Notes on Basic Mathematics

Introduction to Computing Foundations

Foo Yong Qi

11 January 2024

Preface

This set of lecture notes was prepared for incoming graduate students enrolled in some of the various graduate programmes offered by the National University of Singapore (NUS) School of Computing (SoC), such as the Graduate Certificate in Computing Foundations, the Master of Computing (General Track) etc. These programmes generally do not formally require any mathematical maturity, though as one would very quickly discover, Computer Science is a fundamentally mathematical discipline. We thus attempt to bridge the gap of mathematical understanding through these notes in hopes that readers will find the pursuit of further study of Computer Science in NUS and beyond slightly less painful than it already is.

This text does not demand very much in mathematical and analytical thinking skills from the reader, and thus has to make total omission of a wide corpus of interesting results from many domains of pure and applied mathematics and computer science. In fact, in the interest of time and space we even omit many topics in basic mathematics that some students would have learnt in high school or undergraduate studies. Instead, we only cover some core mathematical topics that are immediately needed in an introductory study of computing, and defer the acquisition of the rest of mathematics fundamentals to further reading of other sources by the reader.

I teach four sections of the mini-course, covering topics in elementary mathematics such as Algebra (Chapter 1) (used in virtually all domains of computer science including cybersecurity, algorithm analysis, signal processing), Calculus (Chapter 2) (used extensively in numerical analysis, computer graphics, computer vision, machine learning, data mining, and bioinformatics) and Geometry (Chapter 3) (used many industry domains like media, games, optimizations). I will also introduce one of the most fundamental mathematical structures underpinning all of mathematics, known as sets (Chapter 4), and some definitions and properties of numbers.

On mathematical maturity, the general consensus among mathematicians (hobbyists and professionals alike) is that the best (arguably, the *only*) way to get better at mathematics is to *do* mathematics. Hence, at the end of each chapter are exercises which we encourage readers to attempt. Solutions to all exercises are also provided in Appendix A, we encourage readers to at least make some attempt on the exercises before refining their answers using the solutions manual.

Finally, we hope that readers will find these notes useful as they embark on their lifelong journey of computer science study.

Foo Yong Qi
11 January 2024

Contents

Chapter 1

Algebra

Algebra is the abstract key which unlocks the door to the mysteries of the universe.

Carl Friedrich Gauss

Algebra[1] may be regarded as among the most foundational and fundamental branches of mathematics. It deals with symbols and the rules for manipulating these symbols to solve problems and to understand the relationships between quantities. It is so fundamental that it is used by virtually all domains of science, and it would be almost sacrilegious for any introductory text to omit this topic. Thus, this chapter provides a recap on the applications of some fundamental results from algebra for problem solving. Section 1.1 describes how we can solve systems of equations and inequalities, Section 1.2 describes mathematical functions, specifically reasoning about a special class of functions known as polynomials, Section 1.3 recaps exponentials, logarithms and absolutes, and finally Section 1.4 defines some common notation for describing operations on series and how we can work with them.

1.1 EQUATIONS AND INEQUALITIES

We shall take the liberty of assuming that our very first mathematical *statement* or *proposition* in this text is legible to virtually all readers:

$$x = 5001$$

We explain the meaning of this statement out of an abundance of caution. The statement above is an *equation*, which states that the left-hand side (abbreviated *LHS*) of the equation (x) is *equal*[2] to the right-hand side (abbreviated *RHS*) (5001). In words, the statement declares that the *variable* x is *equal* to the number 5001.

[1] Mathematicians from the Islamic Golden Age, notably Al-Khwarizmi, made significant contributions to algebra. His book "Al-Kitab al-Mukhtasar fi Hisab al-Jabr wal-Muqabala" (The Compendious Book on Calculation by Completion and Balancing) provided systematic methods for solving linear and quadratic equations. Many words like 'algebra' and 'algorithm' were derived from Arabic names (*algebra* from *al-Jabr*, *algorithm* from *Al-Khwarizmi*).

[2] *Equality* is the strictest notion of *sameness* in mathematics. Setting aside some other ontological considerations of equality and *identity*, other domains of mathematics also present other notions of *sameness*; most prominently, the concept of *isomorphism* (*iso* for *same*, *morphism* for *shape*).

The following equation should likewise be simple to understand:

$$y = 500 + 1$$

That is, the variable y is equal to the quantity of $500 + 1$, i.e. 501.

Before moving on with some basic notation we must remind readers of different kinds of numbers that are frequently discussed in this text. We define the *natural numbers* as nonnegative whole numbers, i.e. 0, 1, 2 and so on. The *integers* are an extension of the natural numbers, i.e. they are *any* whole number, including negative numbers. *Rational numbers* are fractions where the *numerator* and *denominator* are integers, i.e. they take the form of $\frac{x}{y}$ for integers x and y (x is the numerator and y is the denominator). A *real number* is any number on the number line. We might on occasion, write a number in *standard form*, that is a real number $1 \leq n < 10$, multiplied by a power of 10. For example, the number 1234.56 can be written as 1.23456×10^4. The collection of all natural numbers, integers, rational numbers and real numbers will be denoted as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ respectively. Natural numbers are integers, integers are rational numbers, and rational numbers are real numbers. In other words, if x is a member of $\mathbb{N}$ then it is a member of $\mathbb{Z}$, and so on. We may occasionally choose to distinguish among the reals between the rationals and the non-rationals like $\sqrt{2}$ which we term *irrational*. Collections of mathematical objects like numbers are usually known as *sets* which we describe in more detail in Chapter 4; at least in this chapter, the terms *collection* and *set* will be used interchangeably, to simply mean *some collection*. $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are all sets. We may, on occasion, make use of the notation $x \in A$ to say that x is a member of the collection A—to write $y \in \mathbb{N}$ is to say that y is a natural number.

1.1.1 Basic Mathematical Operations

Aside from the most trivial mathematical notation such as those seen earlier, it is perfectly reasonable for readers to find mathematical syntax rather daunting. It is for this reason we shall make explicit the notations and terms used in this text, and start with only the most simple, before working up from there. For now, we concern ourselves with the following mathematical operations: $+$ (*addition*), $-$ (*subtraction*), $\times$ (*multiplication*), $\div$ (*division*) and a^b (*exponentiation*, i.e. taking a to the b^{th} power). We elide descriptions of these operations since they should be second nature to most. Division by 0 is undefined, and so is 0^0.[3] We may also write x/y as $x \div y$, sort of like 'writing a fraction horizontally'. We also want to be able to systematically and unambiguously evaluate more complicated mathematical expressions, such as when there are more than one occurrences of an operator (such as $a \times b \times c$). In this case, the first four operations *associate to the left*, or are *left-associative*. For example, $a - b - c = ((a - b) - c)$. Exponentiation is *right-associative*, for example, $a^{b^c} = (a^{(b^c)})$.

Food for thought 1.1. Is a left-associative notation for exponentiation useful? In other words, is it useful for $a^{b^c} = (a^b)^c$?

If a mathematical expression contains a mix of different operators like so:

$$1 + 2 \times 3 - 4 \div 5^2$$

we group sub-expressions by the *order of precedence* given by *PEMDAS*[4]: Parentheses, Exponentiation, Multiplication and Division, Addition and Subtraction, where parentheses have the

[3] We may later define $0^0 = 1$.
[4] Some schools teach BODMAS, which gives the same order.

highest precedence. Therefore:

$$\begin{aligned}1 + 2 \times 3 - 4 \div 5^2 &= 1 + (2 \times 3) - (4 \div (5^2)) \\ &= 1 + 6 - \frac{4}{25} \\ &= 6.84\end{aligned}$$

In fact, addition and multiplication of real numbers are *associative* and *commutative*, and multiplication is *distributive* over addition.

PROPOSITION 1.1 (Associativity of addition and multiplication). *Addition (+) and multiplication (×) are* ***associative****, i.e.* $a + (b + c) = (a + b) + c$ *and* $a \times (b \times c) = (a \times b) \times c$ *for all real numbers* a, b *and* c.

PROPOSITION 1.2 (Commutativity of addition and multiplication). *Addition (+) and multiplication (×) are* ***commutative****, i.e.* $a + b = b + a$ *and* $a \times b = b \times a$ *for all real numbers* a, b *and* c.

PROPOSITION 1.3 (Distributivity of multiplication over addition). *Multiplication is* ***distributive*** *over addition, i.e.*

$$a \times (b + c) = a \times b + a \times c$$

Finally, you might have frequently seen expressions of multiplication or products without the $\times$ symbol, e.g. $2(5+3) = 2\times(5+3)$. This is known as *implicit multiplication by juxtaposition*, which we shall define to have a higher precedence (with respect to PEMDAS) than multiplication and division, but less than exponentiation. Therefore, the mathematical expression $y \div 2(x + z)$ should be evaluated as $y \div (2 \times (x + z))$, and $3z^4$ to be $3 \times (z^4)$.

With these in mind, we will omit many parentheses and symbols unless only to be explicit.

1.1.2 Systems of Equations

Equations relating variables are incredibly powerful for solving problems of determining unknown quantities. Take, for example, the following problem statement.

Example 1.1. Alice has 1.5 times more oranges than Bob. Bob has 5 more oranges than Charlie. Finally, Alice has 15 more oranges than Charlie. How many oranges do Alice, Bob and Charlie each have?

We can express this statement as a *system of equations* that relate the quantities of oranges possessed by Alice, Bob and Charlie. For this, we first let a, b and c be the number of oranges that Alice, Bob and Charlie have respectively. Then, we express what we know from the problem as Equation 1.1, Equation 1.2 and Equation 1.3.

$$a = 1.5b \tag{1.1}$$

$$b = c + 5 \tag{1.2}$$

$$a = c + 15 \tag{1.3}$$

We can manipulate this system of equations via some rules:

1. If in one equation we have $X = Y$, if X (or Y) occurs in another equation, we can substitute that occurrence of X for Y (or Y for X)

2. We can do any operation to both sides of any equation (the operation on each side must be exactly the same).

For example, since we have by Equation 1.3 that $a = c + 15$ and a occurs in Equation 1.1, we can replace a with $c + 15$ in Equation 1.1, giving us:

$$c + 15 = 1.5b \tag{1.4}$$

Also, we can multiply both sides of Equation 1.4[5] by 2, giving us:

$$\begin{aligned}
& & c + 15 &= 1.5b & & (1.4)\\
&\Rightarrow & 2(c + 15) &= 2(1.5b) & &\triangleright \text{ multiply both sides by 2}\\
&\Rightarrow & 2c + 30 &= 3b & & (1.5)
\end{aligned}$$

Repeated application of these rules gives us the quantities a, b and c. We first express c using b by manipulating Equation 1.5:

$$\begin{aligned}
& & 2c + 30 &= 3b & & (1.5)\\
&\Rightarrow & (2c + 30) - 30) &= 3b - 30 & &\triangleright \text{ subtract 30 from both sides} \quad (1.6)\\
&\Rightarrow & 2c &= 3b - 30 & & (1.7)\\
&\Rightarrow & \frac{2c}{2} &= \frac{3b - 30}{2} & &\triangleright \text{ divide both sides by 2} \quad (1.8)\\
&\Rightarrow & c &= \frac{3b - 30}{2} & & (1.9)
\end{aligned}$$

Now that we have done so, we now express b without any other variables by substituting Equation 1.9 into Equation 1.2:

$$\begin{aligned}
& & b &= c + 5 & & (1.2)\\
& & c &= \frac{3b - 30}{2} & & (1.9)\\
&\Rightarrow & b &= \frac{3b - 30}{2} + 5 & &\triangleright \text{ by substitution of (1.9) onto (1.2)} \quad (1.10)\\
&\Rightarrow & 2b &= 3b - 30 + 10 & &\triangleright \text{ multiply both sides by 2} \quad (1.11)\\
&\Rightarrow & b &= 20 & &\triangleright \text{ add } 20 - 2b \text{ to both sides} \quad (1.12)
\end{aligned}$$

Finally, derivation of $a = 30$ and $c = 15$ should be trivial from Equation 1.12[6].

As we have seen, solving systems of equations typically involves isolating a variable on one side without having it occur in the other (such as in Equation 1.9), so that we can perform

[5]Note that $\Rightarrow$ stands for *implication*, i.e. $c + 15 = 1.5b \Rightarrow 2(c + 15) = 2(1.5b)$ tells us that if $c + 15 = 1.5b$, then $2(c + 15) = 2(1.5b)$. We may also use $\Leftrightarrow$ to mean implication in both directions, i.e. $a \Leftrightarrow b$ means that if a is true, then b is true, and likewise, if b is true then a is true. Most texts might say a is true *if and only if* b is true, which means the same thing.

[6]You might have noticed that we should obtain $20 = b$ instead of $b = 20$. It turns out that these two statements are both equally true since equality (=) is *symmetric*, i.e. if $20 = b$ then $b = 20$. Readers who are somehow unconvinced of this fact can derive $b = 20$ from $20 = b$ by subtracting $b + 20$ on both sides then multiply both sides by -1.

a substitution in another equation (such as in Equation 1.2) such that the variable no longer occurs in the equation entirely (such as in Equation 1.10). In addition, there is in fact more than one way to derive a, b and c using the same rules, and we leave it up to the reader to verify that these lead to the same values.

Food for thought 1.2. We shall attempt to prove that $1 = 2$.
Let $a = b$. Then[a]:

$$a = b \tag{1.13}$$

$$\Rightarrow \quad a^2 = ab \quad \triangleright \text{ multiply both sides by } a \tag{1.14}$$

$$\Rightarrow \quad a^2 - b^2 = ab - b^2 \quad \triangleright \text{ subtract } b^2 \text{ from both sides} \tag{1.15}$$

$$\Rightarrow \quad (a+b)(a-b) = b(a-b) \quad \triangleright \text{ factorize both sides} \tag{1.16}$$

$$\Rightarrow \quad a + b = b \quad \triangleright \text{ divide both sides by } (a-b) \tag{1.17}$$

$$\Rightarrow \quad b + b = b \quad \triangleright \text{ substitute } a \text{ using (1.13)} \tag{1.18}$$

$$\Rightarrow \quad 2b = b \tag{1.19}$$

$$\Rightarrow \quad 1 = 2 \quad \triangleright \text{ divide both sides by } b \tag{1.20}$$

We know very clearly that $1 \neq 2$. That must either mean that our proof is wrong, or that all of mathematics is fundamentally flawed. Which is more likely, and if our proof is wrong, which step(s) of the proof is incorrect?

[a] Recall from algebra that $a^2 - b^2 = (a+b)(a-b)$.

1.1.3 Inequalities

We hope that readers do not find the following statement unfamiliar:

$$1 \leq 2$$

This is an *inequality* stating that 1 is less than or equal to 2. Similarly recall that $x < y$ means x is *strictly* less than (and not equal to) y, $x \geq y$ means x is greater than or equal to y, and $x > y$ means that x is *strictly* greater than (and not equal to) y. We can combine two inequalities in a single statement, for example $x < y < z$ means $x < y$ and $y < z$. Alternatively for real numbers x, y and z we may rewrite $x < y < z$ as $y \in (x, z)$; (x, z) is known as an *open interval* containing all real numbers *strictly* between x and z, i.e. (x, z) contains all real numbers that are both strictly greater than x and strictly less than z. To write $x \leq y \leq z$ is to mean that y is greater than (or equal to) x but less than (or equal to) z; for this we can re-write this as $y \in [x, z]$ where $[x, z]$ is the *closed interval* bounded by x and z. In which case, the notation $y \in [x, z)$ or $y \in (x, z]$ should be intuitive; the first statement claims that $x \leq y < z$ and the second claims that $x < y \leq z$. Note that (a, b) can mean several things as you shall see later; we presume that the meaning of (a, b) can be understood based on context alone.

Just like equations, we can perform operations on both sides of an inequality and it will still hold, except that multiplication of both sides by -1 *reverses* the direction of inequality:

Proposition 1.4 (Multiplication by -1 on inequalities). *For all real numbers x and y, if $x < y$ then $-x > -y$, and if $x \leq y$ then $-x \geq -y$. Effectively, multiplication (or division) by -1 reverses the direction of inequality.*

Proof Sketch. First observe that for any real number k, if $x < y$ then we have $x + k < y + k$, i.e. addition (and therefore subtraction) preserves the direction of the inequality. Then, by subtracting $x + y$ from both sides of $x < y$ we get $x - x - y < y - x - y$, therefore $-y < -x$ and $-x > -y$. Argue similarly for $\leq$.

□

We show an incredibly simple application of this statement.

CLAIM 1.5. $x < 0 \Rightarrow 4x > 5x$.

Proof. First, observe from PROPOSITION 1.4 that $x < 0 \Rightarrow -x > 0$. Then,

$$\begin{array}{lcl} & 4 < 5 & \\ \Rightarrow & 4(-x) < 5(-x) & \triangleright \text{ multiplication by } -x \\ \Rightarrow & 4(-x)(-1) > 5(-x)(-1) & \triangleright \text{ multiplication by } -1 \\ \Rightarrow & 4x > 5x & \end{array}$$

□

Systems of inequalities typically come from formulations of optimization problems that can be solved by *linear programming*, which we will not cover in these notes. We give one example formulation here since problems like these are of interest to operations research (used widely in business and engineering optimization) and artificial intelligence. Readers may opt to read more about linear programming at their own time.

Example 1.2 (Optimization problem formulated as a system of linear inequalities). Toyco manufactures two kinds of soft toys: ducks and teddy bears; currently, two people work at toyco, Alice and Bob. Alice is willing to work up to 40 hours per week and is paid \$5 per hour, and Bob is willing to work up to 50 hours per week and is paid \$6 per hour. Ducks sell for \$25 each, and requires Alice to work 1 hour and Bob 2 hours to make; the raw materials for manufacturing ducks cost \$5 each. Teddy bears sell for \$22 each, takes Alice 2 hours and Bob 1 hour to make, and the raw materials for manufacturing teddy bears cost \$4. For Toyco to maximize its weekly profits, it must solve the following maximization problem:
Let d and t be the number of ducks and teddy bears produced in a week by Toyco. Notice that the profit earned from each duck is $\$25 - \$5 - \$2(6) - \$5 = \$3$ and that of each teddy bear is $\$22 - \$2(5) - \$6 - \$4 = \$2$:

$$\begin{array}{lcl} \max & \mathit{profit} = 3d + 2t & \triangleright \text{ total profits} \\ \text{s.t.} & d + 2t \leq 40 & \triangleright \text{ constraint on Alice's working hours} \\ & 2d + t \leq 50 & \triangleright \text{ constraint on Bob's working hours} \\ & d, t \geq 0 & \end{array}$$

1.2 FUNCTIONS

We now show more notation that we hope is familiar to readers.

$$f(x) = 2x + 3$$

Here, f is a *function*. Functions receive input(s) and produce some output, similar to a machine. The *arity* of a function is the number of *parameters* that it receives; f defined above has arity 1 (it only receives x) and is also called *unary*. If we define another function $g(x, y) = 3x - y$ then g has arity 2 and is also called *binary*.

1.2.1 Unary Functions

When discussing notation earlier, we have implicitly (in some cases, explicitly) stated the kinds of mathematical objects that can inhabit the stated variables. For example, we have stated in Proposition 1.1 that the associativity of addition and multiplication only applies to the real numbers; in all other discussions we have implicitly assumed that variables only contain real numbers. Functions are no different, and we may opt to specify the kinds of objects that functions receive and produce as output. The collection of all permissible inputs to a function is known as its *domain*. We can more or less assume that f defined above can receive any real number, and therefore the domain of f is $\mathbb{R}$. We should also see that given any real number x, $f(x)$ is also a real number; this therefore tells us that the *codomain* of f, the collection of mathematical objects that can be produced by f, is also $\mathbb{R}$. To give a simple analogy, a washing machine receives dirty clothes and produces clean clothes; then washing machines are functions whose domain is the collection of all dirty clothes, and codomain is the collection of all clean clothes.

To make explicit the domain and codomain of a function, we write $f : \mathbb{R} \to \mathbb{R}$ to say that the function f has domain and codomain $\mathbb{R}$. We shall frequently implicitly assume that the domain and codomain of functions discussed in this text can be understood based on context, and occasionally write the domain and codomain of functions only to be explicit.

If X is the domain of a function f then the collection of *all* and *only* the possible outputs of the function given members of X is known as its *image*, or the *image of X under f*, in other words, for each member x of the domain X of f the image of X under f is the collection containing only all $f(x)$. For example if we allow the domain of f defined above to be $\mathbb{N}$ then clearly the codomain of f is also $\mathbb{N}$, but the image of $\mathbb{N}$ under f is not $\mathbb{N}$, it is instead a sub-collection of $\mathbb{N}$, excluding 0, 1 and 2, because 0, 1 and 2 cannot be produced by f on any natural number input; in fact the image of $\mathbb{N}$ under f are the positive *odd* numbers greater than or equal to 3. We may also say *image* to mean a particular output of f under a particular input x, i.e. given input x, the image of x is $f(x)$.

Example 1.3. Let g be the following function:

$$\begin{aligned} g &: \mathbb{R} \to \mathbb{N} \\ g(x) &= 1 \end{aligned}$$

g is a unary function that maps real numbers to natural numbers, i.e. its domain is $\mathbb{R}$ and its codomain is $\mathbb{N}$. The image of $\mathbb{R}$ under g, or the image of g, is the collection containing only the number 1.

Functions work by receiving input and substituting the name of the parameter with the input everywhere in its definition. For example, given $f(x) = 2x + 3$, given input 2, we substitute x with 2 everywhere in the function definition, giving us $f(2) = 2(2) + 3$, therefore $f(2) = 7$. Therefore, functions relate inputs to outputs, or members of the domain to their respective images. This relationship can be plotted on a graph where the horizontal axis represents the input values and the vertical axis represents the output values, as shown in Figure 1.1.

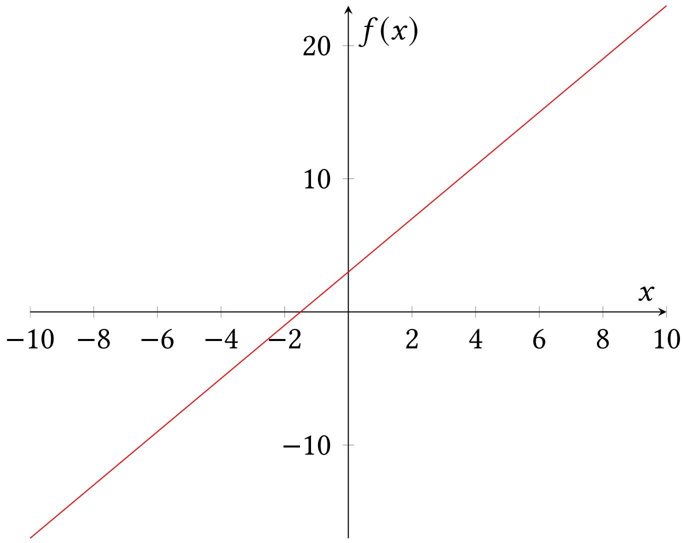


Figure 1.1: Graph of $f(x) = 2x + 3$.

An interesting question to ask is, at what values do the function intersect the x-axis? In words, for what values of x is $f(x) = 0$? Let us investigate this for $f(x) = 2x + 3$:

$$
\begin{aligned}
& & 2x + 3 &= 0 & & \\
&\Rightarrow & 2x &= -3 & &\triangleright \text{ subtract 3 from both sides} && (1.21)\\
&\Rightarrow & x &= -\frac{3}{2} & &\triangleright \text{ divide both sides by 2} && (1.22)
\end{aligned}
$$

Equation 1.22 shows that f intersects the x-axis when $x = -\frac{3}{2}$; a quick glance at Figure 1.1 also shows this to be the case. This value of x is also known as the *root* of the equation/function.

1.2.2 Univariate Polynomials

Our definition of the function f from earlier is a specific instance of a *polynomial function*, in particular, it is a *linear function*.

Definition 1.1 (Polynomial function). A (uni-variate) *polynomial function* of degree n is a function on a variable, producing the sum of (a finite number of) terms where each term is the variable raised to a natural number exponent not exceeding n, multiplied by a real number *coefficient*. In more precise terms, a polynomial $f(x)$ of degree n is of the form

$$f(x) = c_n x^n + c_{n-1} x^{n-1} + \ldots c_1 x + c_0$$

where n is a natural number and c_n to c_0 are real numbers.

As we can see from this definition, $f(x) = 2x + 3$ is a polynomial of degree 1, where $c_1 = 2$ and $c_0 = 3$. In general, the root of a linear function is $-\frac{c_0}{c_1}$. One question you might ask, is "why are finding roots important?" The simple answer is, given any polynomial f, determining the value of the input that produces a specific output corresponds to the root of another polynomial; for some value y, in trying to find a value of x where $f(x) = y$, we can simply find the root of $g(x) = f(x) - y$.

Example 1.4. Let us define another polynomial $f(x) = -4x + 5$. For what value of x will $f(x) = 10$? One way to go about this is forming another polynomial $g(x) = f(x) - 10 = -4x - 5$ and finding the root of g. The root of g is $\frac{5}{4}$, and we can see that $f(\frac{5}{4}) = 10$!

As such, let us find more root-finding formulae for higher-ordered polynomials. A *quadratic equation* is a polynomial with degree 2, i.e. it is of the form $f(x) = c_2x^2 + c_1x + c_0$ with nonzero c_2. For example, we can define a polynomial $f(x) = x^2 + 2x - 5$. Then, the relationship between input numbers and their corresponding outputs can likewise be graphed, as shown in Figure 1.2.

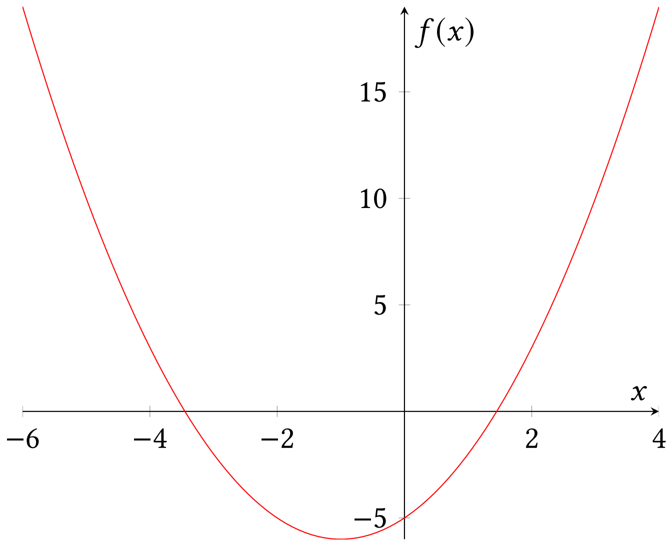


Figure 1.2: Graph of $f(x) = x^2 + 2x - 5$.

How can we find the root(s) of a quadratic equation? Readers might recall the famous quadratic formula, stating: the roots of a quadratic equation $ax^2 + bx + c$ is given by[7][8]

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Let us show that this is indeed the case. First, recall that for all x and y, $(x+y)^2 = x^2 + 2xy + y^2$. Then,

$$\begin{aligned}
& ax^2 + bx + c = 0 & \\
\Rightarrow \quad & x^2 + \frac{b}{a}x + \frac{c}{a} = 0 & \triangleright \text{ divide by } a \\
\Rightarrow \quad & x^2 + \frac{b}{a}x = -\frac{c}{a} & \triangleright \text{ subtract } \frac{c}{a} \\
\Rightarrow \quad & x^2 + 2\frac{b}{2a}x = -\frac{c}{a} & \triangleright \ \frac{x}{y} = 2\frac{x}{2y} \\
\Rightarrow \quad & x^2 + 2\frac{b}{2a}x + \left(\frac{b}{2a}\right)^2 = -\frac{c}{a} + \left(\frac{b}{2a}\right)^2 & \triangleright \text{ add } \left(\frac{b}{2a}\right)^2
\end{aligned}$$

[7] a is nonzero.
[8] Recall that $\sqrt{a} = a^{1/2}$.

$$
\begin{aligned}
&\Rightarrow & \left(x+\frac{b}{2a}\right)^2 &= -\frac{c}{a}+\left(\frac{b}{2a}\right)^2 && \triangleright \text{ factorize LHS}\\
&\Rightarrow & \left(x+\frac{b}{2a}\right)^2 &= -\frac{c}{a}+\frac{b^2}{4a^2} && \triangleright \left(\frac{x}{y}\right)^2 = \frac{x^2}{y^2}\\
&\Rightarrow & \left(x+\frac{b}{2a}\right)^2 &= -\frac{4ac}{4a^2}+\frac{b^2}{4a^2} && \triangleright \frac{x}{y} = \frac{4ax}{4ay}\\
&\Rightarrow & \left(x+\frac{b}{2a}\right)^2 &= \frac{b^2-4ac}{4a^2} && \triangleright \frac{a}{c}+\frac{b}{c} = \frac{a+b}{c}\\
&\Rightarrow & x+\frac{b}{2a} &= \pm\sqrt{\frac{b^2-4ac}{4a^2}} && \triangleright \text{ take square roots}\\
&\Rightarrow & x+\frac{b}{2a} &= \pm\frac{\sqrt{b^2-4ac}}{2a} && \triangleright \sqrt{\frac{a}{b}} = \frac{\sqrt{a}}{\sqrt{b}}\\
&\Rightarrow & x &= -\frac{b}{2a}\pm\frac{\sqrt{b^2-4ac}}{2a} && \triangleright \text{ subtract} \frac{b}{2a}\\
&\Rightarrow & x &= \frac{-b\pm\sqrt{b^2-4ac}}{2a} && \triangleright \frac{a}{c}+\frac{b}{c} = \frac{a+b}{c}
\end{aligned}
$$

As such, we have that the roots of f is $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ where $a = 1$, $b = 2$, $c = -5$, giving us two roots:

$$\frac{-2+\sqrt{4+20}}{2} = \sqrt{6}-1 \qquad\qquad \frac{-2-\sqrt{4+20}}{2} = -1-\sqrt{6}$$

From what we have done, two questions naturally come to mind: (1) if a linear equation has 1 root and a quadratic equation has two roots, does an n-degree polynomial have n roots? (2) can we find formulae to obtain the roots of an n-degree polynomial for all n? The answer to the first question is yes! The fact that an n-degree polynomial has (up to) n roots is known as the *fundamental theorem of algebra*. In fact, it states that all n-degree polynomials have precisely n *complex*[9] roots counted with their multiplicity. On the other hand, the answer to the second is no; for degree $n \geq 5$, there is no *general* formula expressed with the mathematical operations we have just described earlier that obtains the roots of any polynomial of degree n. In fact, there are polynomials (with degree at least 5) whose roots cannot even be expressed with our mathematical operations[10]! Finally, while there are indeed formulae for finding the roots of *cubic* (degree 3) and *quartic* (degree 4) polynomials, we shall omit them here due to their complexity. Instead, we offer examples of cubic and quartic functions in Figure 1.3.

