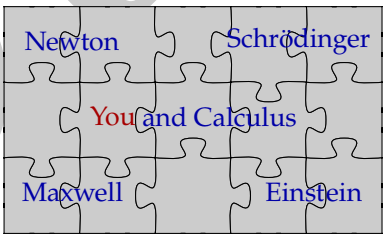
Calculus Done Right

From Arithmetic to Spacetime

By

Dayeol Choi



2025

https://dayeolchoi.com

Contents

Preface

And you will realize the truth, and it shall make you free.

— John 8:32

This book provides an efficient introduction to the essentials of calculus needed for further studies in mathematics (real analysis, abstract algebra, point set topology), physics (especially quantum mechanics), and other quantitative fields such as economics or computer science. The only necessary background is proficiency in grade school arithmetic. If you can work through the arithmetic review in Chapter 1, then you are ready to tackle the rest of the book.

$$\overbrace{\text{Calculus}}^{\textit{what}}\ \underbrace{\text{Done Right}}_{\textit{how}}$$

The emphasis throughout is on exposing the **simplicity** that lies at the core of the subject. In particular, calculus really is the simplest thing one could come up with to create an arithmetic of velocities and displacements!

$$i\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = H|\Psi\rangle \qquad \nabla\cdot\boldsymbol{E} = \rho/\epsilon_0 \qquad \nabla\times\boldsymbol{E} = -\partial_t\boldsymbol{B} \qquad \nabla\cdot\boldsymbol{B} = 0 \qquad \nabla\times\boldsymbol{B} = \mu_0\boldsymbol{J} + \mu_0\epsilon_0\partial_t\boldsymbol{E}$$

Now, a language embeds aspects of its background and culture. Concepts that are intimately expressed in a single word in one language may need several tomes to spell out in another. Calculus is no different, and we will see that making the simplest of extensions will lead us to Schrödinger's equation and the probabilistic quantum world. When we discuss *multivariable* calculus, we will similarly be lead to Maxwell's equations, where our attempt to make sense of its most straightforward consequence will lead us to special relativity and a radical rethinking of space and time.

So simple that **you could have made these discoveries yourself**. Let us begin!

Dayeol Choi

1

Arithmetic

Let's start at the very beginning. If you are confident in your abilities in arithmetic and basic algebra, feel free to jump ahead to Chapter 2. What is 1 + 1? It doesn't get any easier than that. Of course the answer to 1 + 1 is 2. Now, these numbers must mean something. For example, we might be counting the number of apples in a pantry, and we observe that there is one apple next to another apple, and so we conclude that there are 2 apples.

Very good, 1 + 1 = 2 and in particular 1 apple + 1 apple = 2 apples. If 1 + 1 = 2, what is the answer to 1 apple + 1 orange? Since 1 + 1 = 2, do we conclude the answer is 2? No, because we are trying to add apples to oranges. When we say 1 + 1 = 2, we are assuming that each quantities are compatible. Thus the answer to 1 apple + 1 orange is that the sum is unresolvable.[1] An analogous question would be: what is 1 meter plus 1 second? Once again, such questions cannot be answered as their units do not match. Units matter, and we will draw on this key insight over and over again.

1.1 Units

All physical theories must have something to say quantitatively about the world around us. In order to communicate coherently about real world objects, we must agree on a set of units. For example, the distance from one café to another might be 50 meters. Or is it 164 feet?

This is one case where trying to please everyone turns out to be helpful. In order to make everyone happy, let us agree to refer to all sorts of distance measurements as a **L**ength. Thus the height of a building and the distance from the earth to the sun are both instances of **L**engths.

Now, in order to indicate speed, we usually divide something by time. For example, 6 slices of pizzas per hour might mean the speed at which pizza slices were consumed. Similarly, the distance from the earth to the sun, divided by the time it takes for light to hit the earth from the sun indicates the speed of light. Thus dividing a length by time gives us speed:

$$\frac{\textbf{L}\text{ength}}{\textbf{T}\text{ime}} = \text{Speed}.$$

Some like to use seconds to measure time, others like to use hours; we will call all time measurements **T**ime. Suppose I ate 6 slices of pizza per hour for 2 hours. Then, I ate a total of: 6 slices/hour

[1] If you think there is another possible answer, you are right! We will return to this point later.

$\times$ 2 hours = 12 slices. If we multiplied the speed of light by 1 year, which is a **T**ime, then

$$\underbrace{\frac{\textbf{L}\text{ength}}{\cancel{\textbf{T}\text{ime}}}}_{\text{speed of light}} \times \underbrace{\cancel{\textbf{T}\text{ime}}}_{\text{1 year}} = \underbrace{\textbf{L}\text{ength.}}_{\text{one lightyear}}$$

Another fundamental type of measurement is mass, which for now we will use interchangeably with weight. Some folks use kilograms, others use pounds. We will refer to these as **M**ass.

