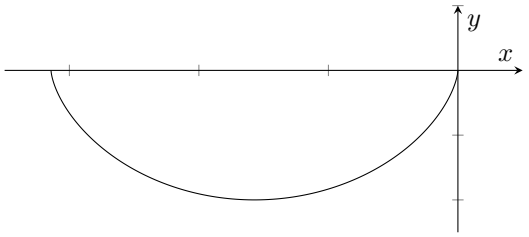
Finishing up the calculation for x, we finally have

$$x = x_0 - R\,(\theta - \sin\,(\theta)) \,, \tag{16.7}$$

and solving similarly for y,

$$y = y_0 - R\,(1 - \cos\,(\theta)) \,. \tag{16.8}$$

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 $\theta \to -\theta$ 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.

6.2 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

$$T = \int_{x_0}^{x_1} \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \frac{dx}{\sqrt{2g\left(y_0 - y\left(x\right)\right)}} ,$$

where x_1 is located at the bottom of the cycloid (upturned as sketched above).

Knowing the solution to $y\left(\theta\right)$, we can write

$$y_0 - y = R\left(\cos\left(\theta_0\right) - \cos\left(\theta\right)\right) ,$$

where θ_0 characterizes the initial position of a body sliding from rest. Then, the equation for velocity as a function of θ reads

$$v\left(\theta\right) = \sqrt{2gR\left(\cos\left(\theta_0\right) - \cos\left(\theta\right)\right)} .$$

The remaining quantities in the integral must also be expressed in terms of θ. Doing so carefully, the T-integral becomes

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sqrt{1 - \cos\left(\theta\right)}\, d\theta}{\sqrt{\cos\left(\theta_0\right) - \cos\left(\theta\right)}} .$$

Note that the change of variables $\theta \to -\theta$ 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

$$\sin\left(\frac{\theta}{2}\right) = \sqrt{\frac{1}{2} - \frac{\cos\left(\theta\right)}{2}}$$
$$\cos\left(\theta\right) = 2\cos^2\left(\frac{\theta}{2}\right) - 1 ,$$

so now

$$T = \sqrt{\frac{R}{g}} \int_{\theta_0}^{\pi} \frac{\sin\left(\theta/2\right) d\theta}{\sqrt{\cos^2\left(\theta_0/2\right) - \cos^2\left(\theta/2\right)}} .$$

Proceed with the u-substitution

$$u = \cos\left(\theta/2\right) / \cos\left(\theta_0/2\right)$$
$$du = -d\theta \sin\left(\theta/2\right) / 2\cos\left(\theta_0/2\right) ,$$

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.

7 Lagrange Multipliers

An item from the calculus tool chest, not particularly related to the Euler-Lagrange equation or Lagrange mechanics, is the idea of *Lagrange multiplier*. A Lagrange multiplier is used when solving an optimization problem subject to a particular constraint. For instance, if we have a two-variable function $f\left(x, y\right)$, there may be reason to extremize f subject to a different function $g\left(x, y\right) = c$, where c characterizes a level curve on g.

7.1 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:

$$\vec{\nabla} f\left(x, y\right) = \lambda \vec{\nabla} g\left(x, y\right)$$

Said a different way, a constrained optimization problem minimizes a functional $\vec{F}$

$$\vec{F}\left[x, y, \lambda\right] = \vec{\nabla} f\left(x, y\right) - \lambda \vec{\nabla} g\left(x, y\right)$$

at critical points $\left(x_0, y_0\right)$.

When there are multiple constraint functions, each introduces a new and different λ. The gradient of the original function and all constraints remain proportional:

$$\vec{\nabla} f\left(x, y\right) = \lambda_1 \vec{\nabla} g_1\left(x, y\right) + \lambda_2 \vec{\nabla} g_2\left(x, y\right) + \cdots$$

Worked Example

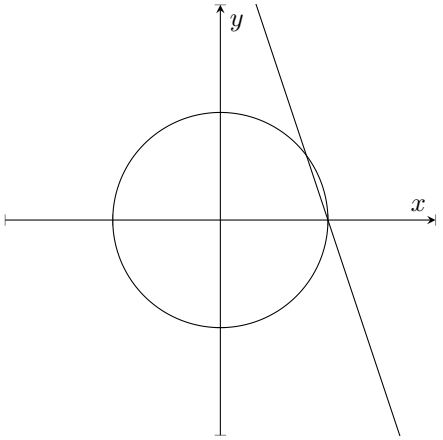
For an example, suppose we have a function

$$f\left(x, y\right) = x^2 + y^2 ,$$

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 $f\left(x, y\right)$.



Constructing the functional $\vec{F}[\lambda]$, we have

$$\begin{aligned}\vec{F}[\lambda] &= \vec{\nabla} f(x,y) - \lambda \vec{\nabla} g(x,y) \\ &= \langle 2x, 2y \rangle - \lambda \langle 3, 1 \rangle \ .\end{aligned}$$

Setting the left side to zero, we gain two new equations

$$\begin{aligned}2x_0 &= 3\lambda \\ 2y_0 &= \lambda \ ,\end{aligned}$$

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

$$\begin{aligned}\lambda &= 3/5 \\ x_0 &= 9/10 \\ y_0 &= 3/10 \ ,\end{aligned}$$

and the problem is solved.

Practice

Problem 1

Identify all critical points of the function $f(x,y) = x^2 + 2y^2$ that coincide with the unit circle $x^2 + y^2 = 1$.

Problem 2

Find the largest rectangle that fits inside the first quadrant of the ellipse $x^2/a^2 + y^2/b^2 = 1$. Answer:

$$\begin{aligned}x_0 &= a/\sqrt{2} \\ y_0 &= b/\sqrt{2}\end{aligned}$$

Problem 3

Find the largest square that fits inside the first quadrant of the ellipse $x^2/a^2 + y^2/b^2 = 1$. Hint:

$$\vec{\nabla}(xy) = \lambda_1 \vec{\nabla}\left(\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1\right) + \lambda_2 \vec{\nabla}(x - y)$$

Problem 4

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:

$$\vec{\nabla}\left(\pi R^2 L + \frac{2}{3}\pi R^2 H\right) = \lambda \vec{\nabla}\left(2\pi RL + 2\pi R\sqrt{R^2 + H^2}\right)$$

Answer:

$$V = \frac{AR}{3} = A^{3/2}(2\pi)^{-1/2}\frac{5^{-1/4}}{3}$$

7.2 Constrained Systems

Let us return to Equation (16.5) 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:

$$\frac{\partial \Lambda}{\partial f^{(j)}} - \frac{d}{dx}\left(\frac{\partial \Lambda}{\partial f_x^{(j)}}\right) = \sum_{k=1}^{n} \lambda_k \frac{\partial g_k}{\partial f^{(j)}} \qquad (16.9)$$

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.

7.3 Generalized Force

From a mechanical point of view, the right side of Equation (16.9) 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 (16.9) reads

$$\frac{\partial L}{\partial q^{(j)}} - \frac{d}{dt}\left(\frac{\partial L}{\partial q_t^{(j)}}\right) = Q_{(j)} = \sum_{k=1}^{n} \lambda_k \frac{\partial g_k}{\partial f^{(j)}} \, .$$

8 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.

8.1 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

$$F[y] = \int_{x_0}^{x_1} (\rho g y(x) - \lambda) \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \, dx \, .$$

8.2 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

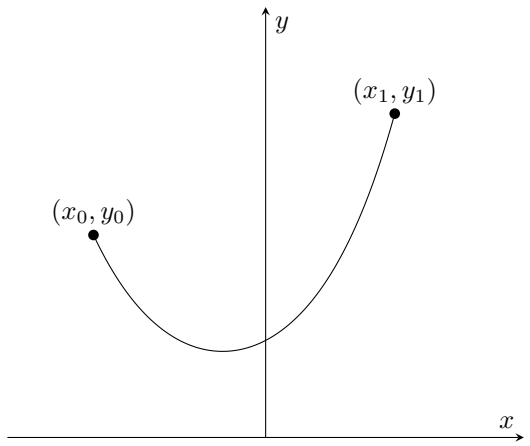
$$x = \frac{C_0}{\rho g} \cosh^{-1}\left(\frac{u}{C_0/\rho g}\right) + C_1 \, ,$$

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:



8.3 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

$$L = \frac{C_0}{\rho g} \sinh\left(\frac{x - C_1}{C_0/\rho g}\right)\Bigg|_{x_0}^{x_1} .$$

8.4 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

$$2q = \sinh(q) ,$$

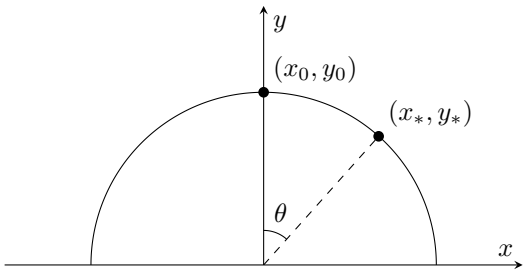
approximately solved by

$$\frac{x_0}{C_0/\rho g} = q \approx 2.177 .$$

With this, H comes out to

$$H \approx x_0 \left(\frac{\cosh(2.177) - 1}{2.177}\right) \approx 1.59\, x_0 .$$

9 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.

9.1 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:

$$\theta_* = \arccos\left(\frac{2}{3}\right) \approx 0.841 \text{ rad} \approx 48.2^\circ$$

9.2 Constrained Motion Analysis

It it 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 (16.9) yields two items:

$$\frac{\partial L}{\partial r} - \frac{d}{dt}\left(\frac{\partial L}{\partial r_t}\right) = \lambda\frac{\partial}{\partial r}(r - R)$$
$$\frac{\partial L}{\partial \theta} - \frac{d}{dt}\left(\frac{\partial L}{\partial \theta_t}\right) = \lambda\frac{\partial}{\partial \theta}(r - R)$$

Carrying these calculations through gives a pair of equations

$$mr^2\omega_t = -2mrr_t\omega + mgr\sin(\theta)$$
$$mr_{tt} = mr\omega^2 - mg\cos(\theta) - \lambda,$$

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 the 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

$$\omega_t = \frac{g}{R}\sin(\theta)$$
$$\omega_*^2 = \frac{g}{R}\cos(\theta_*).$$

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.

10 Maximal Area

10.1 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:

$$A[y] = \int_{x_0}^{x_1} y\, dx$$
$$L[y] = \int_{x_0}^{x_1} \sqrt{1 + y_x^2}\, dx$$

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 (16.4) 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

$$q = \lambda^2 - w^2$$
$$dq = -2w\, dw \,.$$

From here the integration results in

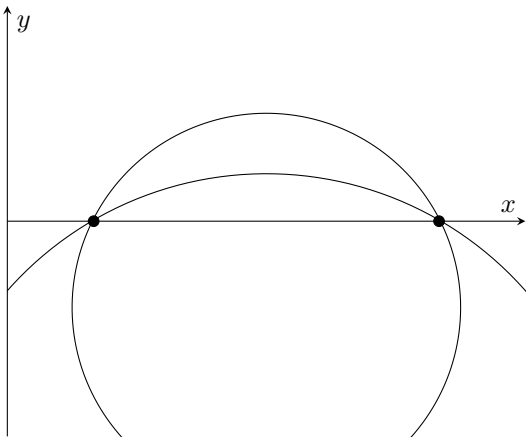
$$x = -\sqrt{q} + C_0 \,,$$

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:



10.2 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.

10.3 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:

$$A[x, y] = \oint \frac{1}{2}(x\, dy - y\, dx)$$
$$L[x, y] = \oint \sqrt{dx^2 + dy^2}$$

To proceed, frame each of x and y as parametric equations in the variable t. Then, using the chain rule, the above become

$$A[x, y] = \oint \frac{1}{2}(x\, y_t - y\, x_t)\, dt$$
$$L[x, y] = \oint \sqrt{x_t^2 + y_t^2}\, dt \,,$$

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:

$$\frac{d}{dt}\left(\frac{\partial \Lambda}{\partial x_t}\right) - \frac{\partial \Lambda}{\partial x} = 0$$
$$\frac{d}{dt}\left(\frac{\partial \Lambda}{\partial y_t}\right) - \frac{\partial \Lambda}{\partial y} = 0$$

Substituting Λ into each gives

$$\frac{d}{dt}\left(\frac{-1}{2}y - \frac{\lambda x_t}{\sqrt{x_t^2 + y_t^2}}\right) = \frac{1}{2}y_t$$
$$\frac{d}{dt}\left(\frac{1}{2}x - \frac{\lambda y_t}{\sqrt{x_t^2 + y_t^2}}\right) = \frac{-1}{2}x_t \,,$$

both of which are easily integrated. Introducing respective integrations constants C_0, C_1, we find, after simplifying:

$$x - C_0 = \frac{\lambda y_t}{\sqrt{x_t^2 + y_t^2}}$$
$$y - C_1 = \frac{-\lambda x_t}{\sqrt{x_t^2 + y_t^2}}$$

Square both sides and add to recover the formula for a circle:

$$(x - C_0)^2 + (y - C_1)^2 = \lambda^2$$

10.4 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

$$A[r, \theta] = \oint \frac{1}{2} r^2 \, d\theta$$
$$L[r, \theta] = \oint r \sqrt{1 + \left(\frac{1}{r} r_\theta\right)^2} \, d\theta \, ,$$

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 (16.4), 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

$$d\vec{s} = \langle dr, r \, d\theta \rangle \, ,$$

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

$$\frac{1}{2} r_-^2 - C = \lambda r_- \sin\left(\frac{-\pi}{2}\right)$$
$$\frac{1}{2} r_+^2 - C = \lambda r_+ \sin\left(\frac{\pi}{2}\right) \, ,$$

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

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta) \ , \end{aligned}$$

we swiftly make out the fingerprint of a circle:

$$\begin{aligned} r^2 &= 2\lambda r \sin(\theta) \\ x^2 + y^2 &= 2\lambda y \\ x^2 + (y-\lambda)^2 &= \lambda^2 \end{aligned}$$

Part V

Applications

Chapter 17

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

$$x = \mathrm{Re}\,(z) \ ,$$

and the y-component is the 'imaginary part' of z:

$$y = \mathrm{Im}\,(z)$$

Complex Conjugate

Any complex number z has a complex conjugate $\overline{z} = z^*$, 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

$$z_1 * z_2 = z_2 * z_1$$
$$\overline{z_1} * \overline{z_2} = \overline{z_1 * z_2} \,,$$

where the generalized operator represents either complex addition (+) or complex multiplication (·).

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:

$$(z_1 z_2)\, z_3 = z_1\,(z_2 z_3)$$
$$z_1\,(z_2 + z_3) = (z_1 z_2) + (z_1 z_3)$$

Complex Magnitude

The complex magnitude

$$|z| = \sqrt{z\overline{z}} = \sqrt{(x + iy)\,(x - iy)} = \sqrt{x^2 + y^2}$$

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:

$$\frac{z_1}{z_2} = \frac{z_1\,\overline{z_2}}{|z_2|^2}$$

One can readily check that this definition readily satisfies $\overline{z_1} * \overline{z_2} = \overline{z_1 * z_2}$, where the generalized arithmetic operator ($*$) is replaced by the division symbol (/).

1.3 Complex Plane

Polar Representation

Complex numbers lend naturally to polar representation

$$x = r\cos(\phi)$$
$$y = r\sin(\phi)$$
$$r = |z| = \sqrt{x^2 + y^2}$$
$$\phi = \arctan(y/x) \ ,$$

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

$$z = r\left(\cos\left(\phi\right) + i\sin\left(\phi\right)\right) \ .$$

Rotations

The special complex number z_θ with $|z_\theta| = 1$ and any phase θ is a rotation operator for complex numbers:

$$z_\theta = \cos\left(\theta\right) + i\sin\left(\theta\right)$$

This is because the product $z_\theta z$ for any $z\left(r,\phi\right)$ results in:

$$z_\theta z = r\left(\cos\left(\phi+\theta\right) + i\sin\left(\phi+\theta\right)\right)$$

That is multiplying by z_θ has the effect of a change of phase $\phi \to \phi + \theta$.

1.4 Euler's Formula

Repeated Rotations

Making repeated use of the rotation operator z_θ, suppose we start with a complex number $z\left(r,\phi\right)$ and make n identical rotations by the angle θ:

$$z_\theta^n z = r\left(\cos\left(\phi+n\theta\right) + i\sin\left(\phi+n\theta\right)\right)$$

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 $h \to \infty$, the quantity θ/h tends to zero, warranting the small-angle approximation to replace both trigonometry terms:

$$z_\theta = \lim_{h\to\infty}\left(1 + \frac{i\theta}{h}\right)^h$$

The right side is precisely the definition of Euler's constant e raised to the power $i\theta$. In summary, we have found

$$z_\theta = e^{i\theta} = \cos\left(\theta\right) + i\sin\left(\theta\right) \ ,$$

one of the msot useful relationships in mathematics.