1.2.3 N-nary Functions

Let us generalize our understanding further to functions that accept an arbitrary number of parameters. A function that accepts n parameters (a.k.a. has arity n) is described as *n-nary*. For example, a binary function may accept two parameters x and y and may be defined as $f(x, y) = 2x+3y$. When describing the domain of n-nary functions, we frequently denote them as sets containing *n-tuples*. In the case of binary functions, this would be the set of 2-tuples

[9]Complex numbers are an extension of the real numbers with $i = \sqrt{-1}$.

[10]This is known as the Abel-Ruffini Theorem; proof of which is usually shown with Galois theory nowadays. Galois theory, while incredibly abstract, actually has widespread applications in cybersecurity and program analysis!

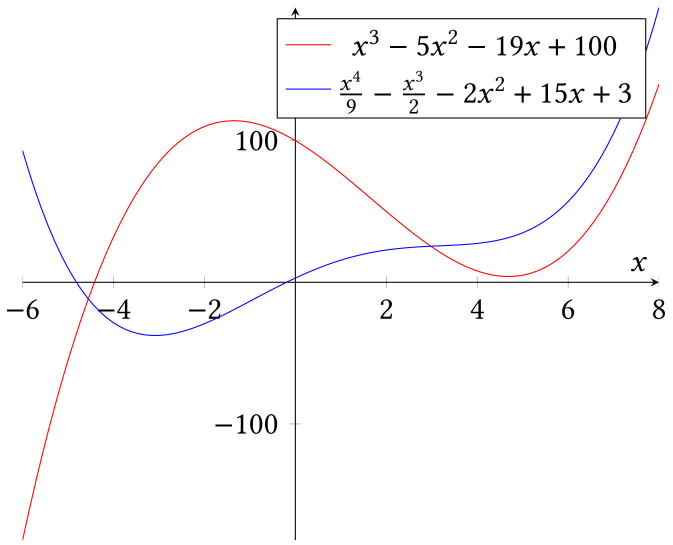


Figure 1.3: Examples of cubic and quartic functions.

a.k.a. pairs like $(2, 3)$ or $(4, 1)$, which is precisely the *cartesian product* of two sets[11]. We define cartesian products formally in Subsection 4.1.3; simply put, the cartesian product of two sets A and B, denoted $A \times B$, is a collection of pairs (a, b), where a belongs to A and b belongs to B. For example, if we let the domain of x and y be $\mathbb{R}$, then the domain of f is $\mathbb{R} \times \mathbb{R}$ which contains pairs of real numbers. If we define a function $g(x, y, z) = x + y - z$ such that x and z are real numbers but y can only be natural numbers, then we can describe the domain of g as $\mathbb{R} \times \mathbb{N} \times \mathbb{R}$. Thus, we have defined $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as $f(x, y) = 2x + 3y$.

Remark 1.3. Binary functions are interesting syntactically, because they can be written as an 'operation' between two terms. For example, the usual arithmetic operation $+$ is precisely a function on two numbers, which can also be written as $+(2, 3) = 5$. This form of notation is known as *prefix notation*; putting $+$ in the middle as per usual, is known as *infix notation*[a]. Notice that $+$ is just a name, as is $\times$, $-$ and $\div$. Thus, given our binary function f, writing $4\ f\ 5$ can be understood to mean $f(4, 5)$. We can in fact interpret $-x$ as the 'negative function' applied to x, i.e. it is $-(x)$.
Hopefully, you might see that many syntactically different terms (i.e. different by appearance) actually represent the same thing: the function $f(x) = 2x + 3$ is the exact same function as $f(y) = 2y + 3$;[b] if we let $g(x, y) = 2x^2 - 3y$ and let $* = g$ then $4 * 2 = *(4, 2) = g(4, 2) = 2(4)^2 - 3(2) = 26$. This phenomenon also has formal definitions which we shall elide, but we make note of this since being able to identify syntactically different but *semantically* identical constructs (i.e. they mean the same thing) is crucial to learning programming.

[a]There is also *postfix* notation, where $+$ is written at the end, as in $2\ 3\ + = 5$.
[b]In the λ-calculus, which is a theoretical model of computation, this is known as α-congruence.

Let us observe the relationship that is provided by f. Observe that we have two parameters

[11]There are, in fact, other interpretations of n-nary functions. In programming, we do not typically describe functions as acting on sets of objects; instead, they are functions that receive an object of some domain *type* and correspondingly produce an object of its codomain type. The domain of n-nary function types are also n-tuples, which correspond to product types; however as you might later find, we can convert all n-nary functions f to a *unique* unary function via something known as *currying*—this curried function produces an object in the *exponential* set/type, which happens to also be a function.

x and y that are independent of each other, and the result $f(x, y)$ depends on these values. As such, graphing f requires a 3-dimensional space with three axes—f is then represented as a surface in this space. In this case, since f is linear in x and y, f is a plane in the graph. We may also define more complex binary functions like $g(x, y) = x^2 + y^2$, both of which we plot in Figure 1.4.

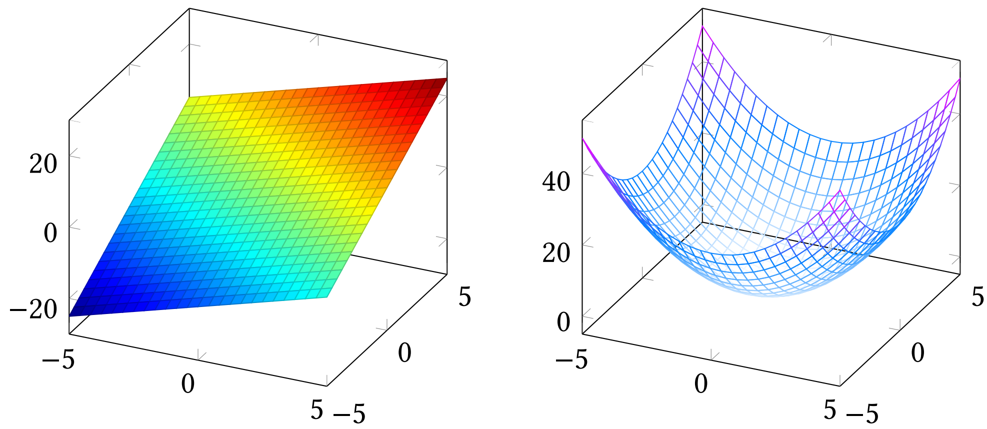


Figure 1.4: Example binary functions: $f(x, y) = 2x + 3y$ (left) and $g(x, y) = x^2 + y^2$ (right).

In general there is nothing special about n-nary functions; it is simply an extension of unary functions. You can even think of an n-nary function as a unary function that receives a single object as an n-tuple, and the elements of the tuple are then extracted into the corresponding parameters of the function. For example, given $f(x, y) = 2x + 3y$, f can be seen as a function that receives a single pair p of two numbers, x and y. These numbers are then extracted from the pair, and the result is $2x + 3y$. It then stands to reason that functions may also produce pairs (and any tuple) as output, as is the case for a function $h : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ given by $h(x) = (2x, 3x)$.

1.2.4 Function Composition

In many branches of mathematics, and certainly in computing, we wish to pass the result of a function into another function. This is known as *function composition*. Let us define a binary operation $\circ$ that describes this phenomenon.

Definition 1.2. Suppose we have functions f and g where the codomain of f (at the very least, its image) is equal to (or a subset of) the domain of g. Then, the composite of f and g, denoted $g \circ f$ (read as g after f) is the function $(g \circ f)(x) = g(f(x))$ whose domain is equal to the domain of f and codomain is equal to the codomain of g.

Again let us simplify this construction by thinking of functions as machines. Suppose we have two machines f and g where f receives eggs and produces chicks, and g receives chicks and produces chickens, we can compose these machines to get a new machine $g \circ f$ that receives an egg, passes it through f giving us a chick, then directly passing that chick through g, giving us a chicken. From this analogy it should be clear when two functions are *not* composable; if f produces chicks but g receives ducklings, $g \circ f$ cannot be defined.

Example 1.5. Suppose we have two functions $h : \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ and $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by $h(x) = (2x, 3x)$ and $k(x, y) = 2x + 3y$. Then, $k \circ h : \mathbb{R} \to \mathbb{R}$ is the function given by

$$(k \circ h)(x) = k(h(x)) = k(2x, 3x) = 4x + 9x = 13x$$

1.3 OTHER FUNCTIONS

1.3.1 Exponentials

Earlier, we have described polynomials, which are a special class of functions. However, in general, functions on a variable x do not require that each term be a power of x; we can also have functions that take a constant value (that is greater than 0 and not equal to 1) raised to the x^{th} power! These are known as *exponential functions*. For example, the exponential function $f(x) = 2^x$ is shown in Figure 1.5. You might be able to see that exponential functions have several properties: (1) $f(x+y) = f(x)f(y)$ (2) $f(x-y) = f(x)/f(y)$, (3) $f(x)^y = f(xy)$. In other words, for all real numbers a, m and n where $a > 0$, (1) $a^{m+n} = a^m a^n$, (2) $a^{m-n} = a^m/a^n$, (3) $(a^m)^n = a^{mn}$.

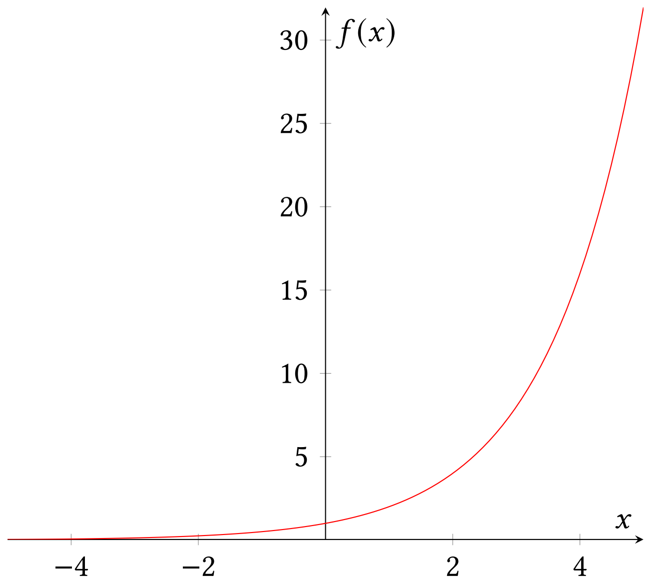


Figure 1.5: Graph of $f(x) = 2^x$.

Exponential functions are incredibly useful in real-world scenarios. For example, if we started with \$1 in a bank that pays 10% interest per year compounded annually, the amount of money we have in the bank after 5 years is given by $\$(1 + 1/10)^5$.

A natural question would then be to ask, what if we had compounded more frequently (instead of per annum)? Even still, what if we had *continuously* compounding interest? To formalize this problem, let's say we had \$1 in a bank giving 100% interest annually. If in a year, the bank compounds our interest n times, we would have $\$(1 + 1/n)^n$ after one year. How much money would we have if the bank compounds our interest 'infinity' times in a year? It has been shown that this quantity is not infinitely large, but rather, a fixed constant value 2.71828..., which many of you would know is known as *Euler's constant* e! As we know, e is one of the most frequently occurring constants, and we shall see some important characteristics of it later.

1.3.2 Inverses and Logarithms

In defining functions earlier a natural construction to have would be the *inverse* of a function, i.e. given a function $f(x)$, the inverse of f, denoted f^{-1}, would be such that $(f^{-1} \circ f)(x) = x$ and $(f \circ f^{-1})(x) = x$. For example, given $f(x) = 2x$ then its inverse is $f^{-1}(x) = x/2$ (you should be able to clearly see why). As you might recall, the inverse of an exponential function is known as its *logarithm*. However, inverses do not exist in general as we shall see later[12], and in the case of exponential functions, its inverse is not define for some constants, for example, the inverse of $f(x) = 1^x$ and $f(x) = 0^x$ is not defined, or when the constants are negative. In all other cases, we have:

$$f(x) = a^x = b \quad \Leftrightarrow \quad f^{-1}(b) = \log_a b = x$$

For example, if $f(x) = 2^x$, then $f^{-1}(x)$ can be defined as $\log_2 x$. By definition, $(f^{-1} \circ f)(x) = x$, so $\log_a a^x = x$.

There are more laws of exponentiation and logarithms which we shall list here. Assume that the base of the exponentiation and logarithms are positive real numbers that are not equal to 1:

$$a^b a^c = a^{b+c} \qquad \left(a^b\right)^c = a^{bc} \qquad a^c b^c = (ab)^c$$

$$\frac{a^b}{a^c} = a^{b-c} \qquad \frac{a^c}{b^c} = \left(\frac{a}{b}\right)^c \qquad a^0 = 1$$

$$\log_a b + \log_a c = \log_a bc \qquad \log_a b - \log_a c = \log_a \frac{b}{c} \qquad \log_a b = \frac{\log_c b}{\log_c a}$$

$$\log_a \left(b^x\right) = x \log_a b \qquad a^{\log_a x} = x \qquad \log_a a = 1$$

$$x = y \Leftrightarrow a^x = a^y \Leftrightarrow \log_a x = \log_a y$$

Let us graph our inverse function given by $f^{-1}(x) = \log_2 x$, shown in Figure 1.6.

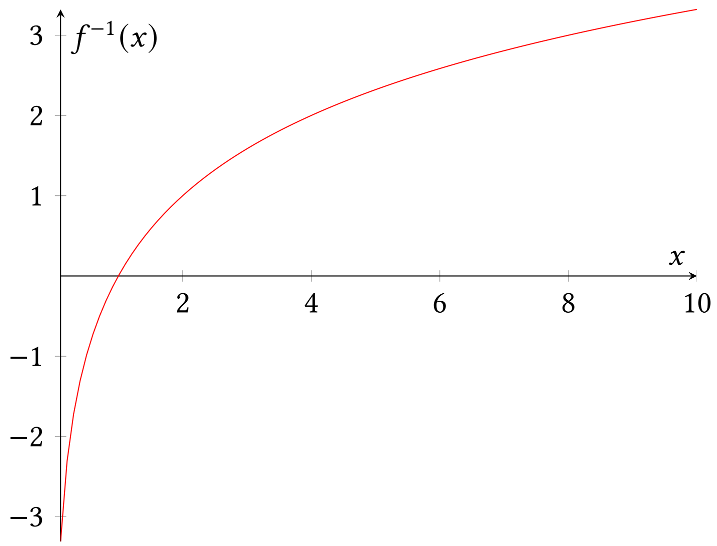


Figure 1.6: Graph of $f^{-1}(x) = \log_2 x$.

You might observe from Figure 1.5 that $f(x) > 0$ for all x, and dually, from Figure 1.6 that $f^{-1}(x)$ is only defined on $x > 0$. This result is occasionally useful.

Let us show some uses of exponentials and logarithms frequently found in introductory computing.

[12]For the impatient, a function only has an inverse if it is a one-to-one function, i.e. a *bijection* or *bijective function*. In some cases, we only demand a *left inverse*, which is a less restrictive version of the inverse. The analogue of an injective function in other areas of algebra is known as a *monomorphism* or *monic*.

Example 1.6. Suppose we start with one cell today and every day each cell goes through cell division, dividing itself into two, how many do we get after n days? The answer is, of course, $f(n) = 2^n$. For example, three days from now, we would have $f(3) = 8$ cells. We can depict how the number of cells grow as a *tree* (that is upside down), shown in Figure 1.7. In the first layer, we have our single cell S, which is our only cell in the beginning. After one day, S splits into two cells, A and B, as shown in the second layer, meaning that after one day, we now have two cells. Each of those cells split further into two cells, then each of those split again, giving us a total of eight cells after three days, which is the width of the bottom-most layer. The dual question would then be, how many days is required to get at least n cells if we started from 1 cell? The answer would naturally be $g(n) = \log_2 n$; more specifically, it is actually $g(n) = \lceil \log_2 n \rceil$ where $\lceil x \rceil$ is the *ceiling* of x, which is x rounded *up* (towards ∞) to the nearest integer. The dual of the *ceiling* is the *floor* of x denoted $\lfloor x \rfloor$, which is rounding *down* (towards $-\infty$) to the nearest integer. Clearly, we can see that getting eight cells starting from one requires $\log_2 8 = 3$ days, corresponding to the height of the tree in Figure 1.7 (the number of layers $-$ 1). We can also generalize this result, saying that if a cell divides itself into m other cells in a day, then after n days we would have $f(n) = m^n$ cells, and similarly getting n cells starting from one requires $g(n) = \log_m n$ days.

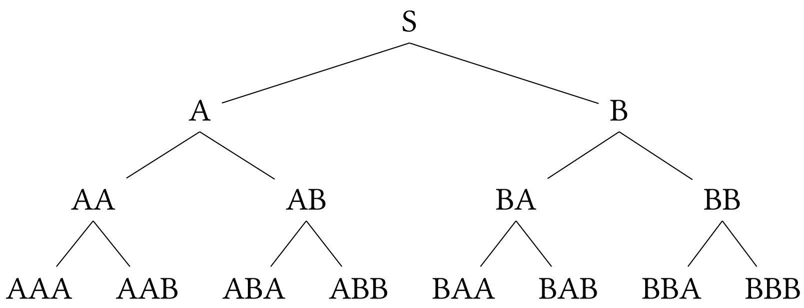


Figure 1.7: Tree showing cell division after 3 days.

Example 1.7. Suppose we have some paper that only has space for writing n digits. What is the largest natural number we can write on this paper? Again, the answer is $f(n) = 10^n - 1$, i.e. if we had space for writing 5 digits, the largest natural number we could write is $f(5) = 10^5 - 1 = 99999$. Conversely, given n what is the minimum amount of space required on that piece of paper? Yet again, the answer is $g(n) = \lceil \log_{10}(n+1) \rceil$ space; writing the number 12345 requires $\lceil \log_{10}(12346) \rceil = 5$ digits worth of space!

1.3.3 Absolutes

Given any real number x, the *absolute* value[13] of x, denoted $|x|$, is given by the following rule:

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{otherwise} \end{cases}$$

[13] You might have learnt this to be the *modulus* of x. We shall avoid this term because the word is very similar to the term *modulo*, which has an entirely different meaning.

What we have just described is a *piece-wise function*, i.e. the function is composed of multiple 'pieces', each piece coming from one of the rules in the equation. The function is defined by multiple rules; the actual rule for deciding the output of the function on a particular input is the first rule whose condition is satisfied. For example, $|7| = 7$ because 7 satisfies the first rule ($7 \geq 0$ is true), whereas $|-2| = -(-2) = 2$ because -2 does not satisfy the first rule (it satisfies the second rule, since 'otherwise' is a 'catch-all'). Graphing the function $f(x) = |x|$ shows precisely the two pieces that are defined by f, as seen in Figure 1.8.

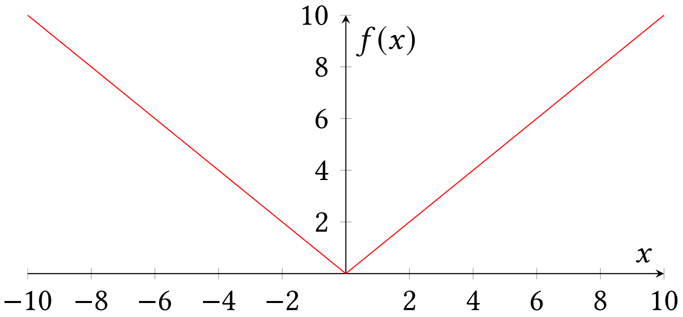


Figure 1.8: Graph of $f(x) = |x|$.

This 'absolute function' has several properties which can quite easily be shown. For all real numbers x and y and natural numbers n, and functions $f : \mathbb{R} \to \mathbb{R}$:

$$|x| \geq 0 \qquad |xy| = |x||y| \qquad |-x| = x$$

$$|x^n| = |x|^n \qquad |f(x)| = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ -f(x) & \text{otherwise} \end{cases} \qquad |x| = y \Leftrightarrow x = y \text{ or } -x = y$$

To describe the fifth rule, in general, $-f(x)$ produces the same graph as $f(x)$, except reflected along the x-axis. For example, the function $g(x) = f(x) = x^3 - 5x^2 - 19x$ is graphed in red in Figure 1.9 (left), while $g(x) = -f(x) = -x^3 + 5x^2 + 19x$ is graphed in blue in the same graph. Then, $|f(x)|$ is simply taking the pieces of $f(x)$ and $-f(x)$ that are above the x-axis, as shown in Figure 1.9 (right).

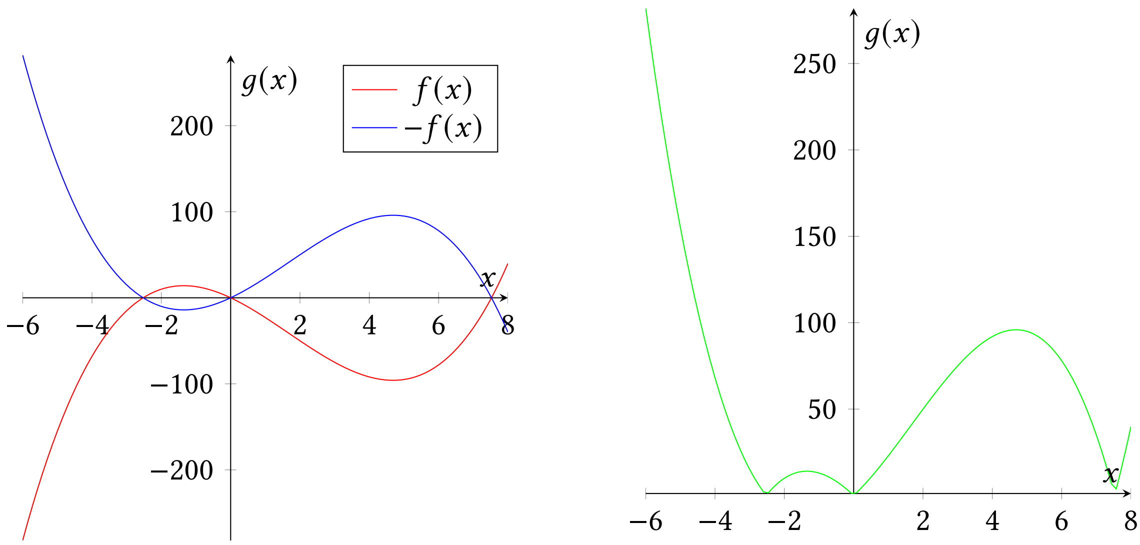


Figure 1.9: An example of $f(x)$ graphed together with $-f(x)$ and $|f(x)|$.

1.4 SERIES

Given $f(x) = 2x + 3$ on natural numbers, we can obtain the *series* $f(0), f(1), f(2), \dots$ which is the sequence $3, 5, 7, \dots$. We may then ask what the sum or product of the terms of this sequence is. For this, let us introduce new notation describing the sum or product of terms.

Sum of terms. Suppose we wanted to express the sum $f(0) + f(1) + \cdots + f(n)$. For this, we write

$$\sum_{x=0}^{n} f(x)$$

This means that we shall sum all terms in the form of $f(x)$, starting with the term where $x = 0$, up to and including the term where $x = n$. Dually, we can denote the product of a sequence of terms.

Product of terms. Suppose we wanted to express the product $f(0) \times f(1) \times \cdots \times f(n)$. For this, we write

$$\prod_{x=0}^{n} f(x)$$

Example 1.8. The *factorial* of n, also known as n factorial, denoted $n!$, is the product of all positive integers up to and including n, i.e.

$$n! = n \times (n-1) \times \cdots \times 2 \times 1 = \prod_{i=1}^{n} i$$

For example, $4! = 4 \times 3 \times 2 \times 1 = 24$. A special case of factorial is $0!$, which is defined to be $0! = 1$.

Example 1.9. We can in fact write infinite sums/products. For example, e can be expressed as

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} + \frac{1}{1 \times 2 \times 3 \times 4} + \dots$$

Now let us look at some simple series and some of their properties.

1.4.1 Arithmetic Series

Definition 1.3. A series of terms $u_0, u_1, u_2, \dots$ is an *arithmetic progression* if the difference between adjacent terms is constant, i.e. for any natural number i, we have $u_{i+1} - u_i = c$ for a constant c. In other words, the function producing each term can be described as a linear polynomial $u_x = f(x) = cx + u_0$.

For example, the series $123, 128, 133, 138, \dots$ is an arithmetic progression; the i^{th} term is given by $u_i = 5i + 123$, where i is a natural number. What is the sum of, say, the first 10 terms of this series? Notice that the i^{th} term and the $(9-i)^{\text{th}}$ can always be paired up to give the same sum, i.e. for all i from 0 to 9, we have $u_i + u_{9-i} = 123 + 5i + (123 + 5(9-i)) =$

$123 + 5i + 123 + 45 - 5i = 2(123) + 45$, and the number of pairs we have is $n/2 = 5$. As such, we get that the sum of the first 10 terms of our arithmetic progression is $5(2(123) + 45)$.

In general, suppose we have an arithmetic progression where each of the x^{th} terms (starting from 0) is defined by the polynomial $ax + b$, for constants a and b; this is the series $b, a + b, 2a + b, \dots$. Then, the sum of the first n terms can be found by computing:

- The i^{th} term and $(n - 1 - i)^{\text{th}}$ terms sum to a constant for all i. This is $u_0 + u_{n-1} = b + a(n-1) + b = a(n-1) + 2b$
- There are $n/2$ of such sums in the sum of the first n terms (it doesn't matter if n is odd; the middle term can be seen as two copies of halves of itself)

Thus, the sum of the first n terms is

$$\sum_{x=0}^{n-1} (ax + b) = \frac{n}{2}\left(a(n-1) + 2b\right) = \frac{an^2 - an + 2bn}{2}$$

Remark 1.4. From this arithmetic progression whose terms are determined by the linear polynomial $f(x) = ax + b$, we may obtain a new series where the i^{th} term is the *sum of the first i terms of the arithmetic progression*, i.e. it is the series

$$\sum_{i=0}^{0} f(i), \sum_{i=0}^{1} f(i), \sum_{i=0}^{2} f(i), \dots$$

Since we know how to find the sum of the first n terms of this arithmetic progression, each term in the series can be expressed as a quadratic function $g(n) = (an^2 + (2b - a)n)/2$:

$$\sum_{i=0}^{0} f(i), \sum_{i=0}^{1} f(i), \sum_{i=0}^{2} f(i), \dots = g(0), g(1), \dots$$

Notice that each term is now expressed as a quadratic equation. While not immediately apparent, this result is useful especially when studying data structures and algorithms. This new series also the series of *partial sums* of our arithmetic progression, which is a *divergent* series. This implies that the *sum of all terms in our arithmetic progression* is not finite[a].

[a]You may have heard that the sum of all natural numbers is equal to $-1/12$. This is not mathematically precise, obviously. The more mathematically rigorous statement would be that 'we can assign $-1/12$ to the sum of all natural numbers'. This can be done by *analytic continuation*, which is a technique of *complex analysis.*

1.4.2 Geometric Series

Definition 1.4. A series of terms $u_0, u_1, u_2, \dots$ is a *geometric progression* if the ratio between adjacent terms is constant, i.e. for any natural number i, we have $u_{i+1}/u_i = c$ for a constant c. In other words, the function producing each term is the product of an exponential function, in the form of $u_x = f(x) = u_0 c^x$.

For example, the series $3, 6, 12, 24, \dots$ is a geometric progression; the i^{th} term is given by $u_i = 3(2^i)$ where i is a natural number.

Let us skip right ahead to find the sum of the first n terms of any geometric progression, whose terms are given by ab^i. The first n terms is thus $a, ab, ab^2, \ldots, ab^{n-1}$. Let us denote this sum as $S(n)$, given by

$$S(n) = \sum_{i=0}^{n-1} ab^i = a + ab + ab^2 + \cdots + ab^{n-1} \tag{1.23}$$

Now observe that:

$$b \times S(n) = ab + ab^2 + \cdots + ab^{n-1} + ab^n \tag{1.24}$$

Thus, if we subtract Equation 1.23 from Equation 1.24 (and vice-versa), we have

$$\begin{aligned}
b \times S(n) - S(n) &= ab^n - a && \triangleright (1.24) - (1.23)\\
&= a(b^n - 1)\\
&= (b-1)(S(n))\\
S(n) &= \frac{a(b^n-1)}{b-1} && (1.25)\\
S(n) - S(n) \times b &= a - ab^n && \triangleright (1.23) - (1.24)\\
&= a(1-b^n)\\
&= (1-b)(S(n))\\
S(n) &= \frac{a(1-b^n)}{1-b} && (1.26)
\end{aligned}$$

As such,

$$\sum_{i=0}^{n-1}(ab^i) = \frac{a(b^n-1)}{b-1} = \frac{a(1-b^n)}{1-b}$$

Food for thought 1.5. We cannot divide by 0, as such the equation is invalid for geometric progressions where $b = 1$. However, when $b = 1$, is it not a geometric progression anymore? How can we find the sum of the first n terms of such a series?