Einstein told us that Energy is mass times the speed of light squared. In symbols, this is $E = mc^2$, where E is energy, m is mass, and c is the speed of light. Notice that when using symbols, we omit the $\times$ symbol. Thus $E = mc^2$ means $E = m \times c^2$, which in turn means $E = m \times c \times c$. Since m is a **M**ass and c is **L**ength divided by **T**ime (speed),

$$\begin{aligned}\text{Energy} &= \textbf{M}\text{ass} \times \left(\frac{\textbf{L}\text{ength}}{\textbf{T}\text{ime}}\right)^2 = \textbf{M}\text{ass} \times \frac{\textbf{L}\text{ength}}{\textbf{T}\text{ime}} \times \frac{\textbf{L}\text{ength}}{\textbf{T}\text{ime}} \\ &= \textbf{M}\text{ass} \times \frac{\textbf{L}\text{ength}^2}{\textbf{T}\text{ime}^2}.\end{aligned}$$

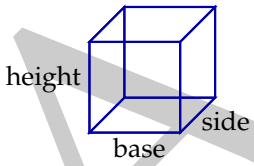


Figure 1.1: A cube with a base, side, and height.

Here is my first challenge for you. The main challenge with this one is getting your pencil and paper out. Later ones may not be this easy!

Challenge 1

(a) We can simplify arithmetic expressions using cancellation. For example,

$$\frac{5 \times 10 \times 3}{3 \times 2} \times 2 = \frac{5 \times 10 \times \cancel{3}}{\cancel{3} \times \cancel{2}} \times \cancel{2} = 5 \times 10.$$

Try simplifying the following expressions. By convention, $2^2 = 2 \times 2$ and $5^3 = 5 \times 5 \times 5$.

$$5^5 \times \frac{3}{5^3} \qquad\qquad \frac{5^5 \times \frac{3}{5^3}}{2^2}$$

(b) A volume of a cube is the base of the cube multiplied by the side of the cube and the hight of the cube (see Figure 1.1). A density of a substance is its **M**ass divided by volume. Use cancellation to simplify the following expression as much as possible. Do you recognize it?

$$\frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2}.$$

(c) Remember that we cannot add apples to oranges. Thus the mathematical expression

$$2 \text{ oranges} + 3 \text{ apples}$$

makes no sense. Similarly, we cannot add a **L**ength to **T**ime. Identify which of the following arithmetic expressions are valid:

$$\frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2} + \frac{\textbf{L}\text{ength}}{\textbf{T}\text{ime}}, \quad \frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2} + \text{Energy}, \quad \frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2} + \frac{\text{Energy}^2}{\textbf{T}\text{ime}}.$$

What we have found in Challenge 1 is that the laws of nature are constrained rather strongly. For example, if Einstein told us $E = mc + m/c$ or $E = m + c^2$, there's no way either could be true because the units cannot match. Let's put this idea into use.

Below is an image of the Trinity nuclear test, the first detonation of a nuclear weapon in history.

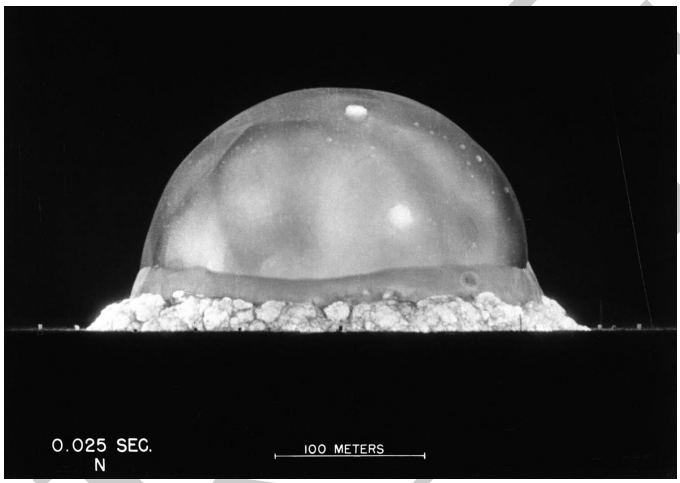


Figure 1.2: Courtesy of US Government Defense Threat Reduction Agency, successor agency to the Manhattan Project.

As we can see from the test image, the energy from the bomb is released in what appears to be a spherical blast. The **radius** of a sphere is the distance from the center of the sphere to its boundary. Thus radius is a **L**ength. The *radius* of the blast (let's label this with the letter R) will be proportional to the *energy* of the bomb (we'll refer to this with the letter E), and the **T**ime since blast, t.[2] If the bomb was surrounded by dense material, such as concrete and steel, we'd imagine the blast radius will be smaller. On the other hand, if the bomb was surrounded by less dense material like air, the blast radius will be larger. We will refer to the *density* of the surrounding material with the Greek letter ρ (rho). Below is a table summarizing what we have. Later on, we will use the shorthand appearing in the fourth column to save on space. As is customary when using symbols, $\mathrm{ML}^2/\mathrm{T}^2$ omits the '×' symbol.