Calculus-based Derivation

Begin with the polar representation of a complex number

$$z = r\left(\cos\left(\phi\right) + i\sin\left(\phi\right)\right) \ ,$$

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

$$\ln\left(z\right) = \ln\left(r\right) + i\phi + \cancel{C} \ ,$$

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 $z_1\left(r_1,\phi_1\right)$, $z_2\left(r_2,\phi_2\right)$, we always have

$$z_1 z_2 = r_1 r_2\,e^{i(\phi_1+\phi_2)}$$
$$\frac{z_1}{z_2} = \frac{r_1}{r_2}\,e^{i(\phi_1-\phi_2)}$$

Complex Logarithm

Consider a general complex number

$$z\left(r,\phi\right) = r\,e^{i\phi} \ ,$$

which admits a logarithmic form

$$\ln\left(z\right) = \ln\left(r\right) + i\phi \ .$$

Branches

Unlike $z\left(r,\phi\right)$, 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:

$$\begin{aligned} z^w &= \left(r\,e^{i\phi}\right)^{\alpha+i\beta} \\ &= r^\alpha r^{i\beta}\,e^{i\phi\alpha}\,e^{-\phi\beta} \\ &= e^{\ln(r)\alpha}\,e^{\ln(r)i\beta}\,e^{i\phi\alpha}\,e^{-\phi\beta} \\ &= e^{\alpha\ln(r)-\beta\phi}\,e^{i(\beta\ln(r)+\alpha\phi)} \\ &= \exp\left(\left(\ln\left(r\right)+i\phi\right)\left(\alpha+i\beta\right)\right) \\ &= e^{w\ln(z)} \end{aligned}$$

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

$$z\left(t\right) = r\left(t\right)e^{i\theta(t)}\,,$$

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:

$$\begin{aligned} \frac{d}{dt}z\left(t\right) &= \dot{r}e^{i\theta} + ie^{i\theta}r\omega \\ \frac{d^2}{dt^2}z\left(t\right) &= e^{i\theta}\left(\ddot{r} - r\omega^2\right) + ie^{i\theta}\left(2\dot{r}\omega + r\dot{\omega}\right) \end{aligned}$$

The results for $\dot{z}$ and $\ddot{z}$ are each complex numbers, carrying real and imaginary components. By associating

$$\begin{aligned} e^{i\theta} &\to \hat{r} \\ ie^{i\theta} &\to \hat{\theta}\,, \end{aligned}$$

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 $x\left(t\right)$ 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 $w\left(t\right)$:

$$\ddot{w} + \omega_0^2\,w = 0$$

As a complex number w, decomposes to

$$w\left(t\right) = x\left(t\right) + iy\left(t\right)\,,$$

or in terms of real and imaginary components

$$\begin{aligned} x\left(t\right) &= \mathrm{Re}\left(w\left(t\right)\right) \\ y\left(t\right) &= \mathrm{Im}\left(w\left(t\right)\right)\,, \end{aligned}$$

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 $w\left(t\right)$ as

$$w\left(t\right) = A\,e^{\lambda t}\,,$$

where A is the amplitude constant and λ is a frequency constant. Substitution of $w\left(t\right)$ 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

$$w\left(t\right) = A_1\,e^{i\omega_0 t} + A_2\,e^{-i\omega_0 t}$$

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

$$w(t) = a_1 \cos(\phi_1 + \omega_0 t) + a_2 \cos(\phi_2 - \omega_0 t) + i a_1 \sin(\phi_1 + \omega_0 t) + i a_2 \sin(\phi_2 - \omega_0 t) ,$$

or, splitting the real from the imaginary parts:

$$x(t) = a_1 \cos(\phi_1 + \omega_0 t) + a_2 \cos(\phi_2 - \omega_0 t)$$
$$y(t) = a_1 \sin(\phi_1 + \omega_0 t) + a_2 \sin(\phi_2 - \omega_0 t)$$

These solutions are identical up to the phase constants ϕ_1, ϕ_2. For instance, let $\phi_2 \to \phi_2 + \pi/2$ to transform the sines into cosines. Proceeding with the $x(t)$-equation, let

$$a = a_1 \cos(\phi_1) + a_2 \cos(\phi_2)$$
$$b = -a_1 \sin(\phi_1) + a_2 \sin(\phi_2) ,$$

and we get

$$x(t) = a \cos(\omega_0 t) + b \sin(\omega_0 t) .$$

Or, to make the solution even tighter, let

$$a = R \cos(\phi_0)$$
$$b = R \sin(\phi_0) ,$$

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 camping 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

$$\frac{d}{dt}x\left(t\right)=\dot{x}=d_{t}x\left(t\right)\ ,$$

the differential equation of the damped oscillator can be written

$$\left[d_{tt}+bd_{t}+\omega_{0}^{2}\right]x\left(t\right)=0\,,$$

where the quantity in square brackets is an object that operates on $x\left(t\right)$. Proceed by 'completing the square' within the operator to land at

$$\left[d_{t}+\frac{b}{2}\right]^{2}x\left(t\right)=\left(\frac{b^{2}}{4}-\omega_{0}^{2}\right)x\left(t\right)\ .$$

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

$$D^{2}x\left(t\right)=-\tilde{\omega}^{2}x\left(t\right)\ .$$

Next, let

$$x\left(t\right)=e^{-bt/2}\tilde{x}\left(t\right)\ ,$$

and the above becomes

$$e^{-bt/2}D^{2}\tilde{x}\left(t\right)+\tilde{x}\left(t\right)D^{2}e^{-bt/2}=-\tilde{\omega}^{2}e^{-bt/2}\tilde{x}\left(t\right)\ .$$

Dealing with the middle term first, notice

$$\begin{aligned}D^{2}e^{-bt/2}&=\left[d_{tt}+bd_{t}+\frac{b^{2}}{4}\right]e^{-bt/2}\\&=\left(\frac{b^{2}}{4}-\frac{b^{2}}{2}+\frac{b^{2}}{4}\right)e^{-bt/2}=0\,,\end{aligned}$$

and the above simplifies to

$$D^{2}\tilde{x}\left(t\right)=-\tilde{\omega}^{2}\tilde{x}\left(t\right)\ .$$

At this point, we finally apply the case of critical damping, in where $\tilde{\omega}=0$. This reduces the above to

$$D^{2}\tilde{x}\left(t\right)=0\,,$$

having general solution

$$\tilde{x}\left(t\right)=a_{1}+a_{2}t\,.$$

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:

$$x\left(t\right)=e^{-bt/2}\left(a_{1}+a_{2}t\right)$$

Driven Harmonic Oscillator

It's also possible to analyze the so-called driven oscillator, generally described by

$$\ddot{x}+b\dot{x}+\omega_{0}^{2}x=f\left(t\right)\ .$$

Due to the linearity in the left hand side, we know already that the solution to the above takes the form

$$x\left(t\right)=x_{h}\left(t\right)+x_{p}\left(t\right)\ ,$$

i.e. the sum of a homogeneous part $x_{h}\left(t\right)$ and a particular part $x_{p}\left(t\right)$. Knowing $x_{h}\left(t\right)$ already to be

$$x_{h}\left(t\right)=R\,e^{-bt/2}\cos\left(\omega_{0}t-\phi_{0}\right)\ ,$$

the task is reduced to finding a particular solution $x_{p}\left(t\right)$.

In general, finding the particular solution to the the above with arbitrary $f\left(t\right)$ is as difficult as it sounds, so we make the job easier by considering a sinusoidal driving function

$$f\left(t\right)=\gamma\cos\left(\alpha t\right)\ .$$

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 $w\left(t\right)$,

$$\ddot{w}+b\dot{w}+\omega_{0}^{2}w=\gamma\,e^{i\alpha t}\,.$$

Next, we postulate complex solutions of the form

$$w\left(t\right)=A\,e^{i\alpha t}$$

for some arbitrary constant A, and the above becomes

$$A\left(-\alpha^{2}+ib\alpha+\omega_{0}^{2}\right)\cancel{e^{i\alpha t}}=\gamma\,\cancel{e^{i\alpha t}}\,.$$

Solving for A gives

$$A=\frac{\gamma\left(\omega_{0}^{2}-\alpha^{2}\right)-i\gamma b\alpha}{\left(\omega_{0}^{2}-\alpha^{2}\right)^{2}+b^{2}\alpha^{2}}\ ,$$

where if we let

$$u=\omega_{0}^{2}-\alpha^{2}$$
$$v=\beta\alpha\,,$$

a complex number q can be written as

$$q=u-iv=q_{0}\,e^{-i\phi_{0}}\ ,$$

with

$$q_0 = \sqrt{\left(\omega_0^2 - \alpha^2\right)^2 + b^2\alpha^2}\,.$$

Rewriting A, we have:

$$A = \frac{\gamma q_0\, e^{-i\phi_0}}{q_0^2} = \frac{\gamma\, e^{-i\phi_0}}{\sqrt{\left(\omega_0^2 - \alpha^2\right)^2 + b^2\alpha^2}}$$

Finally, the particular solution to the driven oscillator reads

$$w\left(t\right) = \frac{\gamma\, e^{i(\alpha t - \phi_0)}}{\sqrt{\left(\omega_0^2 - \alpha^2\right)^2 + b^2\alpha^2}}$$
$$\phi_0 = \tan^{-1}\left(\frac{b\alpha}{\omega_0^2 - \alpha^2}\right)$$

Note that this particular solution lacks a term like $\exp\left(bt/2\right)$, which, much unlike the homogeneous solution, doesn't decay over long times.

The amplitude A depends on how ω_0 relates to α. Calculating

$$\frac{d}{d\alpha}\left|A\left(\alpha\right)\right| = 0$$

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 $\vec{B}$. The force incident on the particle is given by

$$\vec{F} = q\,\vec{v}\left(t\right) \times \vec{B}\,,$$

where $\vec{v}\left(t\right)$ is the instantaneous velocity. Equations of motion are determined by applying Newton's second law

$$\vec{F} = m\frac{d}{dt}\vec{v}\left(t\right)\,.$$

Being a three-dimensional problem, let us choose to align the magnetic field with the positive z-axis. Then, by eliminating $\vec{F}$, we have

$$m\frac{d}{dt}\vec{v}\left(t\right) = qB\,\vec{v}\left(t\right) \times \hat{z}\,,$$

or

$$\frac{d}{dt}\vec{v}\left(t\right) = \omega_0\,\vec{v}\left(t\right) \times \hat{z}$$
$$\omega_0 = \frac{qB}{m}\,.$$

The unpacks into three equations:

$$\frac{d}{dt}v_x\left(t\right) = \omega_0 v_y\left(t\right)$$
$$\frac{d}{dt}v_y\left(t\right) = -\omega_0 v_x\left(t\right)$$
$$\frac{d}{dt}v_z\left(t\right) = 0$$

Noting that $v_z\left(t\right)$ is constant in time, we immediately know the solution for $z\left(t\right)$, namely

$$z\left(t\right) = z_0 + v_z t\,,$$

and we may focus entirely on what occurs in the xy-plane. Defining a complex variable

$$w\left(t\right) = v_x\left(t\right) + iv_y\left(t\right)\,,$$

we capture the information written above using the derivative

$$\frac{d}{dt}w\left(t\right) = \omega_0 v_y\left(t\right) - i\omega_0 v_x\left(t\right)\,,$$

simplifying to

$$i\frac{d}{dt}w\left(t\right) = \omega_0 w\left(t\right)\,.$$

This first-order differential equation is easily solved by

$$w\left(t\right) = Ae^{-i\omega_0 t} = \left|A\right| e^{i(\phi_0 - \omega_0 t)}\,,$$

where A is an arbitrary complex constant. Taking the real and imaginary parts of the above, the equations for $v_x\left(t\right)$, $v_y\left(t\right)$ emerge:

$$v_x\left(t\right) = \mathrm{Re}\left(w\left(t\right)\right) = \left|A\right|\cos\left(\omega_0 t - \phi_0 t\right)$$
$$v_y\left(t\right) = \mathrm{Im}\left(w\left(t\right)\right) = -\left|A\right|\sin\left(\omega_0 t - \phi_0\right)$$

While the above can be integrated separately to attain equations of motion $x\left(t\right)$, $y\left(t\right)$, we can integrate $w\left(t\right)$ directly by introducing a variable

$$R = x\left(t\right) + iy\left(t\right)\,,$$

whose derivative is $w\left(t\right)$:

$$\frac{d}{dt}R\left(t\right) = w\left(t\right)$$

Then, the integral of the above can be written

$$R\left(t\right) = R_0 + \left|A\right|\int_0^t e^{i\left(\phi_0 - \omega_0 t'\right)}dt'\,,$$

which can be solved with relative ease:

$$\begin{aligned} R(t) &= R_0 + |A| e^{i\phi_0} \int_0^t e^{-i\omega_0 t'} dt' \\ &= R_0 + |A| e^{i\phi_0} \frac{1}{-i\omega_0} e^{-i\omega_0 t'} \Big|_0^t \\ &= R_0 + |A| e^{i\phi_0} \frac{i}{\omega_0} \left(e^{-i\omega_0 t} - 1\right) \\ &= R_0 + |A| e^{i\phi_0 - i\omega_0 t} \frac{i}{\omega_0} - |A| e^{i\phi_0} \frac{i}{\omega_0} \end{aligned}$$

Repacking the integration constants together, we write

$$\tilde{R}_0 = R_0 - |A| e^{i\phi_0} \frac{i}{\omega_0} ,$$

the solution reads

$$\begin{aligned} R(t) &= \tilde{R}_0 + |A| e^{i(\phi_0 - \omega_0 t)} \frac{i}{\omega_0} \\ &= \tilde{R}_0 + \frac{|A|}{\omega_0} \left(i \cos(\omega_0 t - \phi_0) + \sin(\omega_0 t - \phi_0)\right) . \end{aligned}$$

Finally, we find

$$\begin{aligned} x(t) &= \mathrm{Re}(R(t)) = \tilde{x}_0 + \frac{|A|}{\omega_0} \sin(\omega_0 t - \phi_0) \\ y(t) &= \mathrm{Im}(R(t)) = \tilde{y}_0 + \frac{|A|}{\omega_0} \cos(\omega_0 t - \phi_0) , \end{aligned}$$

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 the 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

$$\frac{dw(z, \bar{z})}{dz} = \lim_{\Delta z \to 0} \frac{\partial w}{\partial z} + \frac{\partial w}{\partial \bar{z}} \frac{\Delta \bar{z}}{\Delta z} ,$$

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 $\Delta z \to 0$.

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:

$$\frac{dw}{dz} = \frac{\partial w}{dz}$$
$$\frac{\partial w}{\partial \bar{z}} = 0$$

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:

$$f_1(z) = x^2 - y^2$$
$$f_2(z) = x^2 + iy^2$$
$$f_3(z) = r^2(\cos(\theta) + i\sin(\theta))$$

Check this by calculating derivatives of each:

$$\frac{\partial}{\partial \bar{z}} f_1(z) = \frac{\partial}{\partial \bar{z}}(z\bar{z}) = z \neq 0$$
$$\frac{\partial}{\partial \bar{z}} f_2(z) = \frac{1}{2}\left(\frac{\partial f_2}{\partial x} + i\frac{\partial f_2}{\partial y}\right) = x - y \neq 0$$
$$\frac{\partial}{\partial \bar{z}} f_3(z) = \frac{1}{2\bar{z}}\left(r\frac{\partial f_3}{\partial r} + i\frac{\partial f_3}{\partial \theta}\right) = \frac{2rz - irz}{2\bar{z}} \neq 0$$

On the other hand, the following functions are analytic for all z:

$$f_4(z) = x^2 + 2ixy - y^2 = z^2$$
$$f_5(z) = \ln(r) + i\theta = \ln(z)$$
$$f_6(z) = r^\alpha e^{i\alpha\theta} = z^\alpha$$

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*:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$$
$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

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

$$u(x, y) = c_1$$
$$v(x, y) = c_2$$

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

$$\begin{aligned}\nabla^2 f &= \nabla \cdot \nabla f(x, y) \\ &= \frac{\partial^2 u}{\partial_{xx}} + \frac{\partial^2 v}{\partial_{xx}} + \frac{\partial^2 u}{\partial_{yy}} + \frac{\partial^2 v}{\partial_{yy}} \\ &= \frac{\partial^2 v}{\partial_{xy}} - \frac{\partial^2 u}{\partial_{xy}} - \frac{\partial^2 v}{\partial_{xy}} + \frac{\partial^2 u}{\partial_{xy}} = 0\,,\end{aligned}$$

or more strongly,

$$\nabla^2 u(x, y) = 0$$
$$\nabla^2 v(x, y) = 0\,.$$

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 $\vec{A}$ and $\vec{B}$ defined in terms of harmonic functions $u(x,y)$, $v(x,y)$:

$$\vec{A} = u\,\hat{x} - v\,\hat{y}$$
$$\vec{B} = v\,\hat{x} + u\,\hat{y}$$

Imposing the Cauchy-Riemann equations onto $\vec{A}$ and $\vec{B}$, we see that the divergence and curl of each field resemble Maxwell's equations in charge-free two-dimensional space:

$$\vec{\nabla}\cdot\vec{A} = 0$$
$$\vec{\nabla}\times\vec{A} = 0$$
$$\vec{\nabla}\cdot\vec{B} = 0$$
$$\vec{\nabla}\times\vec{B} = 0$$

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

$$x - x_0 = \rho\cos\theta$$
$$y - y_0 = \rho\sin\theta\,,$$

$f(z)$ my be written

$$\begin{aligned} f(z) &= \frac{x - x_0}{(x-x_0)^2 + (y-y_0)^2} \\ &\quad - i\frac{y - y_0}{(x-x_0)^2 + (y-y_0)^2} \\ &= \frac{1}{\rho}(\cos\theta - i\sin\theta) \\ &= u + iv\,, \end{aligned}$$

where

$$u = \rho^{-1}\cos\theta$$
$$v = -\rho^{-1}\sin\theta\,.$$

Then, the fields $\vec{A} = u\,\hat{x} - v\,\hat{y}$ and $\vec{B} = v\,\hat{x} + u\,\hat{y}$ respectively tell us

$$\vec{A} = \frac{\cos\theta\,\hat{x} + \sin\theta\,\hat{y}}{\rho} = \frac{\hat{r}}{\rho}$$
$$\vec{B} = \frac{-\sin\theta\,\hat{x} + \cos\theta\,\hat{y}}{\rho} = \frac{\hat{\theta}}{\rho}\,.$$

Explicitly, $\vec{A}$ is proportional to the electric field vector due to a line of charge, whereas $\vec{B}$ 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 $\vec{A} = u\hat{x} - v\hat{y}$, $\vec{B} = v\hat{x} + u\hat{y}$, write the integral in vector notation

$$\int_C f(z)\ dz = \int_C \vec{A}\cdot d\vec{l} + i\int_C \vec{B}\cdot d\vec{l}\,,$$

where $d\vec{l} = dx\,\hat{x} + dy\,\hat{y}$. Evidently, the the integral of a function $f(z)$ in the complex plane decomposes into a pair of integrals involving the fields $\vec{A}$, $\vec{B}$.