Just like arithmetic progressions we now have a general formula for obtaining the sum of the first n terms of a geometric progression. For example, for our series $3, 6, 12, 24, \ldots$, the sum of the first 10 terms is given by $3(2^{10} - 1)$.

What is the sum of *all* terms in a geometric progression? As stated earlier, there is no finite sum of all terms in an arithmetic progression, but this is not necessarily the case for geometric progressions! Intuitively, notice that if $|b| < 1$, as n gets incredibly large, b^n gets incredibly small, in fact it gets closer to 0! As such, the term $(1 - b^n)$ is close to 1. Thus, when as n gets larger causing $(1 - b^n)$ to get closer to 1, we get that the sum of the first n terms get closer to $a/(1 - b)$. This in fact is the sum of the geometric progression when $|b| < 1$.

However, what does it mean that *as n gets incredibly large, b^n is close to 0?* This is known as the *limit* of b^n as n *tends* towards infinity, written

$$\lim_{n\to\infty} b^n = 0$$

Definition 1.5 (Limit)**.** Let $f(0), f(1), f(2), \ldots$ be a sequence of real numbers. If the *limit* of this sequence exists, the real number L is the limit of this sequence, i.e.

$$\lim_{n\to\infty} f(n) = L$$

if for any real number $\epsilon > 0$, there exists natural number N such that for all $n > N$,

$$|f(n) - L| < \epsilon$$

Thus, let us construct a sequence $b^0, b^1, \ldots, b^n, \ldots$. Then, to say $\lim_{n\to\infty} b^n = 0$ means that for any real number $\epsilon > 0$, there will always exist some N such that for all $n > N$, $|b^n - 0| < \epsilon$. For example, let $b = 0.5$ and $\epsilon = 0.01$. 0 is indeed the limit of b^n as $n \to \infty$ because we can let $N = 6$ and show that for all $n > 8$, $b^n < 0.01$. Indeed, $0.5^7 = 0.0078125$, and for any $n > 7$, $0.5^n < 0.0078125 < 0.01$. In fact, we can make ϵ arbitrarily small, and we can still find some N such that for all $n > N$, $0.5^n < \epsilon$.

Example 1.10. Consider the series $1, 1/2, 1/4, 1/8, \ldots$. Clearly, each term can be written as $(1/2)^i$. Since $|b| = 1/2 < 1$, the sum of all terms of the series is thus $1/(1/2) = 2$.

1.4.3 Recurrence Relations

In some interesting cases, the terms of a series depend on earlier terms in the series. For example, the factorial of n is the product of numbers from 1 to n. However, notice that for all $n \geq 1$:

$$\begin{aligned} n! &= n \times (n-1) \times \cdots \times 2 \times 1 \\ (n+1)! &= (n+1) \times \quad n \times (n-1) \times \cdots \times 2 \times 1 \\ &= (n+1) \times n! \end{aligned}$$

Thus, we can describe each term of the series $0!, 1!, 2!, \ldots$ as being:

$$n! = \begin{cases} 1 & \text{if } n = 0 \\ n \times (n-1)! & \text{otherwise} \end{cases}$$

This series is known as a *recurrence relation*, which is a sequence where the n^{th} term depends on earlier terms.

Another famous example of recurrence relations is the classic *Fibonacci series*, given by $1, 1, 2, 3, 5, 8, \ldots$, where each term is the sum of the previous two terms (the first two terms are equal to 1). This can thus be written as

$$\text{fib}_n = \begin{cases} 1 & \text{if } n = 0 \text{ or } 1 \\ \text{fib}_{n-1} + \text{fib}_{n-2} & \text{otherwise} \end{cases}$$

Reasoning about recurrence relations, and in general, *recursive* functions (functions that are defined by some expression containing itself), is incredibly important in programming and computer science.

EXERCISES

1.1. Solve the system of equations (1) $a = 2b$ and (2) $b = a + 3$.

1.2. Suppose we define a binary function $f(x, y) = x/y$. Would $\mathbb{N} \times \mathbb{N}^+ \to \mathbb{N}$ (where $\mathbb{N}^+$ represents *positive* natural numbers) be a correct description of f? If not, why?

1.3. Find the root of the linear equation $-5x + 7$.

1.4. Find the root(s) of the quadratic equation $2x^2 + 3x - 5$.

1.5. Let $f(x) = 2x^2$ and $g(y) = y + 4$. Write the functions $g \circ f$ and $f \circ g$.

1.6. Show that if $f(n) = \log_a n$ and $g(n) = \log_b n$ for some $a, b > 1$, then $g(n) = k \times f(n)$ for some constant k.

1.7. Find the sum of the first 10 terms of $6, 2, -2, -6, -10, \ldots$

1.8. Find the limit of

$$\lim_{n\to\infty} \frac{n-1}{n}$$

1.9. Is the following sequence a geometric progression?

$$\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$$

1.10. Let $f(x) = \frac{1}{x^2}$. Is the sum of all terms in the following series $f(0), f(1), f(2), \ldots$ finite?

1.11. Re-write $f(n) = n^2$ as a recurrence relation for all natural numbers n.

Chapter 2

Calculus

> *If I have been able to see further, it was only because I stood on the shoulders of giants.*
>
> Isaac Newton

Having studied (unary) functions on real numbers, two questions may arise:

1. How do we determine the rate of change of the function at a specific point? Is there a formula or function to describe the rate of change of a function at any point?

2. How do we determine the area under a function between two points? Is there a formula or function to describe the area under a function between any two points, or perhaps, the area under the entire function?

The study of these two questions is known as the *calculus of infinitesimals*, or better known as just *calculus*. There are two branches of calculus, *differential* calculus answers the first question, while *integral* calculus answers the second. This chapter provides a recap on some fundamental results of calculus; Section 2.1 describes derivatives and rates of change, while Section 2.2 describes integrals and areas under curves.

One of the key findings in calculus is that *the rate of change of an exponential function* $f(x) = a^x$ *at a point* n *is directly proportional to* $f(n)$. In particular, there is an exponential function such that the rate of change of the function at a point n is *precisely equal* to $f(n)$; this function, as you might expect, is $\exp(x) = e^x$, in other words, for all (real numbers) x, the rate of change of e^x is precisely equal to e^x. Since e is so crucial in calculus, we will make frequent uses of the exp function, and the *natural logarithm* $\log_e x$ with base e, which will be termed $\ln x$. We may from this fact, along with others, describe the rate of change, and area under, other functions.

In addition, in general, not all functions are amenable to differentiation and integration at all points[1]. We omit going into detail on what constitutes differentiability and integrability. Just note that the functions that we have studied (except absolutes) are differentiable at all points, and that continuous functions are all integrable. You may assume in this chapter that the functions that we will study have well-defined notions of derivatives and integrals.

[1] You might have heard that all differentiable functions are continuous, and that all the functions we've seen are differentiable, therefore functions that are continuous are therefore differentiable. While it is true that differentiable functions are continuous, the converse is not always true.

2.1 DIFFERENTIATION

The *derivative* of a function f, denoted f', is the function describing the rate of change of f. If we re-write $f(x)$ as y, i.e. $y = f(x)$, we may describe the rate of change of f as dy/dx as we shall see later. The process of finding the derivative of a fuction is known as *differentiation*.

Let us build some intuition behind derivatives. Suppose we have a function $f(x)$ and we want to find how much $f(x)$ changes between two values, n, and $n+a$. Doing so is incredibly simple; this is $f(n+a) - f(n)$. Now, the *rate* at which this changes can simply be obtained by dividing over the change in x, i.e. $n + a - n = a$. As such, the *gradient* of the line intersecting f at n and $n+a$ is $(f(n+a) - f(n))/a$.

Now suppose we are attempting to find the rate of change of a function $y = f(x) = x^2$ at the point where $x = 4$ and $f(4) = 16$. As before we can approximate the rate of change at $x = 4$, by adding some number, say 4, to x, giving us $x + 4 = 8$. Then, the value of the function at $x + 4$ is then $f(x+4) = f(8) = 64$. Like before, we can construct a line intersecting f at 4 and $4+4$, which as gradient $(f(8) - f(4))/4 = 12$. This is our approximate rate of change of $f(x)$ at $x = 4$. We visualize this in Figure 2.1 (left).

Let us make our approximation more precise. Instead of adding 4 to x, let us add an even smaller number, say 2, to x, and see what that yields us. Basically, we now have two pairs of numbers, $(x, f(x))$ and $(x+2, f(x+2))$, i.e. $(4, 16)$ and $(6, 36)$. The line intersecting these two points has gradient $(f(x+2) - f(x))/(x+2-x) = (36-16)/2 = 10$. This can be seen in Figure 2.1 (right).

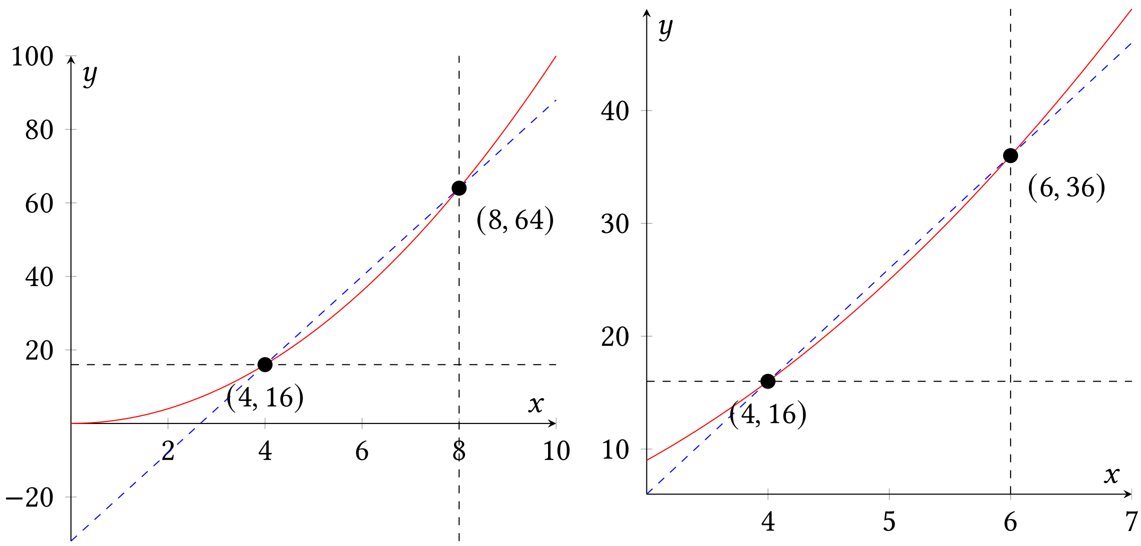


Figure 2.1: Graphs of $y = x^2$ plotted showing intersections with lines at $x = 4$.

As we move less far away from x, our approximation of the rate of change at $x = 4$ gets more precise. That is, for an infinitesimally small change in x, which we shall call dx, the formula

$$\frac{f(x+dx) - f(x)}{dx}$$

gives us the rate of change at any x. In the case where $f(x) = x^2$, we get

$$\begin{aligned}\frac{f(x+dx)-f(x)}{dx} &= \frac{(x+dx)^2 - x^2}{dx} \\ &= \frac{x^2 + 2xdx + (dx)^2 - x^2}{dx} \\ &= \frac{dx(2x+dx)}{dx} \\ &= 2x + dx\end{aligned}$$

Since dx is infinitesimally small, we get that the rate of change of $f(x)$ at all x is given by $2x$, so the rate of change at $x = 4$ is 8. Thus, (loosely speaking) if we have $y = f(x)$ then the rate of change of a small dy over an infinitesimally small change dx, given by dy/dx, is $(f(x+dx) - f(x))/dx$. Of course, dy/dx is not a fraction, but what we hope to have done is to give you some intuition behind the notation for derivatives, and how it can be calculated.

2.1.1 How Differentiation Works

One definition of the derivative of a function f at x is given by limits:

Definition 2.1. Assume that f is *differentiable* at x. Then, the derivative of f at x is given by

$$f'(x) = \frac{df}{dx}(x) = \lim_{n \to 0} \frac{f(x+n) - f(x)}{n}$$

This definition states that the derivative of f at x is given by $(f(x+n) - f(x))/n$ as n *approaches* 0. However, working out such limits is a nontrivial task for many functions. As such, mathematicians throughout history have managed to discover some *rules* of differentiation that apply to many functions. Many of these proofs are too complicated to cover here, as such, we shall simply provide some of these rules and see some applications of them. Where the rules are simple to derive, we will describe.

2.1.2 Exponential Functions and the Chain Rule

We shall begin discussing differentiation techniques with the very first rule, known as the *chain rule.*

Proposition 2.1 (Derivative of composites).

$$(g \circ f)' = (g' \circ f) \cdot f'$$

where $\cdot$ represents multiplication; in more familiar terms:

$$h(x) = g(f(x)) \Leftrightarrow h'(x) = g'(f(x)) \cdot f'(x)$$

In Leibniz notation (with $\frac{dy}{dx}$ like above), if we have $y = f(x)$ and $z = g(y)$ then

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

Next, recall at the beginning of this chapter, an incredibly famous result (along with a result for natural logarithms):

PROPOSITION 2.2 (Derivative of natural exponentiation and logarithms). *The rate of change of $f(x) = e^x$ at any point x is $f(x)$ itself, and that of $f(x) = \ln x$ is the reciprocal of x:*

$$\frac{d}{dx}e^x = e^x \qquad \frac{d}{dx}\ln x = \frac{1}{x}$$

Further recall that by definition, $a = e^{\ln a}$ for $a > 0$. Now, let us try to express any exponential function a^x as an exponential function with e and ln:

$$\begin{aligned} & a^x = y \\ \Leftrightarrow \quad & \ln a^x = \ln y \\ \Leftrightarrow \quad & x \ln a = \ln y \\ \Leftrightarrow \quad & e^{x\ln a} = e^{\ln y} \\ \Leftrightarrow \quad & e^{x\ln a} = y = a^x \end{aligned}$$

As such, we have just shown (virtually by definition) that for any $a > 0$, $a^x = e^{x\ln a}$.

THEOREM 2.3 (Derivative of Exponential with General Base). *Let $a > 0$ be a constant. $\frac{d}{dx}a^x = a^x \ln a$.*

Proof. Let $u = x\ln a$ and $y = e^u = e^{x\ln a} = a^x$ where a is a constant greater than 0. Notice that since $u = x\ln a$ is just a linear function with a constant gradient (rate of change) $\ln a$, by the chain rule,

$$\begin{aligned} \frac{dy}{dx} &= \frac{dy}{du}\cdot\frac{du}{dx} \\ &= e^u \ln a \\ &= e^{x\ln a}\ln a \\ &= a^x \ln a \end{aligned}$$

□

We have just used the chain rule to prove our first result: if $f(x) = a^x$ for any constant $a > 0$, we get $f'(x) = a^x \ln a$. Notice that if $f(x) = e^x$ then we get that $f'(x) = e^x \ln e = e^x$ which is exactly as we expect.

Example 2.1. Let us find the derivative of $f(x) = 3^{x^2}$. Let us allow $g(x) = x^2$ and $h(u) = 3^u$. This allows us to show that $f(x) = h(g(x))$. By the chain rule,

$$f'(x) = h'(g(x))g'(x)$$

. We know that $h'(u) = 3^u \ln 3$. Later on, we will also show that $g'(x) = 2x$. As such, $f'(x) = \ln 3 \cdot 2x \cdot 3^{x^2}$.

2.1.3 Polynomials

Products

Let us consider the derivative of the function $h(x) = f(x)g(x)$, i.e. the product of two functions.

Proposition 2.4 (Derivative of Products). *Let $h(x) = f(x)g(x)$. Then, $h'(x) = f'(x)g(x) + f(x)g'(x)$.*

Proof Sketch. We shall approach this by analysis on limits. We should get, that by definition:

$$\begin{aligned}
h'(x) &= \lim_{n\to 0} \frac{h(x+n) - h(x)}{n} \\
&= \lim_{n\to 0} \frac{f(x+n)g(x+n) - f(x)g(x)}{n} \\
&= \lim_{n\to 0} \frac{f(x+n)g(x+n) - f(x)g(x+n) + f(x)g(x+n) - f(x)g(x)}{n} \\
&= \lim_{n\to 0} \frac{(f(x+n) - f(x))g(x+n) + f(x)(g(x+n) - g(x))}{n} \\
&= \lim_{n\to 0} \left(\frac{f(x+n) - f(x)}{n} g(x+n) + f(x)\frac{g(x+n) - g(x)}{n}\right) \\
&= \lim_{n\to 0} \frac{f(x+n) - f(x)}{n} \cdot \lim_{n\to 0} g(x+n) + \lim_{n\to 0} f(x) \cdot \lim_{n\to 0} \frac{g(x+n) - g(x)}{n} \\
&= \lim_{n\to 0} \frac{f(x+n) - f(x)}{n} \cdot g(x) + f(x) \cdot \lim_{n\to 0} \frac{g(x+n) - g(x)}{n} \\
&= f'(x)g(x) + f(x)g'(x)
\end{aligned}$$

□

A simple corollary to this is the case where one of the functions is a constant. The rate of change of a constant function $f(x) = a$ is 0, as such, if we have $h(x) = f(x)g(x)$ where $f(x) = a$, then $h'(x) = f'(x)g(x) + f(x)g'(x) = 0g(x) + a \cdot g'(x) = a \cdot g'(x)$.

Another corollary is on the derivative of functions that are quotients of two functions.

Corollary 2.5 (Derivative of Quotients). *If $h(x) = f(x)/g(x)$, then $h'(x) = (f'(x)g(x) - f(x)g'(x))/(g(x)^2)$.*

Proof. By definition, $f(x) = g(x)h(x)$. Thus, by the product rule, $f'(x) = g'(x)h(x) +$

$g(x)h'(x)$.

$$\begin{aligned}
& & f'(x) &= g'(x)h(x) + g(x)h'(x) & \\
&\Leftrightarrow & g(x)h'(x) &= f'(x) - g'(x)h(x) & \triangleright \text{ rearranging terms} \\
&\Leftrightarrow & g(x)h'(x) &= f'(x) - g'(x)\frac{f(x)}{g(x)} & \triangleright \text{ by definition of } h(x) \\
&\Leftrightarrow & g(x)h'(x) &= \frac{f'(x)g(x) - g'(x)f(x)}{g(x)} & \\
&\Leftrightarrow & h'(x) &= \frac{f'(x)g(x) - g'(x)f(x)}{g(x)^2} &
\end{aligned}$$

□

Powers

Now let us consider a function in the form of $f(x) = x^n$ where $n \in \mathbb{R}$.

PROPOSITION 2.6 (Derivative of Powers). *If* $f(x) = x^n$, $f'(x) = nx^{n-1}$.

Proof Sketch. Let us temporarily consider the case where $x > 0$. By definition, $\ln x$ and e^x are inverses of each other, so we should get that

$$\begin{aligned}
x^r &= e^{\ln x^n} \\
&= e^{n \ln x}
\end{aligned}$$

Thus, Let $f(x) = e^{n \ln x}$. Then,

$$\begin{aligned}
f'(x) &= e^{n \ln x} \cdot \frac{n}{x} & \triangleright \text{ by chain rule and product rule} \\
&= x^n \frac{n}{x} & \\
&= nx^{n-1} &
\end{aligned}$$

The cases where $x < 0$ and $x = 0$ give the same result, if defined.

□

Sums

Now, let us consider a function in the form of $h(x) = f(x) + g(x)$, i.e. it is the sum of two functions.

THEOREM 2.7 (Derivative of Sums). *If* $h(x) = f(x) + g(x)$, *then* $h'(x) = f'(x) + g'(x)$.

Proof Sketch. We shall approach this by analysis on limits. We should get that, by definition,

$$\begin{aligned}
h'(x) &= \lim_{n\to 0} \frac{h(x+n)-h(x)}{n} \\
&= \lim_{n\to 0} \frac{f(x+n)+g(x+n)-f(x)-g(x)}{n} \\
&= \lim_{n\to 0} \frac{f(x+n)-f(x)+g(x+n)-g(x)}{n} \\
&= \lim_{n\to 0} \left(\frac{f(x+n)-f(x)}{n} + \frac{g(x+n)-g(x)}{n} \right) \\
&= \lim_{n\to 0} \frac{f(x+n)-f(x)}{n} + \lim_{n\to 0} \frac{g(x+n)-g(x)}{n} \\
&= f'(x)+g'(x)
\end{aligned}$$

□

Derivatives of Polynomials

With the results above, we have an incredibly simple systematic approach to obtaining the derivative of polynomials. Suppose we have $y = c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$. Then,

$$\begin{aligned}
\frac{dy}{dx} &= \frac{d}{dx}c_n x^n + \frac{d}{dx}c_{n-1}x^{n-1} + \cdots + \frac{d}{dx}c_1 x + \frac{d}{dx}c_0 && \triangleright \text{ sum rule} \\
&= nc_n x^{n-1} + (n-1)c_{n-1}x^{n-2} + \cdots + c_1 && \triangleright \text{ power rule}
\end{aligned}$$

Example 2.2. Let us find the derivative of $f(x) = (x^2+2x+4)/(x-5)$. We can express this as $f(x) = g(x)/h(x)$ where $g(x) = x^2+2x+4$ and $h(x) = x-5$. First, we compute g' and h':

$$\begin{aligned}
g'(x) &= 2x+2 \\
h'(x) &= 1
\end{aligned}$$

As such,

$$\begin{aligned}
f'(x) &= \frac{g'(x)h(x)-g(x)h'(x)}{g(x)^2} \\
&= \frac{(2x+2)(x-5)-(x^2+2x+4)(1)}{(x-5)^2} \\
&= \frac{2x^2-10x+2x-10-x^2-2x-4}{x^2-10x+25} \\
&= \frac{x^2-10x-14}{x^2-10x+25} \\
&= \frac{x^2-10x+25-39}{x^2-10x+25} \\
&= 1 - \frac{39}{x^2-10x+25}
\end{aligned}$$

2.1.4 Higher-Ordered Derivatives

Let us have $y = f(x)$. As we know, $dy/dx = f'(x)$. We can take the derivative of the derivative of f, that is a *second derivative* of f, denoted $d^2y/dx^2 = f''(x)$. Then the derivative of *that* is the third derivative of f, denoted $d^3y/dx^3 = f^{(3)}(x)$. In general, we can get the n^{th} derivative of f denoted $d^ny/dx^n = f^{(n)}(x)$ which can be obtained by repeated differentiation.

Example 2.3. Let us show a general equation describing $f^{(n)}(x)$ where $f(x) = 4e^{-2x}$. First, let us show up to the third derivative of $f(x)$:

$$\begin{aligned} f(x) &= 4e^{-2x} \\ f'(x) &= 4e^{-2x}(-2) \\ f''(x) &= \frac{d}{dx}(4e^{-2x}(-2)) = 4e^{-2x}(-2)(-2) = 4e^{-2x}(-2)^2 \\ f^{(3)}(x) &= \frac{d}{dx}(4e^{-2x}(-2)^2) = 4e^{-2x}(-2)^3 \end{aligned}$$

We should be able to see that for all natural numbers n, $f^{(n)}(x) = (-2)^n f(x)$.

Derivatives (and second derivatives) allow us to classify points of interest of our function. Suppose we have a function $f(x) = x^2 - 4x + 10$, graphed in Figure 2.2 (left). One question we can ask is, "at what value of x do we get 0 as the rate of change?" In other words, "at what value of x do we get a *stationary point*?" The equivalent formulation would be, "at what value of x do we get $f'(x) = 0$?"

To perform this calculation, we get $f'(x) = 2x - 4$, graphed in Figure 2.2 (right). We find the root of this polynomial, which we get $x = 2$. This implies that the rate of change of $f(x)$ when $x = 2$ is 0.

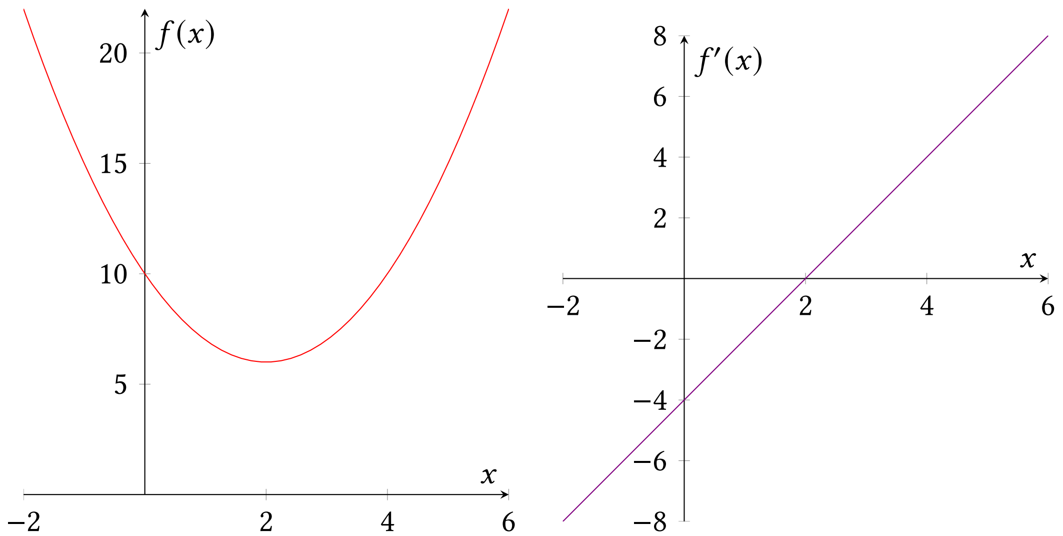


Figure 2.2: Graph of $f(x) = x^2 - 4x + 10$ (left) and its derivative $f'(x) = 2x - 4$ (right).

Let us repeat this process of finding *stationary points* for another function, say $g(x) = -2x^2 + 5x + 7$ graphed in Figure 2.3 (left). We have that $g'(x) = -4x + 5$ graphed in Figure 2.3 (right), and finding the root of $g'(x)$ we get that $g'(5/4) = 0$. Thus, $x = 5/4$ is a stationary point in $g(x)$.

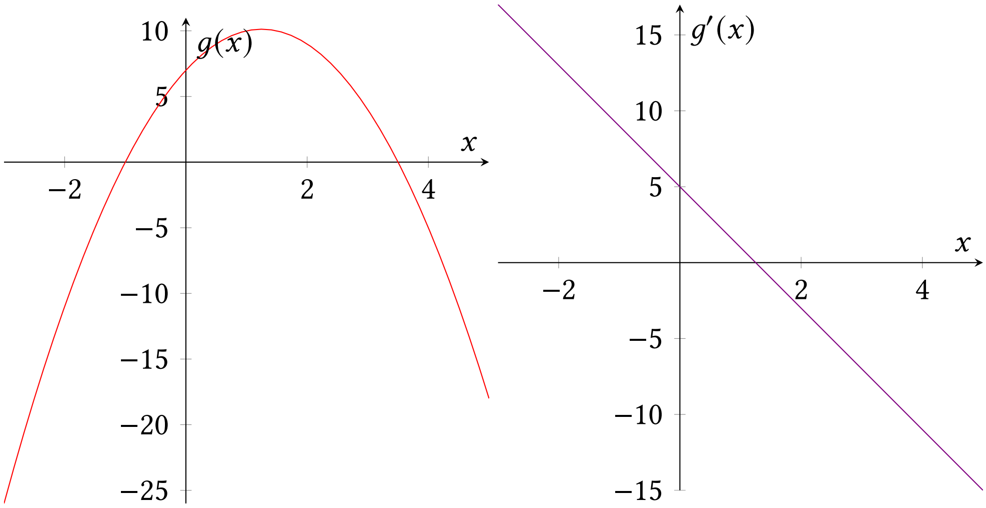


Figure 2.3: Graph of $g(x) = -2x^2 + 5x + 7$ (left) and its derivative $g'(x) = -4x + 5$ (right).

You might notice that the stationary points in $f(x)$ and $g(x)$ are very different; the stationary point in f is a *minimum* point, while that in g is a *maximum* point.

Definition 2.2. Given function f, the *global maximum* of f is x_0 if for all x, $f(x_0) \geq f(x)$. Conversely, the *global minimum* of f is x_0 if for all x, $f(x_0) \leq f(x)$.
x_0 is a *local maximum* of f if there exists $\epsilon > 0$ such that for all $x \in [x_0 - \epsilon, x_0 + \epsilon]$, $f(x_0) \geq f(x)$. Similarly, x_0 is a *local minimum* of f if there exists $\epsilon > 0$ such that for all $x \in [x_0 - \epsilon, x_0 + \epsilon]$, $f(x_0) \leq f(x)$.