[2]Within seconds of the blast, a larger value of t will lead to a larger blast radius R.

variable	meaning	unit type	unit shorthand
R	Radius (of a Blast)	**L**ength	L
E	Energy (of a Bomb)	**M**ass $\times \frac{\text{Length}^2}{\text{Time}^2}$	ML^2/T^2
t	Time passed	**T**ime	T
ρ	Density (of the surrounding)	**M**ass/**L**ength^3	M/L^3

A **function** is object that takes in an input and yields an output. For example, if $f : x \mapsto 2x$, then the function f takes in a number x and returns $2x$. We need some predictability, thus whenever we input the same value to a function, the function is required to return the same output. A function can sometimes be represented as a formula. For example, our function f can also be written as $f(x) = 2x$, where the left side of the equation is a formula to the function on the right side.

What we would like to know is the simplest formula for the energy of a bomb E, given that we know its blast radius R at time t with surrounding material density ρ (there could be other contributing factors, but the ones we have written down look like they are the most important). From Challenge 1, we know that there is a simple formula to express such cases. Since

$$\text{Energy} = \frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2}$$

we see that energy E must be proportional to $\frac{R^5\rho}{t^2}$, by considering the units involved. I say proportional to, because the simplest formula for energy E could be

$$E = 3\frac{R^5\rho}{t^2}, \text{ or } E = 3.14\frac{R^5\rho}{t^2}, \text{ or } E = 5\frac{R^5\rho}{t^2}, \text{ or } \ldots.$$

We cannot rule out any such possibilities because a number itself has no units. 5 meters is a Length with a unit of meter, but the number 5 has no units. We call the numbers 0, 1, 2, 3, 4, ... that we use to count, the **natural numbers**. The **integers** are those numbers consisting of the natural numbers and its negative counterparts $-1, -2, -3, \ldots$ (the convention is that $-0 = 0$). The numbers $1, 2, 3, \ldots$ are also called positive integers.

We make explicit our ignorance by including a number β, as shown below. Without additional information, we cannot know β, only that it has no units.

$$E = \beta\frac{R^5\rho}{t^2} \tag{1.3}$$

Challenge 2

(a) Since we do not know what β is, let us assume $\beta = 1$ for now. Does Equation 1.3 make sense? Is an increase in blast radius associated with more energy? If we had a very dense surrounding material (thus a high density ρ), what would that tell us about the energy? What if the time to reach a specific blast size was smaller, what would that tell us about energy E?

(b) Using a calculator, the nuclear test image, and Equation 1.3 with $\beta = 1$, estimate the energy released by the trinity experiment. We will use meters (m) for radius R, kg/m^3 for density ρ, and seconds (s) for time t. Thus E has the unit $\text{kg}\cdot\text{m}^2/\text{s}^2$.[3] Just eyeball the value for radius R, and use the fact that air density ρ is about $1\text{kg}/\text{m}^3$.

[3] The $\cdot$ is a shortened form of $\times$. We need a multiplication symbol because units aren't always single letters.

(c) The unit of energy $\text{kg}\cdot\text{m}^2/\text{s}^2$ is called a **joule** (symbol J). The standard convention for explosive energy released by a fission weapon like Trinity is thousands of tons of TNT, called **kilotons**. Using the fact that 4.2×10^9 joules is about 1 ton of TNT, convert your answer in part (b) into kilotons. This is just an estimate, feel free to round to the nearest kiloton.

(d) Look up the yield of the Trinity nuclear test online and compare with your result from (c). Use it to find the number β, rounding to the nearest integer.[4]

I hope that Challenge 2 gives a first indication that arithmetic is more than what we punch into calculators. Our next step is to figure out how we can obtain the formula

$$E = \frac{R^5 \rho}{t^2}$$

in the first place. To do this, we will need to review a bit of multiplication.

1.2 Exponentiation

The multiplication $13 \cdot 9$ can be done relatively easily in our heads if we remember that

$$13 \cdot 9 = (10 + 3) \cdot 9 = 10 \cdot 9 + 3 \cdot 9 = 90 + 27.$$[5]

Since $13 \cdot 9 = 9 \cdot 13$, an entirely equivalent calculation is

$$13 \cdot 9 = 9 \cdot 13 = 9 \cdot (10 + 3) = 9 \cdot 10 + 9 \cdot 3 = 90 + 27.$$

In order to make general mathematical statements, we will almost always use symbols in place of numbers, just like we used R to refer to a radius (of a blast). Suppose we have three numbers on hand which we denote by the letters a, b, and c. Then the above calculations may be expressed as

$$(a + b) \cdot c = a \cdot c + b \cdot c \quad \text{and} \quad a \cdot (b + c) = a \cdot b + a \cdot c.$$

We will usually skip the '$\cdot$' when using symbols and we will write the above as $(a + b)c = ac + bc$ and $a(b + c) = ab + ac$. Thus $a(b + c + d) = ab + ac + ad$ and $(i + j + k + l)z = iz + jz + kz + lz$.

Challenge 3

(a) Use the fact that $(a+b)(c+d) = ac+ad+bc+bd$ to show that $(10+x)(10+y) = 10\cdot(10+x+y)+xy$. We will apply it and a slightly tweaked version of it in part (b).

(b) Do the multiplication $16 \cdot 14$ in your head. Next, do the multiplication $116 \cdot 114$ in your head.