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 $\hat{k}$, we have

$$\begin{aligned} \oint_C f(z)\ dz &= \int_\Omega \hat{k}\cdot\left(\vec{\nabla}\times\vec{A}\right)\,dx\,dy \\ &\quad + i\int_\Omega \hat{k}\cdot\left(\vec{\nabla}\times\vec{B}\right)\,dx\,dy\,, \end{aligned}$$

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

$$z(\theta) = \lim_{r\to 0} z_0 + re^{i\theta} .$$

The act of taking $r \to 0$ 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:

$$I_n = \oint_C \sum_{m=-\infty}^{\infty} \frac{a_m (z-z_0)^m\, dz}{(z-z_0)^n}$$
$$= \sum_{m=-\infty}^{\infty} a_m \oint_C \frac{dz}{(z-z_0)^{n-m}}$$

The remaining integral is simply $I^{(0)}_{n-m}$, so we have

$$\begin{aligned} I_n &= \sum_{m=-\infty}^{\infty} a_m I^{(0)}_{n-m} \\ &= \sum_{m=-\infty}^{\infty} a_m 2\pi i\, \delta\left(n-m-1\right) \ , \end{aligned}$$

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:

$$\begin{aligned} a_n &= \frac{1}{2\pi i}\oint_\Gamma \frac{f\left(z\right)dz}{\left(z-z_0\right)^{n+1}} \\ n &= 0, \pm 1, \pm 2, \ldots \end{aligned}$$

Note that Γ is a contour topologically equivalent to C_0.

Example: Annulus

Consider a function $f\left(z\right)$ that is analytic in the annulus

$$R_1 \leq \left|z-z_0\right| \leq R_0$$

centered on z_0. Begin by writing the $n=0$ case of the Cauchy integral formula

$$2\pi i\, f\left(z\right) = \oint_C \frac{f\left(\tilde{z}\right)d\tilde{z}}{\tilde{z}-z} \ ,$$

where the function $f\left(z\right)$ 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:

$$a_{n\geq 0} = \frac{1}{2\pi i}\oint_{C_0} \frac{f\left(z\right)dz}{\left(z-z_0\right)^{n+1}}$$

Along contour C_1, the fraction $1/\left(\tilde{z}-z\right)$ may be expanded via geometric series

$$\begin{aligned} \frac{1}{\tilde{z}-z} &= \frac{1}{\tilde{z}-z_0+z_0-z} \\ &= \frac{1}{z_0-z}\,\frac{1}{1+\frac{\tilde{z}-z_0}{z_0-z}} = \sum_{m=0}^{\infty}\frac{\left(z_0-\tilde{z}\right)^m}{\left(z_0-z\right)^{m+1}} \ , \end{aligned}$$

which guarantees convergence as

$$\left|z-z_0\right| > \left|\tilde{z}-z_0\right| = R_1 \ .$$

To discover the restriction on a_n along C_1, replace $\tilde{z}-z$ and $f\left(\tilde{z}\right)$ in the integral formula as follows:

$$\begin{aligned} 2\pi i\, f\left(z\right) &= \oint_{C_1}\frac{f\left(\tilde{z}\right)\,d\tilde{z}}{\tilde{z}-z} \\ &= \oint_{C_1}\sum_{n=-\infty}^{\infty} a_n\left(\tilde{z}-z_0\right)^n \\ &\quad\times\sum_{m=0}^{\infty}\frac{\left(z_0-\tilde{z}\right)^m}{\left(z_0-z\right)^{m+1}}\,d\tilde{z} \end{aligned}$$

Keep condensing terms to write

$$\begin{aligned} 2\pi i\, f\left(z\right) &= \sum_{n=-\infty}^{\infty} a_n \sum_{m=0}^{\infty}\left(z-z_0\right)^{-(m+1)} \\ &\quad\times\oint_{C_1}\left(\tilde{z}-z_0\right)^{m+n}\,d\tilde{z} \ , \end{aligned}$$

and further

$$\begin{aligned} \cancel{2\pi i}\, f\left(z\right) &= \cancel{2\pi i}\sum_{n=-\infty}^{\infty} a_n \\ &\quad\times\sum_{m=0}^{\infty}\left(z-z_0\right)^{-(m+1)}\delta\left(-\left(m+n\right)-1\right) \ . \end{aligned}$$

The delta functions tells us $m+n=-1$, and we find

$$f\left(z\right) = \sum_{n=-1}^{-\infty} a_n\left(z-z_0\right)^n \ .$$

Evidently, only the negative n-terms have survived on contour C_1. We conclude that

$$a_{n<0} = \frac{1}{2\pi i}\oint_{C_1}\frac{f\left(z\right)dz}{\left(z-z_0\right)^{n+1}} \ .$$

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 $f\left(z\right)$ integrated along a closed contour C free of singularities always resolves to zero.

Isolated singularities (poles) $z_0^{(p)}$ are handled by expanding $f\left(z\right)$ as a Laurent series

$$f\left(z\right) = \sum_{n=-\infty}^{\infty} a_n^{(p)}\left(z-z_0^{(p)}\right)^n \ ,$$

where the coefficients a_n were found to be

$$\begin{aligned} a_n^{(p)} &= \frac{1}{2\pi i}\oint_{C_0^{(p)}}\frac{f\left(z\right)\,dz}{\left(z-z_0^{(p)}\right)^{n+1}} \\ n &= 0, \pm 1, \pm 2, \ldots \ . \end{aligned}$$

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

$$\lim_{z \to z_0} \frac{z - z_0}{q(z)} = \frac{1}{q'(z)}\,,$$

and then eliminate $z - z_0$ using

$$g(z) = (z - z_0) f(z) = (z - z_0) \frac{p(z)}{q(z)}$$

to get

$$\lim_{z \to z_0} g(z) = \frac{p(z)}{q'(z)} = \text{Res}[f(z_0)]\,.$$

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

$$I = 2\pi i \operatorname{Res}\left[f\left(i\right)\right] = \pi$$

Example 1
Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{dx}{\left(x^2+1\right)\left(x^2+4\right)}$$

Contour contains two poles.

$$I = \oint_C \frac{dz}{\left(z+i\right)\left(z-i\right)\left(z+2i\right)\left(z-2i\right)}$$

$$g_1\left(z\right) = \frac{\cancel{\left(z-i\right)}}{\cancel{\left(z-i\right)}\left(z+i\right)\left(z^2+4\right)}$$

$$g_2\left(z\right) = \frac{\cancel{\left(z-2i\right)}}{\left(z^2+1\right)\cancel{\left(z-2i\right)}\left(z+2i\right)}$$

$$I = 2\pi i\left(g_1\left(i\right)+g_2\left(2i\right)\right) = \frac{\pi}{6}$$

(Or use partial fractions.)

Example 2
Evaluate:

$$I = \int_0^{\infty} \frac{x^2\,dx}{\left(x^2+4\right)\left(x^2+9\right)}$$

Contour contains two poles.

$$2I = \oint_C \frac{z^2\,dz}{\left(z+2i\right)\left(z-2i\right)\left(z+3i\right)\left(z-3i\right)}$$

$$g_1\left(z\right) = \frac{z^2\cancel{\left(z-2i\right)}}{\left(z+2i\right)\cancel{\left(z-2i\right)}\left(z^2+9\right)}$$

$$g_2\left(z\right) = \frac{z^2\cancel{\left(z-3i\right)}}{\left(z^2+4\right)\left(z+3i\right)\cancel{\left(z-3i\right)}}$$

$$I = \frac{2\pi i}{2}\left(g_1\left(2i\right)+g_2\left(3i\right)\right) = \frac{\pi}{10}$$

Example 3
Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{dx}{\left(x^2+1\right)^2}$$

Contour contains one order-two pole.

$$I = \oint_C \frac{dz}{\left(\left(z+i\right)\left(z-i\right)\right)^2}$$

$$g\left(z\right) = \frac{\cancel{\left(z-i\right)^2}}{\left(\left(z+i\right)\cancel{\left(z-i\right)}\right)^2}$$

$$I = 2\pi i\left(\frac{d}{dz}g\left(z\right)\Big|_{z=i}\right) = \frac{\pi}{2}$$

Example 4
Evaluate:

$$I = \int_0^{\infty} \frac{dx}{\left(4x^2+1\right)^3}$$

Contour contains one order-3 pole.

$$2I = \oint_C \frac{dz}{4^3\left(z+\frac{i}{2}\right)^3\left(z-\frac{i}{2}\right)^3}$$

$$g\left(z\right) = \frac{\cancel{\left(z-\frac{i}{2}\right)^3}}{4^3\left(z+\frac{i}{2}\right)^3\cancel{\left(z-\frac{i}{2}\right)^3}}$$

$$I = \frac{2\pi i}{2}\left(\frac{1}{2}\frac{d^2}{dz^2}\,g\left(z\right)\Big|_{z=i/2}\right) = \frac{3\pi}{32}$$

5.3 Jordan's Lemma

Starting with the Fourier transform integral

$$I = \int_{-\infty}^{\infty} f\left(x\right)e^{ikx}\,dx\,,$$

we carry the problem to the complex plane under three assumptions: (i) $k > 0$ and is a real number, (ii) $f\left(z\right)$ is analytic in the upper-half plane with the exception of simple poles, (iii) $\lim_{|z|\to\infty} f\left(z\right) = 0$.

Jordan's lemma for Fourier transform states that the integration path can be closed by an infinite semi-circle in the upper-half plane. For th $k < 0$ case, the path would enclose the lower half-plane.

5.4 Sine and Cosine in Polynomial

For a real variable $a > 0$, the integral

$$I = \int_{-\infty}^{\infty} \frac{\cos kx}{x^2+a^2}\,dx$$
$$= \operatorname{Re}\int_{-\infty}^{\infty} \frac{e^{ikx}\,dx}{x^2+a^2} = \oint_C \frac{e^{ikz}\,dz}{z^2+a^2}$$

is easily recast as a contour integral using Jordan's lemma.

Using $g\left(z\right) = e^{ikz}/\left(z+ia\right)$, we quickly find $I = \left(\pi/a\right)e^{-ka}$. Generalizing this, one finds:

$$\int f\left(x\right)\cos\left(kx\right)\,dx = \operatorname{Re}\int f\left(x\right)e^{ikx}\,dx$$
$$\int f\left(x\right)\sin\left(kx\right)\,dx = \operatorname{Im}\int f\left(x\right)e^{ikx}\,dx$$

Example 5
Evaluate:

$$I = \int_0^{\infty} \frac{\cos 2x}{9x^2+4}\,dx$$

Contour contains one pole.

$$I = \frac{1}{18}\,\mathrm{Re}\oint_C \frac{e^{2iz}\,dz}{\left(z+\frac{2i}{3}\right)\left(z-\frac{2i}{3}\right)}$$

$$g\left(z\right) = \frac{e^{2iz}}{z+\frac{2i}{3}}$$

$$g\left(2i/3\right) = \frac{3}{4i}\,e^{-4/3}$$

$$I = \frac{2\pi i}{18}\frac{3}{4i}\,e^{-4/3} = \frac{\pi}{12}\,e^{-4/3}$$

Example 6

Evaluate:

$$I = \int_0^\infty \frac{\cos 2x}{\left(9x^2+4\right)^2}\,dx$$

Contour contains one order-two pole.

$$I = \frac{1}{2\cdot 9^2}\,\mathrm{Re}\oint_C \frac{e^{2iz}\,dz}{\left(z+\frac{2i}{3}\right)^2\left(z-\frac{2i}{3}\right)^2}$$

$$g\left(z\right) = \frac{e^{2iz}\cancel{\left(z-\frac{2i}{3}\right)^2}}{\left(z+\frac{2i}{3}\right)^2\cancel{\left(z-\frac{2i}{3}\right)^2}}$$

$$g^{(1)}\left(2i/3\right) = -i\left(\frac{14}{3}\cdot\frac{3^3}{4^3}\right)e^{-4/3}$$

$$I = 2\pi i\left(g^{(1)}\left(2i/3\right)\right) = \frac{7\pi}{288}\,e^{-4/3}$$

Example 7

Evaluate:

$$I = \int_{-\infty}^\infty \frac{\cos kx}{\left(x^2+a^2\right)^2}\,dx$$

Contour contains one order-two pole.

$$I = \mathrm{Re}\oint_C \frac{e^{ikz}\,dz}{\left(z^2+a^2\right)^2}$$

$$g\left(z\right) = \frac{e^{ikz}}{\left(z+ia\right)^2}$$

$$g^{(1)}\left(ia\right) = \frac{i}{4\,e^{ka}}\frac{ka+1}{a^3}$$

$$I = \frac{\pi}{2\,e^{ka}}\frac{ka+1}{a^3}$$

Example 8

Evaluate:

$$I = \int_{-\infty}^\infty \frac{x\sin kx}{x^2+a^2}\,dx$$

Contour contains one pole.

$$I = \mathrm{Im}\oint_C \frac{z\,e^{ikz}\,dz}{z^2+a^2}$$

$$g\left(z\right) = \frac{z\,e^{ikz}}{z+ia}$$

$$g\left(ia\right) = \frac{1}{2}e^{-ka}$$

$$I = \mathrm{Im}\,\frac{2\pi i}{2\,e^{ka}} = \frac{\pi}{e^{ka}}$$

Example 9

Evaluate:

$$I = \int_{-\infty}^\infty \frac{x\sin kx}{\left(x^2+a^2\right)^2}\,dx$$

Exploit a previous example to ease calculations.

$$\begin{aligned} I &= -\frac{\partial}{\partial k}\int_{-\infty}^\infty \frac{\cos kx}{\left(x^2+a^2\right)^2}\,dx \\ &= -\frac{\partial}{\partial k}\left(\frac{\pi}{2\,e^{ka}}\frac{ka+1}{a^3}\right) = \frac{\pi k}{2a\,e^{ka}} \end{aligned}$$

or, by standard means:

$$I = \mathrm{Im}\oint_C \frac{z\,e^{ikz}\,dz}{\left(z^2+a^2\right)^2} = \mathrm{Im}\,\frac{2\pi ik}{4a\,e^{ka}} = \frac{\pi k}{2a\,e^{ka}}$$

Example 10

Evaluate:

$$I = \int_{-\infty}^\infty \frac{x\sin x}{x^2+4x+5}\,dx$$

Contour contains one pole.

$$I = \mathrm{Im}\oint_C \frac{z\,e^{ikz}\,dz}{\left(z+2+i\right)\left(z+2-i\right)}$$

$$g\left(z\right) = \frac{z\,e^{ikz}\cancel{\left(z+2-i\right)}}{\left(z+2+i\right)\cancel{\left(z+2-i\right)}}$$

$$g\left(-2+i\right) = \frac{\left(-2+i\right)\,e^{i(-2+i)}}{2i}$$

$$I = \frac{\pi}{e}\left(2\sin 2+\cos 2\right)$$

5.5 Trigonometric Functions

Integrals of the form

$$I = \int_0^{2\pi} f\left(\cos\left(\theta\right), \sin\left(\theta\right)\right) d\theta$$

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 $dz = ire^{i\theta}\, d\theta$, the above becomes

$$I = -i \oint_{C_0} f\left(\frac{z + z^{-1}}{2}, \frac{z - z^{-1}}{2i}\right) \frac{dz}{z}\,.$$

Example 11

Evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{13 + 5\sin\theta}$$

Contour contains one pole.

$$I = -i \oint_{C_0} \frac{dz}{z} \frac{1}{13 + (5/2i)\left(z - \bar{z}\right)}$$
$$= \oint_{C_0} \frac{dz}{\frac{5}{2}\left(z + \frac{i}{5}\right)\left(z + 5i\right)}$$
$$g\left(z\right) = \frac{\cancel{\left(z + \frac{i}{5}\right)}}{\frac{5}{2}\cancel{\left(z + \frac{i}{5}\right)}\left(z + 5i\right)}$$
$$g\left(\frac{-i}{5}\right) = \frac{-i}{12}$$
$$I = 2\pi i \left(\frac{-i}{12}\right) = \frac{\pi}{6}$$

Example 12

Evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{5 - 4\sin\theta}$$

Contour contains one pole.

$$I = -i \oint_{C_0} \frac{dz}{z} \frac{1}{5 + 2i\left(z - \bar{z}\right)}$$
$$= -\oint_{C_0} \frac{dz}{2\left(z - \frac{i}{2}\right)\left(z - 2i\right)}$$
$$g\left(z\right) = \frac{-1 \cdot \cancel{\left(z - \frac{i}{2}\right)}}{\cancel{\left(z - \frac{i}{2}\right)} 2\left(z - 2i\right)}$$
$$g\left(\frac{i}{2}\right) = \frac{-i}{3}$$
$$I = 2\pi i \left(\frac{-i}{3}\right) = \frac{2\pi}{3}$$

Example 13

For $a > |b|$, evaluate:

$$I = \int_0^{2\pi} \frac{d\theta}{a + b\cos\theta}$$

Contour contains one pole.

$$I = -i \oint_{C_0} \frac{2\, dz}{2az + b\left(1 + z^2\right)}$$
$$= -i \oint_{C_0} \frac{2\, dz}{b\left(z + \frac{a}{b} - \frac{\sqrt{a^2 - b^2}}{b}\right)\left(z + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}\right)}$$
$$z_0 = -\frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}$$
$$g\left(z\right) = \frac{\cancel{2\left(z - z_0\right)}}{\cancel{b\left(z + \frac{a}{b} - \frac{\sqrt{a^2 - b^2}}{b}\right)}\left(z + \frac{a}{b} + \frac{\sqrt{a^2 - b^2}}{b}\right)}$$
$$g\left(z_0\right) = \frac{1}{\sqrt{a^2 - b^2}}$$
$$I = 2\pi i \left(\frac{-i}{\sqrt{a^2 - b^2}}\right) = \frac{2\pi}{\sqrt{a^2 - b^2}}$$

Example 14

For $a > 1$, evaluate:

$$I = \int_0^{\pi} \frac{d\theta}{\left(a + \cos\theta\right)^2}$$

Contour contains one order-two pole.