Now we might ask, how do we determine if the stationary points of a function are (local) maxima or minima? The answer is to use second derivatives! Let us graph the second derivatives of f and g, i.e. $f''(x) = 2$ and $g''(x) = -4$ in Figure 2.4:

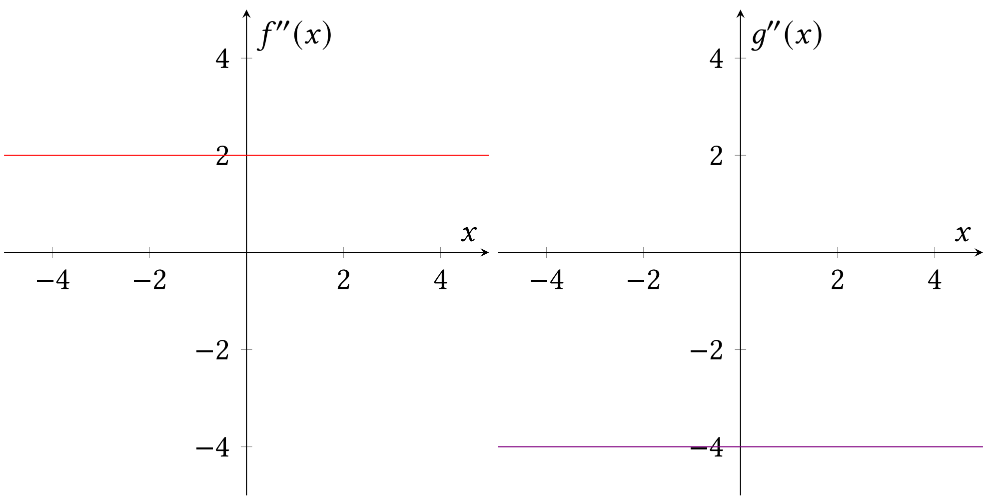
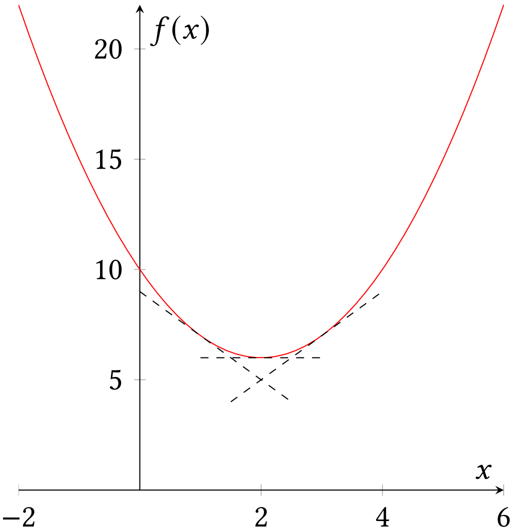


Figure 2.4: Graphs of $f''(x) = 2$ (left) and $g''(x) = -4$ (right).

Observe that $f''(2) = 2$ and $g''(5/4) = -4$. That means that at their respective stationary points, the rate of change of the gradient of f at $x = 2$ is increasing, while that of g at $x = 5/4$ is decreasing. That means that slightly to the left of $x = 2$, the gradient is negative (it must be the case since $f'(2) = 0$) meaning $f(x)$ was sloping downwards there, and slightly to the right of $x = 2$, the gradient is positive so $f(x)$ was sloping upwards, making $x = 2$ a local minimum of f! Similarly, since $g''(5/4)$ is negative, it must mean that slightly to the left of $x = 5/4$, the gradient is positive so g was sloping upwards, and to the right of $x = 5/4$ the gradient is negative and g was sloping upwards, making $x = 5/4$ a local maximum of g! This can be seen by the dashed gradient lines in Figure 2.5.



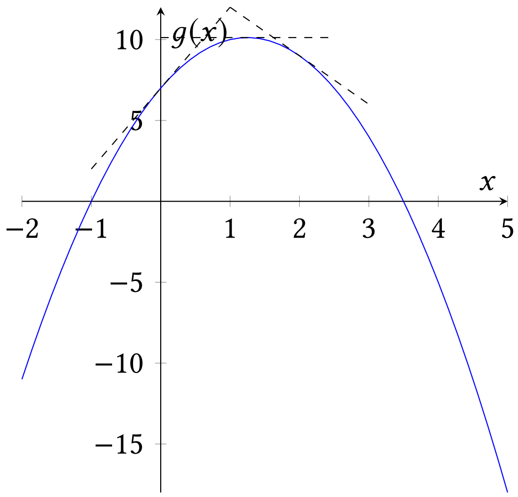


Figure 2.5: Graphs of $f(x) = x^2 - 4x + 10$ (left) and $g(x) = -2x^2 + 5x + 7$ (right) with their gradient lines surrounding stationary points.

As such, we have that: if x_0 is a stationary point at f and $f''(x_0) > 0$, then x_0 is a local minimum, and if x_0 is a stationary point at f and $f''(x_0) < 0$, then x_0 is a local maximum! Note that the inverse is not necessarily true; x_0 being a local minimum does not imply that $f''(x_0) > 0$, and vice-versa. Additionally, not all stationary points are local maxima or minima. For example, the function $h(x) = x^3 + 5$ has derivative $h'(x) = 3x^2$, clearly a stationary point exists at $x = 0$. This stationary point is neither locally (or even globally) minimal or maximal, as seen by the graph in Figure 2.6.

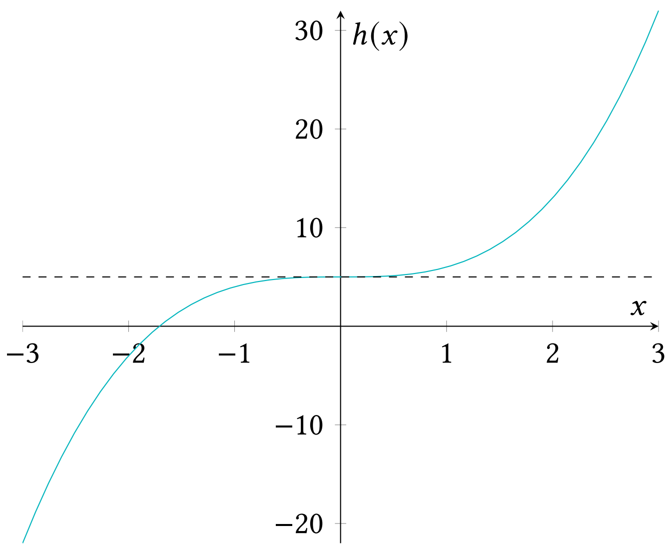


Figure 2.6: Graph of $h(x) = x^3 + 5$.

2.1.5 Partial Derivatives

Throughout our discussion on differentiation, we have only looked at unary functions. What about binary functions, or n-nary functions in general? For this, we have something known as *partial derivatives*, which loosely, describes the rate of change of the function based on one variable, while fixing the others. This is sometimes called the derivative of f in that variable's direction.

Definition 2.3. Suppose we have $f(x_0, x_1, \ldots, x_n)$. The partial derivative of f in the direction of x_i at the point $(a_0, \ldots, a_n)$ is given by

$$\frac{\partial f}{\partial x_i}(a_0, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_0, \ldots, a_i + h, \ldots, a_n) - f(a_0, \ldots, a_i, \ldots, a_n)}{h}$$

Simply put, to calculate the partial derivative of a function in the direction of some variable, fix the other variables as constants and perform differentiation as per usual.

Example 2.4. Suppose we have $f(x, y) = x^2 + 3xy + 4y^2$. Performing partial differentiation in the direction of x we get (fixing y as a constant):

$$\frac{\partial f}{\partial x}(x, y) = 2x + 3y$$

and performing partial differentiation in the direction of y we get (fixing x as a constant):

$$\frac{\partial f}{\partial y}(x, y) = 3x + 8y$$

After perfoming partial differentiation on an n-nary function, we can obtain the gradient of the function at a multi-dimensional variable, denoted ∇f:

Definition 2.4. Suppose we have $f(x_0, x_1, \ldots, x_n)$. Then,

$$\nabla f(a_0, \ldots, a_n) = \left(\frac{\partial}{\partial x_0}(a_0, \ldots, a_n), \ldots, \frac{\partial}{\partial x_n}(a_0, \ldots, a_n) \right)$$

Vector calculus concerns itself very much with these gradients, and has important implications in a variety of computing disciplines, such as analysis, graphics and machine learning. You will learn more about vectors in the future; the main takeaway from this section is that computing gradients of higher-dimensional functions can be done by obtaining partial derivatives on the direction of each variable; doing so is done by standard differentiation on unary functions as we have done.

2.2 INTEGRATION

As described in the beginning of this chapter, calculus attempts to answer two questions relating rates of change and areas under curves. The second branch of calculus, known as *integral* calculus, concerns itself with the latter.

Differentiation and *integration* are tightly linked. As with the previous section, we shall attempt to build some intuition behind integration first. Suppose we had a function $f(x)$ and we want to calculate the area under f between $x = 0$ and $x = a$ for some positive real value a. Notationally, this quantity is written as

$$\int_0^a f(x)dx$$

Let this quantity be expressed as a function $F(x)$: area under f from 0 to x. We can approximate the area by drawing a giant rectangle of width a and height $f(a)$, whose area is $a \times f(a)$.

To refine our approximation, instead of only drawing one rectangle, we can draw two rectangles! Let us split the horizontal line between 0 to a into two equal parts, from 0 to b and b to a where $b = a/2$; each line has length $a/2$. Then, we have two rectangles, one from 0 to b with height $f(b)$, and the other from b to a of height $f(a)$, so the total area of the two rectangles is $a/2 \times f(a/2) + a/2 \times f(a)$. If we let $f(x) = x^2$ and $a = 8$, our description so far can be seen in Figure 2.7.

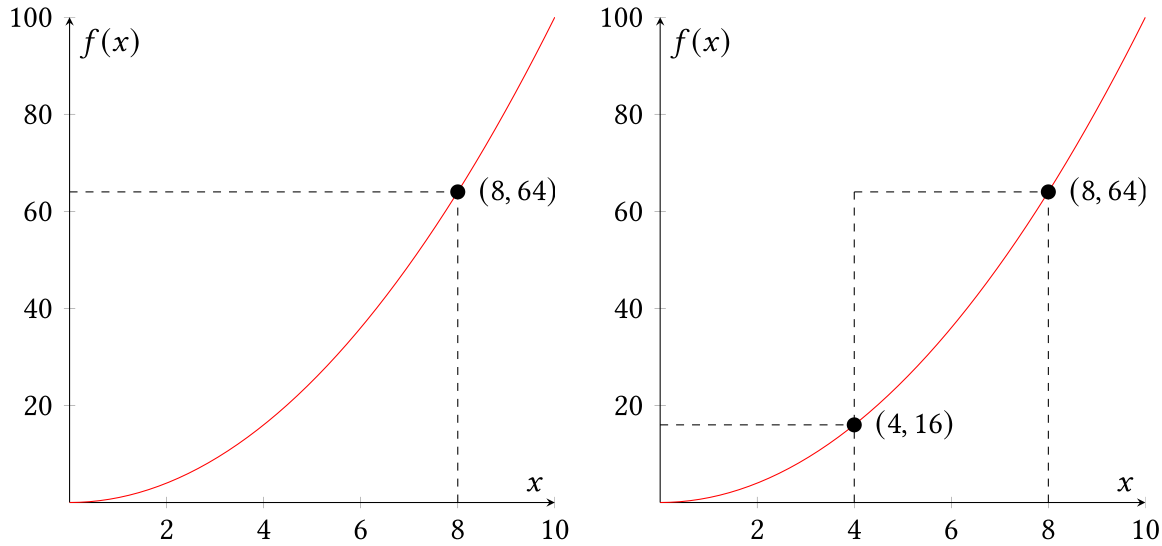


Figure 2.7: Graphs of $f(x) = x^2$ plotted with approximations of $F(x)$.

As you can see, as we decrease the width of each rectangle, the 'excess' amount of our approximation starts to decrease too, therefore, as the width of each rectangle decreases towards 0, our approximation becomes the true value of $F(a)$!

In fact, by the (first) *fundamental theorem of calculus.* the integral (area under a function) and the derivative are closely linked[2]. Let us look at another interpretation of the integral. Suppose we have $F(x)$ that already perfectly computes the area under f from 0 to x. Then, we can estimate the value of $F(x + a) - F(x)$ by $a \times f(x)$. Again, as we decrease the value of a, our approximation becomes more precise, as depicted in Figure 2.8.

We now have the approximation

$$F(x + a) - F(x) \approx a \times f(x)$$

If we include the excess amount as the quantity ϵ, then we get the equality

$$F(x + a) - F(x) = a \times f(x) + \epsilon$$

$$\Leftrightarrow \qquad f(x) = \frac{F(x + a) - F(x)}{a} - \frac{\epsilon}{a}$$

[2]The proof of this theorem is significantly more rigorous than what we are about to present.

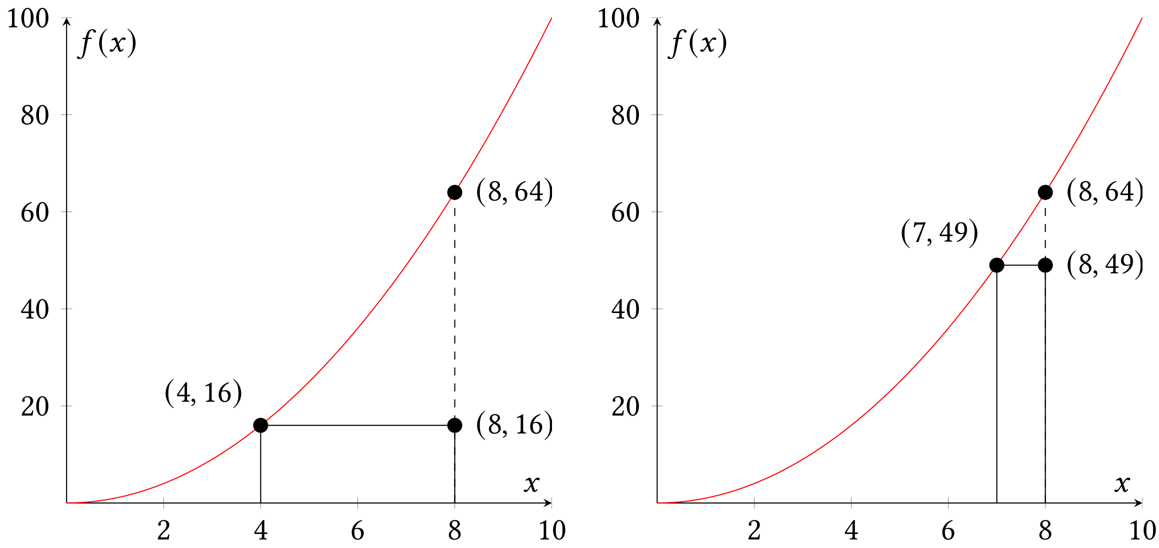


Figure 2.8: Graphs of $f(x) = x^2$ plotted with approximations of $F(x+a) - F(x)$.

Notice that as a tends towards 0, the error term ϵ/a also goes to 0. Thus, taking limits on a, we get

$$\lim_{a\to 0} f(x) = f(x) = \lim_{a\to 0} \frac{F(x+a) - F(x)}{a} = F'(x)$$

We have just recovered the definition of the derivative; essentially, what we have is that differentiating the integral of f gives us f back, so the act of integration and differentiation are inverse operations of each other! This gives us a nice way to obtain a function that gives us the area of another function.

Moreover, the (second) fundamental theorem of calculus states that if we have functions $f(x)$ and an antiderivative $F(x)$ such that $F'(x) = f(x)$, then the area under f between two values a and b can be computed by the difference of $F(b)$ and $F(a)$:

THEOREM 2.8 (Second fundamental theorem of calculus). *If $F'(x) = f(x)$,*

$$\int_a^b f(x)dx = F(b) - F(a)$$

This theorem is great because as long as we have $f(x) = F'(x)$, then $F(x)$ is able to give us the area of $f(x)$ in any region!

2.2.1 How Integration Works

Since we have shown the correspondence between differentiation and integration, we similarly show the inverses of the rules discussed earlier so that we can more easily obtain the area under a function. If we have $F'(x) = f(x)$, then F is known as the *integral* of f. We can write F as an *indefinite integral* over f:

$$F(x) = \int f(x)dx$$

We may also denote *definite integrals* as the area under f in between two x-values, a and b, written earlier:

$$\int_a^b f(x)dx = F(b) - F(a)$$

And clearly,

$$\int_a^b f(x)dx = \int_c^b f(x)dx + \int_a^c f(x)dx = F(b) - F(c) + F(c) - F(a) = F(b) - F(a)$$

We make note of a special observation: if we had a f and its derivative is f', then by the sum and power rule, the derivative of the function $f(x) + c$ for some constant c is also f'! As such, there are infinitely many antiderivatives of any function f (if the integral exists), so in many cases we will see $F(x) = g(x) + c$ for a constant c; just note that this c denotes 'some arbitrary constant' and is not important. In many instances, when calculating the area under a function between a specific interval $[a, b]$, by the second fundamental theorem of calculus we get this constant cancels out, since $F(b) - F(a) = g(b) + c - g(a) - c = g(b) - g(a)$.

2.2.2 Exponentials and Reciprocals

Recall that if we have $f(x) = e^x$ and $g(x) = \ln x$ then we have $f'(x) = e^x$ and $g'(x) = 1/x$. Then it is straightforward to see that

$$F(x) = \int e^x dx = e^x + c \qquad G(x) = \int \frac{1}{x} dx = \ln x + c$$

Furthermore, we have by the chain rule,

$$\frac{d}{dx} e^{f(x)} = e^{f(x)} f'(x) \qquad \frac{d}{dx} \ln f(x) = \frac{f'(x)}{f(x)}$$

Thus, in general, we get that

$$\int e^{f(x)} f'(x) dx = e^{f(x)} + c \qquad \int \frac{f'(x)}{f(x)} dx = \ln f(x) + c$$

Example 2.5. Let us find the integral of $f(x) = 2\ln a \cdot x a^{x^2}$. As before, we can express an exponential function with any arbitrary base as one with e as its base, i.e. $a^x = e^{x \ln a}$. As such, we can re-write f as:

$$f(x) = 2\ln a \cdot x e^{x^2 \ln a}$$

Notice that $\frac{d}{dx}\left(x^2 \ln a\right) = 2\ln a \cdot x$ As such, let $g(x) = x^2 \ln a$ so $g'(x) = 2\ln a \cdot x$. We can now re-express f as

$$f(x) = e^{g(x)} g'(x)$$

so the integral of f is now

$$F(x) = \int e^{g(x)} g'(x) dx = e^{g(x)} + c = e^{x^2 \ln a} + c = a^{x^2} + c$$

Example 2.6. Let us now find the integral of

$$f(x) = \frac{6x-9}{3x^2-9x-3}$$

Instead of factoring the fraction or performing long division, we can immediately see that the numerator of the fraction is the derivative of the denominator. Thus, we can immediately show that

$$F(x) = \int \frac{6x-9}{3x^2-9x-3}dx = \ln\left(3x^2-9x-3\right)+c$$

2.2.3 Integration by Substitution

In full generality, the inverse of the chain rule states that if we have $h(x) = g'(f(x)) \cdot f'(x)$, then

$$H(x) = \int g'(f(x)) \cdot f'(x)dx = g(f(x)) + c$$

Notice that if we let $u = f(x)$ then we get that $\frac{du}{dx} = f'(x)$, so $du = f'(x)dx$. Thus,

$$\int g'(f(x)) \cdot f'(x)dx = \int g'(u)du = g(u) + c = g(f(x)) + c$$

Finding another variable u that captures $f(x)$ so that the integrand can be re-written is known as *integration by substitution*. This technique helps us integrate more complex functions more easily. In fact, *Example* 2.5 and *Example* 2.6 make use of *integration by substitution* as well!

2.2.4 Polynomials

Products with Constants

One of the corollaries to the product rule states that if we have $y = cf(x)$, then $dy/dx = cf'(x)$. As such, if we have $f(x) = cg(x)$,

$$\int cg(x)dx = c\int g(x)dx = cG(x) + d$$

Sums

By the sum rule, if we have $h(x) = f(x) + g(x)$ then $h'(x) = f'(x) + g'(x)$. Thus,

$$H(x) = F(x) + G(x) + c$$

This rule, together with the rule on products with constants, shows the *linearity* of integration, that is,

$$\int af(x) + bg(x)dx = \int af(x)dx + \int bg(x)dx = a\int f(x)dx + b\int g(x)dx = aF(x) + bG(x) + c$$

Power Rule

From the power rule we have that if $f(x) = x^n$ then $f'(x) = nx^{n-1}$. Thus, if $f(x) = nx^{n-1}$ then

$$F(x) = \int nx^{n-1}dx = x^n + c$$

In combination with our rule on products with constants, we get that

$$\int ax^n dx = \frac{a}{n+1}\int (n+1)x^n dx = \frac{a}{n+1}x^{n+1} + c$$

With these observations, we are able to easily find the integral of polynomials:

$$\begin{aligned}\int c_n x^n + \cdots + c_0 dx &= \int c_n x^n dx + \cdots + \int c_0 dx \\ &= \frac{c_n}{n+1}x^{n+1} + \ldots c_0 x + c\end{aligned}$$

Example 2.7. Let us try to find the integral of $f(x) = -3x^2 + 4x + 9$:

$$\begin{aligned}F(x) &= \int \left(-3x^2 + 4x + 9\right) dx \\ &= \frac{-3}{3}x^3 + \frac{4}{2}x^2 + 9x + c \\ &= -x^3 + 2x^2 + 9x + c\end{aligned}$$

2.2.5 Integration by Parts

Recall from the product rule that if we have $h(x) = f(x)g(x)$, then $h'(x) = f'(x)g(x) + f(x)g'(x)$. As such,

$$\int (f'(x)g(x) + f(x)g'(x))dx = f(x)g(x) + c$$

Rearranging terms, we get that

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$$

or equivalently,

$$\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx$$

In other words, if we have two functions $f(x)$ and $g(x)$, then if we are able to find $F(x)$ and $g'(x)$, we are able to compute

$$\int f(x)g(x)dx = F(x)g(x) - \int F(x)g'(x)dx$$

To clearly show this correspondence, let us have $h(x) = F(x)g(x)$ so via the product rule, $h'(x) = f(x)g(x) + F(x)g'(x)$, so $\int (f(x)g(x) + F(x)g'(x))dx = F(x)g(x)$. Rearranging terms as we have done before gives us this correspondence.

This technique is known as *integration by parts*, which allows us to reexpress the integral of the product of two functions into another integral that can be solved more easily based on the rules we already know.

Example 2.8. Suppose we are trying to obtain the integral of $f(x) = xe^x$. Obtaining F directly is nontrivial, as such, we can attempt to break down f into two functions and see if we can find the integral of one function, and the derivative of the other easily. Let us try letting $g(x) = e^x$ and $h(x) = x$. As such, we obtain $G(x) = e^x + c$ and $h'(x) = 1$, so

$$\begin{aligned} F(x) &= G(x)h(x) - \int G(x)h'(x)dx \\ &= (e^x + c)(x) - \int (e^x + c)dx \\ &= xe^x + cx - e^x - cx \\ &= (x-1)e^x \end{aligned}$$

Alternatively, we may have $g(x) = x$ and $h(x) = e^x$. Then we obtain $G(x) = x^2/2 + c$ and $h'(x) = e^x$, so

$$\begin{aligned} F(x) &= G(x)h(x) - \int G(x)h'(x)dx \\ &= \left(\frac{x^2}{2} + c\right)e^x - \int \left(\frac{x^2}{2} + c\right)e^x dx \end{aligned}$$

However, the integral $\int (\frac{x^2}{2} + c)e^x dx$ is harder to obtain than the original integral, so this choice is not ideal f or integration by parts.

2.2.6 Volume by rotating functions around axes

We may decide that we want to obtain the volume of a function rotated around the x-axis. For example, find the volume of the cone obtained by rotating $y = x$ around the x-axis, shown in Figure 2.9.

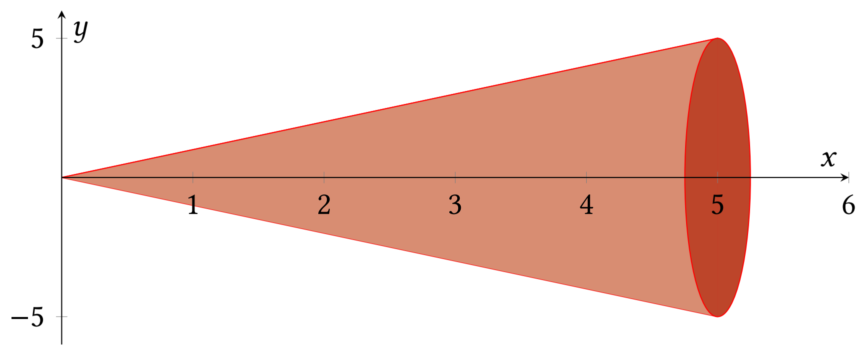


Figure 2.9: $y = x$ rotated around the x-axis.

What you might see is that we are summing over an infinite number of small circles, each with radius x and thus with area[3] πx^2. As such, we can express this volume as the integral

$$\int \pi x^2 dx = \pi \int x^2 dx = \frac{\pi}{3}x^3 + c$$

[3]Recall that the area of a circle with radius R is πR^2.

Geometrically, we get that the volume of the cone is $\pi/3 \times r^2 \times h$, where r is the radius of the cone at its base, and h is its height. In this case, our cone has $r = h = x$, and thus the integral we have obtained corresponds to the geometric volume.

2.2.7 Integrals over more than one variable

Just like in differentiation, we may perform integration on functions with multiple variables, however, *indefinite* integrals does not extend readily into the multi-variate case. However, we can certainly do so for *definite* integrals of multi-variable functions over the real numbers. For example, given a function $f(x, y)$ and ranges for x and y, for example $x \in [a, b]$ and $y \in [c, d]$, we can obtain the double integral of $f(x, y)$ by $\int_c^d \left(\int_a^b f(x, y)dx\right) dy$, the series of two integrals over single variables (keeping the other variable constant), to calculate the *volume* under the surface of $f(x, y)$.

Example 2.9. Suppose we have $f(x, y) = x^2 + 2y$. Then, let us attempt to find the volume under this surface where $x \in [0, 3]$ and $y \in [0, 2]$. To do so, we can compute the double integral as follows (arbitrary constants will be omitted since they will be cancelled out):

$$\int_0^2 \left(\int_0^3 \left(x^2 + 2y\right) dx\right) dy = \int_0^2 G(3) - G(0)dy$$

where $G(x) = x^3/3 + 2yx$. Clearly, $G(3) = 27/3 + 6y = 9 + 6y$, and $G(0) = 0$. Then to continue the integration,

$$\int_0^2 (9 + 6y)dy = H(2) - H(0)$$

where $H(y) = 9y + 3y^2$. Clearly, $H(2) - H(0) = 18 + 12 - 0 = 30$, so the area under the surface is 30.

In fact, by Fubini's theorem, we can swap the order of integration. Let us perform the same calculations in *Example* 2.9 but integrating with respect to y first, i.e. we first get

$$\int_0^3 \left(\int_0^2 (x^2 + 2y)dy\right) dx = \int_0^3 (G(2) - G(0))\, dx$$

where $G(y) = x^2y + y^2$, so we get $G(2) - G(0) = 2x^2 + 4 - 0$. Continuing with the second integration,

$$\int_0^3 \left(2x^2 + 4\right) dx = H(3) - H(0)$$

where $H(x) = 2x^3/3 + 4x$. As such, we get $H(3) - H(0) = 18 + 12 = 30$, which is the same answer.

EXERCISES

2.1. Find the derivative of $f(x) = 4^{(2x+3)}$.

2.2. Find the derivative of $f(x) = 4x^2e^{2x}$.

2.3. Find the derivative of

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

2.4. Find all stationary points of $f(x) = x^2 + 2x + 3$ and determine if these are local minima, maxima, or neither.

2.5. Find the second derivative of $f(x) = x^4$ and determine the value of $f''(0)$. Is $x = 0$ a minimum point? What does this say about the statement "If x is a minimum point of f then $f''(x) > 0$."

2.6. Find the integral of $f(x) = \frac{x^2-3}{x^3-9x}$.

2.7. Find the integral of $f(x) = x^2e^{2x}$

Chapter 3

Geometry

What has been affirmed without proof can also be denied without proof.

Euclid

Perhaps one of the primary motivations of mathematics was, and still is, to study the real world: lines, planes, angles, shapes, curves, have always been of keen interest to mathematicians. The area of mathematics that is concerned with this endeavour is *geometry*, which has applications and a wide variety of domains even in computing, from graphics, even to optimization algorithms. Geometry even shows up in seemingly unrelated areas of mathematics, such as arithmetic.

The most influential *geometer* of all time is *Euclid*, an ancient Greek mathematician who laid out the foundations of geometry. In his treatise, *Elements*, Euclid laid out five postulates for geometry that remain influential for over two millenia—these include relatively uncontroversial statements such as "all right angles are equal". However, Euclid's fifth postulate, known as the *parallel* postulate, states that two non-parallel lines must intersect if extended long enough. While it appears to be superfluous, this axiom was required to proof certain well-known statements in geometry. However, geometers then discovered (much later) that removing this postulate gave rise to an incredibly interesting (consistent) area of geometry in what is now known as non-Euclidean geometry. Ironically, while (Euclidean) geometry was essentially set out to study the real-world, as Albert Einstein showed with general relativity, space is non-Euclidean. However, Euclidean geometry still has incredible amounts of use today.

For a more modern treatment of the subject, *analytic geometry* lays the foundation for geometry as we know it today. Typically, analytic geometry studies Euclidean spaces, such as the 2-dimensional plane (denoted $\mathbb{R}^2$) or the 3-dimensional Euclidean space (denoted $\mathbb{R}^3$), and points, lines, curves, surfaces etc. can all be described as equations or coordinates in this space. All throughout this text, we have seen examples of analytic geometry in action, for example, by stating that the function $f(x) = x^2 - 4$ intersects the x-axis at $(2, 0)$ and $(-2, 0)$—this pair of numbers are known as *cartesian coordinates* of a point in $\mathbb{R}^2$.

In this chapter, we shall revisit some of the fundamental results in geometry that you might find useful for various areas of computing, such as graphics, bioinformatics and numerical analysis.