Now that we have reviewed the multiplication of two numbers, let us review the multiplication of a finite collection of numbers. We know that $1000 = 10 \cdot 10 \cdot 10$ and $10000 = 10 \cdot 10 \cdot 10 \cdot 10$. As a convenient notation, let us agree to write $1000 = 10^3$ and $10000 = 10^4$ instead. Similarly, $0.1 = 1/10$ is written as 10^{-1}, which means that $0.0001 = 0.1 \cdot 0.1 \cdot 0.1 \cdot 0.1 = 10^{-4}$. This bookkeeping convention is called **exponentiation** and we typically indicate this using the word **power**. For example, 10^4 is 10 to the power of 4 and 10^{-4} is 10 to the power of -4. The number we are exponentiating is called the **base**. Thus 10 is the base of both 10^4 and 10^{-4}.

Below are the exponentiation rules. The letters a and d are positive integers which we use as bases. The letter b and c are the powers and they could be integers or fractions of nonzero integers.

[4] G.I. Taylor was one of the first outside the Manhattan Project's core group to estimate the yield of Trinity based on blast photos. This was in 1950 when not only was Trinity's yield a *Top Secret*, but only one country in the world had any nuclear arsenal. G.I. Taylor did not use dimensional analysis to obtain his results.

[5] From now on, we will prefer using the symbol $\cdot$ instead of $\times$.

- $\boxed{a^0 = 1}$ as in $3^0 = 1$ and $\mathbf{L}\text{ength}^0 = 0$.
- $\boxed{a^{-b} = \dfrac{1}{a^b}}$ as in $8^{-1} = \frac{1}{8}$ and $\mathbf{T}\text{ime}^{-2} = \frac{1}{\mathbf{T}\text{ime}^2}$.[6]
- $\boxed{a^b \cdot a^c = a^{b+c}}$ as in $6^2 \cdot 6^5 = 6^7$ and $\mathbf{M}\text{ass}^3 \cdot \mathbf{M}\text{ass}^{-4} = \mathbf{M}\text{ass}^{-1}$.
- $\boxed{(a \cdot d)^b = a^b \cdot d^b}$ as in $(2 \cdot 5)^4 = 2^4 \cdot 5^4$ and $(\mathbf{M}\text{ass} \cdot \mathbf{T}\text{ime})^2 = \mathbf{M}\text{ass}^2 \cdot \mathbf{T}\text{ime}^2$.
- $\boxed{(a^b)^c = a^{bc}}$ as in $(2^{-3})^4 = 2^{(-3)\cdot 4} = 2^{-12}$ and $(\mathbf{L}\text{ength}^2)^4 = \mathbf{L}\text{ength}^8$.

When the fraction $\frac{1}{2}$ is used as a power, the number $a^{1/2}$ is usually written $\sqrt{a}$. Thus, the exponentiation rules allow us to write the following.

$$5^{3/2} = (5^3)^{1/2} = \sqrt{5^3}$$

More generally, for each positive integer n, one sees $a^{1/n}$ written as $\sqrt[n]{a}$, called the n**th root**. For example, $3^{1/5} = \sqrt[5]{3}$. The 2nd root is usually called the **square root**, thus $\sqrt{5^3}$ is the square root of 5^3. The key take away from fractional powers is that using the fifth exponentiation rule:

$$\left(a^{p/q}\right)^q = a^{(p/q)\cdot q} = a^p$$

where a is not a negative number and p, q are positive integers. For example, $\left(54^{5/3}\right)^3 = 54^5$.

Simultaneous equations

We are now ready to obtain a formula for the nuclear blast yield. For reasons that will be clear much later, we will first calculate a formula for the radius of a nuclear blast. Hence, we will first find how the radius of a nuclear blast R is related to the energy of a bomb E, time since blast t, and surrounding material density ρ.

As we have seen before, equations such as

$$R = E + t + \rho \quad \text{or} \quad R = E + t \cdot \rho$$

are not possible because the units don't match. For example, in the former case we know that it makes no sense to add energy to time and density. On the other hand, the simplest formula that could work is

$$R = d \cdot E^a \cdot t^b \cdot \rho^c, \tag{1.4}$$

where a, b, c, and d are unknown numbers. The first three are the ones we use to make the units match in both sides of the equation. The number d on the other hand has no units; such numbers are called **dimensionless constants**. The number β in Equation 1.3 was a dimensionless constant.

[6] From this rule it follows that if a is nonzero, then $\frac{b}{1/a} = ab$. As an example, if we divide three pizzas, each into eight slices, then there will be twenty four slices: $\frac{3}{1/8} = 3 \cdot 8 = 24$.

For instance, notice that R is a length, so it has no units of time T.[7] However, on the right side of Equation 1.4, the variables E and t include the unit T. This means we need to find numbers a and b such that the unit T cancels out. Similarly, the radius R is independent of mass M. But the variables E and ρ have unit M. So we will have to find a and c that cancels out the unit M.