$$I = -i \oint_{C_0} \frac{2z\, dz}{\left(2az + 1 + z^2\right)^2}$$
$$= -i \oint_{C_0} \frac{2z\, dz}{\left(z + a - \sqrt{a^2 - 1}\right)^2 \left(z + a + \sqrt{a^2 - 1}\right)^2}$$
$$z_0^+ = -a + \sqrt{a^2 - 1}$$
$$z_0^- = -a - \sqrt{a^2 - 1}$$
$$g\left(z\right) = \frac{2z\cancel{\left(z - z_0^+\right)^2}}{\cancel{\left(z - z_0^+\right)^2}\left(z - z_0^-\right)^2}$$
$$g^{(1)}\left(z\right) = \frac{-2\left(z + z_0^-\right)}{\left(z - z_0^-\right)^3}$$
$$g^{(1)}\left(z_0^+\right) = \frac{a}{2\left(a^2 - 1\right)^{3/2}}$$
$$I = 2\pi i \left(\frac{-ia}{2\left(a^2 - 1\right)^{3/2}}\right) = \frac{\pi a}{\left(a^2 - 1\right)^{3/2}}$$

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:

$$I\left(1-e^{2i\pi a}\right) = \oint_C \frac{e^{az}\,dz}{1+e^z}$$

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,

$$I\left(1-e^{2i\pi a}\right) = 2\pi i\frac{e^{i\pi a}}{e^{i\pi}}\,,$$

and

$$I = \frac{\pi}{\sin(\pi a)}\,.$$

Example 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

$$I = \int_0^{\infty} \frac{dx}{1+x^n}\,.$$

Contour contains one pole.

$$I\left(1-e^{2\pi i/n}\right) = \oint_C \frac{dz}{1+z^n}$$

$$z_0^n = -1 \to z_0 = (-1)^{-n} = e^{i\pi/n}$$

$$g(z) = \frac{\left(z-e^{i\pi/n}\right)}{1+z^n}$$

$$g(z_0) \propto \frac{0}{0} \to \text{Need L'hopital.}$$

$$g(z_0) = -\frac{e^{i\pi/n}}{n}$$

$$I\left(1-e^{2\pi i/n}\right) = -2\pi i\frac{e^{i\pi/n}}{n}$$

$$I = \frac{\pi/n}{\sin(\pi/n)}$$

5.7 Regularization

Principal Value

Consider the *principal value* integral

$$I = \mathrm{P}\int_{-\infty}^{\infty} \frac{f(x)\;dx}{x-x_0}\,,$$

where by shifting $x \to z$, we assume that $f(z)$ is analytic except for a finite number of poles, and that $|f| \to 0$ 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

$$\begin{aligned} I &= \mathrm{P}\int_{-\infty}^{\infty} \frac{f(x)\;dx}{x-x_0} \\ &= \lim_{\delta\to 0}\int_{-\infty}^{\infty} \frac{(x-x_0)\,f(x)\;dx}{(x-x_0)^2+\delta^2} \\ &= \lim_{\delta\to 0}\oint_C \frac{(z-x_0)\,f(z)\;dz}{(z-x_0)^2+\delta^2}\,. \end{aligned}$$

After a little complex algebra, find

$$\begin{aligned} I = &\lim_{\delta\to 0}\oint_C \frac{f(z)\;dz}{z-x_0+i\delta} \\ &+ \lim_{\delta\to 0}\oint_C i\delta\frac{f(z)\;dz}{(z-x_0-i\delta)(z-x_0+i\delta)}\,, \end{aligned}$$

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

$$I^+ = \pi i f(x_0) + \lim_{\delta\to 0} \oint_C \frac{f(z)\ dz}{z - x_0 + i\delta},$$

where if we started with $\delta < 0$ instead, the integration contour would flip to the lower-half plane, resulting in

$$I^- = -\pi i f(x_0) + \lim_{\delta\to 0} \oint_C \frac{f(z)\ dz}{z - x_0 - i\delta}.$$

In tighter notation (regardless of path or the sign of δ), one may write

$$I = \mathrm{P}\int_{-\infty}^{\infty} \frac{f(x)\ dx}{x - x_0} = \mathrm{P}\oint_C \frac{f(z)\ dz}{z - x_0},$$

reminding us to *include* x_0 inside integration contour, but take the residue with a factor of 1/2.

Example 16

Evaluate:

$$I = \int_{-\infty}^{\infty} \frac{\sin x}{x}\, dx$$

This is a straightforward principal value integral:

$$I = \mathrm{Im}\left(\pi i\, e^{i\cdot 0} + \lim_{\delta\to 0} \oint_C \frac{e^{iz}\, dz}{z + i\delta}\right) = \pi$$

Example 17

Evaluate each of:

$$I = \mathrm{P}\int_{-\infty}^{\infty} \frac{\cos kx}{(x - x_0)(x^2 + 2)}\, dx$$

$$J = \mathrm{P}\int_{-\infty}^{\infty} \frac{\sin kx}{(x - x_0)(x^2 + 2)}\, dx$$

$$\begin{aligned} K &= \mathrm{P}\oint_C \frac{e^{ikz}}{(z - x_0)(z^2 + 2)}\, dz \\ &= i\pi \frac{e^{ikx_0}}{x_0^2 + 2} + 2\pi i \left(\frac{e^{-k\sqrt{2}}}{\left(\sqrt{2}i - x_0\right)\left(2\sqrt{2}i\right)}\right) \\ &= \left(\frac{-\pi}{x_0^2 + 2}\left(\sin kx_0 + \frac{x_0\, e^{-\sqrt{2}k}}{\sqrt{2}}\right)\right) \\ &\quad + i\left(\frac{\pi}{x_0^2 + 2}\left(\cos kx_0 - e^{-\sqrt{2}k}\right)\right) \\ &= I + iJ \end{aligned}$$

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:

$$u(x_0) = \frac{1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty} \frac{v(x)}{x - x_0}\, dx$$

$$v(x_0) = \frac{-1}{\pi}\,\mathrm{P}\int_{-\infty}^{\infty} \frac{u(x)}{x - x_0}\, dx$$

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

$$\theta_0 + 2\pi N \le \theta < \theta_0 + 2\pi(N + 1)$$

for an integer N, called the *branch*, that indexes multiples of 2π.

Example 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.

$$\bar{z}\frac{\partial}{\partial \bar{z}}\left(z^a\right) = \frac{a}{2}\, e^{ia\theta}\left(r^a - r^a\right) = 0$$
$$\bar{z}\frac{\partial}{\partial \bar{z}}\left(\ln z\right) = \left(\frac{r}{r} + i^2\right) = 0$$

Products with Powers

We next consider the integral

$$I = \int_0^\infty f\left(x\right) x^a \, dx \, ,$$

where $f\left(x\right)$ 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 $R \to \infty$, 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 $\delta \to 0$ we have:

$$I = \int_{C_+} f\left(z\right) z^a \, dz$$
$$-e^{2i\pi a}\, I = \int_{C_-} f\left(z\right) z^a \, dz$$

Solving for I, the final answer pops out:

$$\int_0^\infty f\left(x\right) x^a \, dx = \frac{1}{1 - e^{2i\pi a}} \oint_C f\left(z\right) z^a \, dz$$

Note that the integration contour surrounds the whole complex plane minus the branch cut. Don't forget to include all poles in residue calculations.

Example 19

Evaluate:

$$I = \int_0^\infty \frac{x^a \, dx}{\left(1 + x\right)^2}$$

Contour contains one order-two pole.

$$\oint_C \frac{z^a \, dz}{\left(1 + z\right)^2} = 2\pi i \sum_p \mathrm{Res}\left[f\left(z_0^{(p)}\right)\right]$$

$$g\left(z\right) = \frac{z^a \cancel{\left(1 + z\right)^2}}{\cancel{\left(1 + z\right)^2}}$$
$$g^{(1)}\left(z\right) = a\, z^{a-1}$$
$$g^{(-1)}\left(z\right) = a\, 1^{a-1}\, e^{i\pi a}\, e^{-i\pi} = -a\, e^{i\pi a}$$
$$I = \frac{-2\pi i a\, e^{i\pi a}}{1 - e^{2i\pi a}} = \frac{\pi a}{\sin\left(\pi a\right)}$$

Positive Real Domain

The general problem

$$I = \int_0^\infty f\left(x\right) \, dx$$

exploits a branch cut spectacularly. We begin by considering a different integral

$$\tilde{I} = \int_0^\infty f\left(x\right) \ln\left(x\right) \, dx$$

on the same contours C_+ and C_- used above, on the branch $0 \le \theta < 2\pi$. This gives us

$$\int_{C_+} f\left(z\right) \ln\left(z\right) dz = \tilde{I}$$
$$\int_{C_-} f\left(z\right) z^a \, dz = -\tilde{I} - 2\pi i \int_0^\infty f\left(x\right) dx \, ,$$

which sum together to perfectly cancel $\tilde{I}$, provided the usual assumptions that allow the integral with $R \to \infty$ to vanish. Solving for the original I, we find:

$$\int_0^\infty f\left(x\right) \, dx = \frac{i}{2\pi} \int_C f\left(z\right) \ln\left(z\right) dz$$

Example 20

Evaluate:

$$I = \int_0^\infty \frac{dx}{\left(x + 2\right)\left(x + 1\right)^2}$$

Contour contains two poles.

$$I = \frac{i}{2\pi} \cdot 2\pi i \left(\frac{d}{dz}\frac{\ln z}{z+2}\bigg|_{z=-1} + \frac{\ln z}{\left(z+1\right)^2}\bigg|_{z=-2} \right)$$
$$= 1 - \ln 2$$

Chapter 18

Iterative Methods

1 Matrix Tools

1.1 Linear Systems Review

The linear system $A\vec{x} = \vec{b}$ is ubiquitous in multivariate systems, and rears its face in a majority of numerical analysis applications.

We'll restrict analysis to systems that are can actually be 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.

Calculating the Inverse

Often, we're interested in 'solving' for $\vec{x}$, which requires having the inverse of A, namely A^{-1} such that

$$\left(A^{-1}A\right)\vec{x} = \vec{x} = A^{-1}\vec{b}\,.$$

Pursuing a popular technique, define the submatrix S_{jk} that removes the jth row and the kth 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 jkth component of the matrix B is equal to a constant times the entire determinant of the S_{kj} sub-matrix.

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:

$$A^{-1} = \frac{1}{\det A} B$$

Calculating the RREF

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 $\vec{b}$ 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 t.he original and its transformed version.

Starting with the upper triangular form, a 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 jth 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 $\vec{x}$ as the right-most column of the augmented matrix $I|x$. This, of course, is Gaussian elimination.

Gauss-Jordan 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.

1.2 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.

For example, 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, kth component of matrix B, such as running the loops backwards.

1.3 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
```

1.4 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 jkth 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 can 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
```

1.5 Sub-Matrix

For a matrix A, the sub-matrix S_{jk} that removes the jth row and kth 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
```

1.6 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:

$$\det A = \sum_{\substack{j \le n \\ k=1}}^{n} (-1)^{j+k} A_{jk} M_{jk}$$

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
```