3.1 TRIGONOMETRY

Perhaps the most important shape in all of geometry is the right-angled triangle, shown in Figure 3.1: This is clearly a triangle, since it has three sides. Three line segments make up

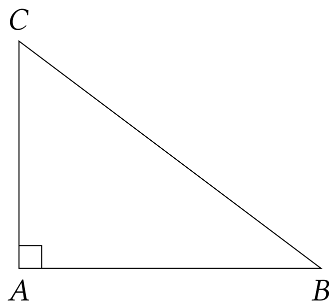


Figure 3.1: A right-angled triangle.

this triangle: AC (the line from point A to C, which we can also denote CA), AB and BC. We denote the length of a line segment using the 'absolute' notation, i.e. $|AB|$ represents the length of the line segment AB. The longest side BC is known as the *hypotenuse* of the triangle. To describe the *angles* at certain points we write them in relation to other points on the triangle, for example, $\angle ACB$ is the angle made at C between the two lines AC and CB. Here, $\angle CAB = 90°$ is a right angle. Recall that a full revolution (the angle made by going around once) is $360°$, so $\angle CAB$ is the angle of a quarter-turn. There is another unit of measure for angles, known as *radians*, where $360° = 2\pi$ rad, which will be useful for later[1]. As such, the right angle can also be defined to be $\pi/2$ rad. Additionally it might be occasionally useful to state the unit of measurement of length, e.g. centimetres, inches etc. In our case, we shall omit these units since they are generally interchangeable; if $|AB|$ being 5 centimetres shows that $|BC|$ is 10 centimetres, then if $|AB|$ is 5 inches, $|BC|$ will also be 10 inches.

Perhaps the most important result in all of geometry (certainly the most well-known ones) is Pythagoras' Theorem.

THEOREM 3.1 (Pythagoras' Theorem). *Let a, b and c be the lengths of the three sides of a right-angled triangle, where c is the length of its hypotenuse. Then,*

$$a^2 + b^2 = c^2$$

Example 3.1. Suppose we know that AC (of the triangle ABC in Figure 3.1) has length 3 and BC has length 5. By Pythagoras' Theorem, $|AB| = \sqrt{5^2 - 3^2} = \sqrt{16} = 4$.

3.1.1 Similarity

Let us take our original triangle from Figure 3.1 and superimpose it with the same triangle but scaled down in half, shown in Figure 3.2. We now have two triangles, ABC and $AB'C'$. You might find that even though the lengths of the two triangles are different, (1) the angles made by the triangles are identical, and (2) coinciding lengths have the same ratio, i.e. $|AC'| \div |AC| = |B'C'| \div |BC| = |AB'| \div |AB|$.

Several insights can be made from this fact. Firstly, this justifies our earlier rationale for omitting units of length. Secondly, it tells us that it is perhaps not as important to know the

[1]Some people prefer to use $360° = \tau$ rad where $\tau = 2\pi$. We shall use π since it is much more widely used.

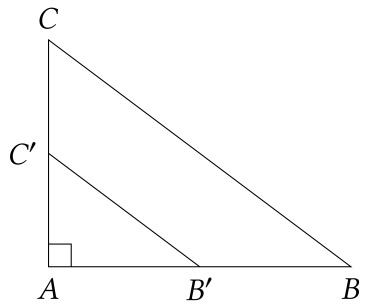


Figure 3.2: Two similar right-angled triangles.

exact lengths of a triangle, and instead, the relationships between the lengths and the angles are much more important.

As such, these are what is known as *similar* triangles. In fact, this concept of *similarity* can be extended to all shapes.

Definition 3.1 (Similarity)**.** Two shapes are *similar* if one shape can be obtained from the other by uniform scaling (shrinking or enlarging), rotation, translation or reflection.

Geometry consist of all sorts of shapes, like circles, ellipses, and so on. Currently we are studying triangles which is a *polygon*[2]. Polygons are shapes obtained by connecting line segments to form a closed polygonal chain. Polygons can be formed by an arbitrary number of line segments, no less than three (which is the triangle). Squares, rectangles, pentagons (regular shape with five sides) are all polygons.

In the case of polygons, two polygons are similar if they have the same number of sides, their corresponding angles are equal, and their corresponding sides have the same ratio. Clearly, our two triangles ABC and $AB'C'$ have the same number of sides, have equal corresponding angles, and their corresponding sides have the same ratio, which makes them similar.

In fact, for the case of triangles, the AAA Theorem states that two triangles are similar if each of the three angles they make are the same. And even stronger claim is that as long as *two* of the angles they make are the same, they are also similar. This is because the sum of the angles made by *any* triangle is always $180° = \pi$ rad, so knowing that two corresponding angles are equal imply that the third corresponding angle must also be equal.

Proposition 3.2. *The sum of all (internal) angles made by an n-sided polygon is equal to* $(n-2)\pi$ *rad* $= (n-2) \times 180°$.

Let us attempt to show why this is the case. Pretend you were a turtle and you were trying to draw the triangle ABC in an anti-clockwise fashion, shown in Figure 3.3. Starting at A, you would move towards B, and at B, you would turn left by some amount (indicated in peach color). Then you would walk to C, turning left, and then back to A, and once you have turned left again (facing B), you should realize that you have rotated precisely $360°$. This tells us that the total *external angle* made by any polygon is $360° = 2\pi$ rad. This is the sum of the angles in peach color in Figure 3.3.

We also know that a line segment essentially forms an angle of $180° = \pi$ rad, so it means at each point, the sum of external angle (in peach) and *internal angle* (in teal) is always $180°$. In this triangle, we should have three of such sums, and in an n-sided polygon, we would have

[2]We limit our discussion of polygons to *simple polygons*, which are polygons whose edges do not intersect.

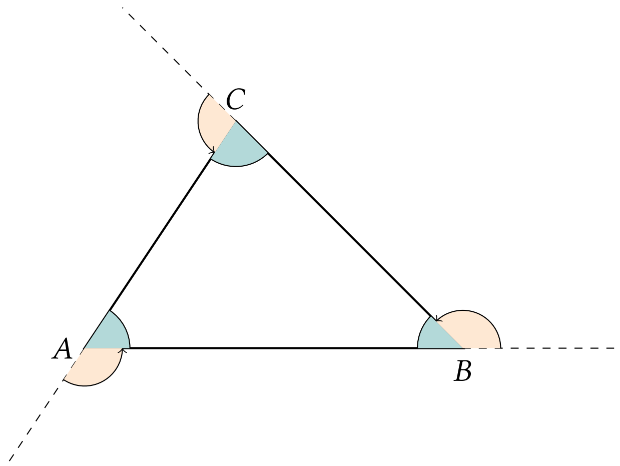


Figure 3.3: Internal and external angles of a triangle.

n of such sums. Thus, the sum of all the teal and peach coloured angles in Figure 3.3 would be $180 \times 3 = 540°$, and in general, for an n-sided polygon such a sum would be $180n°$. Since the sum of the external angles is always $360°$, we get that the internal angles always sum to $180n - 360 = 180n - 2(180) = 180(n-2)°$!

Example 3.2. The sum of the internal angles made by a rectangle is $180(4-2) = 360°$, and the sum of the internal angles made by a polygon is $180(5-2) = 540°$.

Proposition 3.2 tells us that if we know two of the angles of a triangle, the other angle is also known, i.e. if two angles of a triangle are a and b, and the third angle is $(180-a-b)°$. Thus, knowing just two of the corresponding angles of two triangles are equal, immediately tells us that all three corresponding angles are equal, implying that the two triangles are similar.

3.1.2 sin, cos and tan

By now we should be convinced that when two triangles have the same corresponding angles, they are similar, so their exact dimensions are not so important. Instead, the relationship between the sides of a triangle and the angles they make might be. One natural question to ask is, suppose we have a right-angled triangle (thereby fixing one angle to be $\pi/2$ rad), and one of the other angles of the triangle is α (therefore the other angle is $\beta = (\pi/2 - \alpha)$ rad). What is the relationship between α and the sides of the triangle?

Let us redraw our right angled triangle from Figure 3.1 in Figure 3.4, this time denoting the angle made by $\angle ABC$ as α. Then, we can describe the relationship between the sides of the triangle and α by three functions, sin (sine), cos (cosine) and tan (tangent), which is defined as

$$\sin\alpha = \frac{|AC|}{|BC|} \qquad \cos\alpha = \frac{|AB|}{|BC|} \qquad \tan\alpha = \frac{|AC|}{|AB|}$$

Defining these functions in this way is likely somewhat confusing. As such, let us denote the length of the hypotenuse BC as H, the length of the opposite edge of α AC as O, and the length of the adjacent edge of α AB as A. Then, we can more succinctly describe sin, cos and tan, by

$$\sin\alpha = \frac{O}{H} \qquad \cos\alpha = \frac{A}{H} \qquad \tan\alpha = \frac{\sin\alpha}{\cos\alpha} = \frac{O}{A}$$

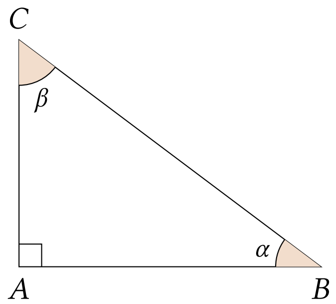


Figure 3.4: A right-angled triangle where $\angle ABC$ is labelled α.

As a reminder, the exact lengths of the edges of the triangle is unimportant, since all right-angled triangles that have one other angle equal to α is similar to the triangle that we have drawn. Additionally, we are assuming that all three edges have nonzero lengths.

We list some common angles θ and their corresponding outputs from the three trigonometric functions in Table 3.1.

θ	0° (0 rad)	30° $\left(\frac{\pi}{6}\text{ rad}\right)$	45° $\left(\frac{\pi}{4}\text{ rad}\right)$	60° $\left(\frac{\pi}{3}\text{ rad}\right)$	90° $\left(\frac{\pi}{2}\text{ rad}\right)$
$\sin\theta$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
$\cos\theta$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
$\tan\theta$	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	undefined

Table 3.1: Special values for sin, cos and tan.

Notice that if we defined $\beta = (\pi/2 - \alpha)$ rad as the angle $\angle ACB$ as shown in Figure 3.4, we immediately get the following correspondences:

$$\sin\beta = \sin\left(\frac{\pi}{2} - \alpha\right) = \frac{|AB|}{|BC|} = \cos\alpha \qquad \cos\beta = \cos\left(\frac{\pi}{2} - \alpha\right) = \frac{|AC|}{|BC|} = \sin\alpha$$

These trigonometric functions also have inverses, where arctan, arccos and arcsin are the inverses of tan, cos and sin respectively[3], i.e. we have $\arctan(\tan\alpha) = \alpha$, and so on. For example, $\arcsin(1/2) = \pi/6$ rad since $\sin(\pi/6) = 1/2$.

These functions relate all three edges of the triangle to α. Of course, we may also obtain the reciprocals of our functions giving us the 'inverse' relationships between the edges based on the angle α, which gives us three more functions, sec (secant), csc (cosecant) and cot (cotangent), defined in the obvious way:

$$\sec\alpha = \frac{1}{\cos\alpha} = \frac{H}{A} \qquad \csc\alpha = \frac{1}{\sin\alpha} = \frac{H}{O} \qquad \cot\alpha = \frac{1}{\tan\alpha} = \frac{A}{O}$$

[3]Sometimes these functions are denoted as $\sin^{-1}\alpha$. While this notation is very intuitive for inverses, unfortunately, the convention is that for all other n, $\sin^n\alpha = (\sin\alpha)^n$, which makes things incredibly confusing, for example, $\sin^2\alpha = (\sin\alpha)^2$ and not $\sin(\sin\alpha)$.

Let us show the graphs of $\sin x$, $\cos x$ and $\tan x$ and observe some of their properties in Figure 3.5, Figure 3.6 and Figure 3.7. For the sine and cosine waves, these have a *period* (horizontal length of the wave until it repeats) of 2π (as we expect, since x rad $= (2\pi + x)$ rad) and an *amplitude* (distance from the highest/lowest points of the wave and center of the wave) of 1. The tangent curve also repeats with period π (i.e. $\tan x = \tan(\pi + x)$) and has *asymptotes* at $\pi/2$ and $3\pi/2$, this is because at these points, $\cos x = 0$, so as x gets closer to these values, $\tan x$ gets larger and larger (or more negative, depending on which direction).

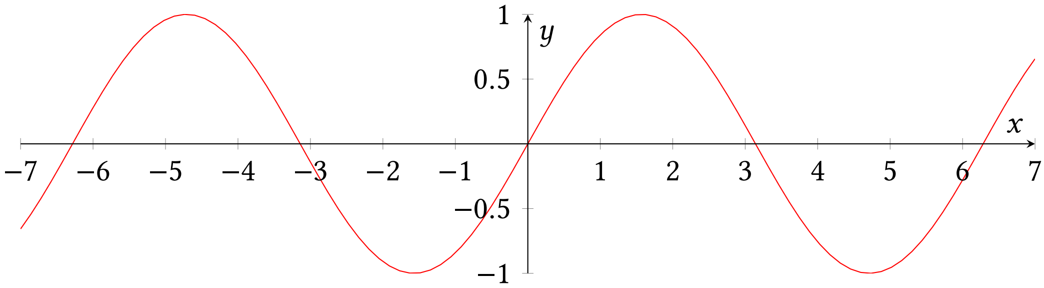


Figure 3.5: Graph of $y = \sin(x)$.

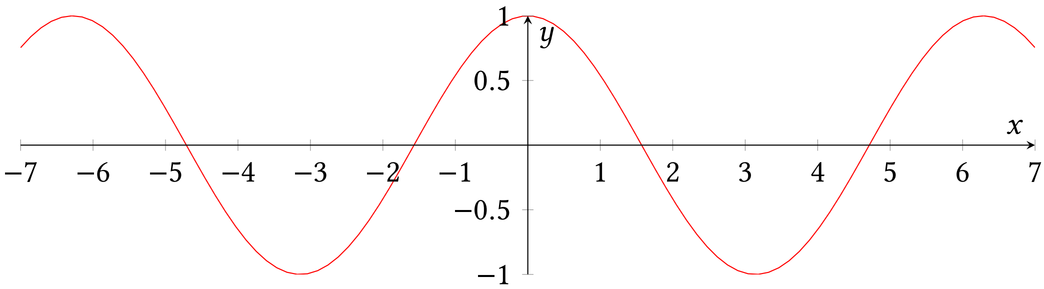


Figure 3.6: Graph of $y = \cos(x)$.

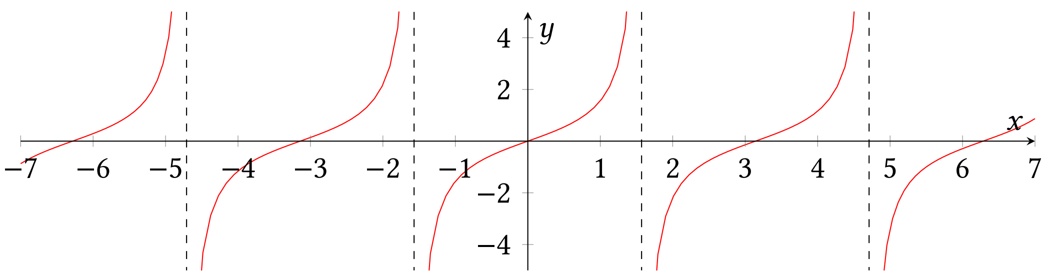


Figure 3.7: Graph of $y = \tan(x)$.

3.1.3 Trigonometric and Triangle Identities

The three trigonometric functions described earlier are among the most important in all of mathematics. For now, we shall note some of their significance in elementary geometry. For the following identities we shall make reference to the triangle in Figure 3.8, with the angles at points A, B and C, and their corresponding lengths of opposite edges a, b and c respectively.

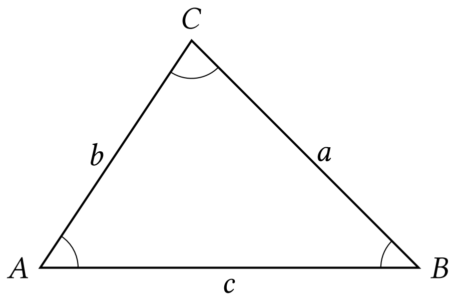


Figure 3.8: An arbitrary triangle with points with angles A, B and C whose corresponding opposite edges have lengths a, b and c respectively.

Sine Rule

Given any arbitrary triangle,

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}$$

The rule tells us that just by knowing the lengths of two sides and the opposite angle of one of the known sides, we can obtain the dimensions of the triangle.

Example 3.3. Suppose we have a triangle ABC where $A = \pi/3$ rad, and $a = 4$ and $b = 3$. Our goal now is to find c and angles B and C. From the sine rule, we get that

$$\frac{a}{\sin A} = \frac{4}{\sin \frac{\pi}{3}} = \frac{8}{\sqrt{3}}$$

As such, we get that

$$\frac{b}{\sin B} = \frac{c}{\sin C} = \frac{8}{\sqrt{3}}$$

We can use this fact, together with the fact that $A + B + C = \pi$ to solve for c, B and C:

$$\begin{aligned}
\frac{3}{\sin B} &= \frac{8}{\sqrt{3}} \\
\sin B &= \frac{3\sqrt{3}}{8} \\
B &= \arcsin \frac{3\sqrt{3}}{8} \\
C &= \frac{2\pi}{3} - \arcsin \frac{3\sqrt{3}}{8} \\
\frac{c}{\sin\left(\frac{2\pi}{3} - \arcsin \frac{3\sqrt{3}}{8}\right)} &= \frac{8}{\sqrt{3}} \\
c &= \frac{8 \sin\left(\frac{2\pi}{3} - \arcsin \frac{3\sqrt{3}}{8}\right)}{\sqrt{3}}
\end{aligned}$$

Cosine Rule

As an extension to Pythagoras' Theorem, given any arbitrary triangle,

$$c^2 = a^2 + b^2 - 2ab\cos C$$

Area of triangles

We should be very comfortable with finding the area of any arbitrary triangle; given a triangle of width a and height h, the area of the triangle is $\frac{1}{2}ah$. However, let us draw our triangle showing the relationship between the height of the triangle and the adjacent edge in Figure 3.9:

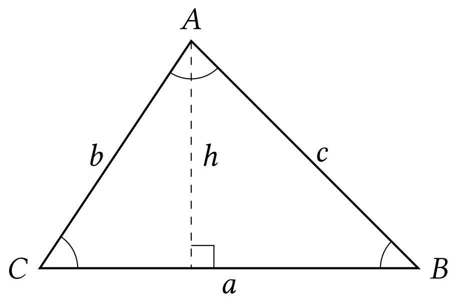


Figure 3.9: Finding the area of a triangle.

We know from this diagram, by definition, we have $\sin C = h/b$ and $\sin B = h/c$, or equivalently, $h = b\sin C = c\sin B$. As such, substituting this equivalence into our original equation, we get that the area of any arbitrary triangle is also given by

$$\frac{1}{2}ab\sin C$$

3.2 ANALYTIC GEOMETRY

Now that we have reviewed the basics of trigonometry and geometry, we can move on to discussing geometry in and analytic context, such that points, lines and shapes can all be described as equations in a space. For this section we will focus primarily on $\mathbb{R}^2$, that is formed by two axes, the x-axis, and the y-axis, both of them being perpendicular to each other.

A point on this plane is composed of two parts forming its coordinates: the x part, and the y part. This is written as the cartesian coordinates (a, b), where the x part is a and the y part is b. This is the same as representing a point as the intersection of two equations $x = a$ and $y = b$. The point $(0, 0)$ is frequently denoted O, the origin.

For example, let us plot three points of a triangle, $A = (2, 3)$, $B = (-1, 4)$ and $C = (-\frac{1}{2}, 1)$, and three line segments joining these points in Figure 3.10, and observe some properties of this triangle from an analytic viewpoint.

Let us now observe some properties of this triangle from an analytic viewpoint.

3.2.1 Line Segments

As seen in Chapter 1, a line can be drawn by the linear equation $y = mx + c$, where m is the *slope* or *gradient* of the line, and c is the value of y when $x = 0$, i.e. the y-value when the line

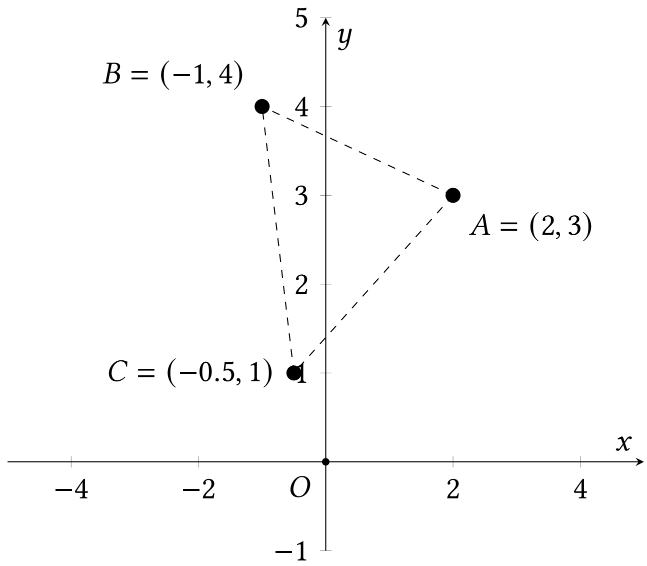


Figure 3.10: A triangle *ABC*.

intersects the y-axis. There are two questions we might ask. First, given two points, what is the equation that describes the line? And secondly, what is the length of the line segment and angle that it makes?

Equation of line segment

Let us answer the first question. To obtain the equation describing the line segment CA, we construct a system of equations that represent the requirements of this line, so that once we have solved it, we would have the parameters m and c.

In general, finding m is incredibly straightforward. Suppose we have two points (a, b) and (a', b') and we wanted to find the gradient of the line segment that intersects these two points. Let us write a system of two equations that describe this fact:

$$\begin{aligned} b &= am + c \\ b' &= a'm + c \end{aligned}$$

From the first equation we can see that $c = b - am$, which we shall substitute into the second equation.

$$\begin{aligned} && b' &= a'm + b - am \\ \Leftrightarrow && b' - b &= a'm - am \\ \Leftrightarrow && b' - b &= m(a' - a) \\ \Leftrightarrow && m &= \frac{b' - b}{a' - a} \end{aligned}$$

This should be completely unsurprising since as shown in Chapter 2 the gradient of a straight line is constant and can be found just by dividing the difference in y values with the difference in x values.

Proposition 3.3. *Given two points (a, b) and (a', b'), the gradient m of the line segment* *intersecting these two points is given by*

$$m = \frac{b' - b}{a' - a}$$

Once we have found m, finding c should be incredibly straightforward by substituting the newly found m and one (x, y) pair from one of the points into the equation of a line.

Remark 3.1. From the equation that obtains the value of m, we can tell that m would be undefined if $a' - a = 0$ i.e. $a = a'$. However, this is not an issue because such a line has equal x-values, and therefore is simply a vertical line described by the equation $x = a$.

Example 3.4. Let us find the equation representing the line AC in Figure 3.10. The line intersects points A and C, so the gradient of this line is

$$\frac{3 - 1}{2 - (-\frac{1}{2})} = \frac{4}{5}$$

Then, we know that this line has equation $y = mx + c$ where $m = 4/5$. Thus, substituting $(x, y) = (2, 3)$ we get

$$3 = \frac{8}{5} + c$$

so we get $c = 1.4$. Thus, the line segment AC is described by the equation

$$y = \frac{4}{5}x + 1\frac{2}{5}$$

Length and angle of line segment

Let us isolate the line segment AC from Figure 3.10 and show it in Figure 3.11, and attempt to find its length. Specifically, we have constructed a right-angled triangle with w and h being the length of its sides, and $|AC|$ is the length of the hypotenuse.

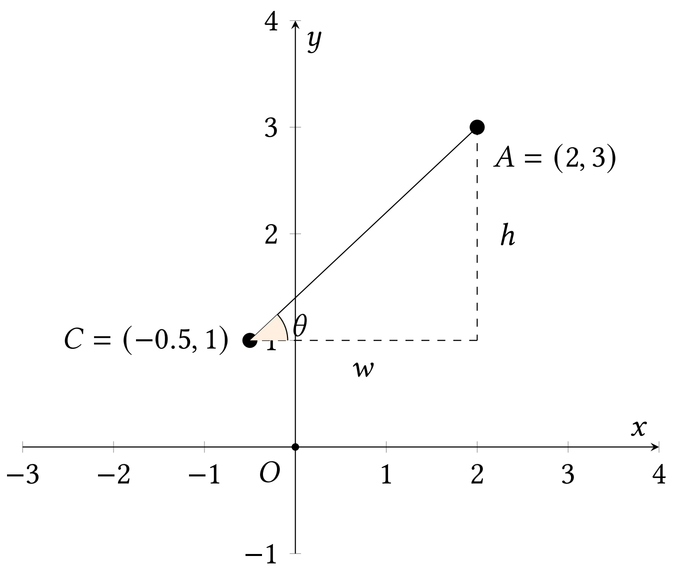


Figure 3.11: The line segment AC.

You might notice that w is the difference in x values of the two points, and h is the difference in y values of the two points. In addition, by Pythagoras' Theorem[4], we see that $w^2+h^2 = |AC|^2$! Thus, for any two points (a, b) and (a', b'), the length of the line segment intersecting these two points is simply $\sqrt{(a' - a)^2 + (b' - b)^2}$.

Proposition 3.4. *Given two points (a, b) and (a', b'), the line segment connecting these two points has length*

$$\sqrt{(a' - a)^2 + (b' - b)^2}$$

This is sometimes called the Euclidean distance *or* Pythagorean distance *between points (a, b) and (a', b').*

Example 3.5.

$$|AC| = \sqrt{(2 + 0.5)^2 + (3 - 1)^2} \approx 3.2$$

Next, we might ask for the angle that this line makes at one of the points. Of course, a line segment on its own does not make any angles, we must specify *another* line segment, and ask of the angle made at the intersection of the lines. The convention is that the angle that a line makes is the angle *with respect to* the x-axis, meaning, we draw an imaginary horizontal line that intersects the (leftmost) point on the line segment, and ask for the angle that this makes. By convention, lines that go towards the top-right or top-left have angles between 0 and π rad, and lines that go towards the bottom-right or bottom-left have angles between 0 and $-\pi$ radians[5].

Since we have formed a right-angled triangle earlier in Figure 3.11, we should clearly see that the angle made by AC at the point C, denoted θ, can be precisely described by θ where $\tan\theta = h/w$; thus in general, if we have two points (a, b) and (a', b'), the angle made by the line segment intersecting these two points is described by $\arctan\left((b' - b)/(a' - a)\right)$. In other words, if the gradient of the line segment is m, then the angle made by the line segment can be described by $\arctan m$!

Proposition 3.5. *Given a line AB with gradient m, the angle that AB makes with respect to the x-axis is* $\arctan m$.

Remark 3.2. Similar to *Remark* 3.1, when $a' - a = 0$ i.e. $a = a'$ then this equation is undefined. However this case is easily avoidable by observing that such a line segment would be completely vertical and thus have angle $\pi/2$ or $-\pi/2$.

Example 3.6. We know that the gradient of the line segment AC in Figure 3.11 is $4/5$, thus, the angle made by this line, denoted θ, is

$$\arctan\frac{4}{5} \approx 38.7^\circ$$

The natural question to ask would then be, what about the angle made by line AC with respect to the horizontal line interesecting point A? In other words, this is the angle made by

[4] You might notice that the order of points does not matter, i.e. $\sqrt{(a' - a)^2 + (b' - b)^2} = \sqrt{(a - a')^2 + (b - b')^2}$. This is because $(a - a')^2 = (-1 \times (a' - a))^2 = (-1)^2(a' - a)^2 = (a' - a)^2$.

[5] A horizontal line going towards the left make an angle π and $-\pi$ radians. It should be obvious that these two are the same. As such, by convention, such lines will make an angle of π.

the line CA but going in the other direction, thus the angle made would simply be $(\theta+\pi)$ rad. As a reminder, conventionally angles are in the range of $(-\pi, \pi]$, but correcting our result for this convention is straightforward. For example, we know from *Example* 3.6 that $\theta \approx 0.215\pi$ rad, thus the angle going the other way would be roughly 1.215π rad. Since this exceeds the conventional range, we just subtract 2π from it, giving us -0.785π rad.

However, having to convert angles around is rather cumbersome. As such, let us define a new function[6] $\text{arctan2}(y, x)$ that, given a line going from (a, b) to (a', b') (the direction matters this time), finds the angle made by the line segment with respect to the x-axis at the point (a, b) with $\text{arctan2}(b' - b, a' - a)$:

$$\text{arctan2}(y, x) = \begin{cases} \frac{\pi}{2} & \text{if } y > 0 \text{ and } x = 0 \text{ (vertical upwards)} \\ -\frac{\pi}{2} & \text{if } y < 0 \text{ and } x = 0 \text{ (vertical downwards)} \\ \pi & \text{if } y = 0 \text{ and } x < 0 \text{ horizontal left} \\ \arctan \frac{y}{x} & \text{if } y, x > 0 \text{ up-right} \\ \arctan \frac{y}{x} & \text{if } x > 0 \text{ and } y \leq 0 \text{ down-right} \\ \pi + \arctan \frac{y}{x} & \text{if } x < 0 \text{ and } y > 0 \text{ up-left} \\ \arctan \frac{y}{x} - \pi & \text{if } y, x < 0 \text{ down-left} \end{cases}$$

Thus for convenience sake, we will represent the gradient of the line CA (the line going from C to A) as $\frac{4}{5}$, while the gradient of the line AC (the line going in the opposite direction, from A to C) as having gradient $\frac{-4}{-5}$. This is so that the angle made by CA at point C is given by $\text{arctan2}(4, 5) \approx 0.215\pi$ rad, while the angle made by AC at point A is given by $\text{arctan2}(-4, -5) \approx -0.785\pi$ rad.