To proceed, let us convert Equation 1.4 into an equation consisting soley of units. Since the number d has no units, we'll put it aside for now. Using our table of units from earlier, we can write

$$\mathrm{L} = \left(\frac{\mathrm{ML}^2}{\mathrm{T}^2}\right)^a \cdot \mathrm{T}^b \cdot \left(\frac{\mathrm{M}}{\mathrm{L}^3}\right)^c.$$

We use the exponentiation rules from before to simplify the right side of the equation above as

$$\left(\frac{\mathrm{ML}^2}{\mathrm{T}^2}\right)^a \cdot \mathrm{T}^b \cdot \left(\frac{\mathrm{M}}{\mathrm{L}^3}\right)^c = \frac{\mathrm{M}^a\mathrm{L}^{2a}}{\mathrm{T}^{2a}} \cdot \mathrm{T}^b \cdot \frac{\mathrm{M}^c}{\mathrm{L}^{3c}} = \mathrm{M}^a \cdot \mathrm{M}^c \cdot \frac{\mathrm{T}^b}{\mathrm{T}^{2a}} \cdot \frac{\mathrm{L}^{2a}}{\mathrm{L}^{3c}} = \mathrm{M}^{a+c} \cdot \mathrm{T}^{b-2a} \cdot \mathrm{L}^{2a-3c}.$$

Thus

$$\mathrm{L} = \mathrm{M}^{a+c} \cdot \mathrm{T}^{b-2a} \cdot \mathrm{L}^{2a-3c} \text{ or equivalently, } \mathrm{M}^0 \cdot \mathrm{T}^0 \cdot \mathrm{L}^1 = \mathrm{M}^{a+c} \cdot \mathrm{T}^{b-2a} \cdot \mathrm{L}^{2a-3c}.$$

In order to make this equality hold, we need to set the power of M at $a + c = 0$, the power of T at $b - 2a = 0$, and the power of L at $2a - 3c = 1$.

The first requirement tells us that $a = -c$. Hence $2a = -2c$, and plugging this into the third requirement, we have $1 = 2a - 3c = -5c$. Thus $c = -1/5$ and $a = 1/5$. The only thing left is to find b, so let us look at the second requirement: $b - 2a = 0$, which is equivalent to $b = 2a$ (by adding $2a$ to both sides). Since $a = 1/5$, we have $b = 2a = 2/5$. Therefore,

$$L = \left(\frac{\mathrm{ML}^2}{\mathrm{T}^2}\right)^{1/5} \cdot \mathrm{T}^{2/5} \cdot \left(\frac{\mathrm{M}}{\mathrm{L}^3}\right)^{-1/5},$$

or in our original equation form

$$R = d \cdot E^{1/5} \cdot t^{2/5} \cdot \rho^{-1/5}.$$

We now know the relationship between, say, the energy contained in a nuclear bomb and its blast radius. Let us invert the relationship so that we have energy E expressed as a combination of R, t and ρ. Taking the power of 5 to both sides, we get

$$R^5 = d^5 \cdot \frac{E \cdot t^2}{\rho}.$$

Now multiply each side by $\frac{\rho}{d^5 \cdot t^2}$ and let $\beta := 1/d^5$ to get

$$E = \beta \frac{R^5 \cdot \rho}{t^2}, \text{ where } \beta \text{ is a dimensionless constant.} \tag{1.5}$$

Because we are doing arithmetic with units, unit-less numbers (dimensionless constants) cannot be determined by this procedure. Using some additional information in Challenge 2, we found out that β rounds to 1.

[7] We are using the unit shorthand: **M**ass is M, **L**ength is L, and **T**ime is T.

Although our process for finding Equation 1.5 was fairly long, the main problem was that of finding three unknown numbers a, b, and c such that the equations

$$a + c = 0,\ b - 2a = 0,\ \text{and } 2a - 3c = 1$$

are all satisfied simultaneously.[8]

The act of taking a problem, determining the relevant factors and their corresponding units, and using these to investigate relationships between the factors is called **dimensional analysis**.[9] This is a useful skill, and I will be counting on you to do your own dimensional analysis later on. Rest assured, all dimensional analysis we will encounter in this book are *much* simpler than the Trinity problem.

[8] There are three equations that must be satisfied, because we need to make sure that the units of mass, length, and time each match up. Furthermore, there are three unknown numbers (called a, b, c here) because there are three variables (energy E, time t, and density ρ) that form what we want (radius R).

[9] Why not call it unit analysis? Because unlike meters, kilograms, and seconds, **L**ength, **M**ass, and **T**ime are not strictly speaking, units. We will call these **dimensions**, hence the name: *dimensional* analysis.

2

Differentiation

2.1 Arithmetic of Velocities

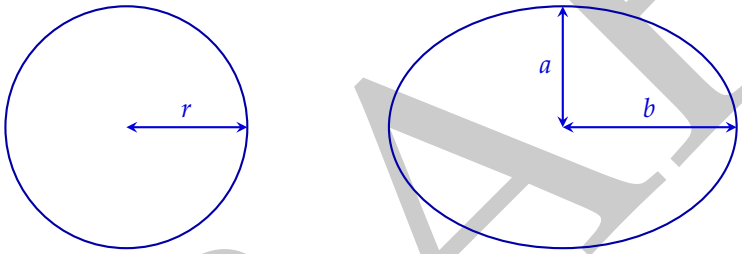


Figure 2.1: A circle of radius r, and an ellipse of height $2a$ and width $2b$.