1.7 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
```

1.8 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 $A|I \to I \to A^{-1}$ system with n rows and $2n$ columns. The structures b and I are interchangeable.

1.9 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 $A|I \to I|A^{-1}$, the upper triangular and lower triangular calculations can be done in any order.

1.10 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 Approximating Integrals

2.1 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

$$f(x_n) - f(x_0) \approx S = \sum_{j=0}^{n-1} f'\left(x_j^*\right) \Delta x \,,$$

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:

$$x_j^* = \begin{cases} x_j & \text{Left sum} \\ x_{j+1} & \text{Right sum} \\ \left(x_j + x_{j+1}\right)/2 & \text{Midpoint sum} \end{cases}$$

Denoting S_{Left}, S_{Right} as the left and right sums respectively, the average of these yields the trapezoid rule:

$$S_{\text{Trap}} = \frac{1}{2}\left(S_{\text{Left}} + S_{\text{Right}}\right)$$

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.

2.2 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

$$I(h) = \frac{2h}{3}\left(A\left(3x_j^2 + h^2\right) + 3Bx_j + 3C\right)\ .$$

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 $3/h$:

$$\begin{aligned}\int_{x_j-h}^{x_j+h} f(x)\,dx &\approx \int_{x_j-h}^{x_j+h} g(x)\,dx \\ &= \frac{h}{3}\left(g(x_j - h) + 4g(x_j) + g(x_j + h)\right)\end{aligned}$$

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

$$\int_a^b f(x)\,dx \approx \sum_{j=0}^{n-1} \frac{h}{3}\left(f(x_j - h) + 4f(x_j) + f(x_j + h)\right)\ , \tag{18.1}$$

which we may take as a final answer.

Pseudocode

As pseudocode, the Equation (18.1) 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 $0 \le x \le 4$ 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)
# Initialize 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:

$$\int_0^4 \left(4x - x^2\right) = \frac{32}{3} = 10.666\overline{6}$$

Weighted Average Identity

Starting with Equation (18.1), for Simpson's rule identify $2h = \Delta x$ and keep simplifying:

$$\int_a^b f(x)\,dx \approx \frac{1}{6}\sum_{j=0}^{n-1}\left(f\left(x_j - \frac{\Delta x}{2}\right) + f\left(x_j + \frac{\Delta x}{2}\right)\right)\Delta x$$
$$\frac{1}{6}\sum_{j=0}^{n-1} 4f(x_j)\,\Delta x\,.$$

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:

$$\int_a^b f(x)\,dx \approx \frac{1}{3}(S_T + 2S_M) \qquad (18.2)$$

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:

$$\int_{-2}^{3}\left(5x^2 - x\right)dx = \frac{335}{6} = 55.8333\overline{3}$$

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.

3 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.

3.1 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 $j = 1, 2, \ldots, n$ 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

$$\frac{\partial F}{\partial m} = 0 = \sum_{j=1}^{n} 2x_j (mx_j + b - y_j)$$
$$\frac{\partial F}{\partial b} = 0 = \sum_{j=1}^{n} 2 (mx_j + b - y_j) \ .$$

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 $(X)(Y) \neq XY$. Note for this example we have α, β never exceeding one.

In terms of the sums X, Y, etc., the minimization of $F(m,b)$ gives a system of two equations with two unknowns:

$$0 = mX^2 + bX - XY$$
$$0 = mX + bn - Y$$

The solution is straightforward using matrix methods or traditional:

$$m = \frac{(n)(XY) - (X)(Y)}{(n)(X^2) - (X)(X)}$$
$$b = \frac{(Y)(X^2) - (X)(XY)}{(n)(X^2) - (X)(X)}$$

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:

$$m \approx 0.479$$
$$b \approx 1.082$$

3.2 Exponential Fit

Consider another set of points $\{(x_j, y_j)\}$ in the Cartesian plane that would be best approximated by an exponential fit

$$y = A\, e^{mx} \ ,$$

where the scaling constant A and exponential parameter m are to be determined form the data given.

For this problem, let $A = \ln(b)$, where b is another constant, and the equation becomes $y = e^{mx+b}$. Take the natural log of both sides to find

$$\ln(y) = mx + b \ .$$

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 logarithm to solve for A.

3.3 Polynomial Fit

For any set of 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:

$$y = A_0 + A_1 x + A_2 x^2 + \cdots + A_m x^m$$

Similar to the straight-line fit, define a vertical distance $h_j(\{A_m\})$ such that

$$h_j = A_0 + A_1 x_j + A_2 x_j^2 + \cdots + A_m x_j^m - y_j \ .$$

Square this distance and sum over all n data points:

$$F(\{A_m\}) = \sum_{j=1}^{n} h_j^2$$
$$= \sum_{j=1}^{n} \left(A_0 + A_1 x_j + \cdots + A_m x_j^m - y_j\right)^2$$

The best-fitting polynomial is the one that that minimizes F with respect to all A_j simultaneously. Writing these out, one finds

$$\frac{\partial F}{\partial A_0} = 2 \sum_{j=1}^{n} \left(A_0 + A_1 x_j + \cdots + A_m x_j^m - y_j\right)$$
$$\frac{\partial F}{\partial A_1} = 2 \sum_{j=1}^{n} x_j \left(A_0 + A_1 x_j + \cdots + A_m x_j^m - y_j\right)$$
$$\frac{\partial F}{\partial A_2} = 2 \sum_{j=1}^{n} x_j^2 \left(A_0 + A_1 x_j + \cdots + A_m x_j^m - y_j\right) \ ,$$

availing the pattern

$$\frac{\partial F}{\partial A_k} = 2 \sum_{j=1}^{n} x_j^k \left(A_0 + A_1 x_j + \cdots + A_m x_j^m - y_j\right) \ ,$$

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:

$$\begin{bmatrix} n & X & X^2 & \cdots & X^m & Y \\ X & X^2 & X^3 & \cdots & X^{m+1} & YX \\ X^2 & X^3 & X^4 & \cdots & X^{m+2} & YX^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ X^m & X^{m+1} & X^{m+2} & \cdots & X^{2m} & YX^m \end{bmatrix}$$

4 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.

4.1 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

$$f_0\left(x_j \leq x < x_{j+1}\right) = y_j$$

Of course, we can also draw lines to the left instead of the right by a shift of index:

$$f_0\left(x_j \leq x < x_{j+1}\right) = y_{j+1}$$

4.2 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.

4.3 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:

$$f_2(x) = A_0 + A_1 x + A_2 x^2$$

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

$$f_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x),$$

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

$$f''(x) = \lim_{h \to 0} \frac{f(x-h) - 2f(x) + f(x+h)}{h^2}.$$

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

$$Q_n(x) = \prod_{j \neq n} (x - x_j)$$
$$R_n(x) = \prod_{æ \neq n} (x_n - x_j) \,,$$

the nth function is

$$L_n(x) = \frac{Q_n(x)}{R_n(x)} \,.$$

In therms of he Lagrange interpolating basis functions, the corresponding $y_n(x)$ is

$$y_n(x) = \sum_{j=0}^{n} y_j L_j(x) \,.$$

4.4 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 $y_2(x)$.

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

$$f_2(x) = A_0 + A_1 x + A_2 x^2 \,,$$

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:

$$f_3(x) = A_0 + A_1 x + A_2 x^2 + A_3 x^3$$

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:

$$\frac{d}{dx}f_5(x) = \sum_{j=1}^{5} jA_j x^{j-1}$$
$$\frac{d^2}{dx^2}f_5(x) = \sum_{j=2}^{5} j(j-1)A_j x^{j-2}$$

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 foll information of the problem is specified in the augmented matrix

$$\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_1^5 & y_1 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_2^5 & y_2 \\ 0 & 1 & 2x_1 & 3x_1^2 & 4x_1^3 & 5x_1^4 & y'_1 \\ 0 & 1 & 2x_2 & 3x_2^2 & 4x_2^3 & 5x_2^4 & y'_2 \\ 0 & 0 & 2 & 3\cdot 2x_1^2 & 4\cdot 3x_1^3 & 5\cdot 4x_1^3 & y''_1 \\ 0 & 0 & 2 & 3\cdot 2x_2^2 & 4\cdot 3x_2^3 & 5\cdot 4x_2^3 & y''_2 \end{bmatrix}.$$

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 18.1.

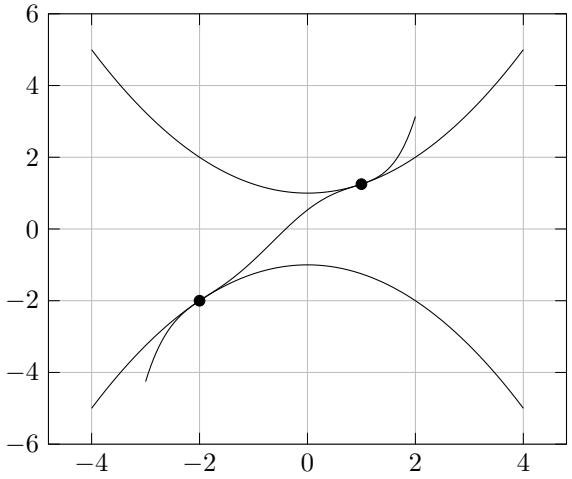


Figure 18.1: Three roads.

4.5 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 $j = 1, 2, 3, \ldots, n$, 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

$$\begin{aligned} f_{j-1}(x_j) &= f_j(x_j) \\ f_j(x_{j+1}) &= f_{j+1}(x_{j+1}) \, , \end{aligned}$$

which says the same thing twice. (Substitute $j \to j+1$ in the first equation to recover the second.)

For continuity in derivative of f, differentiate each of the above once with respect to x

$$\begin{aligned} f'_{j-1}(x_j) &= f'_j(x_j) \\ f'_j(x_{j+1}) &= f'_{j+1}(x_{j+1}) \, , \end{aligned}$$

and then once more:

$$\begin{aligned} f''_{j-1}(x_j) &= f''_j(x_j) \\ f''_j(x_{j+1}) &= f''_{j+1}(x_{j+1}) \, , \end{aligned}$$

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

$$\begin{aligned} f_1(x_1) &= y_1 \\ f_{n-1}(x_n) &= y_n \, . \end{aligned}$$

With the leftover freedom to impose two more equations, choose the second derivative to be zero at each endpoint:

$$\begin{aligned} f''_1(x_1) &= 0 \\ f''_{n-1}(x_n) &= 0 \end{aligned}$$

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

$$\begin{aligned} f'_j(x) = &\frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^2}{2} \\ &+ \frac{k_{j+1}}{(x_{j+1} - x_j)} \frac{(x - x_j)^2}{2} + C \, . \end{aligned}$$

Integrate again to get an equation for $f_j(x)$

$$\begin{aligned} f_j(x) = &\frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^3}{3 \cdot 2} \\ &+ \frac{k_{j+1}}{(x_{j+1} - x_j)} \frac{(x - x_j)^3}{3 \cdot 2} + Cx + D \, , \end{aligned}$$

where D is another integration constant.

The substitutions

$$\begin{aligned} C &= A - B \\ D &= -Ax_{j+1} + Bx_j \end{aligned}$$

makes the above be slightly easier to work with:

$$\begin{aligned} f_j(x) = &\frac{k_j}{(x_j - x_{j+1})} \frac{(x - x_{j+1})^3}{3 \cdot 2} \\ &- \frac{k_{j+1}}{(x_j - x_{j+1})} \frac{(x - x_j)^3}{3 \cdot 2} \\ &+ A(x - x_{j+1}) - B(x - x_j) \end{aligned}$$

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

$$\alpha_j = -\Delta x_{j-}$$
$$\beta_j = 2(\Delta x_{j+} - \Delta x_{j-})$$
$$\gamma_j = \Delta x_{j+}$$
$$\delta_j = 6\left(\frac{\Delta y_{j-}}{\Delta x_{j-}} - \frac{\Delta y_{j+}}{\Delta x_{j+}}\right)$$

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 $j = 2, 3, 4, \ldots, n-1$ need be written out.

Example n=5

Choosing a modest example with $n = 5$ points provided, the above becomes:

$$\alpha_2 0 + \beta_2 k_2 + \gamma_2 k_3 = \delta_2$$
$$\alpha_3 k_2 + \beta_3 k_3 + \gamma_3 k_4 = \delta_3$$
$$\alpha_4 k_3 + \beta_4 k_4 + \gamma_4 0 = \delta_4$$

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 $A\vec{x} = \vec{b}$-like system.

In general, the spline calculation leads to a hefty augmented matrix with $n-2$ rows and $n-1$ columns.

5 Newton's Method

In one dimension, Newton's method is a reliable means for estimating the roots of an equation $g(x)$, which is to say finding the x-value(s) that solve $g(x) = 0$.

The formula for Newton's method comes from a first-order approximation of $g(x)$, namely

$$g_1(x) = g(x_0) + g'(x_0)(x - x_0)\,.$$

Providing an initial guess x_0, Setting $g_1(x_1) = 0$ implicates a new x_1 that should be an an improvement over x_0. This can be continued recursively via the formula

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}\,.$$

Part VI

Advanced Topics

Chapter 19

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.

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:

$$\begin{aligned} A \text{ AND } B &= A \cdot B = A \cap B \\ A \text{ OR } B &= A + B = A \cup B \\ \text{NOT } A &= \bar{A} \\ \text{Empty set} &= \emptyset \end{aligned}$$

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:

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N} \tag{19.1}$$

$P(A)$ 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:

$$N_A + N_{\bar{A}} = N \tag{19.2}$$

$$P\left(\bar{A}\right) + P(A) = 1 \tag{19.3}$$

Classical Probability

A definition that skirts around the invocation of $N \to \infty$ 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.

Example 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?

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:

$$P\left(7766\right) = \frac{1}{N} = \frac{1}{10000}$$

Example 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?

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

$$P\left(N_A\right) = \frac{1}{N} = \frac{1}{256}\,.$$

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:

$$P\left(A \text{ or } B\right) = P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) \quad (19.4)$$

Example 3

Calculate the probability of rolling a 3 or a 4 on a six-sided dice. As mutually exclusive events, we simply have

$$P = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}\,.$$

Example 4

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts.

$$P = \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{3}{52}$$

1.4 Non-Exclusivity

Non-mutually exclusive events are those that cause 'double counting' in $P\left(A \cup B\right)$, and are adjusted by subtracting the probability that both occur:

$$P\left(A \cup B\right) = P\left(A\right) + P\left(B\right) - P\left(A \cap B\right) \quad (19.5)$$

Or, in street terms, the above reads:

$$P\left(A \text{ or } B\right) = P\left(A\right) + P\left(B\right) - P\left(A \text{ and } B\right)$$

Example 5

From a 52-card deck, calculate the probability of drawing a Heart or a Face card, or one that is both.

$$P = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52}$$

Example 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

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

$$P(A \cap B) = \lim_{N\to\infty} \frac{1}{N} N_{A\cap B} \,,$$

where in the $N \to \infty$, 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:

$$P(A \cap B) = P(A)\, P(B) \tag{19.6}$$

Example 7

Calculate the probability of two fair coin tosses each landing on 'tails'.

$$P(T \cap T) = P(T) \cdot P(T) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Example 8

From a 52-card deck, what is the probability of randomly drawing (i) a Queen, (ii) a Heart, (iii) the Queen of Hearts?

$$P(Q) = 1/4$$
$$P(H) = 1/13$$
$$P(Q \cap H) = P(Q) \cdot P(H) = \frac{1}{4} \cdot \frac{1}{13} = \frac{1}{52}$$

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

$$P(B|A) = \lim_{N\to\infty} \frac{1}{N_A} N_{A\cap B} \,.$$

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 $P(A \cap B)$:

$$P(A \cap B) = \lim_{N\to\infty} \frac{N_{A\cap B}}{N} \,.$$

Divide the two equations and simplify to derive the statement of conditional probability:

$$P(A \cap B) = P(B|A)\, P(A) \tag{19.7}$$

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.

Example 9

In a 52-card deck, calculate the probability that the first three cards are Kings.

$$\begin{aligned} P(KKK) &= P(K)\, P(K|K)\, P(K|\,(K|K)) \\ &= \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} \approx 0.000181 \end{aligned}$$

Example 10

In a 52-card deck, calculate the probability that the first three cards are KQJ, in that order, with mixed suits allowed.

$$\begin{aligned} P(KQJ) &= P(K)\, P(Q|K)\, P(J|\,(Q|K)) \\ &= \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} \approx 0.000483 \end{aligned}$$

Example 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 $X + Y \leq 5$?

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:

$$P\left(X = 2 \middle| X + Y \leq 5\right) = \frac{3}{10}$$

Inversion Trick

Making use of the normalization condition

$$P\left(\bar{A}\right) + P(A) = 1\,,$$

it often helps in problem solving to use the negated event $\bar{A}$ as the working variable.

Example 12

Calculate the probability that a random three-card hand drawn from a 52-card deck contains the Queen of Hearts. For this, define the event $\bar{A}$ of *not* drawing the Queen of Hearts and use the inversion trick as follows:

$$\bar{A} = \bar{A}_1 \cdot \bar{A}_2 \cdot \bar{A}_3$$

$$\begin{aligned} P\left(\bar{A}_1 \cap \bar{A}_2\right) &= P\left(\bar{A}_1\right) P\left(\bar{A}_2|\bar{A}_1\right) \\ &= \frac{51}{52} \cdot \frac{50}{51} = \frac{50}{52} \end{aligned}$$

$$\begin{aligned} P\left(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3\right) &= P\left(\bar{A}_1 \cap \bar{A}_2\right) P\left(\bar{A}_3|\bar{A}_1 \cap \bar{A}_2\right) \\ &= \frac{50}{52} \cdot \frac{49}{50} = \frac{49}{52} \end{aligned}$$

$$P(A) = 1 - P\left(\bar{A}\right) = 1 - \frac{49}{52} = \frac{3}{52}$$

Radioactive Decay

An unstable atom is one that expels energy by ejecting a subatomic 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

$$P^{\text{decay}}_{\Delta t/t} = \tau^{-1} \Delta t\,,$$

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

$$P^{\text{alive}}_{\Delta t/t} = 1 - \Delta t/\tau\,.$$

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:

$$P^{\text{alive}}(t) = P^{\text{alive}}_{\Delta t/t_1} \cdot P^{\text{alive}}_{\Delta t/t_2} \cdots P^{\text{alive}}_{\Delta t/t_n} = \left(1 - \frac{t}{n\tau}\right)^n$$

Letting $n \to \infty$ permits use of the identity

$$\lim_{n\to\infty} \left(1 + \frac{A}{n}\right)^n = e^A\,.$$

It follows that the probability that a single unstable atom will still be 'alive' obeys

$$P(t) = e^{-t/\tau}\,. \tag{19.8}$$

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:

$$p_1(A) = 1/6$$
$$p_1(B) = 1/6$$

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:

$$p_2(A|B) = (1/6)(1/6) = (1/6)^2$$
$$p_2(B|B) = (1/6)(1/6) = (1/6)^2$$

With event $B|B$ comes another re-roll, and we stack on the conditional probabilities as

$$p_3(A|B|B) = (1/6)^3$$
$$p_3(B|B|B) = (1/6)^3 \ ,$$

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

$$P(A) = p_1(A) + p_2(A|B) + p_3(A|B|B) + p_4(A|B|B|B) + \cdots \ ,$$

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*:

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)} \tag{19.9}$$

Laptop Repair Shop

Source: CMPSCI240 UMass Amherst 2013