Remark 3.3. What we have presented is a completely informal treatment of lines with direction, which is much more rigorously formalized with *vectors*, which you will soon encounter (not in this text). Hopefully, our characterization will give you some intuition on the subject.

3.2.2 Polygons

Triangles

Now we can go further by discussing polygons such as triangles. Going back to triangle ABC in Figure 3.10, we might ask: (1) What is the length of each edge of ABC? (2) What are the angles at each point? (3) What is the area of ABC? Answering the first question is straightforward as per Subsection 3.2.1. We have by Pythagoras' Theorem that $|BC| = \sqrt{9.25} \approx 3.04$ and $|BA| = \sqrt{10} \approx 3.16$.

To answer the second question, let us first find the equations describing the line segments BC and BA. For BC, we have $y = m'x + c'$ where m' and c' are the gradient and y-intercept of the line respectively. Then,

$$m' = \frac{1-4}{-0.5+1} = \frac{-6}{1}$$

Substituting $(x, y) = (-1, 4)$ and $m' = -6$ into our equation gives us that $c' = -2$; thus the line segment BC can be described by the equation $y = (-6/1)x - 2$. Doing the same for the line

[6] Programmers might notice that this is the definition of the `atan2` functions provided in math libraries of languages like Java or Python.

BA gives us the equation $y = (-1/3)x + 3\frac{2}{3}$. Let us denote the angle made by XY at point X with respect to the x-axis as $\angle XY$. Then, $\angle BC$ and $\angle BA$ is $\text{arctan2}(-6, 1) \approx -0.447\pi$ rad and $\text{arctan2}(-1, 3) \approx -0.102\pi$ rad respectively. Correspondingly, the angles made by $\angle CB$ and $\angle AB$ is $\text{arctan2}(6, -1) \approx 0.553\pi$ rad and $\text{arctan2}(1, -3) \approx 0.898\pi$ rad respectively.

Having obtained these angles, we shall first investigate $\angle ACB$. By observing Figure 3.10 you might notice that this angle can be obtained by $\angle CB - \angle CA$! In general [7], given points X, Y and Z, $\angle YXZ = \angle XY - \angle XZ$. Thus, $\angle ACB = \angle CB - \angle CA \approx 0.553\pi - 0.215\pi = 0.338\pi$ rad. Similarly, $\angle ABC = \angle BA - \angle BC \approx -0.102\pi - -0.447\pi = 0.345\pi$ rad. Finally, $\angle BAC = \angle AC - \angle AB \approx -0.785\pi - 0.898\pi = -1.683\pi = 0.317\pi$ rad. You should see that the three angles sum to $0.338\pi + 0.345\pi + 0.317\pi = \pi$ rad $= 180°$ as we expect.

With all this information, finding the area of ABC should be straightforward and can be done by a variety of methods, one of which is $ab \sin C/2$ rule; using this, we get that the area of ABC is

$$\frac{1}{2}|AB||BC| \sin(\angle ABC) = \frac{1}{2}\sqrt{10}\sqrt{9.25} \sin(0.345\pi) \approx 4.25$$

Polygons in general

Let us ask of the same for any general simple polygon. For example, we can draw a polygon with five sides as shown in Figure 3.12.

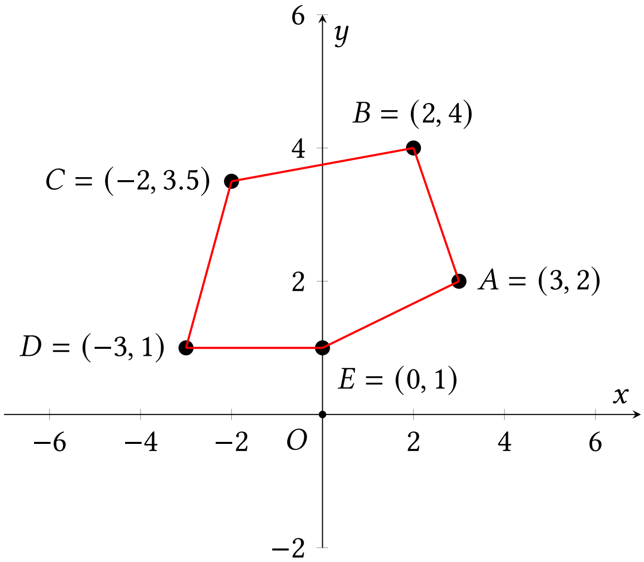


Figure 3.12: A polygon $ABCDE$.

Finding the lengths and equations of the line segments and the angles at each point is no different than before. What is starkly different is, of course, finding the area of such polygons. One strategy that we could employ is to divide the polygon into several triangles, then find the area of each of the individual triangles, and finally take the sum. For example, we can divide $ABCDE$ into three triangles, CDE, CBE and ABE as shown in Figure 3.13.

However, this strategy is rather laborious. Instead, let us observe the following. A *trapezium* is a four-sided polygon where two of the (opposite) sides are parallel. Let us consider a trapezium where two adjacent angles at points intersecting the two parallel lines are right angles, like $WXYZ$ in Figure 3.14, where $W = X = \pi/2$ rad, $|WX| = a$, $|XY| = b$ and $|WZ| = c$.

[7] Depending on which angle you subtract from the other, you may get the larger angle made by the intersection.

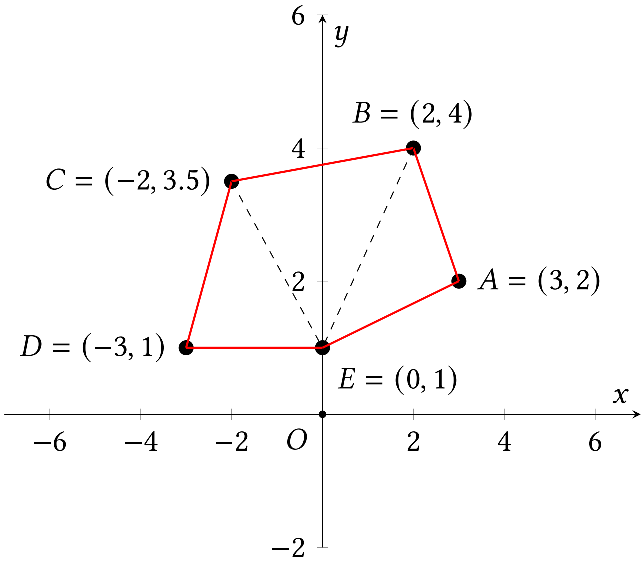


Figure 3.13: A polygon $ABCDE$ divided into three triangles.

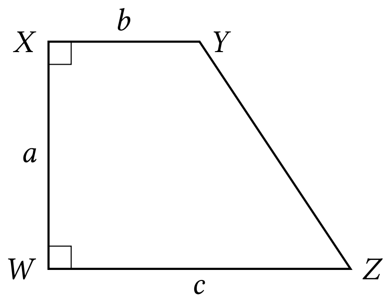


Figure 3.14: A trapezium $WXYZ$ where W and X are right angles.

We could find the area of this trapezium by forming a right-angled triangle by dividing the trapezium from Y to some point on WZ, but that is too cumbersome and requires us to know what the angles at Y and Z are. Instead, observe that if we took the same trapezium, rotated it $180°$ and placed it directly to the right of $WXYZ$, we get a rectangle with height a and width $b + c$, as shown in Figure 3.15. Since the area of this rectangle is $a(b + c)$, it must mean that the area of $WXYZ$ is half of that, which is $a(b + c)/2$!

Proposition 3.6. *Suppose we had a trapezium $WXYZ$ where XY and WZ are parallel, W and X are right angles, and $|WX| = a$, $|XY| = b$ and $|WZ| = c$. Then, the area of $WXYZ$ is*

$$\frac{1}{2}a(b + c)$$

With this observation, let us draw vertical lines from each point in $ABCDE$ down to the x-axis. This makes a trapezium for each side of the polygon, so there are five sides, just like in Figure 3.16. Further suppose that when calculating the area of a trapezium, direction matters. For example, the area of the trapezium going from C to B (left to right) yields a positive value, while the area of the trapezium going from A to E (right to left) yields a negative value. From

Correcting this is easy, if you obtain an angle θ larger than π, then the angle you're looking for is $2\pi - \theta$.

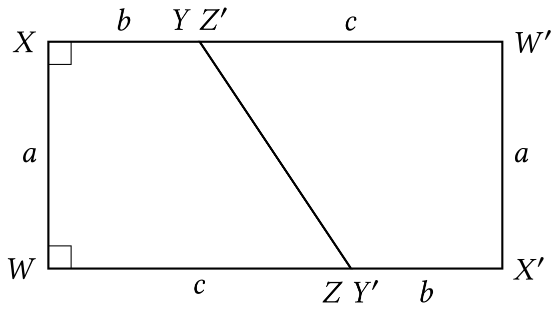


Figure 3.15: Putting two copies of $WXYZ$ together forms a rectangle.

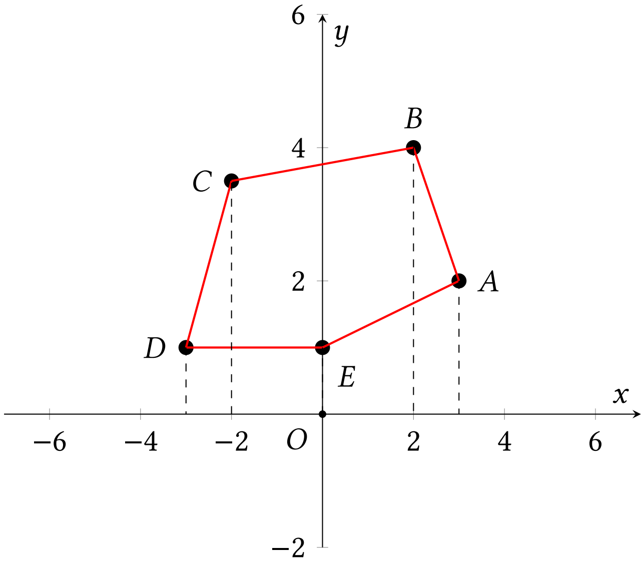


Figure 3.16: Drawing vertical lines from each point of $ABCDE$ yields five trapeziums.

this we can describe the area of the trapezium formed by the line going from (a, b) to (a', b') as

$$\frac{1}{2}(b + b')(a' - a)$$

Now, notice that if we went clockwise along each point (for example, A to E then D then C then B), by summing the area of each trapezium along the way, we get the area of $ABCDE$!

PROPOSITION 3.7. *Suppose we had n points $(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})$ such that the line segments connecting each adjacent pair of points (x_i, y_i) and (x_{i+1}, y_{i+1}) (and (x_{n-1}, y_{n-1}) with (x_0, y_0)) forms a polygon, and the points go clockwise. For simplicity, we let $(x_n, y_n) = (x_0, y_0)$. Then, the area of this polygon is given by*

$$A = \frac{1}{2}(y_1 + y_0)(x_1 - x_0) + \cdots + \frac{1}{2}(y_n + y_{n-1})(x_n - x_{n-1}) = \frac{1}{2}\sum_{i=0}^{n-1}(y_{i+1} + y_i)(x_{i+1} - x_i)$$

This is known as the shoelace formula.

Thus, the shoelace formula shows that the area of $ABCDE$, denoted a, is (starting from A, going clockwise):

$$\begin{aligned} a &= \frac{1}{2}((3)(-3) + (2)(-3) + (4.5)(1) + (7.5)(4) + (6)(1)) \\ &= 12.75 \end{aligned}$$

3.2.3 Circles

We have only looked at polygons thus far, but as we shall see, right-angled triangles and circles come hand-in-hand. To draw a circle centered at the origin with radius r, we would like to draw every point (x, y) such that the distance from (x, y) to the origin is the radius r. You should notice that this is reminiscent of Pythagoras' Theorem, because in fact the equation describing this circle is precisely

$$x^2 + y^2 = r^2$$

For example, we can draw a circle centered at the origin with radius 1 by the formula $x^2+y^2 = 1$, shown in Figure 3.17.

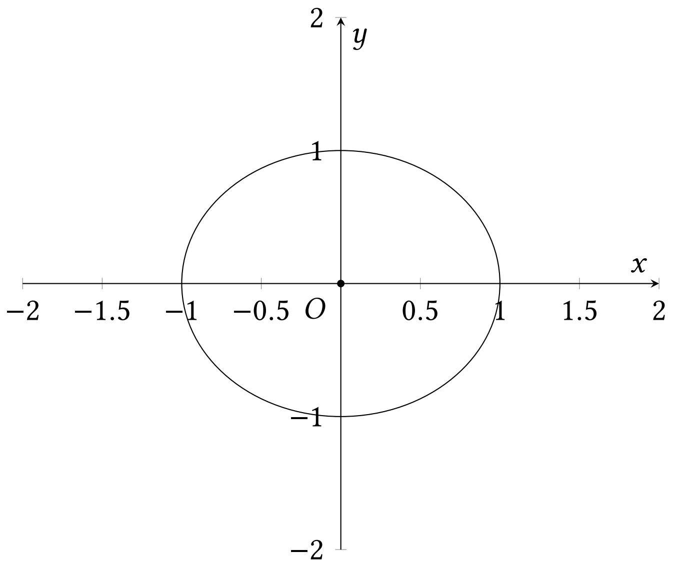


Figure 3.17: Drawing a circle with radius 1 centered at the origin, given by the equation $x^2 + y^2 = 1$.

Some questions may arise from this: (1) How do we draw the same circle centered somewhere else? (2) How do we stretch the circle in the x or y direction? We can extend these questions further. Given *any* graph (in $\mathbb{R}^2$), how do we move, stretch, and even reflect the graph?

Translation

Let us address the first question on moving graphs around, which is known as *translation*. Suppose we have a graph $y = f(x)$, and we wanted to translate the graph upwards by b. Clearly, we can add b to the RHS so that the y-coordinate of each point of the graph moves upwards by b, i.e. $y = f(x) + b$, or $y - b = f(x)$. Basically, by replacing y with $y - b$, our graph moves upwards by b.

Similarly, if we re-express the same graph but as a function of y, i.e. $x = g(y)$, then if we wanted to translate the entire graph to the right by a, we can add a to the RHS so that the x-coordinate of each point of the graph moves to the right by a, i.e. $x = g(y) + a$ or $x - a = g(y)$. Again, by replacing x with $x - a$, our enter graph moves to the right by a.

As such, if we wanted to plot a circle centered at (a, b) with radius r, what we are really doing is to start with a circle centered at $(0, 0)$ given by the equation $x^2 + y^2 = r^2$, then translating this circle rightwards by a, and upwards by b. Thus, based on what we have shown above, replace x with $x - a$ and y with $y - b$, so our new equation becomes $(x-a)^2 + (y-b)^2 = r^2$. For example, if we wanted to plot a circle with radius 2 centered at the point $(1, -2)$, then the equation of the circle is $(x - 1)^2 + (y + 2)^2 = 4$, which we plot in Figure 3.18.

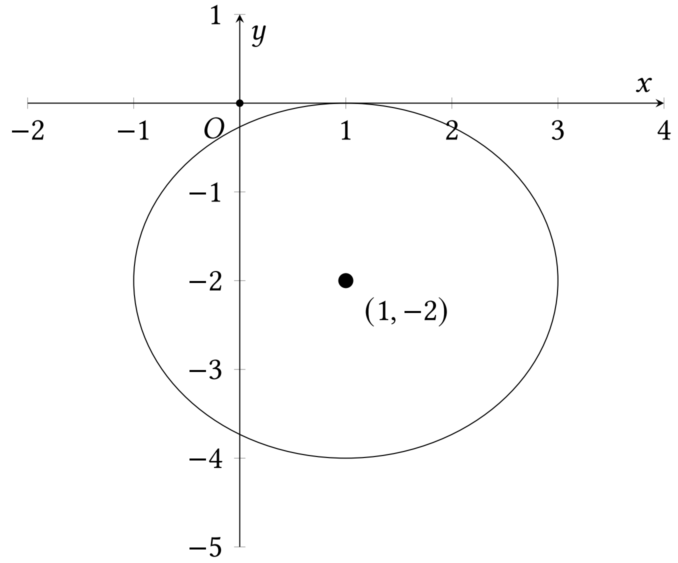


Figure 3.18: Drawing a circle with radius 2 centered at $(1, -2)$, given by the equation $(x - 1)^2 + (y + 2)^2 = 4$.

PROPOSITION 3.8. *The equation representing a circle of radius r centered at the point (a, b) is*

$$(x - a)^2 + (y - b)^2 = r^2$$

Scaling

Let us look at how we can stretch the graph in the x or y direction, also known as *scaling*. To scale a graph in a particular direction, we can argue very similarly as we had before. Suppose we had a graph $y = f(x)$ and we wanted to scale the y-values of each point on the graph by a factor of b. This can be done quite simply by multiplying b to each value of $f(x)$, i.e. we get $y = bf(x)$ or $y/b = f(x)$. As such, if we replaced y with y/b in the equation, the graph is vertically scaled by a factor of b. By similar arguments, we can scale a graph horizontally by a factor of a by replacing x with x/a in the equation. For example, we can double the y values and triple the x values of a unit circle to form an ellipse, as shown in Figure 3.19, given by the equation $(x/3)^2 + (y/2)^2 = 1$.

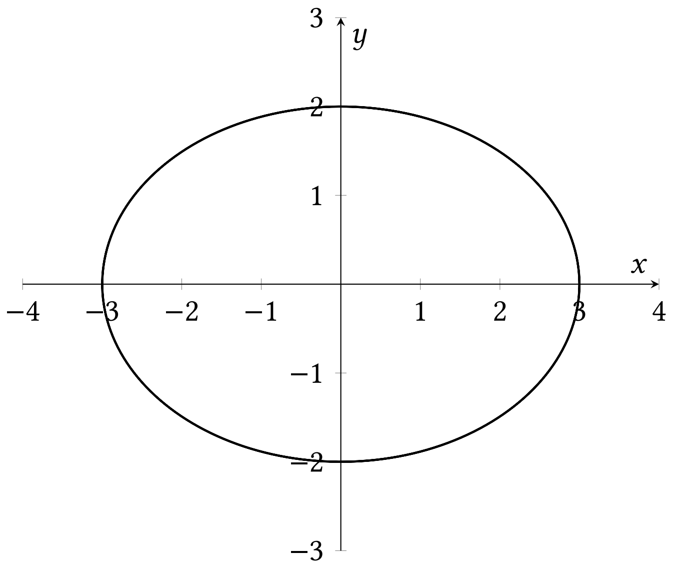


Figure 3.19: An ellipse of width 6 and height 4 centered at the origin given by the equation $(x/3)^2 + (y/2)^2 = 1$.

Reflection

Finally, we might want to *reflect* a graph along the line. In other words, we might imagine placing a mirror on a line, obtaining the graph in its mirror image. Finding reflections on circles may not tell the full story, so instead let us consider reflections on other curves. Let's say we wanted to reflect the graph along the x-axis. Clearly, what we want is for the y-values of the original curve to be on the opposite side of the x-axis, which can be obtained by multiplying by -1, i.e. if $y = f(x)$ then $y = -f(x)$ or $-y = f(x)$ (where replace y by $-y$ would reflect the graph along the x-axis. By similar arguments, we can replace x by $-x$ to reflect the graph along the y-axis. This is shown in the two graphs of Figure 3.20. The original function is $y = x^3 - 5x^2 - 19x + 100$, shown in red on both graphs. On the left, the original graph is reflected along the x axis by replacing y with $-y$, therefore obtaining $y = -x^3 + 5x^2 + 19x - 100$ in blue. On the right, the original graph is reflected along the y axis by replacing x with $-x$, therefore obtaining $y = -x^3 - 5x^2 + 19x + 100$.

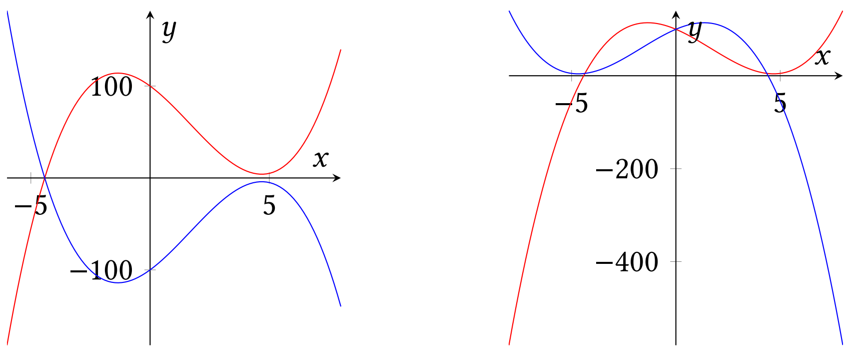


Figure 3.20: Graph of $y = x^3 - 5x^2 - 19x + 100$ (red) reflected along the x-axis (left) and y-axis (right).

Perhaps more interestingly, we may also reflect along the diagonal $y = x$. Observe that what we want is for each (x, y) coordinate of the graph, we should have (y, x) in its reflection.

Thus, by replacing x with y and y with x in our graph, we get its reflection along the diagonal $y = x$.

On a final note, if we wanted to rotate the graph by 180° centered around the origin, we may perform reflection along the x and y axes simultaneously. We show examples of reflection along the diagonal and 180° rotation in Figure 3.21.

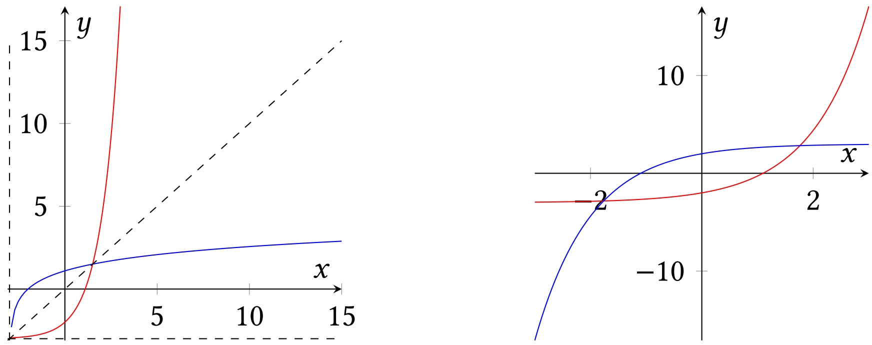


Figure 3.21: Graph of $y = e^x - 3$ (red) reflected along the diagonal (left) and rotated 180° (right).

EXERCISES

3.1. You are given two triangles ABC and DEF. Let A, B, C, D and so on all represent the angles at those points, i.e. they are BAC, ABC and so on, and a, b and so on are their respective opposite sides. If we have $a = 4$, $b = 5$ and $C = 1.4455$ rad, and $d = 12.5$, $e = 15$ and $f = 10$, show that ABC and DEF are similar.

3.2. Find the length, equation and angle of the line segment between the two points $(-1, 1)$ and $(2, -4)$.

3.3. Find the area of the four-sided polygon whose vertices are the points (in clockwise) $A = (-2, 1), B = (-1, 3), C = (2, 4), D = (0, 2)$

3.4. Find the equation that plots a circle of radius 3 centered at the point $(-2, 1)$.

3.5. Find the equation where $y = x^2 - 4x + 5$ is reflected along the line $y = x$.

Chapter 4

Sets and Numbers

The essence of mathematics lies in its freedom.

Georg Cantor

As the foundations of mathematical reasoning started to lean in favour of formalism and away from intuitionism, mathematicians were in the search for a formal system[1] consisting of a finite collection of *axioms* that proves all of mathematics. As of today, it is widely regarded that *set theory*[2], as axiomatized by Ernst Zermelo and Abraham Fraenkel (commonly known as ZF set theory[3]) forms the foundation of all of mathematics.

A wide range of domains in mathematics, especially algebra, is formulated in a set-theoretic manner. For example, monoids are sets equipped with an associative and unital (with respect to an identity element in the set) binary function (which is also a set), groups are monoids admitting inverses, topological spaces are sets endowed with a family of subsets of that set, Turing Machines consist of seven sets, of tape alphabet symbols, states, and so on. As such, knowing a little bit about sets will help you lay the foundation of further study of mathematics and computer science.

Axiomatic set theory itself is a large field, and certainly cannot fit in a one-hour session meant for introductory study. As such, this chapter does not look into it, but only the primitive notions (many of which we have already seen in this text) which is used virtually everywhere. We will then look at a set-theoretic formulation of the natural numbers, before discussing

[1]In 1900, David Hilbert, one of the pioneers of the formalist school, proposed a set of 23 unsolved problems (known as Hilbert's problems) that were important to mathematics. Among these problems are the questions of consistency (can we show that a formal system does not prove that a well-formed statement is true *and* false simultaneously?), completeness (can we show that a formal system can prove all true statements in this system from the axioms?), and decidability (can we show that there is an effective method that can prove a true statement in a finite sequence of steps?). As it turns out, Kurt Gödel and Alan Turing (widely considered the founder of computer science) has shown these to be pretty much false.

[2]George Cantor's results on transfinite numbers and the existence of different kinds of infinities laid much of the foundations of set theory as we know it. However, at the time, it was highly controversial, and Cantor was accused as being a "scientific charlatan" and "corrupter of the youth". David Hilbert defended Cantor's work, saying "No one shall expel us from the paradise that Cantor has created."

[3]Before ZF set theory, Bertrand Russell discovered a paradox in naïve set theory, now known as Russell's paradox, that stems from the *axiom (schema) of unrestricted comprehension.* This paradox, along with other paradoxes discovered by Cantor and other mathematicians, showed that there cannot be a set of all sets. ZF set theory avoids this by replacing this axiom with the *axiom (schema) of specification*, which only allows building sets from subsets.

some properties of numbers to end of this series of notes.

4.1 SETS

4.1.1 The Basic Notions

Set and Multiset

Definition 4.1. A *set* is an *unordered* collection of *unique* objects, known as *elements*. Each element can only occur once in a set, and the order of the elements in the set is irrelevant.

By extension, a *multiset* is a set where the requirement of uniqueness in dropped, so multisets can contain duplicate elements.

The most basic way to describe a set is with *set roster* notation, which is to list the elements of a set enclosed in curly braces. For example, $\{1, 2, 3\}$ is a set containing 1, 2 and 3, $\{4, \{5\}\}$ is a set containing 4 and a set containing 5 (yes, sets can contain sets[4]!) and $\{6, \text{'abc'}\}$ is a set containing 6 and `'abc'` (sets can contain anything!) Because sets are unordered and have unique elements, the set $\{1, 2, 3\}$ is the same as $\{3, 2, 1\}$, and is also the same as the set $\{1, 1, 2, 3\}$. However, the multiset $\{1, 1, 2, 3\}$ is different to the multiset $\{1, 2, 3\}$, since duplicates are considered in multisets.

A set can also not contain any elements as all. This is the unique empty set, denoted $\{\}$ or more commonly, $\emptyset$. Note that $\{\emptyset\} \neq \emptyset$, the former is a set containing one element, the empty set, while the latter is the empty set, which is a set that is empty and does not contain any elements.

Set Membership

The next primitive notion of sets is *set membership*. The proposition $x \in A$ (read x *in* A) states that x is an element of A. For example, $1 \in \{1, 2, 3\}$ is true because 1 is indeed an element of $\{1, 2, 3\}$, but $4 \in \{1, 2, 3\}$ is false because 4 is not an element of $\{1, 2, 3\}$. The negation of set membership is written as $\notin$ (read as *not in*), and therefore $1 \notin \{1, 2, 3\}$ is false and $4 \notin \{1, 2, 3\}$ is true.

Equality of Sets

These notions are all we need to begin describing virtually all of set theory. The first concept we can describe from what we have so far is *equality of sets*, which by the *axiom of extensionality*, states that two sets A and B are equal if they have precisely the same elements.

Axiom 4.1 (Axiom of Extensionality).

$$\forall X \forall Y\ ((\forall z (z \in X \Leftrightarrow z \in Y)) \Rightarrow X = Y)$$

Given any two sets X and Y, if for all objects z we have that z is an element of both X and Y or that z is neither a member of X nor Y, then X is equal to Y.

[4] Ironically, sets cannot contain other sets in Python.

From Axiom 4.1 we get that $\{1, 2, 3\} = \{3, 2, 1\} = \{1, 1, 2, 3\}$ since all the elements of one set are elements of the others, and objects that are *not* in one of the sets are also not elements of the others.

4.1.2 Subsets, Specification and Power Sets

Subsets

We say that A is a *subset* of B, written $A \subseteq B$, if (and only if) every element of A is also an element of B (but not necessarily vice-versa, i.e. we allow an element of B to not be an element of A).

Definition 4.2 (Subset)**.** $A \subseteq B \Leftrightarrow \forall x(x \in A \Rightarrow x \in B)$ A is a *subset* of B if and only if for all objects x if x is an element of A, then x is also an element of B.

We can reverse the $\subseteq$ symbol, to speak of the *superset* of another set, i.e. $B \supseteq A$ (read B superset of A) is equivalent to $A \subseteq B$. Clearly, if we have $A \subseteq B$ and $A \supseteq B$ then we get that $A = B$. To define subsets that are not equal, we can define A as the *proper subset* of B, written $A \subset B$, if A is a subset of B but A is not equal to B, i.e. $A \subset B \Leftrightarrow A \subseteq B$ and $A \neq B$; dually $A \supset B$ states that A is a *proper superset* of B. The negation of each of these symbols is $\nsubseteq$, $\nsupseteq$, $\not\subset$ and $\not\supset$, whose meanings should be easily understood.