Let us begin with a review of dimensional analysis.[1] The formula for an area of a circle of radius r is given by πr^2. What if someone told you that the formula is actually πr^3 or πr? That would not make any sense, because if the circle had its radius measured in meters, we would expect its area to have the units of meter^2, not meter^3 or simply meter. This is the idea behind dimensional analysis: we check to see if the units make sense.

Since there are many different units in use that are interchangeable, we will refer to meters, feet, etc by the generic term “Length", and seconds, hours, etc by the generic term “Time".

Now, it is not possible to simply check the units to get the final answer. For example, we can expect that an area of a circle of radius r will be given by a formula proportional to r^2, but we cannot know the factor π. Thus we cannot rule out the possibility that the area of a circle is given by $2r^2$, to take an example, by only using dimensional analysis. Some additional information must be available. Numbers like 2 and π which have no units, and cannot be figured out with dimensional analysis are called dimensionless constants. The generic terms “Length" and “Time" which represent concrete units of measurement are called dimensions. We use dimensions instead of units because we want the results to be the same, regardless of the exact units we may choose.

[1] If you are looking for more, see Sanjoy Mahajan’s excellent *Street-Fighting Mathematics*.

For example, the formula for an area of a circle should stay the same whether we measure radius in meters or feet.

With this limitation in mind, let us see if we can guess the formula for an area of an ellipse, a shape shown on the right of Figure 2.1. There are two variables we can work with: a and b, each of which we will assign the dimension of Length. The formula of an *area* should have the form of Length^2, and so let us consider the simplest ways we could combine the variables a and b to get such a combination. There are two such simple possibilities: c_1ab, and $c_2a^2 + c_3b^2$, where the numbers c_1, c_2, and c_3 are dimensionless constants.[2] Already we can make a simplification. The area of an ellipse should not depend on the label "a" and "b"—in other words, if we flip the diagram of the ellipse in Figure 2.1 so that the height is the width and vice versa, then the area must remain the same, even though a and b are switched. Therefore, the constants c_2 and c_3 must be the same.

We can rule out candidate formulas by looking at some simple cases. Consider the extreme case where $a := 0$ and $b := 10$. Of course, such an ellipse cannot exist in the physical world, but a formula for an area should capture the fact that such an ellipse will occupy zero area. For the first candidate, $c_1ab = c_1 \cdot 0 \cdot 10 = 0$, which behaves as expected. However, the second candidate fails unless $c_2 = 0$, since $0 = c_2a^2 + c_2b^2 = c_20^2 + c_210^2 = c_210^2$.[3]

With one candidate left, we guess that an area of an ellipse of height $2a$ and width $2b$ is given by the formula c_1ab. This is as far as dimensional analysis will get us. However, we have some extra information: a circle is an example of an ellipse with $a = b$. Thus if $a = b$, our formula for an area of an ellipse should be πa^2. We therefore conclude that $c_1 := \pi$, and our final guess is that the area of an ellipse is given by πab. We will later verify the correctness of this formula using calculus.

Differentiation Rules

Calculus is like a car, it can get us to places we never thought we could be at, with far less effort than we would expect. To get somewhere, we need at least two piece of information—how far away it is, and how long it will take for us to get there. The former pertains to the concept of *displacement* needed, while the latter relates to *velocity*.

Everyone moves about, hence the concept of velocity and displacement are universal. Using dimensional analysis, we can get a huge mileage out of simply applying arithmetic to units. This gives us a strong suspicion that applying arithmetic to other objects may turn out to be fruitful. So here is what we will do. Our goal will be to create an arithmetic of velocities and displacements. We will begin with velocity, because velocity is necessary to exhibit displacement.

To describe velocity, or any kind of motion, we will use functions. The simplest type of functions one could think of are those that keep track of an object's position at each time. The simplest of such functions will be a position function for an object that stays completely still at a location. The next simplest would be a position function for an object that moves at a constant velocity of 1 meter/second in one direction. These two functions are graphed below in Figure 2.2. The **x-axis** of a graph denotes the horizontal line used to represent the input variable's values. In the graphs below, the x-axis is used to represent the input variable "time" t (measured in seconds). The **y-axis**

[2] We could contemplate formulas like $c_4a^3/b + c_5b^{10}/a^8$ or $c_6b^2 + c_7ab$, but these are not the kind of simple formula we are looking for. In any case, these can be ruled out using the methods we use below.

[3] Notice we have replaced the constant c_3 by the constant c_2 because they must have the same value.

of a graph denotes the vertical line used to represent the output variable's values. In the graphs below, the y-axis is used to represent th output variable "position" x (measured in meters).

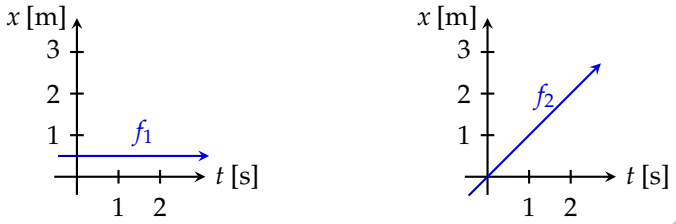


Figure 2.2: (Left) A position function f_1 of a stationary object at position 0.5 m. (Right) A position function f_2 of an object moving at a constant velocity of 1 m/s.