Problem 1

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 is 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 1

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:

$$P(F) = P(F|D)\,P(D) + P(F|S)\,P(S) + P(F|O)\,P(O) = .26$$

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 $P\left(S|F \cap Z^C\right)$, 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:

$$P\left(Z^C|D\right) = 1 - .80 = .20$$
$$P\left(Z^C|S\right) = 1 - .11 = .89$$
$$P\left(Z^C|O\right) = 1 - .09 = .91$$

Proceed by applying Bayes' theorem directly to write

$$P\left(S|F \cap Z^C\right) = \frac{P\left(F \cap Z^C|S\right)P(S)}{P\left(F \cap Z^C\right)}\,.$$

Due to the independence between F and Z^C, the term $P\left(F \cap Z^C|S\right)$ decouples into $P(F|S)\,P\left(Z^C|S\right)$.

The denominator term $P\left(F \cap Z^C\right)$ can be recast as a sum that factors in a similar way:

$$P\left(F \cap Z^C\right) = \sum_{i=D,S,O} P\left(F \cap Z^C|X_i\right)P(X_i) = \sum_{i=D,S,O} P(F|X_i)\,P\left(Z^C|X_i\right)P(X_i)$$

Evaluating the final answer is now straightforward:

$$P\left(S|F \cap Z^C\right) = .7169$$

1.8 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

$$T_{\max} = 4t \qquad T_{\min} = t \times \frac{4}{3}$$

to establish upper and lower estimates of the wall's lifespan. Inserting $t = 8$ years, we find

$$T_{\max} = 32\,\text{yr} \qquad T_{\min} \approx 10.67\,\text{yr}\,,$$

which, when these are added to 1961, produces the pair of results

$$\text{Year}_{\max} \approx 1993 \qquad \text{Year}_{\min} \approx 1972\,.$$

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

$$T_{max} = Nt \qquad T_{min} = \frac{Nt}{N-1}\,.$$

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:

$$p\left(\Delta T\right) = \frac{-2t}{\Delta T} + \sqrt{\left(\frac{2t}{\Delta T}\right)^2 + 1}$$

Example 13

Suppose you encounter a man for the first time in the 42^{nd} 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%.

$$T_{max(33\%)} = 3 \times 42 = 126$$
$$T_{min(33\%)} = \frac{3}{2} \times 42 = 63$$

$$T_{max(66\%)} = 6 \times 42 = 252$$
$$T_{min(66\%)} = \frac{6}{5} \times 42 = 50.4$$

$$T_{max(75\%)} = 8 \times 42 = 336$$
$$T_{min(75\%)} = \frac{8}{7} \times 42 = 48$$

1.9 Dice Stacking

Suppose you are given a pair of distinguishable three-sided dice, abbreviated $2d3$, meant to be rolled simultaneously or in sequence. Denoting the outcomes of either given dice as 1, 2, 3, we can list the possible outcomes for one roll of such a pair of dice:

$$\omega = 11, 12, 21, 13, 22, 31, 23, 32, 33$$

The concatenation of outcomes from $2d3$ is equivalent to a single roll of an *effective* nine-sided dice, abbreviated $d9$.

Effective d5?

One may wonder if other effective dice can be derived from the roll $2d3$. For instance, replacing the concatenation of outcomes with the sum (inserting a plus sign between each digit in the above ω-list), one finds

$$\sigma = 2, 3, 3, 4, 4, 4, 5, 5, 6\ .$$

In a more visual notation, the above is equivalent to:

$$\sigma_2 = \boxdot\boxdot$$
$$\sigma_3 = \boxdot\boxdot, \boxdot\boxdot$$
$$\sigma_4 = \boxdot\boxdot, \boxdot\boxdot, \boxdot\boxdot$$
$$\sigma_5 = \boxdot\boxdot, \boxdot\boxdot$$
$$\sigma_6 = \boxdot\boxdot$$

In total, there are five possible outcomes as sums ranging from 2 to 6. The distribution of outcomes, however, is nonuniform. For instance, one sees that σ_4 can be reached three ways, whereas all other σ_j are less common.

Density Inversion

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.

$$\gamma = \frac{1}{1}\left(2\right), \frac{1}{2}\left(3\right), \frac{1}{3}\left(4\right), \frac{1}{2}\left(5\right), \frac{1}{1}\left(6\right)\ .$$

Multiply by $3 \cdot 2$ to get rid of all denominators:

$$\gamma = 6\left(2\right), 3\left(3\right), 2\left(4\right), 3\left(5\right), 6\left(6\right)$$

Explicitly, γ is a list with twenty items:

$$\gamma = 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 6$$

Qualitatively, we see 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 an gamma ray γ, assign the items in γ to an index via $\gamma\left(k\right)$, 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 of 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 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:

$$x = \texttt{dice}(3) + \texttt{dice}(3)$$

3. If $x = \gamma(k)$, record $x - 1$ as a valid result.
4. Let $k \to k + 1$.
5. If $k > 20$ then Goto 1.
6. Goto 2.

2 Combinatorics

2.1 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 overcounts by a factor of m-factorial, which must be divided out:

$$A_n^m = \frac{n!}{m!} \tag{19.10}$$

Example 1

Consider the set of twelve elements $ACEFGILMNTUX$. What is the probability that a random arrangement of the elements will spell out $MAGNETICFLUX$?

$$P(MAGNETICFLUX) = \frac{1}{12!}$$

Example 2

Consider the word $FLUXELECTRIC$. What is the probability that a random arrangement of the letters will spell out $ELECTRICFLUX$?

$$P(ELECTRICFLUX) = \frac{2!2!2!}{12!}$$

2.2 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:

$$P_n^m = \frac{n!}{(n-m)!} \tag{19.11}$$

Qualitatively, the permutation number tells how many ways there are to choose m unique elements from a set of n total elements.

Example 3

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?

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:

$$P_{10}^4 = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 5040$$

Example 4

In a 52-card deck, calculate the probability that the first three cards are Kings. (This is a repeat of an earlier problem.)

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:

$$P(KKK) = \frac{P_4^3}{P_{52}^3} = \frac{4 \cdot 3 \cdot 2}{52 \cdot 51 \cdot 50} \approx 0.000181$$

Example 5

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.)

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:

$$P(KQJ) = \frac{4^3}{52 \cdot 51 \cdot 50} \approx 0.000483$$

2.3 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.)

Conditional Probability Analysis

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:

$$X(n|n-1) = \frac{365 - (n-1)}{365}$$

$$X(N) = \prod_{n=2}^{N} X(n|n-1)$$

Permutation Analysis

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.

2.4 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.

Example 6

From a 52-card deck, a five-card hand is drawn at random. How many five-card hands are possible?

$$C_{52}^{5} = \frac{52!}{5!\,(52-5)!}$$

Example 7

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.

$$P(RF) = \frac{4}{C_{52}^{5}} = \frac{1}{649740} \approx 0.00000154$$

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.

$$P(5,1) = \frac{1}{C_{75}^5 \cdot C_{15}^1} = \frac{1}{258,890,850}$$
$$P(4,1) = \frac{C_5^4 \cdot C_{70}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{739,688}$$
$$P(3,1) = \frac{C_5^3 \cdot C_{70}^2}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{10,720}$$

$$P(2,1) = \frac{C_5^2 \cdot C_{70}^3}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{473}$$
$$P(1,1) = \frac{C_5^1 \cdot C_{70}^4}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{56}$$
$$P(0,1) = \frac{C_5^0 \cdot C_{70}^5}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{21}$$

$$P(5,0) = \frac{C_5^5 \cdot C_{70}^0 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{18,492,204}$$
$$P(4,0) = \frac{C_5^4 \cdot C_{70}^1 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{52,835}$$
$$P(3,0) = \frac{C_5^3 \cdot C_{70}^2 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{766}$$

$$P(2,0) = \frac{C_5^2 \cdot C_{70}^3 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{34}$$
$$P(1,0) = \frac{C_5^1 \cdot C_{70}^4 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{1}{4}$$
$$P(0,0) = \frac{C_5^0 \cdot C_{70}^5 \cdot C_{14}^1}{C_{75}^5 \cdot C_{15}^1} \approx \frac{2}{3}$$

3 Variables and Expectations

3.1 Normalization

For an event A, the probability $P(A)$ of the event occurring has a trivial yet important relationship to $P\left(\overline{A}\right)$, via the normalization condition

$$1 = P(A) + P\left(\overline{A}\right) .$$

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 $\overline{A}$ and all subsequent $\overline{A}_k$ into the coefficients A_k to write a more general normalization condition:

$$1 = \sum_{k=1}^{n} P(A_k) \tag{19.12}$$

3.2 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:

$$\langle A \rangle = \sum_{k=1}^{n} A_k \cdot P(A_k) \tag{19.13}$$

3.3 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:

$$\langle f \rangle = \sum_{k=1}^{n} f(A_k) \cdot P(A_k) \tag{19.14}$$

With the above, we may also calculate $\left\langle f^2 \right\rangle$ without hesitation:

$$\left\langle f^2 \right\rangle = \sum_{k=1}^{n} \left(f(A_k)\right)^2 \cdot P(A_k) \tag{19.15}$$

Example 1

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.

$$\langle A \rangle = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} + \frac{4}{6} + \frac{5}{6} + \frac{6}{6} = \frac{21}{6} = 3.5$$

Example 2

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.

$$\langle A \rangle = \frac{1}{6} + \frac{0}{6} + \frac{3}{6} + \frac{4 \cdot 2}{6} + \frac{5}{6} + \frac{6}{6} = \frac{23}{6} = 3.83$$

3.4 Standard Deviation

Further insight into f can be gained by inserting $(f(A_k) - \langle f \rangle)^2$ as the argument in Equation (19.15). By doing so, and then taking the square root of the entire result, we arrive at the *standard deviation* in the system:

$$\sigma_f = \sqrt{\sum_{k=1}^{n} (f(A_k) - \langle f \rangle)^2 \cdot P(A_k)} \tag{19.16}$$

Using only the definitions above, it's easy to show that the standard deviation is equivalent to

$$\sigma_f = \sqrt{\langle f^2 \rangle - 2\langle f \rangle \langle f \rangle + \langle f \rangle^2}$$
$$\sigma_f = \sqrt{\langle f^2 \rangle - \langle f \rangle^2}\,. \tag{19.17}$$

3.5 Continuous Distributions

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

$$\sum_{k=1}^{n} P(A_k) \to \int dP(A_k) = 1\,.$$

Probability Distribution Function

At this point, we abbreviate $A_k \to k$, 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, $\omega(k)$ 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:

$$\langle k \rangle = \int_n k \cdot w(k)\, dk \tag{19.18}$$

$$\langle f \rangle = \int_n f(k)\, w(k)\, dk \tag{19.19}$$

$$\sigma_f = \sqrt{\int_n (f(k) - \langle f \rangle)^2\; w(k)\, dk} \tag{19.20}$$

Note that Equation (19.17) still holds in the continuous distribution.

Example 3

What is the expected area of a right triangle with a hypotenuse of k whose non-right angles are uniformly distributed over the interval $(0, \pi/2)$?

$$\langle A \rangle = \frac{k^2/2}{\pi/2} \int_0^{\pi/2} \cos(\theta) \sin(\theta)\; d\theta$$
$$= \frac{k^2/2}{\pi/2} \int_0^1 x\, dx = \frac{k^2}{2\pi}$$

Example 4

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?

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:

$$P = N \int_{L/2}^{L} \frac{L - x}{x}\, dx$$
$$= \frac{(L \ln x - x)\,\big|_{L/2}^{L}}{L/2} = 2\ln 2 - 1 \approx 38.6\%$$

3.6 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

$$\langle A \rangle = \int A \cdot w(A_1, \ldots, A_n)\, dA_1 \ldots dA_n\,.$$

Replace A in the above with its sum representation:

$$\langle A \rangle = \sum_{k=1}^{n} \int A_k \cdot w(A_1, \ldots, A_n)\, dA_1 \ldots dA_n\,,$$

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

$$\langle A \rangle = \langle A_1 \rangle + \langle A_2 \rangle + \cdots + \langle A_n \rangle \ ,$$

which, strictly translated, means *the average of the sum is the sum of the averages*:

$$\langle A \rangle = \sum_{k=1}^{n} \langle A_k \rangle \tag{19.21}$$

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

$$w(A_1, A_n, \ldots, A_n) = w(A_1) w(A_2) \cdots w(A_n)$$

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

$$B = \prod_{k=1}^{n} B_k = B_1 \cdot B_2 \cdots B_n \ ,$$

and let us calculate the average value $\langle B \rangle$. Going by definition, this amounts to

$$\langle B \rangle = \prod_{k=1}^{n} \int B_k \cdot w(B_1) \cdots w(B_n) \, dB_1 \ldots dB_n \ ,$$

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

$$\langle B \rangle = \langle B_1 \rangle \cdot \langle B_2 \rangle \cdots \langle B_n \rangle \ ,$$

which, strictly translated, means *the average of the product is the product of the averages*:

$$\langle B \rangle = \prod_{k=1}^{n} \langle B_k \rangle \tag{19.22}$$

3.7 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 $\left\langle A^2 \right\rangle$ by exploiting the the independence among A_k, resulting in

$$\left\langle A^2 \right\rangle = \sum_{k=1}^{n} \left\langle A_k^2 \right\rangle + \sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle \ .$$

Taking the difference $\left\langle A^2 \right\rangle - \langle A \rangle^2$, the cross terms cancel and we arrive at a simple relation connecting A to its members:

$$\left\langle A^2 \right\rangle - \langle A \rangle^2 = \sum_{k=1}^{n} \left\langle A_k^2 \right\rangle - \langle A_k \rangle^2 + \cancel{\sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle} - \cancel{\sum_{i \neq j} c_{ij} \langle A_i \rangle \langle A_j \rangle}$$

The square root of $\left\langle A^2 \right\rangle - \langle A \rangle^2$ is defined as the *variance* in A:

$$\mathrm{Var}(A) = \sqrt{\left\langle A^2 \right\rangle - \langle A \rangle^2} \tag{19.23}$$

As we've built it, the variance has some more handy expressions:

$$\mathrm{Var}(A) = \sqrt{\sum_{k=1}^{n} \left\langle A_k^2 \right\rangle - \langle A_k \rangle^2} = \sqrt{\sum_{k=1}^{n} \left(\mathrm{Var}(A_k) \right)^2}$$

3.8 Dispersion

A variation in the sum A of independent variables, denoted ΔA, is also known as *dispersion*, defined as:

$$\Delta A = A - \langle A \rangle = \sum_{k=1}^{n} \left(A_k - \langle A_k \rangle \right) \tag{19.24}$$

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

$$\begin{aligned}\Delta A^2 &= ((A_1 - \langle A_1 \rangle) + (A_2 - \langle A_2 \rangle) + \cdots)^2 \\ &= \sum_{k=1}^{n} (A_k - \langle A_k \rangle)^2 + \sum_{i \neq j} c_{ij} \Delta A_i \Delta A j \,,\end{aligned}$$

so then:

$$\langle \Delta A^2 \rangle = \sum_{k=1}^{n} \left\langle (A_k - \langle A_k \rangle)^2 \right\rangle + \cancel{\sum_{i \neq j} c_{ij} \langle \Delta A_i \rangle \langle \Delta A j \rangle}$$

This is result is perhaps not surprising, telling us the total ΔA^2 is the sum of its constituents:

$$\langle \Delta A^2 \rangle = \sum_{k=1}^{n} \langle \Delta A_k^2 \rangle \tag{19.25}$$

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

$$\frac{\sqrt{\langle \Delta A^2 \rangle}}{\langle A \rangle} \approx \frac{1}{\sqrt{n}} \to 0 \,,$$

telling us that fluctuations in A become negligibly small.

3.9 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 $X_n \to 0$, 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 $X_n \to 0$.

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 who 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 $X_n \to \infty$ 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:

$$p \approx 2.718\ldots$$

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:

$$\ln(X_n) = \ln(\tilde{x}_1) + \ln(\tilde{x}_2) + \ln(\tilde{x}_3) + \cdots$$

In the limit $n \to \infty$, 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

$$\ln(X) = \lim_{n\to\infty} \sum_{j=1}^{n} \ln\left(p\frac{j}{n}\right) .$$

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

$$\ln(X) = \lim_{\Delta x\to 0} \frac{1}{p} \sum_{x>0}^{1} \ln(px)\,\Delta x .$$

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:

$$p\ln(X) = (x\ln(p) + x\ln(x) - x)\Big|_0^1$$
$$p\ln(X) = \ln(p) - 1$$

Now comes the final argument. By avoiding $\ln(0) \to -\infty$ and also $\ln(\infty) \to \infty$, we're asking for X to be a finite number. In the infinite limit, it can only be that $X \to 1$:

$$p\ln(1) = 0 = \ln(p) - 1$$

The only solution to $\ln(p) = 1$ is $p = e$ and we're done.

Problem 2

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 $n \approx \ln(A)$.

3.10 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 $0 \le r_k \le 1$ 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 hypervolume in four-dimensions, and so on. The space enclosed by each 'volume' is defined by

$$\sum_{j=1}^{n} r_j \le 1 .$$

For convenience, let us label $r_1 \to z$, $r_2 \to y$, $r_3 \to x$, $r_4 \to t$, $r_5 \to u$.

Examining $n = 2$, the line $z + y = 1$ encloses half of the unit square, formally shown via

$$V_2 = \int_0^1 \int_0^{1-z} dy\,dz$$
$$= \int_0^1 (1-z)\,dz = \left(z - \frac{z^2}{2}\right)\Big|_0^1 = \frac{1}{2} .$$

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:

$$\langle n \rangle = e$$

4 Systems and Distributions

4.1 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':

$$C_n^m = \Omega(m, n) = \binom{m}{n}$$

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:

$$2^n = \sum_{m=0}^{n} \frac{n!}{m!\,(n-m)!}$$

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:

$$P(m, n) = \frac{1}{2^n} \frac{n!}{m!\,(n-m)!} \tag{19.26}$$

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

$$C_1^0 = 1 \qquad C_1^1 = 1$$

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

$$C_2^0 = 1 \qquad C_2^1 = 2 \qquad C_2^2 = 1$$

Similarly, a game of three tosses has

$$\omega_3 = (TTT, TTH, THT, THH, HTT, HTH, HHH)$$

with combinations

$$C_3^0 = 1 \qquad C_3^1 = 3 \qquad C_3^2 = 3 \qquad C_3^3 = 1\,.$$

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.

4.2 Binomial Distribution