Example 4.1. $\{1, 2\} \subseteq \{1, 2, 3\}$ and $\{6, 5, 4\} \supseteq \{4, 5, 6\}$. Also, $\{1, 2\} \subset \{1, 2, 3\}$, but $\{6, 5, 4\} \not\supset \{4, 5, 6\}$ because these two sets are equal.

Set Builder Notation

By the *axiom schema of specification*, we can construct a subset of a set where all elements of this subset satisfies a desired property.

Axiom 4.2 (Axiom (Schema) of Specification (Adapted)).

$$\forall Z \exists Y \left(\forall x(x \in Y \Leftrightarrow x \in Z \text{ and } \phi(x))\right)$$

Given a formula ϕ on x (this formula must not contain Y), for all sets Z, there exists a set Y such that for all objects x, x is an element of Y if and only if x is also in Z and $\phi(x)$ is satisfied. In other words, Y contains all the elements of Z that satisfies ϕ.

For example, let's say we have a set $A = \{1, 2, 3, 4\}$. If we let $\phi(x)$ be the sentence 'x is even', then there is a subset of A that only contains the even numbers in A, this is $B = \{2, 4\}$. This is more concisely described using *set builder* notation, which is to build a set from another set and a desired property. Typically, it looks like this:

$$Y = \{x \in Z \mid \phi(x)\}$$

Example 4.2. From earlier, let us once again construct the subset of $A = \{1, 2, 3, 4\}$ only containing the even numbers. We can denote this set with set builder notation as

$$B = \{x \in A \mid x \text{ is even}\}$$

This is the set $B = \{2, 4\}$, since only 2 and 4 are elements of A and satisfies the formula that x is even.

We may sometimes move some of the formulae around, for example, the notation $\{x \mid x \in Z, \phi(x)\}$ denotes the set of all objects x such that x is an element of Z and $\phi(x)$ is true. This is incredibly useful if we want to denote the *image* of a subset under a function, which is enabled by the *axiom (schema) of replacement*. For example, a formula $\{f(x) \mid x \in Z, \phi(x)\}$ describes the set obtained by first taking the subset $\{x \mid x \in Z, \phi(x)\}$, then pass each element x in that set through f, giving us $f(x)$, then collecting all of those into a set.

Example 4.3. Let us say we have a function $f(x) = x + 1$. The image of $B = \{2, 4\}$ under f is $C = \{f(2), f(4)\} = \{3, 5\}$. Since C is the image of B under f, and B is the subset of $A = \{1, 2, 3, 4\}$ containing all the even numbers, then we can denote C as

$$\{f(x) \mid x \in A, x \text{ is even}\}$$

or equivalently,

$$\{x + 1 \mid x \in A, x \text{ is even}\}$$

Power Sets

The set containing all (and only) the subsets of X is known as the *power set* of X, denoted $\mathcal{P}(X)$.

Axiom 4.3 (Axiom of Power Sets).

$$\forall X \exists Y \forall Z (Z \subseteq X \Rightarrow Z \in Y)$$

For all sets X, there exists a set Y such that for all sets Z, if Z is a subset of X, then Z is an element of Y.

However, given set X, this set Y may be too large (it may contain objects that are not subsets of X). As such, to describe $\mathcal{P}(X)$ which contains *only* the subsets of X, we may use the axiom schema of specification:

$$\mathcal{P}(X) = \{Z \in Y \mid Z \subseteq X\}$$

Example 4.4. Let $A = \{1, 2, 3\}$. Then, $\mathcal{P}(A)$ contains all the subsets of A, i.e. it is the set

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$$

4.1.3 Operations on Sets

Having learnt some of the basic underlying notions of sets, we can now begin to describe some of the operations on sets.

Intersection

Definition 4.3. The *intersection* of two sets A and B, written $A \cap B$, consists of all the objects that are both elements of A and elements of B, i.e. it is the set of all common elements between A and B:

$$A \cap B = \{x \in A \mid x \in B\}$$

Example 4.5. Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$. Then, $A \cap B = \{3, 4\}$.

Furthermore, given a set of sets (also known as a *family of sets*) $\mathcal{F} = \{S_1, \ldots, S_n\}$ where each S_i is a set, we can denote the intersection of all the elements (sets) of $\mathcal{F}$ as

$$\bigcap \mathcal{F} = \bigcap_{A \in \mathcal{F}} A = \bigcap_{i=1}^{n} S_i = \{x \in S_1 \mid x \in S_2, x \in S_3, \ldots, x \in S_n\} = S_1 \cap S_2 \cap \cdots \cap S_n$$

Every term in the equality is notationally equivalent.

Example 4.6. Let $S_1 = \{1, 2, 3, 4\}$, $S_2 = \{3, 4, 5, 6\}$, $S_3 = \{2, 3\}$ and $\mathcal{F} = \{S_1, S_2, S_3\}$. Then, since 3 is the only common element in all three sets S_1 to S_3,

$$\bigcap \mathcal{F} = \{3\}$$

To visualize the concept of the intersection of two sets, we can use *Venn diagrams*. If we let A and B be two sets, visualized as two circles and the elements of the sets fall inside these circles, then $A \cap B$ will be space that overlaps the two circles. We show the Venn diagram of $A \cap B$ in Figure 4.1.

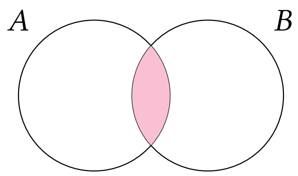


Figure 4.1: Venn diagram where the shaded portion represents $A \cap B$.

Union

Definition 4.4. The *union* of two sets A and B, written $A \cup B$, consists of all the objects that are elements of A and/or elements of B, i.e. it is the set of all the elements of A and B combined:

$$A \cup B = \{x \mid x \in A \text{ and/or } x \in B\}$$

Example 4.7. Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$. Then, $A \cup B = \{1, 2, 3, 4, 5, 6\}$.

However, this definition of the union of two sets does not follow from the axioms we have presented, because $A \cup B$, unlike $A \cap B$, is neither a subset of A or a subset of B. Thus, there is the *axiom of union* which asserts the existence of the union of a family of sets.

Axiom 4.4 (Axiom of Union).

$$\forall \mathcal{F} \exists U \forall Y \forall x (x \in Y \in \mathcal{F} \Rightarrow x \in U)$$

For any family of sets $\mathcal{F}$, there exists a set U such that for all sets x and Y, if x is an element of Y and Y itself is an element (set) of $\mathcal{F}$, then x is also in U.

Similar to the axiom of power sets, this set U we have constructed may be too large, thus we can once again invoke the axiom of specification to specify the smallest set that satisfies the description of the union.

$$\bigcup \mathcal{F} = \{x \in U \mid x \in Y \in \mathcal{F}\}$$

Just like with intersections over a family of sets, the union of a family of sets $\mathcal{F} = \{S_1, \dots, S_n\}$ where each S_i is a set can be written as

$$\bigcup \mathcal{F} = \bigcup_{A \in \mathcal{F}} A = \bigcup_{i=1}^{n} S_i = S_1 \cup S_2 \cup \cdots \cup S_n$$

Example 4.8. Let $S_1 = \{1, 2, 3, 4\}$, $S_2 = \{3, 4, 5, 6\}$, $S_3 = \{7, 8\}$ and $\mathcal{F} = \{S_1, S_2, S_3\}$. Then, taking the union of $\mathcal{F}$ is simply to construct a set containing all the elements that occur in S_1, S_2 and/or S_3.

$$\bigcup \mathcal{F} = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

Finally, just like with intersections, we can depict the union of two sets as a Venn diagram, shown in Figure 4.2.

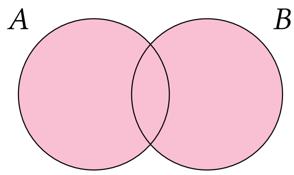


Figure 4.2: Venn diagram where the shaded portion represents $A \cup B$.

The binary operation $\cap$ and $\cup$, just like addition and multiplication, are *associative*, *commutative* and *distributive*[5], as claimed in Proposition 1.1, Proposition 1.2 and Proposition 1.3.

Proposition 4.5. $\cap$ *and* $\cup$ *are associative, commutative and distributive:*

$$(A \cap B) \cap C = A \cap (B \cap C) \qquad (A \cup B) \cup C = A \cup (B \cup C)$$

$$A \cap B = B \cap A \qquad A \cup B = B \cup A$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Complements

The *complement* of a set A, denoted $A^{\complement}$, is the set of all objects that are not elements of A. Of course, *the set of all objects* needs to be defined, otherwise, we do not know what the complement of A should contain. Usually this is implicit; if A is the set of all even natural numbers, then the set of all objects U would be $\mathbb{N}$. In fact, this could be any arbitrary set, as long as it is a superset of A. We shall leave this 'universal set' implicit and assume it can be understood by context.

[5]This is partially why what we have described so far is known as the *algebra of sets*.

Definition 4.5. The complement of a set A with respect to a set U such that $U \supseteq A$, is given by
$$A^{\complement} = \{x \in U \mid x \notin A\}$$

Example 4.9. Suppose the set of all objects is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and $A = \{2, 3, 5, 6, 7\}$. Then,
$$A^{\complement} = \{1, 4, 8, 9, 10\}$$

The complement of a set is depicted as a Venn diagram in Figure 4.3.

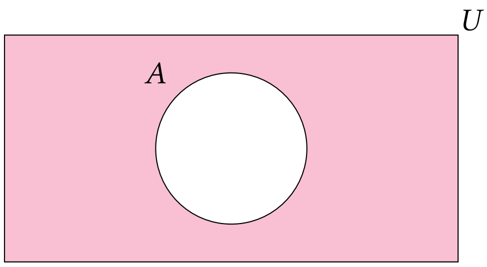


Figure 4.3: Venn diagram where the shaded portion represents $A^{\complement}$.

Difference

From what we have defined so far, we can now describe two notions of *difference* between two sets: *set difference*, and *symmetric difference*.

Definition 4.6 (Set difference)**.** The difference of two sets A and B, denoted $A \backslash B$ (sometimes denoted $A - B$), is the set of all elements in A that are not in B:
$$A \backslash B = \{x \in A \mid x \notin B\} = A \cap B^{\complement}$$

Example 4.10. Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5\}$. Then, since 3 and 4 occurs in both A and B, $A \backslash B = \{1, 2\}$.

Definition 4.7 (Symmetric difference)**.** The symmetric difference of two sets A and B, denoted $A \,\Delta\, B$, is the set of all elements either in A, or in B, but not both:
$$A \,\Delta\, B = (A \cup B) \backslash (A \cap B) = (A \backslash B) \cup (B \backslash A)$$

Example 4.11. Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5\}$. Then, since 3 and 4 occurs in both A and B, $A \,\Delta\, B = \{1, 2, 5\}$.

Set difference and symmetric difference is depicted in the Venn diagrams in Figure 4.4.

Products and Disjoint Unions

At this point we shall temporarily depart from defining operations directly from the axioms ZF which is too cumbersome[6]. In either case, understanding what these operations are is

[6] This is frequently the case for many branches of mathematics—it is usually assumed that we could, *in principle*, derive much of mathematics from these axioms.

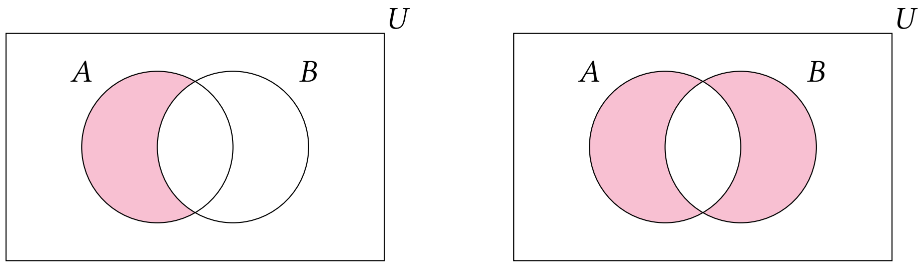


Figure 4.4: Venn diagrams where the shaded portions represents $A \backslash B$ (left) and $A \Delta B$ (right).

significantly more important than their full set-theoretic derivation of their definitions.

We shall start with a binary operation which we have briefly talked about in Section 1.2, the *cartesian product* of two sets, which is a set of all *ordered* pairs, where the first object of each pair is an element of the first set, and the second object of each pair is an element of the second set. A pair of two objects a and b where a is the first and b is the second object, is denoted as (a, b)—we shall not rigorously formalize this notion[7].

Definition 4.8. The cartesian product of two sets A and B, denoted $A \times B$, is defined as

$$\{(a, b) \mid a \in A, b \in B\}$$

Remark 4.1. This is why we have described binary functions on real numbers having domain $\mathbb{R} \times \mathbb{R}$, this is the set of all pairs of real numbers! We described the 2-dimensional Euclidean space as $\mathbb{R}^2$ with points denoted by (x, y)-coordinates for this exact same reason.

Example 4.12. Let $A = \{1, 2, 3\}$ and $B = \{2, 5\}$. Then,

$$A \times B = \{(1, 2), (1, 5), (2, 2), (2, 5), (3, 2), (3, 5)\}$$

We can visualize the cartesian product of two sets by way of a 2-dimensional table: for example, if we had $A = \{x, y, z\}$ and $B = \{1, 2, 3\}$ then $A \times B$ is shown in Figure 4.5.

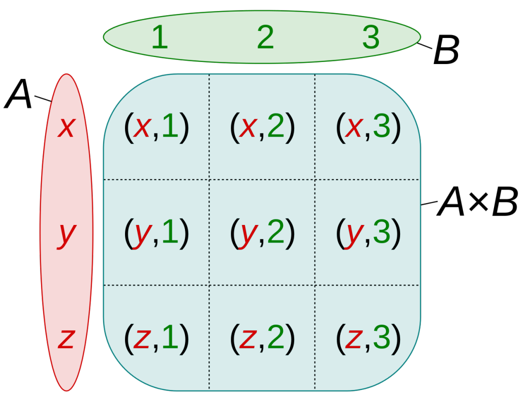


Figure 4.5: Visualization of the cartesian product of two sets. Source: `https://en.wikipedia.org/wiki/Cartesian_product`

[7] If you want to know, Kuratowski's definition of ordered pairs, $(a, b) = \{a, \{a, b\}\}$ is the most commonly used construction of ordered pairs, which allows us to define the cartesian product based on specification of $\mathcal{P}(\mathcal{P}(A \cup B))$ with the same formula described in our definition (i.e. $a \in A, b \in B$).

We can also generalize cartesian products to an arbitrary number of sets, for example, the product of three sets A, B and C, denoted $A \times B \times C$ contains all *triples*[8] (a, b, c) such that $a \in A, b \in B$ and $c \in C$. Alternatively, we can use product notation: if $\mathcal{F} = \{A, B, C\}$ then

$$\prod_{F \in \mathcal{F}} F = A \times B \times C$$

The *dual*[9] of this notion is the *disjoint union* of two sets. Loosely, the disjoint union of two sets A_a and B_b with indices a and b, is the set of all pairs where the first object in the pair is an element of A or B, and the second element is the index of the set where the first object came from. For example, if we have $A_0 = \{1, 2, 3\}$ and $A_1 = \{2, 3\}$, then the disjoint union of A_0 and A_1, denoted $A_0 \uplus A_1$ or $A_0 + A_1$, is equal to

$$\{(1, 0), (2, 0), (3, 0), (2, 1), (3, 1)\}$$

For example, the pair $(1, 0)$ is the number 1 that came from A_0, and the pair $(2, 1)$ is the number 2 that came from A_1.

As before, the disjoint union of an *indexed family of sets* $\mathcal{A} = (A_i : i \in I)$ can be written as (any of these notations are generally accepted)

$$\coprod_{i \in I} A_i = \biguplus_{i \in I} A_i = \sum_{i \in I} A_i$$

As before, Wikipedia has a great illustration of disjoint unions, which we show in Figure 4.6.

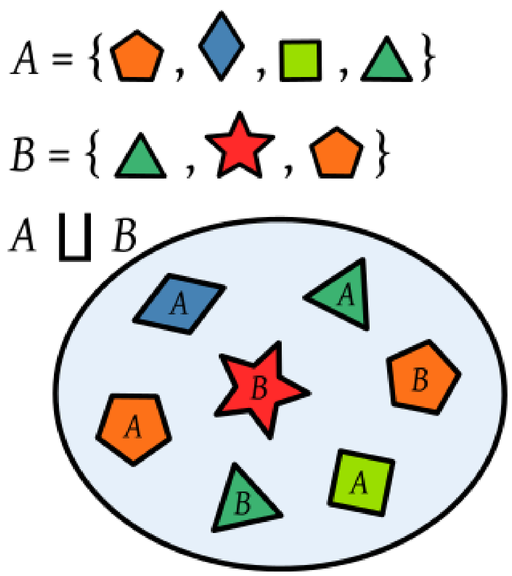


Figure 4.6: Visualization of the cartesian product of two sets. Source: `https://en.wikipedia.org/wiki/Disjoint_union`

[8]You could claim that $\times$ is left-associative so $A \times B \times C = ((A \times B) \times C)$ so objects of this set will look more like $((a, b), c)$ and not (a, b, c). This distinction is not particularly meaningful at this juncture. Particularly, $(A \times B) \times C)$ and $A \times B \times C$ as we have defined are *naturally isomorphic*, so their representation doesn't really matter in many cases.

[9]In the *category* of sets, the disjoint union of two sets corresponds to the *coproduct* of two objects, which is the dual of the product of two objects, corresponding to the cartesian product of two sets.

Remark 4.2. Disjoint unions are only occasionally useful for our purposes; I do want to introduce them here because disjoint unions or coproducts correspond to *sum/either* types in programming.

4.2 FUNCTIONS AND RELATIONS

Now that we have described a lot about sets, we may now proceed to describe functions and relations on sets. Intuitively we might think of sets and functions as being separate entities, where sets are objects and functions and relations are different kinds of objects that relate two or more sets. It turns out that the set-theoretic definition of relations and sets states that these are themselves, sets.

4.2.1 Relations

Definition 4.9 (Relation). Let X be a set. A set R is said to be a *relation* on X if R contains ordered pairs where both objects of each pair are from X, i.e. $R \subseteq X \times X$. If $(x, y) \in R$, we write $R(x, y)$, or in infix notation, $x\ R\ y$, to say that x is R-related to y.

Example 4.13. Suppose we defined our own relation $D \subseteq \mathbb{N} \times \mathbb{N}$ on the *set* of natural numbers $\mathbb{N}$, where D consists of pairs (a, b) where *a divides b*, i.e. $b \div a = c$ where $c \in \mathbb{N}$. Then, D would be the relation

$$\{(1, 0), (1, 1), (1, 2), \dots, (2, 2), (2, 4), \dots, (3, 3), (3, 6), (3, 9), \dots\}$$

Clearly, $4\ D\ 8$ or $D(4, 8)$ is a true proposition but it is not true that $3\ D\ 7$ or $D(3, 7)$.

4.2.2 Properties of relations

Several relations are of interest because of properties they hold. Of keen interest, we have three properties that we shall discuss, *reflexivity*, *symmetry* and *transitivity*.

Reflexivity

Definition 4.10 (Reflexivity). A relation R on X is reflexive if for all $x \in X$, $x\ R\ x$ is true.

Example 4.14. The relation $\leq$ on $\mathbb{N}$ is reflexive; for all natural numbers x and y, we know that $x \leq x$ is true.

Symmetry

Definition 4.11 (Symmetry). A relation R on X is symmetric if for all $x, y \in X$, $x\ R\ y \Leftrightarrow y\ R\ x$, i.e. if $x\ R\ y$, then $y\ R\ x$ (and clearly, the other way round)

Example 4.15. Let us define the relation $x\ D\ y$ on $\mathbb{N}$ as all (x, y) pairs such that $|x - y| = 1$. Clearly, for any $x\ D\ y$ we also get $y\ D\ x$, since $|x - y| = |(-1)(y - x)| = |y - x|$. D is *not* reflexive; for example, it is not true that $1\ D\ 1$. On the other hand, continuing from

Example 4.14, $\leq$ is not symmetric, for example, we know that $1 \leq 2$, but it is not true that $2 \leq 1$.

Transitivity

Definition 4.12 (Transitivity)**.** A relation R on X is transitive if for all $x, y, z \in X$, $x \mathrel{R} y$ and $y \mathrel{R} z \Rightarrow x \mathrel{R} z$, i.e. if $x \mathrel{R} y$ and $y \mathrel{R} z$, then $x \mathrel{R} z$.

Example 4.16. Continuing from *Example* 4.14 and *Example* 4.15, $\leq$ is transitive; if $x \leq y$ and $y \leq z$ then clearly, $x \leq z$. However, D is not transitive. We know that $5 \mathrel{D} 4$ and $4 \mathrel{D} 3$, but it is not the case that $5 \mathrel{D} 3$.

Equivalence Relation

Many mathematical structures may have incredibly similar properties and, for all intents and purposes, are equal. For example, the sets $\{1, 1, 2, 3\}$ and $\{3, 2, 1\}$ may *look* different, but actually represent the same object. As such, we may abstract the notion of equality as relations satisfying the criteria described above.

Definition 4.13. A relation R on X is said to be an *equivalence relation* if it is a reflexive, symmetric and transitive relation.

Example 4.17. Equality between sets, $=$, is an equivalence relation. This should be incredibly intuitive.

Closures

Sometimes we have a relation R on a set X that only describes the relationships between some pairs of objects. We may want to extend this relation into the smallest one that satisfies a desired property. This is known as the *closure*[10] of a relation.

Definition 4.14. The reflexive (or symmetric or transitive) closure of a relation R on X is the smallest relation containing R that is reflexive (or symmetric or transitive). In notation, the reflexive closure of R is

$$R \cup \{(x, x) \in X \times X \mid x \in X\}$$

The symmetric closure of R is

$$R \cup \{(y, x) \in X \times X \mid (x, y) \in R\}$$

And finally, the transitive closure of R is (this is not a completely rigorous definition):

$$R \cup \{(x, z) \in X \times X \mid \exists y_1 \exists y_2 \ldots \exists y_n (x \mathrel{R} y_1, y_1 \mathrel{R} y_2, \ldots, y_n \mathrel{R} z)\}$$

Example 4.18. Suppose we define a relation R on $\mathbb{N}$ as all pairs (x, y) such that $x + 1 = y$. This relation is neither reflexive ($1 \mathrel{R} 1$ is not true), symmetric ($1 \mathrel{R} 2$ is true but $2 \mathrel{R} 1$ isn't)

[10]Closures in computing refer to a different concept in programming.

nor transitive ($1\ R\ 2$ and $2\ R\ 3$ are true but $1\ R\ 3$ isn't). The reflexive and transitive closure of R is $\leq$.

N-nary Relations

We can easily generalize relations on $n > 0$ elements of X. The simplest definition of an n-nary relation R on X is a set that satisfies $R \subseteq X^n$.

We may also generalize relations to relations between objects of different from sets. Earlier, we have defined (binary) relations to be sets $R \subseteq X^2$, but we can also define binary relations R on $X \times Y$, i.e. $R \subseteq X \times Y$.

4.2.3 Functions (again)

We have already looked at functions in Chapter 1, but we shall look at them again in a set-theoretic view.

A function relates two sets A and B. As before, we denote $f : A \longrightarrow B$ as a function f that goes from A to B, where A is known as its domain and B its codomain. For each[11] $a \in A$, f assigns a to *exactly* one object $b \in B$, which we denote $f(a)$. This object $f(a) \in B$ is known as the image of a under f, and as before, we denote the set $\{f(a) \mid a \in A\}$ as the image of A under f.

This function f is represented as the set $\{(a, b) \mid a \in A, b \in B\}$, and is defined such that if $(a, b) \in f$, then $b = f(a)$. Clearly, $f \subseteq A \times B$, so f is also a (binary) relation, but not all relations are functions. We have earlier specified that a function assigns each object in its domain to exactly one object in its codomain, no more, no less; unlike relations, where an element of the domain can appear as the first object in more than one pair, such as the relation $\leq$ (for example, $(1, 2), (1, 3) \in\leq$).

Semi-formally, f is only a function if it satisfies the following condition:

$$\forall a(f(a) = b \text{ and } f(a) = b' \Rightarrow b = b')$$

In other words, given an input a to f, if what we get as output is b and b', it must mean that $b = b'$. This basically characterizes our requirement that for each element of its domain, f assigns not more than one element of its codomain to it.

Example 4.19. On top of all the functions that we have described in earlier chapters, we can also define a function using set roster notation. Suppose we had $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$, then we can define a function $f : A \longrightarrow B$ given by $f = \{(1, 4), (2, 5), (3, 4)\}$, i.e. $f(1) = 4$, $f(2) = 5$ and $f(3) = 4$. Notice that the image of A under f is $\{4, 5\}$, which is a proper subset of the codomain B.

We may depict f from *Example* 4.19 as a diagram showing the correspondence between each element in its domain and codomain, as shown in Figure 4.7.

Food for thought 4.3. Let's say we defined a function on real numbers $f(x) = \sqrt{x}$, for example, $f(4) = -2$. Is f a function?

[11] Importantly, f must assign *every* $a \in A$ to one $b \in B$. We may, however, define *partial* functions, which do not have to perform this assignment on every element in its domain.

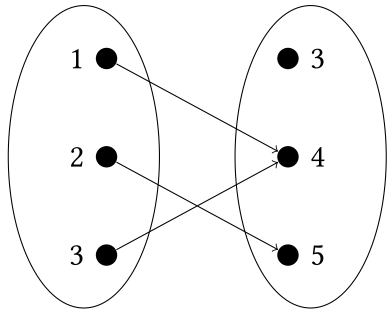


Figure 4.7: A depiction of f from *Example* 4.19.

4.2.4 Properties of functions

There are several properties of functions that are of keen interest to us.

Injections

Definition 4.15 (Injection)**.** A function $f : A \longrightarrow B$ is an *injection* or is *injective* if for every $b \in B$ there is no more than one $a \in A$ such that $f(a) = b$.

Example 4.20. Suppose we have sets $A = \{1, 2\}$ and $B = \{3, 4, 5\}$. The function $g : A \longrightarrow B$ given by $g = \{(1, 3), (2, 5)\}$, i.e. $g(1) = 3$ and $g(2) = 5$ is an injection from A to B. g is shown in Figure 4.8. g is an injection because each element in B only has at most one arrow pointing towards it.

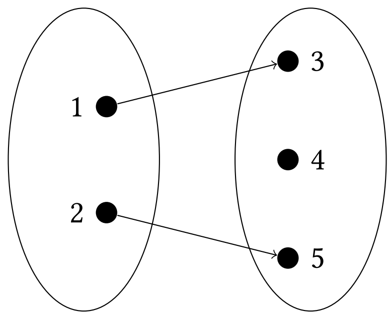


Figure 4.8: An injection $g = \{(1, 3), (2, 5)\}$.

All injective functions $f : A \longrightarrow B$ have a *left inverse* $g : B \longrightarrow A$, which is a function from the image of f back to its domain, such that g 'undoes' f. In other words, if we define the *identity function* on a set A as $\mathrm{id}_A : A \longrightarrow A$ given by $\mathrm{id}_A(a) = a$, then a left inverse g is such that $g \circ f = \mathrm{id}_A$.

Example 4.21. Continuing from *Example* 4.20, the left inverse of g, which we shall call h, is a function from $\{3, 5\}$ (the image of A under g) to A, given by $\{(3, 1), (5, 2)\}$. Clearly for all $a \in A$, $h(g(a)) = a$—$h(g(1)) = 1$, and $h(g(2)) = 2$. We can obtain this function h by inverting all the arrows in Figure 4.8. If we insist that the domain of h be exactly B and not a subset of it, just add an arrow from the number 4 to any number in A e.g. $h(4) = 1$ and this will be a left inverse from B to A.

Surjections

Definition 4.16 (Surjection)**.** A function $f : A \longrightarrow B$ is a *surjection* or is *surjective* if for every $b \in B$ there is at least one $a \in A$ such that $f(a) = b$.

Example 4.22. Suppose we have sets $A = \{1, 2, 3\}$ and $B = \{3, 4\}$. The function $f : A \longrightarrow B$ given by $f = \{(1, 4), (2, 3), (3, 3)\}$, i.e. $f(1) = 4$, $f(2) = 3$ and $f(3) = 3$ is a surjection because each element in B has at least one $a \in A$ such that $f(a) = b$, i.e. each element in B has at least one arrow pointing towards it in Figure 4.9.

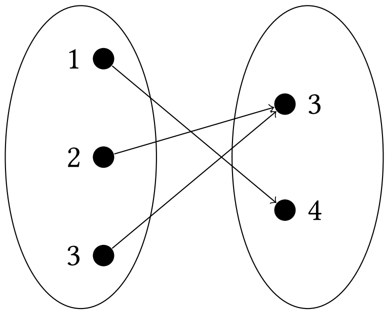


Figure 4.9: A depiction of f from *Example* 4.22.

Surjections are the dual notion of injections, and since injections have left inverses, surjections have the dual, known as *right inverse*s. Given a surjection $f : A \longrightarrow B$, a right inverse $g : B \longrightarrow A$ is such that $f \circ g = \mathrm{id}_B$.

Example 4.23. Continuing from *Example* 4.22, the function $g : B \longrightarrow A$ given by $g = \{(3, 2), (4, 1)\}$ is a right inverse to f. This can be obtained by inverting all the arrows in Figure 4.9, and dropping one of the arrows coming out of the number 3 (this is because a function can only draw one outgoing arrow for each point). We can clearly see that $f \circ g = \mathrm{id}_B$: $f(g(3)) = 3$ and $f(g(4)) = 4$.

Bijections

Definition 4.17 (Bijection)**.** A function $f : A \longrightarrow B$ is a *bijection* or *bijective* function if it is injective and surjective, i.e for every element $b \in B$, there is exactly one $a \in A$ such that $f(a) = b$.