The first position function $f_1 : t \mapsto 0.5$, has velocity 0 m/s for all time, while the second position function $f_2 : t \mapsto t$, has velocity 1 m/s for all time.[4] We will denote the velocity function of an object by adding a $'$ symbol to the object's position function. Thus we write $f_1' : t \mapsto 0$ or equivalently $f_1'(t) = 0$ because there is no motion in our first object, and so the velocity function of f_1 always outputs zero. We say that f_1' is the **zero function**. Even if our stationary object was placed somewhere else, thus shifting our graph of f_1 up or down, it will still be the case that f_1' is the zero function. Thus if a function f is a constant function that outputs the same value for each input t, then

Constant Rule: $f' : t \mapsto 0$.[5]

On the other hand, $f_2'(t) = 1$, because the velocity function of f_2 always outputs 1 (m/s).

Sum rule

We now turn to arithmetic. First, let us try addition. What can dimensional analysis tell us about $(f + g)'$? Taking f and g to be position functions as above, we see that their sum $f + g$ will have outputs of dimension Length. A velocity function $(f + g)'$ will then have outputs of dimension Length/Time. The simplest formula that achieves this is the formula $(f + g)' = c_1 f' + c_2 g'$. The order in which we take the addition should not change the result, so we note that $c_1 = c_2$. Like in the case of the ellipse, we can conjure up an example to help us determine the dimensionless constant c_1. Take f to be the zero function so that $f' = 0$ and $f + g = 0 + g = g$. Hence, $g' = (f + g)' = c_1 f' + c_1 g' = 0 + c_1 g' = c_1 g'$. We see that the constant c_1 is 1, and we have the sum rule.

Sum Rule: $(f + g)' = f' + g'$.

For subtraction, define the function $h : t \mapsto -g(t)$ that flips the sign of the outputs of function g. Applying the sum rule gives $(f - g)' = (f + h)' = f' + h' = f' + (-g') = f' - g'$.

Subtraction Rule: $(f - g)' = f' - g'$.

[4] The notation $f_1 : t \mapsto 0.5$ means the function f_1 turns each input t into 0.5. It is equivalent to writing $f_1(t) = 0.5$. Similarly, $f_2 : t \mapsto t$ means the function f_2 takes each input t and outputs t. It is also written $f_2(t) = t$.

[5] This can also be written $f'(t) = 0$, or equivalently as $f' = 0$.

Product rule

Next, we consider products of position functions f and g. Now, taking a product of position functions is a little weird. For one thing, if function h is the product of the position function f and g, then the ouputs of h will have dimension Length^2 (same as an area function), so h is no longer a position function. This means that it is odd to speak of a velocity function of h. Nevertheless, we can still talk about the rate of change of functions, so instead of speaking about velocity functions, we will speak of **derivatives**. Suppose we have a function that takes inputs with dimension $\diamond$ and outputs quantities of dimension $\heartsuit$. Then the rate of change (the derivative) of the function as we vary inputs (of unit $\diamond$) will have the dimension $\heartsuit/\diamond$. For example, consider a position function whose input is of dimension Time and output is of dimension Length. Its derivative will have dimension Length/Time, just as we expect from a velocity function.[6]

We will write fg to mean the product of functions f and g. That is, the function fg takes an input t and outputs $f(t) \cdot g(t)$.[7] If the dimension of fg is Length^2 and the dimension of the inputs of fg is time, then the derivative $(fg)'$ will have dimension Length^2/Time.

Immediately, we see that the formula for $(fg)'$ cannot be of the form $cf'g'$ for some dimensionless constant c. This is because $cf'g'$ has the dimension Length^2/Time^2, which has an extra division by Time. Instead, the simplest ways we can use the functions f, f', g, g' and combine them to get dimension Length^2/Time are the following three options.

$$(fg)' = c_1(f^2)' + c_2(g^2)' \qquad (fg)' = c_3 ff' + c_4 gg' \qquad (fg)' = c_5 f'g + c_6 fg'$$

The product function fg is the same as the product function gf because the order of multiplication does not matter. Since the labels f and g are interchangeable, we have $c_1 = c_2$, $c_3 = c_4$, and $c_5 = c_6$.

Recall that we were able to narrow down the options when guessing a formula for an area of an ellipse by considering an ellipse with 0 thickness. We can also narrow down our current options by considering the case where f is the zero function. Then $(fg)(t) := f(t)g(t) = 0g(t) = 0$, and since fg is a constant function, the derivative function $(fg)'$ must be the zero function. This fails to be captured by the first two options: $c_1(f^2)' + c_1(g^2)'$ and $c_3 ff' + c_4 gg'$, because we may choose the function g so that each expressions are not the zero function. The only possibility left is the formula $(fg)' = c_5 f'g + c_5 fg'$.