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:

$$\begin{aligned}
\Omega_1 &= T, H \\
\Omega_2 &= TT, TH, HT, HH \\
\Omega_3 &= TTT, TTH, THT, THH, HTT, HTH, HHH
\end{aligned}$$

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:

$$\begin{aligned}
P_1^0(p) &= 1-p \\
P_1^1(p) &= p
\end{aligned}$$

$$\begin{aligned}
P_2^0(p) &= (1-p)^2 \\
P_2^1(p) &= 2\cdot p\,(1-p) \\
P_2^2(p) &= p^2
\end{aligned}$$

$$\begin{aligned}
P_3^0(p) &= (1-p)^3 \\
P_3^1(p) &= 3\cdot p\,(1-p)^2 \\
P_3^2(p) &= 3\cdot p^2\,(1-p) \\
P_3^3(p) &= p^3
\end{aligned}$$

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:

$$P(m,n,p) = \frac{n!}{m!\,(n-m)!}\,(1-p)^{n-m}\,p^m \qquad (19.27)$$

Note there is no need to divide by 2^n. The binomial distribution as written is unit-normalized already.

Analysis

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

$$\left\langle z_k^2\right\rangle = p\cdot 1^2 + (1-p)\cdot 0^2 = p\,.$$

The standard deviation in z, denoted σ_z, is evidently

$$\sigma_z = \sqrt{\left\langle z_k^2\right\rangle - \langle z_k\rangle^2} = \sqrt{p-p^2} = \sqrt{p\,(1-p)}\,.$$

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

$$\left\langle \Delta m^2\right\rangle = \sum_{k=1}^{n}\left\langle \Delta z_k^2\right\rangle = n\left\langle \Delta z^2\right\rangle\,,$$

or, in tighter notation for large-n systems,

$$\sigma_m = \sqrt{n\sigma_z^2} = \sqrt{np\,(1-p)}\,.$$

Example 1

Monique is practicing netball. She knows from past 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%?

This problem is analogous to flipping a weighted coin with bias p. The multiplicity of scoring k shots in N tosses is

$$\Omega\left(k,N,p\right)=\frac{N!}{k!\left(N-k\right)!}\left(1-p\right)^{N-k}p^{k}\,,$$

where summing over k gives the cumulative distribution:

$$99\%=\sum_{k=50}^{N}\frac{N!}{k!\left(N-k\right)!}.3^{N-k}.7^{k}$$

This is best solved by a computer, where one should find

$$N=86\,.$$

Example 2

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 oozes? Repeat the calculation for finding two cups, three cups, etc., up to sixteen cups. Also account for zero cups.

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

$$\begin{aligned}P\left(16,1,1/3\right)&=\frac{16!}{1!\left(16-1\right)!}\left(2/3\right)^{16-1}\left(1/3\right)^{1}\\&=\frac{16}{3}\left(\frac{2}{3}\right)^{15}\approx0.01218\,,\end{aligned}$$

and then for other numbers of gold cups:

$$\begin{aligned}P\left(16,2,1/3\right)&\approx0.04567\\P\left(16,3,1/3\right)&\approx0.1066\\P\left(16,4,1/3\right)&\approx0.1732\\P\left(16,5,1/3\right)&\approx0.2078\end{aligned}$$

$$\begin{aligned}P\left(16,6,1/3\right)&\approx0.1905\\P\left(16,7,1/3\right)&\approx0.1361\\P\left(16,8,1/3\right)&\approx0.07654\\P\left(16,9,1/3\right)&\approx0.03402\end{aligned}$$

$$\begin{aligned}P\left(16,10,1/3\right)&\approx0.01191\\P\left(16,11,1/3\right)&\approx0.003247\\P\left(16,12,1/3\right)&\approx0.0006765\\P\left(16,13,1/3\right)&\approx0.0001041\\P\left(16,14,1/3\right)&\approx0.00001115\\P\left(16,15,1/3\right)&\approx0.0000007434\\P\left(16,16,1/3\right)&\approx0.00000002323\\P\left(16,0,1/3\right)&\approx0.001522\end{aligned}$$

4.3 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:

- $n_A \to$ Number of outcomes A
- $n_B \to$ Number of outcomes B
- $n_C \to$ Number of outcomes C
- $p_A \to$ Probability of of outcome A
- $p_B \to$ Probability of of outcome B
- $p_C \to$ Probability of of outcome C

With this, we can write the probability of the three-state system exhibiting the state (n_A, n_B, n_C, n):

$$P\left(n_A,n_B,n_C,n\right)=\frac{n!}{n_A!n_B!n_C!}\,p_A^{n_A}p_B^{n_B}p_C^{n_C}$$

In the special case $C=0$, the above reduces to the non-normalized binomial distribution.

4.4 Gaussian Distribution

Recall that the probability of generating k results among n total trials in a two-state system is given by

$$P\left(k,n\right)=\frac{1}{2^n}\frac{n!}{k!\left(n-k\right)!}\,.$$

Introduce the shift

$$k\to k+\frac{n}{2}\,,$$

which modifies the above:

$$P\left(k,n\right)=\frac{1}{2^n}\frac{n!}{\left(\frac{n}{2}+k\right)!\left(\frac{n}{2}-k\right)!}$$

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

$$\ln(n!) \approx n\ln(n) - n + \ln\left(\sqrt{2\pi n}\right)$$
$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\,,$$

and the probability density reduces to

$$w(k) = e^{-2k^2/n}\sqrt{\frac{2}{\pi n}}\,. \tag{19.28}$$

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:

$$\langle k\rangle = \int_n k\cdot w(k)\ dk = a$$
$$\left\langle k^2\right\rangle = \int_n k^2\cdot w(k)\ dk = \frac{n}{4} + a^2$$
$$\sigma_k = \sqrt{\left\langle k^2\right\rangle - \langle k\rangle^2} = \sqrt{\frac{n}{4}}$$

4.5 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 $\Delta t \to 0$, and observe that each $P_k(\Delta t)$ relates to its $k-1$ and $k+1$ neighbors as:

$$\lim_{\Delta t\to 0} P_0(\Delta t) \gg P_1(\Delta t) \gg P_2(\Delta t) \gg P_3(\Delta t) \gg \cdots$$

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

$$\frac{n!}{(n-k)!} \approx n^k$$
$$\left(1-\frac{vt}{n}\right)^{n-k} \approx e^{-vt}$$

are valid, and re-casting vt as a dimensionless variable q lands us at the anticipated *Poisson distribution*:

$$P_k(q) = \frac{q^k}{k!}e^{-q} \tag{19.29}$$

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 $\langle k\rangle$, $\left\langle k^2\right\rangle$, and the standard deviation:

$$\langle k\rangle = \sum_{k=0}^{\infty} k\frac{q^k}{k!}e^{-q} = e^{-q}\sum_{k=1}^{\infty}\frac{q^k}{(k-1)!}$$
$$= e^{-q}\sum_{p=0}^{\infty}\frac{q^{(p+1)}}{p!} = e^{-q}qe^{q} = q$$
$$\left\langle k^2\right\rangle = \sum_{k=0}^{\infty} k^2\frac{q^k}{k!}e^{-q} = e^{-q}q\sum_{p=0}^{\infty}(p+1)\frac{q^p}{p!}$$
$$= q + q^2$$
$$\sigma_k = \sqrt{q^2+q-q^2} = \sqrt{q}$$

Chapter 20

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 $\vec{a}$ is defined as a list of numbers or variables. We begin by replacing the 'arrow' notation with *ket* notation popularized by Paul Dirac:

$$\vec{a} = |a\rangle$$

To write out the vector explicitly, list the components inside the $|\rangle$ symbol:

$$\vec{a} = \langle a_1, a_2, a_3 \rangle = |a_1, a_2, a_3\rangle$$

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

$$\phi = \arctan\left(\beta/\alpha\right)$$

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

$$\begin{aligned} z_1 \pm z_2 &= (\alpha_1 + \alpha_2) \pm i\,(\beta_1 + \beta_2) \\ z_1 \cdot z_2 &= (\alpha_1\alpha_2 - \beta_1\beta_2) + i\,(\alpha_1\beta_2 + \alpha_2\beta_1) \\ z_1/z_2 &= z_1 \cdot z_2^* / \left|z_2\right|^2 \end{aligned}$$

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 exits 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

$$\int_a^b \left|f\,(x)\right|^2 \, dx < \infty \,.$$

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:

$$|a\rangle + |b\rangle = |b\rangle + |a\rangle$$
$$|a\rangle + (|b\rangle + |c\rangle) = (|a\rangle + |b\rangle) + |c\rangle$$

Scalar Multiplication

Multiplying a vector by a complex number results in a new vector that is parallel to the original. In particular, this means:

$$\forall\, |a\rangle \in \mathcal{V}\,,\ \forall \alpha \in \mathcal{C}\ :\ \exists\, \alpha\, |a\rangle \in \mathcal{V}$$

As expected, scalar multiplication follows the rules of commutativity and associativity:

$$\alpha\,(\beta\,|a\rangle) = (\alpha\beta)\,|a\rangle$$
$$\alpha\,(|a\rangle + |b\rangle) = \alpha\,|a\rangle + \alpha\,|b\rangle$$
$$(\alpha + \beta)\,|a\rangle = \alpha\,|a\rangle + \beta\,|a\rangle$$

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:

$$\forall\, |a\rangle \in \mathcal{V}\ :\ \exists\, |-a\rangle \in \mathcal{V}\,,$$
$$|a\rangle + |-a\rangle = |0\rangle$$

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:

$$|a\rangle + |0\rangle = |a\rangle$$
$$|b\rangle + |0\rangle' = |b\rangle$$

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:

$$|a\rangle + |-a\rangle = |0\rangle$$
$$|a\rangle + |-a\rangle' = |0\rangle$$

Adding the two equations, we have

$$2\,|a\rangle + |-a\rangle + |-a\rangle' = |0\rangle\ ,$$

which is only true if

$$|-a\rangle + |-a\rangle' = 2\,|-a\rangle$$
$$|-a\rangle' = |-a\rangle\ ,$$

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 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 $\alpha \to -\alpha$:

$$-\alpha\,|a\rangle = |0\rangle$$

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

$$a = (a_1, a_2, \cdots, a_n)$$

obeying the addition rule

$$a + b = (a_1 + b_1, a_2 + b_2, \cdots, a_n + b_n)$$

and the multiplication rule

$$\lambda a = (\lambda a_1, \lambda a_2, \cdots, \lambda a_n)\ .$$

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 $\vec{a} = |a\rangle$, 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 $\alpha = \alpha^* = \overline{\alpha}$, 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 excepting being the zero vector:

$$\langle a|a\rangle = 0 \iff |a\rangle = |0\rangle$$

3.5 Applications

Complex Vector Space

Consider the vector space $\mathcal{C}_n$ whose elements are vectors containing n individual complex numbers, i.e.

$$\exists\, |x\rangle \in \mathcal{C}_n \,:\, |x\rangle = |x_1, x_2, ..., x_n\rangle$$

For two vectors $|a\rangle$ and $|b\rangle$ in $\mathcal{C}_n$, 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 $j = 1, 2, \ldots, n$. 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 subspace $\mathcal{V}' \in \mathcal{V}$. 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 kth 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

$$|a\rangle = |a_1, a_2, \ldots, , a_n\rangle \ .$$

On the right side, we see that the (complex) components form an n-dimensional space of their own, namely $\mathcal{C}_n$.

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 $\mathcal{C}_n$, or

$$\mathcal{V}_{(n)} \cong \mathcal{C}_n \, .$$

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

$$\begin{aligned} |e'_1\rangle &= |\phi_1\rangle \\ |e_1\rangle &= |e'_1\rangle / \sqrt{\langle e'_1|e'_1\rangle} \, . \end{aligned}$$

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

$$\begin{aligned} |e'_n\rangle = |\phi_n\rangle &- \langle e_1|\phi_n\rangle \, |e_1\rangle \\ &- \langle e_2|\phi_n\rangle \, |e_2\rangle - \cdots - \langle e_{n-1}|\phi_n\rangle \, |e_{n-1}\rangle \ , \end{aligned}$$

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:

$$|e'_2\rangle = |\phi_2\rangle - \langle e_1|\phi_2\rangle \, |e_1\rangle = \begin{bmatrix} -1/2 \\ 1/2 \\ 1 \end{bmatrix}$$

$$|e_2\rangle = \frac{1}{\sqrt{6}} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}$$

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

$$P_4(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 \, ,$$

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

$$\begin{aligned} |e'_1\rangle &= |\phi_1\rangle - \cancel{\langle e_0|\phi_1\rangle \, |e_0\rangle} \\ |e_1\rangle &= \sqrt{\frac{3}{2}} \, |\phi_1\rangle = \sqrt{\frac{3}{2}} \, x \, . \end{aligned}$$

Continuing for $|e_2\rangle$, begin with

$$|e_2'\rangle = |\phi_2\rangle - \langle e_0|\phi_2\rangle\,|e_0\rangle - \cancel{\langle e_1|\phi_2\rangle\,|e_1\rangle} = |\phi_2\rangle - \frac{\sqrt{2}}{3}\,|e_0\rangle\ ,$$

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

$$|e_2\rangle = \sqrt{\frac{5}{8}}\left(3x^2-1\right)\ .$$

Turning the same crank, it's straightforwardly shown that the remaining normalized basis vectors resolve to

$$|e_3\rangle = \sqrt{\frac{7}{8}}\left(5x^3-3x\right)$$
$$|e_4\rangle = \frac{3}{8\sqrt{2}}\left(35x^4-30x^2+3\right)\ .$$

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 three calculations to perform:

$$h_0 = \frac{1}{\sqrt{2}}\int_{-1}^{1} x^4\,dx = \frac{\sqrt{2}}{5}$$
$$h_2 = \sqrt{\frac{5}{8}}\int_{-1}^{1}\left(3x^6-x^4\right)dx = \sqrt{\frac{8}{5}}\frac{2}{7}$$
$$h_4 = \frac{3}{8\sqrt{2}}\int_{-1}^{1}\left(35x^8-30x^6+3x^4\right)dx = \frac{1}{35}\frac{8\sqrt{2}}{3}$$

As a reality check, we can readily verify that $|h\rangle$ still corresponds to x^4, as all other x^n terms cancel out:

$$|h\rangle = \sum_{j=0}^{4} h_j\,|e_j\rangle = h_0\,|e_0\rangle + h_2\,|e_2\rangle + h_4\,|e_04\rangle = |\phi_4\rangle$$

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

$$\|a+b\|_m = \max\left(|x_1+x_2|\,,|y_1+y_2|\right)\ .$$

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

$$\begin{aligned}\|a+b\|_m &= \max\left(|x_1+x_2|\,,|y_1+y_2|\right)\\ &\le \max\left(|x_1|+|x_2|\,,|y_1|+|y_2|\right)\\ &\le \max\left(\|a\|_m+\|b\|_m,\|a\|_m+\|b\|_m\right)\\ &\le \|a\|_m+\|b\|_m\,,\end{aligned}$$

satisfying a requirement of a norm.

To complete the job we also establish a linearity relation:

$$\begin{aligned}\|\alpha a\|_m &= \max\left(|\alpha x_1|\,,|\alpha y_1|\right)\\ &= |\alpha|\max\left(|x_1|\,,|y_1|\right)\\ &= |\alpha|\,\|a\|_m\end{aligned}$$

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

$$\begin{aligned}|a\rangle &= (x_1,y_1)\\ |b\rangle &= (x_2,y_2)\ ,\end{aligned}$$

the 'sum' function gives, using the same identities as above,

$$\begin{aligned}\|a+b\|_s &= |a|+|b|\\ &= |x_1+y_1|+|x_2+y_2|\\ &\le |x_1|+|y_1|+|x_2|+|y_2|\\ &\le \|a\|_m+\|b\|_m\,.\end{aligned}$$

To check for linearity, we write

$$\begin{aligned}\|\alpha a\|_s &= |\alpha x_1|+|\alpha y_1|\\ &= |\alpha|\left(|x_1|+|y_1|\right)\\ &= |\alpha|\,\|a\|_s\,.\end{aligned}$$

Unit Ball

In any vector space $\mathcal{V}$, the unit 'ball' $\mathcal{B}_1$ is defined as

$$\mathcal{B} = \{|a\rangle\in\mathcal{V}\ :\ \|a\|\le 1\}\ .$$

Plotting the 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 $(0,\pm1)$ and $(\pm1,0)$, generated by $|x|+|y|\le 1$.

6.3 Identities

Cauchy-Bunyakovsky-Schwarz Inequality

There is a 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

$$\begin{aligned}\|a-\lambda b\|^2 = \|a\|^2 &- \frac{\langle b|a\rangle}{\|b\|^2}\langle a|b\rangle\\ &- \frac{\langle a|b\rangle}{\|b\|^2}\langle b|a\rangle+\frac{\langle a|b\rangle\langle b|a\rangle}{\|b\|^2\cancel{\|b\|^2}}\cancel{|b\|^2}\,.\end{aligned}$$

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:

$$|\langle a|b\rangle|\le\|a\|\|b\|$$

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,

$$\begin{aligned}\|a+\lambda b\|^2 &= \langle a-\lambda b|a+\lambda b\rangle\\ &= \|a\|^2+\lambda\langle a|b\rangle+\lambda^*\langle b|a\rangle+\lambda^2|b\|^2\\ &= \|a\|^2+2\mathrm{Re}\left(\lambda\langle a|b\rangle\right)+\lambda^2|b\|^2\,,\end{aligned}$$

and let $\lambda=1$:

$$\|a+b\|^2=\|a\|^2+2\mathrm{Re}\left(\langle a|b\rangle\right)+|b\|^2$$

By the Cauchy-Bunyakovsky-Schwarz inequality, the middle term is less than $\|a\|\|b\|$:

$$\begin{aligned}\|a+b\|^2 &\le \|a\|^2+2\|a\|\|b\|+|b\|^2\\ &\le \left(\|a\|+|b\|\right)^2\end{aligned}$$

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:

$$\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$$

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:

$$\|a+b\|^2 - \|a-b\|^2 = 4\,\mathrm{Re}\left(\langle a|b\rangle\right)$$

One may imagine what combination on the left yields $\mathrm{Im}\left(\langle a|b\rangle\right)$ on the right. It's straightforward to show that the job is done by:

$$i\|a+ib\|^2 - i\|a-ib\|^2 = 4\,\mathrm{Im}\left(\langle a|b\rangle\right)$$

Polar Identity

The sum of $4\,\mathrm{Re}\left(\langle a|b\rangle\right)$ and $4\,\mathrm{Im}\left(\langle a|b\rangle\right)$, each given above, leads to the *polar identity*. First note that

$$4\,\mathrm{Re}\left(\langle a|b\rangle\right) + 4\,\mathrm{Im}\left(\langle a|b\rangle\right) = 4\left(\langle a|b\rangle\right) \ ,$$

which yields the identity

$$\begin{aligned}4\left(\langle a|b\rangle\right) = \|a+b\|^2 - \|a-b\|^2 \\ + i\|a+ib\|^2 - i\|a-ib\|^2\end{aligned}$$

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

$$\sin(x) \le \sqrt{\sin(x)} \le \sqrt{x}$$

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:

$$|a\rangle = \lim_{k\to\infty} |a_k\rangle$$

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

$$\begin{aligned}\|a - a_m\| &< \alpha\epsilon \\ \|a - a_n\| &< \beta\epsilon\end{aligned}$$

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

$$\|a\| = \sqrt{|a|_1^2 + |a|_2^2 + \cdots + |a|_n^2} \, .$$

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 $\mathcal{C}_{ab}$ of complex-valued functions $f\,(x) \ : \ x \in [a,b]$. Here we define the 'supremum' function

$$\|f\|_{sup} = \max\left\{|f\,(x)| \ : \ x \in [a,b]\right\} \ .$$

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

$$\begin{aligned} \|f\|_{sup} &= 1 \\ \|g\|_{sup} &= \pi \\ \|f+g\|_{sup} &= 1 + \pi \\ \|f-g\|_{sup} &= -1 - \pi \, , \end{aligned}$$

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 nth 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 $m \to \infty$ 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 form 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 ($n \to \infty$), 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 $\mathcal{C}_2[a,b]$ 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 $\mathcal{C}_2[a,b]$ 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 $j = 1, 2, \ldots$ 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:

$$\langle b|L|a\rangle = \langle L^\dagger b|a\rangle$$

$$\left(L^\dagger\right)^\dagger = L$$

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:

$$\langle u|AB|v\rangle = \langle (AB)^\dagger u|v\rangle$$

$$\langle u|AB|v\rangle = \langle A^\dagger u|Bv\rangle = \langle B^\dagger A^\dagger u|v\rangle$$

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