Bijections are really important because they put the elements of two sets into a one-to-one correspondence with each other. If $f : A \longrightarrow B$ is a bijection, then we can see f simply as a 'renaming function', simply changing what the elements of the set look like. As we shall see later, if we obtain B from A via a bijection f, then there is a unique function that brings us back to A, i.e. we can rename the elements of B in a unique way that lets us recover A. More broadly, this concept is known as an *isomorphism*, which is central to the study of mathematics, because it relates two seemingly different mathematical structures that are really just different representations of the same thing.

Example 4.24. Suppose we had $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. The function $f : A \longrightarrow B$ given by $f = \{(1, 4), (2, 5), (3, 3)\}$, i.e. $f(1) = 4$, $f(2) = 5$ and $f(3) = 3$ is a bijection from

A to B. We can clearly see this to be the case since it is both an injection and a surjection. f as shown in Figure 4.10 may shed light on why we say that bijections put elements of the sets in a one-to-one correspondence with each other.

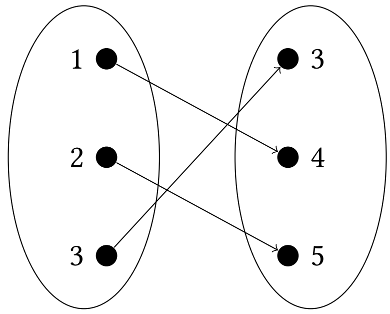


Figure 4.10: A depiction of f from *Example* 4.24.

Since injections have left inverses and surjections have right inverses, bijections have both. In fact, every bijection has a unique function, known as the *inverse*, that reverses the bijection in both ways, i.e. if we have a bijection $f : A \longrightarrow B$, then there will always exist a unique inverse function $g : B \longrightarrow A$ such that $g \circ f = \mathrm{id}_A$ and $f \circ g = \mathrm{id}_B$.

Example 4.25. Continuing from *Example* 4.24, the function $g : B \longrightarrow A$ given by $g(3) = 3$, $g(4) = 1$ and $g(5) = 2$ is the inverse of f. This is done by inverting all the arrows in Figure 4.10. Clearly, there is no other way to obtain another inverse of f, which makes g unique.

Example 4.26. Circling back to Subsection 1.3.2, the exponential function $f(x) = e^x$ is a bijection between $\mathbb{R}$ and the interval $(0, \infty)$. Its inverse is the natural logarithmic function, $g(x) = \ln x$.

Remark 4.4. Before continue our discussion on functions, we shall make a small note that will only be occasionally useful. Since functions themselves are sets, they can also assemble into sets. Thus, we can denote the set of all functions from A to B as B^A.

4.2.5 Cardinality of Sets

An incredibly natural question to ask is, 'how many elements is contained in a set?' This is known as the *cardinality* of a set.

Definition 4.18 (Cardinality)**.** The *cardinality* of a set A, denoted $|A|$, is the number of elements of that set.

Example 4.27. Let $A = \{1, 2, 3\}$. Then, the cardinality of A is $|A| = 3$.

The cardinality of a set can usually be done simply by counting its elements. However, we would like a more systematic way to compute the cardinality of a set that was built from some of the operations seen earlier in Subsection 4.1.3. Let us investigate.

Power Sets

Proposition 4.6.

$$|\mathcal{P}(A)| = 2^{|A|}$$

To create a subset of A, there are two choices for each element: to include it in the subset, or to not include it. Since there are $|A|$ elements, then there are $2^{|A|}$ ways to construct a subset of A.

Intersections, Unions, Complements and Differences

Proposition 4.7.

$$|A \cup B| = |A| + |B| - |A \cap B|$$
$$|A \cap B| = |A| + |B| - |A \cup B|$$
$$|A^{\complement}| = |U| - |A|$$
$$|A \backslash B| = |A \cup B| - |B|$$
$$|A \mathbin{\Delta} B| = |A \cup B| - |A \cap B|$$

Have a look at the Venn diagrams in Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4 to see for yourself.

Products, Disjoint Unions and Exponential Sets

Proposition 4.8.

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

Let us understand why this is the case. In the case of cartesian products, each element of A will be paired up with every element of B, giving us $|B|$ elements. There are $|A|$ elements of A, therefore the cardinality of the cartesian product is $|A| \times |B|$. This is partially why this is known as the cartesian *product*, and is in fact written with the same symbol as multiplication of numbers.

In the case of disjoint unions, every element of A and every element of B will appear in exactly one pair of $|A + B|$ as the first object of the pair. There are $|A| + |B|$ total elements of A and B, making it the cardinality of the disjoint union. This is why the disjoint union is (sometimes) written as $+$.

In the case of exponential sets (B^A is the set of all functions from A to B), because functions map each element of A to exactly one element of B, so each element of A has $|B|$ choices to draw an arrow to. Since there are A elements, the total number of choices (and thus functions) would be $|B|^{|A|}$.

Comparison of cardinalities

Sometimes sets can be too hard to count since they are very large, and even still, in some cases, they could be infinitely large: sets like $\mathbb{N}$, $\mathbb{R}$ and so on are all infinitely large sets. We might want to ask instead, "how do we compare the cardinality of two sets?"

As it turns out, we can use functions to compare the cardinality of two sets. You should re-look the diagrams in Figure 4.8, Figure 4.9 and Figure 4.10 again to give you some intuition.

Proposition 4.9. *Given two sets A and B:*

- $|A| \leq |B|$ *if and only if there is an injection from A to B.*
- $|A| \geq |B|$ *if and only if there is a surjection from A to B.*
- $|A| = |B|$ *if and only if there is a bijection from A to B.*

This proposition should follow naturally from the definitions of injections, surjections and bijections. We can do so by arguing from the fact that every element of B corresponds to at most one, at least one, or exactly one element of A respectively.

Some incredibly counter-intuitive results arise from this fact:

- The set of natural numbers has the same size as the set of nonnegative even numbers. This is shown by the bijection $f(x) = 2x$ and $f^{-1}(x) = x/2$. In other words, there are as many even numbers as there are whole numbers. This is despite the fact that the set of nonnegative even numbers is a proper subset of the set of natural numbers.
- The set of natural numbers has the same size as the set of integers (including negative numbers), i.e. $|\mathbb{N}| = |\mathbb{Z}|$. This is done by constructing a bijection that yo-yos (or snakes) back and forth between the positive numbers and the negative numbers: $f = \{(0,0), (1,1), (2,-1), (3,2), (4,-2), (5,3), (6,-3), \ldots\}$. Such a function f can be defined as $f(x) = \lceil x \rceil$ if x is odd, and $-x/2$ otherwise. Its inverse can be defined as $f^{-1}(x) = -2x$ if $x \leq 0$, and $2x - 1$ otherwise.
- There are more real numbers than integers (even though there are infinitely many of both), i.e. $|\mathbb{R}| > |\mathbb{N}|$. This is done by showing that there is an injection from $\mathbb{N}$ to $\mathbb{R}$: $f(x) = x$ (remember that natural numbers are also real numbers), and then showing that there cannot be a bijection between these two sets, which was shown by George Cantor using a method now known as *diagonalization.* This was incredibly controversial as it suggested different sizes of infinities, some larger than others. As such, sets as large as the set of natural numbers is known to be *countably infinite*, while sets as large (or larger) than the set of real numbers is said to be *uncountably infinite.*
- For all sets A (including infinitely large ones), $|\mathcal{P}(A)| > |A|$. This is also shown by diagonalization.

4.3 NUMBERS

As described in the beginning of this chapter, everything in mathematics should be, in principle, derived from the axioms of set theory. One last axiom that we will consider in this text is the *axiom of infinity.* Let us denote $S(x)$ to abbreviate $x \cup \{x\}$. For example, $S(\{1\}) = \{1, \{1\}\}$. Then, by assuming the empty set exists, we can construct an infinitely large set X such that the empty set is in X, and if a set Y is in X, then so is $S(Y)$.

Axiom 4.10 (Axiom of Infinity).

$$\exists X\left((\exists e(\forall z(z \notin e \textit{ and } e \in X)) \textit{ and } \forall Y(Y \in X \Rightarrow S(Y) \in X)\right)$$

There exists a set X such that:

1. *There exists a set e such that for all sets z, z is not an element of e and e is in X. In other words, the empty set $\emptyset$ exists and is in X, and*
2. *for all sets Y, if Y is in X, then $S(Y)$ is also in X.*

This set X as we have constructed, known as the von Neumann ordinal, is one of the set-theoretic constructions of the set of natural numbers $\mathbb{N}$.

4.3.1 The Natural Numbers

Typically, the natural numbers is understood as the sequence of numbers satisfying the *Peano axioms*:

1. 0 is a natural number.
2. Every natural number n has a successor, denoted $S(n)$, which is also a natural number.
3. 0 is not the successor of any natural number.
4. No two different natural numbers share the same successor.
5. Induction: if a proposition is true for 0, and the truth of the proposition for n implies the truth of the proposition for $S(n)$, then the proposition is true for all the natural numbers.

Meanwhile, we have the von Neumann construction of natural numbers:

1. $\emptyset \in \mathbb{N}$ will be the number 0.
2. If $n \in \mathbb{N}$, then $S(n) = n \cup \{n\}$ is the successor of m and is also in $\mathbb{N}$.

For example $0 = \emptyset$, $S(0) = \{\emptyset\}$, $S(S(0)) = \{\emptyset, \{\emptyset\}\}$, and so on.

We can see that this construction of the natural numbers satisfies the Peano axioms. From now, we shall denote the succesor of 0 as 1, the successor of 1 as 2, and so on.

Once we have rigorously defined the natural numbers, let us look at some constructions of common mathematical operations we have seen in Chapter 1.

$$+ : \mathbb{N} \to \mathbb{N}$$
$$a + b = \begin{cases} a & \text{if } b = 0 \\ S(a) + z & \text{if } b = S(z) \end{cases}$$

For example, $3 + 2 = S(3) + 1 = S(S(3)) + 0 = S(S(3)) = 5$.

$$\times : \mathbb{N} \to \mathbb{N}$$
$$a \times b = \begin{cases} 0 & \text{if } b = 0 \\ (a \times z) + a & \text{if } b = S(z) \end{cases}$$

For example, $4 \times 3 = (4 \times 2) + 4 = ((4 \times 1) + 4) + 4 = (((4 \times 0) + 4) + 4) + 4 = 0 + 4 + 4 + 4 = 12$.

By now, you should be convinced that the natural numbers, and the common arithmetic operations are well defined, and can be naturally extended to other forms of numbers like integers and rational numbers. Real numbers are defined in other ways which is out of the scope of this text. From this point forward, we shall depart from the set-theoretic view of natural numbers and describe properties of natural numbers as we know them.

4.3.2 Prime Numbers

Just like molecules are made of atoms, and atoms are made of particles like protons and electrons, and protons and neutrons are made of quarks, what are natural numbers made of?

The *fundamental theorem of arithmetic* states that every natural number greater than 1 can be expressed as a unique product of prime numbers. Then we must ask, what are the prime numbers?

Divisibility

Going back to *Example* 4.13, we have defined a relation D that relates two numbers of the first divides the other. This is more commonly denoted as $a \mid b$, which is the proposition stating that a divides b, i.e. $b \div a$ is a natural number. For example, $4 \mid 12$ since $12 \div 4 = 3$ but $5 \nmid 6$ since $6 \div 5 = 1.25$ which is not a natural number.

Definition 4.19 (Prime number)**.** A natural number p is a *prime number* if it is greater than 1 and the only numbers that divide p are 1 and p. In notation, it is a natural number p satisfying the following proposition:

$$\forall x(x \in \mathbb{N} \text{ and } x \mid p \Rightarrow x = 1 \text{ or } x = p)$$

In other words, p is a prime number if for all natural numbers x, if x divides p, then $x = 1$ or $x = p$.

The dual notion, numbers that are not prime, are known as *composite* numbers.

Example 4.28. 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31 are all prime numbers.

By the fundamental theorem of arithmetic, every number can be written as a *unique* product of composite numbers. For exanple, $2 = 2$, $4 = 2 \times 2 = 2^2$, $20 = 2^2 \times 5$, and so on. Expressing this product for a number is known as *prime factorization*, and is of incredible importance in several areas like cybersecurity and numerical analysis.

Common Divisors and Multiples

From the notion of divisibility we can describe the common factors (divisors) and multiples of two numbers. For example, the *greatest common divisor* of two numbers a and b, usually denoted as $\gcd(a, b)$, is the greatest number that divides both a and b; the dual, the *lowest common multiple*, usually denoted as $\text{lcm}(a, b)$ is the smallest number that is divisible by a and b.

Definition 4.20 (Greatest Common Divisor)**.** A natural number n is the *greatest common divisor* of two natural numbers a and b defined by:

$$n = \gcd(a, b) \Leftrightarrow n \mid a \text{ and } n \mid b \text{ and } \forall x(x \mid a \text{ and } x \mid b \Rightarrow x \leq n)$$

n is the greatest common divisor of a and b if and only if n divides both a and b and for all other numbers x that divide both a and b, $x \leq n$.

Definition 4.21 (Lowest Common Multiple)**.** A natural number n is the *lowest common multiple* of two natural numbers a and b defined by:

$$\mathrm{lcm}(a, b) = n \Leftrightarrow a \mid n \text{ and } b \mid n \text{ and } \forall x(a \mid x \text{ and } b \mid x \Rightarrow x \geq n)$$

n is the lowest common multiple of a and b if and only if a and b both divide n and for all other numbers x that are divisible by both a and b, $x \geq n$.

Example 4.29. $\gcd(18, 24) = 6$ and $\mathrm{lcm}(4, 6) = 12$.

Two numbers a and b are *coprime* if $\gcd(a, b) = 1$. Primality and coprimality testing is incredibly important in a wide variety of areas, such as cybersecurity.

Computing the greatest common divisor and lowest common multiple can be done in a variety of ways. We shall introduce some computational methods eventually (see Euclidean algorithm). The naive way also works; to find the greatest common divisor of a and b, find the set of all divisors of a and the set of all divisors of b, take the intersection, and the maximum of the intersection. Similarly, find the set of all multiples of a (not exceeding $a \times b$), find the set of multiples of b (similarly not exceeding $a \times b$), find the intersection and take the minimum of the intersection.

Another method of doing so is to treat the prime factorization of a number n as a multiset of primes. For example, we know that $20 = 2^2 \times 5$, so we can treat the prime factors of 20 to be the multiset $\{2, 2, 5\}$. Similarly, $15 = 3 \times 5$ so we can treat its set of prime factors as $\{3, 5\}$.

Taking intersections and unions of multisets is slightly different. When determining the intersection $A \cap B$ of multisets A and B, for each element a in A, if a occurs in A x times and a occurs in B y times, then a will occur in $A \cap B$ $\min(x, y)$ times. For example, if $A = \{2, 2, 2, 3\}$ and $B = \{2, 2, 3, 3\}$, since 2 only occurs twice in B and 3 only occurs once in A, then $A \cap B = \{2, 2, 3\}$. When taking unions, for each element a in A or B, if a occurs in A x times and occurs in B y times, then $A \cup B$ will contain $\max(x, y)$ copies of A. Thus, continuing with our earlier example, $A \cup B = \{2, 2, 2, 3, 3\}$ since 2 occurs three times in A and 3 occurs twice in B.

Then, let $F(n)$ produce the set of prime factors of n, and let $\prod X$ be the product of all the numbers in the set X. We get that

$$\gcd(a, b) = \prod(F(a) \cap F(b))$$

and

$$\mathrm{lcm}(a, b) = \prod(F(a) \cup F(b))$$

Example 4.30. From earlier, $F(20) = \{2, 2, 5\}$ and $F(15) = \{3, 5\}$. Then,

$$\gcd(20, 15) = \prod(F(20) \cap F(15)) = \prod\{5\} = 5$$

and

$$\mathrm{lcm}(20, 15) = \prod(F(20) \cup F(15)) = \prod\{2, 2, 3, 5\} = 60$$

4.3.3 Positional Number Systems

As we have seen earlier, every natural number can be expressed as a unique product of prime factors. However, there are is another fact about natural numbers that is particularly interesting to computing; for every number $x > 1$ known as the *radix* or *base*, every natural number can be expressed a unique (ignoring leading 0 coefficients) polynomial $f(x)$ where all coefficients c_i are natural numbers such that $0 \leq c_i < x$.

Example 4.31. Let our radix x be 10. The number 12345 can be expressed as the unique polynomial

$$1 \cdot 10^4 + 2 \cdot 10^3 + 3 \cdot 10^2 + 4 \cdot 10^1 + 5$$

As you can see, the coefficients of the polynomial forms the digits of 12345! This is precisely because the counting system that we use, known as the *decimal* (or base 10) counting system, has 10 digits (hence radix 10), so naturally, the coefficients must form our number! We will write 12345 (as we know it) as 12345_{10} so that we do not confuse the same sequence of digits in other bases. Note that if we choose a different base, the number 12345 as we know it looks different; however all of these representations *mean* the same number. For a precise meaning of the number 12345, we may rely on the von Neumann construction being a set of 12345 elements constructed in the inductive way as shown above.

As such, not only can we say that all numbers are a unique composition of prime factors, we can say that given a radix, all numbers are a unique composition of digits in that radix!

Two other bases are of keen interest to computer scientists: the base 2 or *binary* and the base 16 or *hexadecimal* counting systems. The binary counting system has two digits, 0 and 1, while the hexadecimal system has 16 digits, `0` to `9` and `A` to `F` where `A` is 10 and `F` is 15.

Example 4.32. Let us write the number 127 as polynomials in bases 10, 2 and 16:

$$127 = 1 \cdot 10^2 + 2 \cdot 10^1 + 7 = 1 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 + 1 = 7 \cdot 16^1 + \mathtt{F}$$

Thus, $127_{10} = 1111111_2 = \mathtt{7F}_{16}$.

How do we re-express a number in a different base? This is quite simple. First, recall that when computing $a \div b$, it could be the case that $b \nmid a$, so $a \div b$ gives a rational number. This number will be in the form $x + (y/b)$ where $y < b$. Here, x is known as the *quotient* and y is known as the *remainder*. For example, $9 \div 4$ gives quotient 2 and remainder 1. Then, let's say we are trying to re-express n in base b. To do so,

1. Let $n \div b$ give quotient q and remainder r.
2. Write (from right to left) r.
3. Update the new value of n to be q.
4. Go back to step 1 if $n > 0$, otherwise, what you have written down is the new number.

Example 4.33. Let us express 11_{10} in binary.

1. $11 \div 2$ gives quotient 5 and remainder 1.
2. We write (from right to left) 1, so what we have written down so far is 1.
3. We let n be 5.
4. Repeat. $5 \div 2$ gives quotient 2 and remainder 1.
5. We write (from right to left) 1, so what we have written down so far is 11.
6. We let n be 2.

7. Repeat. $2 \div 2$ gives quotient 1 and remainder 0.
8. We write (from right to left) 0, so what we have written down so far is 011.
9. We let n be 1.
10. Repeat. $1 \div 2$ gives quotient 0 and remander 1.
11. We write (from right to left) 1, so what we have written down so far is 1011.
12. We let n be 0.
13. We are done, so $1011_2 = 11_{10}$.

EXERCISES

4.1. Let $A = \{0, 1, 2, 3\}$. What is the set $\{x + 2 \mid x \in A, x > 1\}$?

4.2. Let $A = \{\emptyset\}$. What is $\mathcal{P}(A)$?

4.3. Let $A = \{\emptyset, \{\emptyset\}\}$ and $B = \{\emptyset\}$. What is $A \cap B$?

4.4. Let $A = \{\emptyset\}$ and $B = \{1, 2\}$. What is $A \cup B$?

4.5. What is $(\mathbb{N} \setminus \{3, 4, 5\})^{\complement}$?

4.6. Let $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$. What is $(A \Delta B) \cup (A \cap B)$?

4.7. Let $A = \{2, 3\}$ and $B = \{3, 2\}$. What is $A \times B$ and $B \times A$?

4.8. Let $A = \{1, 2, 3, 4, 5\}$ and let R be a relation on A given by $R(x, y)$ if $x + 1 < y$. Is R reflexive, symmetric or transitive, or a combination of either?

4.9. Let $A = \{1, 2\}$ and $B = \{3, 4\}$, and let $f = \{(1, 3), (2, 4), (1, 4)\}$. Is f a function from A to B?

4.10. Let $f(x) = x^2$ be a function $\mathbb{R} \to \mathbb{R}_{\geq 0}$ i.e. the codomain of f are the nonnegative real numbers. Is it an injection, surjection, bijection or neither? If it has a left/right inverse, what is it?

4.11. Express 24200 as the product of prime factors, i.e. find the prime factorization of 24200.

4.12. Find $\gcd(9, 15)$ and $\operatorname{lcm}(9, 15)$.

4.13. Express 300 in binary and hexadecimal.

Appendix A

Solutions to Exercises

A.1 CHAPTER 1: ALGEBRA

Exercise 1.1: We have $a = 2b$ and $b = a + 3$, so we can substitute the first equation into the second, getting $a/2 = a + 3$, so $a = 2a + 6$. This gives us $a = -6$. Substitute this back to either equation gives us $b = -3$.

Exercise 1.2: That is incorrect; dividing two natural numbers could give us rational numbers, which are not natural.

Exercise 1.3: Use the formula to find (rather straightforwardly) that the root of this linear equation is $x = 7/5$.

Exercise 1.4: Use the quadratic formula to find that the roots of this quadratic equation are $x = 1$ and $x = -2.5$.

Exercise 1.5: Substitute the RHS $f(x) = 2x^2$ into g to find $g \circ f$, and vice versa. We should get $(g \circ f)(x) = 2x^2 + 4$ and $(f \circ g)(x) = 2(x + 4)^2$.

Exercise 1.6: By the logarithm rules we get that

$$\log_a n = \frac{\log_b n}{\log_b a}$$

, thus we have $\log_b n = \log_b a \log_a n$. Since $\log_b a$ is a constant we thus shown that $g(n) = k \times f(n)$ where $k = \log_b a$.

Exercise 1.7: We can obtain five pairs of numbers that each sum to -24, so the sum is -120.

Exercise 1.8: The limit is 1. We can observe this happening with the sequence $1/2, 2/3, 3/4, 4/5, 5/6, \ldots$.

Exercise 1.9: No it is not. This is because $(1/2)/(1/1) = 1/2$ but $(1/3)/(1/2) = 2/3$, so the ratio of each consecutive term is not constant.

Exercise 1.10: Yes it is. If we create a new series $g(0), g(1), g(2), \ldots$ where $g(x) = (1/2)^x$, we can see that for all n, $|f(n)| < |g(n)|$ so the series necessarily converges, and has a finite sum.

Exercise 1.11: Observe that $n^2-(n-1)^2 = n(n-1)+n-(n-1)(n-1) = (n-n+1)(n-1)+n = 2n-1$

for all n.

$$f(n) = \begin{cases} 0 & \text{if } n = 0 \\ f(n-1) + 2n - 1 & \text{otherwise} \end{cases}$$

A.2 CHAPTER 2: CALCULUS

Exercise 2.1: Let $g(x) = 2x + 3$ and $f(u) = 4^u$. By the chain rule, we get that $(f \circ g)'(x) = 4^{g(x)} \ln 4 g'(x) = 2 \ln 4 \times 4^{(2x+3)}$.

Exercise 2.2: By the product rule, $f'(x) = \frac{d}{dx}(4x^2) \times e^{2x} + 4x^2 \times \frac{d}{dx}(e^{2x}) = 8xe^{2x} + 8x^2e^{2x} = (1+x)(8xe^{2x})$.

Exercise 2.3: Notice that

$$\frac{d}{dx}(x^2 + 3x + 4) = 2x + 3 \qquad \frac{d}{dx}(2x - 1) = 2$$

Thus, by the quotient rule, $f'(x) = ((2x+3)(2x-1) - (x^2+3x+4)(2))/4 = x^2/2 - x/2 - 11/4$.

Exercise 2.4: We can see that $f'(x) = 2x + 2$, and the root of this polynomial is -1, i.e. $f'(-1) = 0$. Thus $x = 1$ is a stationary point. We also know that $f''(x) = 2 > 0$, so $x = 1$ is a local (global) minimum point.

Exercise 2.5: $f'(x) = 4x^3$ and $f''(x) = 12x^2$. $f''(0) = 0$. However, we know that $x = 0$ is a minimum point because for all $x \neq 0, x^4 > 0$. Thus, it is not true that "if x is a minimum point of f, then $f''(x) > 0$."

Exercise 2.6: We shall integrate by substitution. We can rewrite f as $f(x) = (1/3)(x^2 - 3)/(x^3/3 - 3x)$. Furthermore, we can see that $\frac{d}{dx}(x^3/3 - 3x) = x^2 - 3$. Thus, f is the derivative of ln: i.e. $F(x) = (1/3)\ln(x^3/3 - 3x) + c$.

Exercise 2.7: We shall integrate by parts. Let $f(x) = e^{2x}$ and $g(x) = x^2$. Then, $\int (f(x)g(x)dx) = F(x)g(x) - \int (F(x)g'(x)dx)$. We can see that $F(x) = e^{2x}/2$ and $g'(x) = 2x$, thus, we have

$$\int (e^{2x}x^2dx) = \frac{1}{2}x^2e^{2x} - \int (xe^{2x}dx) \tag{A.1}$$

Let us integrate the integral on the RHS by parts again. Let $h(x) = e^{2x}$ and $k(x) = x$; $H(x) = e^{2x}/2$ and $k'(x) = 1$. Since we have $\int h(x)k(x)dx = H(x)k(x) - \int H(x)k'(x)dx$, we get that

$$\int (xe^{2x}dx) = \frac{1}{2}xe^{2x} - \int \frac{e^{2x}}{2}dx \tag{A.2}$$

Substituting Equation A.2 back into Equation A.1 we get

$$\int (e^{2x}x^2dx) = \frac{1}{2}x^2e^{2x} - \frac{1}{2}xe^{2x} + \int \frac{e^{2x}}{2}dx \tag{A.3}$$

Clearly we can see that $\int e^{2x}/2dx = e^{2x}/4$, so substituting this back into Equation A.3 we get

$$\begin{aligned}\int (e^{2x}x^2dx) &= \frac{1}{2}x^2e^{2x} - \frac{1}{2}xe^{2x} + \frac{e^{2x}}{4} + c \\ &= \frac{e^{2x}}{2}\left(x^2 - x + \frac{1}{2}\right) + c\end{aligned}$$

A.3 CHAPTER 3: GEOMETRY

Exercise 3.1: By the cosine rule we have $c^2 = a^2+b^2-2ab\cos C$, thus $c = \sqrt{16+25-18\cos 1.4455} = 6$. We can see that each $f = 2.5a$, $d = 2.5b$ and $e = 2.5c$, so these triangles are similar.

Exercise 3.2: By Pythagoras' Theorem the length of the line segment = $\sqrt{(-4-1)^2+(2--1)^2} \approx 5.83$. The gradient of the line segment is $(-4-1)/(2--1) = -5/3$. The y-intercept $c = -4-(-5/3)(2) = -2/3$. Thus, this line segment is represented by the equation $y = -5x/3 - 2/3$. The angle made by the line segment is $\arctan -5/3 \approx -1.03$ rad.

Exercise 3.3: The area of the polygon is given by

$$\frac{1}{2}\left((3+1)(-1--2)+(4+3)(2--1)+(2+4)(0--2)+(2+1)(0--2)\right) = 21.5$$

Exercise 3.4: $(x+2)^2+(y-1)^2 = 9$.

Exercise 3.5: Simply swap x and y, giving us $x = y^2-4y+5$.

A.4 CHAPTER 4: SETS AND NUMBERS

Exercise 4.1: The set $\{x \in A \mid x > 1\}$ is $\{2,3\}$, so what we should obtain at the end is $\{4,5\}$.

Exercise 4.2: $\{\emptyset, \{\emptyset\}\}$.

Exercise 4.3: The only common element is $\emptyset$, so $A \cap B = \{\emptyset\}$.

Exercise 4.4: $\{1,2,\emptyset\}$.

Exercise 4.5: Removing 3, 4 and 5 from $\mathbb{N}$ then taking its complement (with respect to $\mathbb{N}$) gives us back $\{3,4,5\}$.

Exercise 4.6: $A \,\Delta\, B = \{1,2,4,5\}$, and $A \cap B = \{3\}$. Thus, the result is actually $A \cup B = \{1,2,3,4,5\}$.

Exercise 4.7: $A \times B = B \times A = \{(2,2),(2,3),(3,2),(3,3)\}$. Note that in general, $A \times B \neq B \times A$. We obtained this result only because $A = B$.

Exercise 4.8: $R = \{(1,3),(1,4),(1,5),(3,5),(2,4)\}$. R is clearly neither reflexive nor symmetric, but it is transitive: we see that $(1,3)$ and $(3,5)$ are in R, and so is $(1,5)$.

Exercise 4.9: It is not a function because $f(1) = 3$ and $f(1) = 4$, which is not permissible in a function.

Exercise 4.10: It is a surjection because for every nonnegative real number its square roots (both positive and negative) will be mapped to it by f. It is not an injection because, for example, $f(2) = f(-2) = 4$. Since it is a surjection, its right inverse is $g(x) = |\sqrt{x}|$. For example, $g(4) = 2$, and $f(g(4)) = f(2) = 4$.

Exercise 4.11: $24200 = 2^3 \times 5^2 \times 11^2$.

Exercise 4.12: $\gcd(9, 15) = 3$, $\operatorname{lcm}(9, 15) = 45$.

Exercise 4.13: $300_{10} = 100101100_2 = \mathtt{12C}_{16}$.

Index