$$AB = \left(B^\dagger A^\dagger\right)^\dagger = (BA)^\dagger \ ,$$

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

$$\langle f|Ag\rangle = \langle A^\dagger f|g\rangle$$

$$\int_\Omega f^* \partial_x g \, dx = \int_\Omega \left(\partial_x\right)^\dagger f^* g \, dx \, .$$

Integrating the left side by parts, we write

$$f^* g\big|_{\partial\Omega} - \int_\Omega \partial_x f^* g \, dx = \int_\Omega \left(\partial_x\right)^\dagger f^* g \, dx \, ,$$

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 lift 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 these established, the next move is to calculate the eigenvectors directly using

$$\begin{aligned} L\,|x^{(j)}\rangle &= \lambda_j\,|x^{(j)}\rangle \\ j &= 1, 2, \ldots, n \end{aligned}$$

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 $j = 1, \ldots, n$. 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:

$$\begin{aligned} &\lambda = 0 \\ &x_j\,(t) = x_{0j} + x_{1j}\,t \end{aligned}$$

For $\lambda \neq 0$, we have a linear combination of exponential terms:

$$\begin{aligned} &\lambda \neq 0 \\ &x_j\,(t) = x_{0j}\,e^{\sqrt{\lambda}t} + x_{1j}\,e^{-\sqrt{\lambda}t} \end{aligned}$$

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:

$$\det\left(A^\dagger\right) = \det\left(\left(A^*\right)^T\right) = \left(\det\left(A\right)\right)^*$$

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:

$$\mathrm{tr}A = A_{11} + A_{22} + \cdots + A_{nn} = \sum_{k=1}^{n} A_{kk}$$

12 Hermitian Matrix

Recall that an operator A that is it's 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

$$\overline{\langle x|A^\dagger y\rangle} = \sum_{ij} \left(A^\dagger_{ji}\right)^* x_j y_i^* \,,$$

telling us that the components of a Hermitian operator obey

$$A^*_{ij} = A^\dagger_{ji} \,.$$

Anti-Hermitian Matrix

The properties of anti-Hermitian operators also extend to matrices. An anti-Hermitian matrix satisfies

$$\begin{aligned} A^\dagger &= -A \\ A^*_{ij} &= -A_{ij} \,. \end{aligned}$$

As a corollary, we note that if a matrix A is Hermitian, then iA is anti-Hermitian, and vise-versa.

Symmetric Matrix

When the components of a Hermitian matrix are all real-valued, the matrix is symmetric. Work this from the identity

$$A^*_{ij} = A^\dagger_{ji} \,,$$

and notice that the complex conjugate A^* is just A again:

$$A_{ij} = A^\dagger_{ji}$$

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:

$$\begin{aligned} A\,|\psi_n\rangle &= a_n\,|\psi_n\rangle \\ B\,|\phi_n\rangle &= b_n\,|\phi_n\rangle \end{aligned}$$

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:

$$\begin{aligned} |\psi_n\rangle &= \sum_m \gamma_{mn}\,|\phi_m\rangle \\ |\phi_n\rangle &= \sum_m \tilde\gamma_{mn}\,|\psi_m\rangle \;. \end{aligned}$$

In the above, γ, $\tilde\gamma$ are matrix coefficients. We gain a restriction on γ, $\tilde\gamma$ 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 ride 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 for $B\,|\phi_n\rangle$ yields

$$\frac{1}{b_n}\sum_m \gamma_{mn}\tilde\gamma_{m'm}b_{m'} = \delta_{m'n} \,.$$

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 $B\left|\psi_n\right\rangle$. Begin by calculating $B\left|\psi_n\right\rangle$ to get

$$\begin{aligned} B\left|\psi_n\right\rangle &= B\sum_m \gamma_{mn}\left|\phi_m\right\rangle \\ &= \sum_m \gamma_{mn} b_m \left|\phi_m\right\rangle \\ &= \sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} \left|\psi_{m'}\right\rangle \ . \end{aligned}$$

The rest goes as

$$\begin{aligned} AB\left|\psi_n\right\rangle &= A\sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} \left|\psi_{m'}\right\rangle \\ &= \sum_m \sum_{m'} \gamma_{mn} b_m \tilde{\gamma}_{m'm} a_{m'} \left|\psi_{m'}\right\rangle \\ &= \sum_{m'} \delta_{nm'} b_n a_{m'} \left|\psi_{m'}\right\rangle \\ &= a_n b_n \left|\psi_n\right\rangle \ . \end{aligned}$$

To finish, let us write the commutator of A and B to conclude

$$\begin{aligned} [AB]\left|\psi_n\right\rangle &= (AB - BA)\left|\psi_n\right\rangle \\ &= (a_n b_n - a_n b_n)\left|\psi_n\right\rangle = 0 \, , \end{aligned}$$

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

$$\begin{aligned} \left|e_1\right\rangle &= \frac{1}{\sqrt{2}} \\ \left|e_2\right\rangle &= \frac{1}{\sqrt{2}}\left(\sin\left(\frac{\pi}{2}x\right) + \cos\left(\pi x\right)\right) \\ \left|e_3\right\rangle &= \frac{1}{\sqrt{2}}\left(\sin\left(\frac{\pi}{2}x\right) - \cos\left(\pi x\right)\right) \ , \end{aligned}$$

the corresponding matrix B is calculated from

$$B = \begin{bmatrix} \left\langle e_1\right| B\left|e_1\right\rangle & \left\langle e_1\right| B\left|e_2\right\rangle & \left\langle e_1\right| B\left|e_3\right\rangle \\ \left\langle e_2\right| B\left|e_1\right\rangle & \left\langle e_2\right| B\left|e_2\right\rangle & \left\langle e_2\right| B\left|e_3\right\rangle \\ \left\langle e_3\right| B\left|e_1\right\rangle & \left\langle e_3\right| B\left|e_2\right\rangle & \left\langle e_3\right| B\left|e_3\right\rangle \end{bmatrix} \ ,$$

where

$$\left\langle e_j\right| B\left|e_k\right\rangle = \int_{-1}^{1} e_j^*(x)\,\partial_{xx} e_k(x)\,dx \, .$$

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

$$\begin{aligned} \left|e_1\right\rangle &= \frac{1}{\sqrt{2}} \\ \left|e_2\right\rangle &= \frac{1}{\sqrt{2}}\sin\left(\pi x\right) \\ \left|e_3\right\rangle &= \frac{1}{\sqrt{2}}\cos\left(\pi x\right) \ , \end{aligned}$$

we first check that the subset is closed under operations A:

$$\begin{aligned} A\left|e_1\right\rangle &= \partial_x\left(\frac{1}{\sqrt{2}}\right) = 0 \\ A\left|e_2\right\rangle &= \partial_x\left(\frac{1}{\sqrt{2}}\sin\left(\pi x\right)\right) \\ &= \frac{\pi}{\sqrt{2}}\cos\left(\pi x\right) = \pi\left|e_3\right\rangle \\ A\left|e_3\right\rangle &= \partial_x\left(\frac{1}{\sqrt{2}}\cos\left(\pi x\right)\right) \\ &= -\frac{\pi}{\sqrt{2}}\sin\left(\pi x\right) = -\pi\left|e_2\right\rangle \end{aligned}$$

Each nontrivial result can be written in terms of the original basis vectors. Therefore $A\left|x\right\rangle$, where $\left|x\right\rangle$ 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 $A_{jk} = \left\langle e_j\right| A\left|e_k\right\rangle$, 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

$$\sigma_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$
$$\sigma_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$$
$$\sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

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

$$\sigma_k^{2m} = I$$
$$\sigma_k^{2m+1} = \sigma_k$$

for integer m. Furthermore, we have

$$\sigma_2 \sigma_3 = i\sigma_1$$
$$\sigma_3 \sigma_1 = i\sigma_2$$
$$\sigma_1 \sigma_2 = i\sigma_3 \, .$$

Eigenvectors and Eigenvalues

Eigenvectors and eigenvalues of the Pauli matrices are determined by

$$\sigma_k \left|x\right\rangle = \lambda_k \left|x\right\rangle \, .$$

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

$$x_2 = \lambda_1 x_1$$
$$x_1 = \lambda_1 x_2 \, ,$$

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:

$$\lambda_1 = +1 \;\rightarrow\; \left|x^{(+)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$
$$\lambda_1 = -1 \;\rightarrow\; \left|x^{(-)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Repeating the exercise on $\sigma_2 \left|y\right\rangle = \lambda_2 \left|y\right\rangle$, and a third time on $\sigma_3 \left|z\right\rangle = \lambda_3 \left|z\right\rangle$, we find the eigenvalues are always $\lambda_k = \pm 1$. The corresponding normalized eigenvectors turn out to be

$$\left|y^{(+)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} i \\ 1 \end{bmatrix}$$
$$\left|y^{(-)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}$$
$$\left|z^{(+)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$\left|z^{(-)}\right\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \, .$$

14 Functions of Operators

14.1 Motivation

Recall that the single time derivative operator $L = \partial_t$ as it appears in the problem $L\left|x\right\rangle = \lambda \left|x\right\rangle$ has exponentially-evolving eigenvectors $\left|x\,(t)\right\rangle = e^{\lambda t}\left|x_0\right\rangle$, where $\left|x_0\right\rangle$ is the initial state in a fixed orthonormal basis. The more general statement

$$\left|x\,(t)\right\rangle = e^{Lt}\left|x_0\right\rangle$$

leads to the same solution, easily shown by taking a time derivative:

$$\partial_t \left|x\,(t)\right\rangle = L\, e^{Lt}\left|x_0\right\rangle = L\left|x\,(t)\right\rangle$$

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

$$e^{Lt} = \sum_{k=0}^{\infty} \frac{1}{k!} (Lt)^k$$
$$= I + (Lt) + \frac{1}{2!}(Lt)^2 + \frac{1}{3!}(Lt)^3 + \cdots \, .$$

Grouping even terms and odd terms separately, the above reads

$$e^{Lt} = \left(I + \frac{1}{2!}(Lt)^2 + \frac{1}{4!}(Lt)^4 + \cdots\right) + \left(Lt + \frac{1}{3!}(Lt)^3 + \frac{1}{5!}(Lt)^5 + \cdots\right) .$$

Making the substitution $L = -i\tilde{H}$, we further have

$$e^{-i\tilde{H}t} = \left(I - \frac{1}{2!}\left(\tilde{H}t\right)^2 + \frac{1}{4!}\left(\tilde{H}t\right)^4 - \cdots\right) - i\left(\tilde{H}t - \frac{1}{3!}\left(\tilde{H}t\right)^3 + \frac{1}{5!}\left(\tilde{H}t\right)^5 - \cdots\right) ,$$

simplifying nicely to

$$e^{-i\tilde{H}t} = \cos\left(\tilde{H}t\right) - i\sin\left(\tilde{H}t\right) ,$$

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:

$$e^{-i\sigma_1\mu t} = \begin{bmatrix} \cos(\mu t) & -i\sin(\mu t) \\ -i\sin(\mu t) & \cos(\mu t) \end{bmatrix}$$

$$e^{-i\sigma_2\mu t} = \begin{bmatrix} \cos(\mu t) & -\sin(\mu t) \\ -\sin(\mu t) & \cos(\mu t) \end{bmatrix}$$

$$e^{-i\sigma_3\mu t} = \begin{bmatrix} e^{-i\mu t} & 0 \\ 0 & e^{i\mu t} \end{bmatrix}$$

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 $\tilde{\mathcal{V}}$, 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

$$U_{jk} = \langle\tilde{e}_j|e_k\rangle$$
$$\tilde{U}_{jk} = \langle e_j|\tilde{e}_k\rangle ,$$

which can be interpreted as a set of coordinate transformation matrices that carry $\{x_j\} \in \tilde{\mathcal{V}} \to \{\tilde{x}_j\} \in \tilde{\mathcal{V}}$, and vice versa.

These matrices are intricately related, which we first notice by the observation

$$\tilde{U}_{jk} = \langle e_j|\tilde{e}_k\rangle = \overline{\langle\tilde{e}_k|e_j\rangle} = U^*_{kj} = U^\dagger_{jk} ,$$

telling us each matrix is the other's Hermitian conjugate (dropping component notation):

$$\tilde{U} = U^\dagger$$
$$U = \tilde{U}^\dagger$$

Now, the double transformation $\{x_j\} \to \{\tilde{x}_j\} \to \{x_j\}$ (and vice versa) tells us that each combination $\tilde{U}U$ and $U\tilde{U}$ is an identity matrix:

$$\tilde{U}U = I$$
$$U\tilde{U} = I$$

Combining the two previous results yields four identities:

$$U^\dagger U = I$$
$$UU^\dagger = I$$
$$\tilde{U}\tilde{U}^\dagger = I$$
$$\tilde{U}^\dagger\tilde{U} = I$$

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

$$\langle Ux|Uy\rangle = \langle U^\dagger Ux|y\rangle = \langle Ix|y\rangle = \langle x|y\rangle ,$$

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 $\{x_j\} \in \tilde{\mathcal{V}}$.

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

$$\begin{aligned} |g_j\rangle &= U\,|e_j\rangle \\ j &= 1, 2, 3, \ldots, N \ . \end{aligned}$$

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 jth vector from the primed basis and churns out the jth vector from the unprimed basis. Or, the $\tilde{U}$ matrix takes the jth 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:

$$\begin{aligned} U_{jk} &= \langle \tilde{e}_j|e_k\rangle = \langle \tilde{e}_j|\, U\,|\tilde{e}_k\rangle \\ \tilde{U}_{jk} &= \langle e_j|\tilde{e}_k\rangle = \langle e_j|\, \tilde{U}\,|e_k\rangle \end{aligned}$$

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:

$$\begin{aligned} |\tilde{e}_1\rangle &= \cos(\phi)\,|e_1\rangle + \sin(\phi)\,|e_2\rangle \\ |\tilde{e}_2\rangle &= -\sin(\phi)\,|e_1\rangle + \cos(\phi)\,|e_2\rangle \end{aligned}$$

Using the above formula for U_{jk} applied to standard two-dimensional geometry, we quickly find the components of U to be

$$\begin{aligned} U_{xx} &= \langle \tilde{e}_1|e_1\rangle = \cos\phi \\ U_{xy} &= \langle \tilde{e}_1|e_2\rangle = -\sin\phi \\ U_{yx} &= \langle \tilde{e}_2|e_1\rangle = \sin\phi \\ U_{yy} &= \langle \tilde{e}_2|e_2\rangle = \cos\phi \ , \end{aligned}$$

and similarly for $U^\dagger = \tilde{U}$:

$$\begin{aligned} \tilde{U}_{xx} &= \langle e_1|\tilde{e}_1\rangle = \cos\phi \\ \tilde{U}_{xy} &= \langle e_1|\tilde{e}_2\rangle = \sin\phi \\ \tilde{U}_{yx} &= \langle e_2|\tilde{e}_1\rangle = -\sin\phi \\ \tilde{U}_{yy} &= \langle e_2|\tilde{e}_2\rangle = \cos\phi \end{aligned}$$

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:

$$0 \le \alpha < 2\pi$$
$$0 \le \beta < \pi$$
$$0 \le \gamma < 2\pi$$

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

$$\sum_j A_{ij} x_j = y_i$$
$$\sum_j \tilde{A}_{ij} \tilde{x}_j = \tilde{y}_i .$$

Substituting

$$x_j = \sum_k \tilde{x}_k \langle e_j | \tilde{e}_k \rangle$$
$$y_i = \sum_k \tilde{y}_k \langle e_i | \tilde{e}_k \rangle$$

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 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 orthornormal 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

$$C_j\,(t) = C_j\,(t=0)\,e^{-iE_jt/\hbar}\,.$$

Note that the initial value of each $C_n\,(0)$ is calculated from the initial condition $|\psi\,(0)\rangle$ according to

$$\langle\phi^{(k)}|\psi\,(0)\rangle = \sum_j C_j\,(t)\,\langle\phi^{(k)}|\phi^{(j)}\rangle = C_k\ ,$$

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

$$|\psi\,(t)\rangle = \sum_j C_j\,(0)\,e^{-iE_jt/\hbar}\,|\phi^{(j)}\rangle\ .$$

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

$$\begin{aligned}|x^{(+)}\rangle &= \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \\ |x^{(-)}\rangle &= \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}\ .\end{aligned}$$

With the eigenvectors for the operator on hand, the vector $|u\,(t)\rangle$ can be expressed as a linear combination

$$\begin{aligned}|u\,(t)\rangle &= C_+\,(t)\,|x^{(+)}\rangle + C_-\,(t)\,|x^{(-)}\rangle \\ &= \sum_j C_j\,(t)\,|x^{(j)}\rangle\ .\end{aligned}$$

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

$$C_j\,(t) = C_j\,(t=0)\,e^{-i\lambda_jt}\,.$$

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 orthonromal 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

$$\begin{aligned}|u\,(t)\rangle = &\frac{1}{\sqrt{2}}\cos\,(\delta t)\begin{bmatrix}C_+ + C_-\\C_+ - C_-\end{bmatrix} \\ &+ \frac{i}{\sqrt{2}}\sin\,(\delta t)\begin{bmatrix}-C_+ + C_-\\-C_+ - C_-\end{bmatrix}\ .\end{aligned}$$

16.3 Damped Harmonic Oscillator

The differential equation that governs the damped harmonic oscillator reads

$$\frac{d^2}{dt^2}x\left(t\right)-b\frac{d}{dt}x\left(t\right)+\omega_0^2x\left(t\right)=0\,,$$

which is a second-order equation. However, this problem can be turned into a system of first-order equations by defining a vector

$$\left|u\left(t\right)\right\rangle=\begin{bmatrix}x\left(t\right)\\dx\left(t\right)/dt\end{bmatrix}$$

along with a matrix

$$A=\begin{bmatrix}0&1\\-\omega_0^2&-b\end{bmatrix}\,,$$

so the problem may be rewritten

$$\frac{d}{dt}\left|u\left(t\right)\right\rangle=A\left|u\left(t\right)\right\rangle\,.$$

Solving the eigenvalue problem

$$A\left|q^{(j)}\right\rangle=\lambda_j\left|q^{(j)}\right\rangle\,,$$

we quickly find

$$\lambda=-\frac{b}{2}\pm\sqrt{\frac{b^2}{4}-\omega_0^2}\,,$$

with corresponding eigenvectors

$$\left|q^{(+)}\right\rangle=\frac{1}{\sqrt{\lambda_+^2+1}}\begin{bmatrix}1\\\lambda_+\end{bmatrix}$$

$$\left|q^{(-)}\right\rangle=\frac{1}{\sqrt{\lambda_-^2+1}}\begin{bmatrix}1\\\lambda_-\end{bmatrix}\,.$$

Note that the eigenvectors $\left|q^{(\pm)}\right\rangle$ are linearly independent but not orthogonal. Regardless of this, the general solution can be written as a linear combination

$$\begin{aligned}\left|u\left(t\right)\right\rangle&=C_+\left(t\right)\left|q^{(+)}\right\rangle+C_-\left(t\right)\left|q^{(-)}\right\rangle\\&=\sum_j C_j\left(t\right)\left|q^{(j)}\right\rangle\,.\end{aligned}$$

Applying the A-operator to both sides of the above produces

$$\begin{aligned}\sum_j\frac{d}{dt}\left(C_j\left(t\right)\right)\left|q^{(j)}\right\rangle&=A\sum_j C_j\left(t\right)\left|q^{(j)}\right\rangle\\&=\sum_j C_j\left(t\right)\lambda_j\left|q^{(j)}\right\rangle\,,\end{aligned}$$

immediately implying

$$\frac{d}{dt}C_j\left(t\right)=C_j\left(t\right)\lambda_j\,,$$

solved by

$$C_j\left(t\right)=C_j\left(t=0\right)e^{\lambda_j t}\,.$$

The general solution now reads

$$\left|u\left(t\right)\right\rangle=\sum_j C_j\left(0\right)e^{\lambda_j t}\left|q^{(j)}\right\rangle\,.$$

Setting $t=0$ in the above gives

$$\left|u\left(0\right)\right\rangle=\sum_j C_j\left(0\right)\left|q^{(j)}\right\rangle\,.$$

Note however that the basis vectors $\left|q^{(k)}\right\rangle$ are not orthogonal, thus the coefficients $C_j\left(0\right)$ cannot be isolated by taking the inner product with $\left\langle q^{(k)}\right|$. To proceed, write the above in component form to get

$$u_k\left(0\right)=\sum_j C_j\left(0\right)q_k^{(j)}\,.$$

This is a linear system of the form $A\left|x\right\rangle=\left|b\right\rangle$ with $\left|u\left(0\right)\right\rangle$ playing the role of $\left|b\right\rangle$, the vector components $q_k^{(j)}$ serving as matrix components, and the components $C_j\left(0\right)$ corresponding to $\left|x\right\rangle$.

Index