Once again we will examine a simple case to find the dimensionless constant c_5. Define f to be the function $t \mapsto t$ and let $g := 1$, the constant function $t \mapsto 1$. Then $(fg)(t) := f(t)g(t) = t \cdot 1 = t$, and so $(fg)' = 1$. On the other hand, since $f' = 1$ and $g' = 0$, we find that

$$1 = (fg)' = c_5 f'g + c_5 fg' = c_5 \cdot 1 \cdot 1 + c_5 \cdot t \cdot 0 = c_5 + 0 = c_5.$$

Therefore, the dimensionless constant c_5 is one, and we have the product rule:

Product Rule: $(fg)' = f'g + fg'$.

Finally, we discuss the division operation. Consider two functions f and g. Suppose $g(0) = 0$; then $f(0)/g(0)$ is undefined, and so f/g cannot be defined. We cannot divide function f by function

[6] We will equate a function's dimension with the dimension of the function's outputs.

[7] More succinctly, $fg : t \mapsto f(t)g(t)$, or equivalently, $(fg)(t) := f(t)g(t)$.

g if g outputs the value 0 at any point in time. To prevent this, we will need to assume that g is always nonzero, so that for each t, the value $1/g(t)$ is defined. By cancellation, the product function $\left(\frac{f}{g}\right) g = f$. Apply the product rule to the product function $\left(\frac{f}{g}\right) g$ to get

$$f' = \left(\left[\frac{f}{g}\right] g\right)' = \left(\frac{f}{g}\right)' g + \left(\frac{f}{g}\right) g'.$$

This gives us an equation $f' = \left(\frac{f}{g}\right)' g + \frac{fg'}{g}$ that we can solve for $\left(\frac{f}{g}\right)'$. Subtract the second term in the right side from both sides of the equation to get

$$f' - \frac{fg'}{g} = \left(\frac{f}{g}\right)' g.$$

Now, multiply both sides by the function $1/g$ and we have

$$\frac{f'}{g} - \frac{fg'}{g^2} = \left(\frac{f}{g}\right)'.$$

Since $\frac{f'}{g} = \frac{f'g}{g^2}$, the left side can be written as one expression: $\frac{f'g - fg'}{g^2}$. The rule for division is then

Quotient Rule: $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$.

Power rule

Next, we examine functions of the form $f : x \mapsto x^k$, where k is a natural number.[8] We are free to choose the dimension of our input variable. To change things up, this time let us assume a dimension of Length for the input x. The outputs of function f will then have dimension Length^k. This means that the derivative of f will have dimension Length^{k-1}.[9] Our simplest guess is then

$$f'(x) = cx^{k-1}.$$

Now let us try a few examples. If $k = 0$, then $f(x) = x^0 = 1$ by convention, and so $f(x) = 1$, with $f'(x) = 0$ by the constant rule. If $k = 1$, then $f(x) = x^1 = x$, and so $f(x) = x$, with $f'(x) = 1x^0$. If $k = 2$, then $f(x) = x^2$, and we apply the product rule to get $f'(x) = (x \cdot x)' = 1 \cdot x + x \cdot 1 = 2x^1$. If $k = 3$, then $f(x) = x^3$, and applying the product rule gives

$$f'(x) = (x \cdot x^2)' = 1 \cdot x^2 + x \cdot (x^2)' = x^2 + x \cdot 2x^1 = x^2 + 2x^2 = 3x^2.$$

We see that the constant c depends on the value of k, so we will take c to be a dimensionless function of k. In particular, $c(0) = 0$, $c(1) = 1$, $c(2) = 2$, and $c(3) = 3$. The pattern appears to be $c(k) := k$ and so our final guess is that

$$(x^k)' = kx^{k-1}. \tag{2.3}$$

[8] Natural numbers are numbers we use to count the number of objects with. They consist of: 0, 1, 2, 3, ...

[9] This is because Length^k/Length equals Length^{k-1}.

There is every possibility that this formula breaks down and fails to work for some value of $k > 3$. So let S denote the collection of natural numbers for which Formula 2.3 above fails to hold.

Ideally the collection S is an empty collection, but if it is not, then there will be a natural number in the collection S which is the smallest.[10] Call this number n. Since n is in the collection S, our formula will fail to hold for the number n. However, because the number $n-1$ is smaller than n, it is not in the collection S. Thus our Formula 2.3 will work for the natural number $n-1$, giving us $(x^{n-1})' = (n-1)x^{n-2}$. Applying the product rule to the identity $x^n = x \cdot x^{n-1}$ and using our formula $(x^{n-1})' = (n-1)x^{n-2}$ gives the following.

$$(x^n)' = (x \cdot x^{n-1})' = 1 \cdot x^{n-1} + x(x^{n-1})' = x^{n-1} + x(n-1)x^{n-2} = x^{n-1} + (n-1)x^{n-1} = nx^{n-1}$$

And we see that $(x^n)' = nx^{n-1}$, but this is simply Formula 2.3 from before! The formula works for the number n, meaning that n could not have been in the collection S. Since the collection of natural numbers S has no smallest element, S must be an empty collection.

We conclude that all natural numbers obey our formula! Therefore, for each natural number k

Power Rule: $(x^k)' = kx^{k-1}$.

And that concludes our introduction to the differentiation rules. Things may have gotten hairy here and there, but the main point is that (i) differentiation rules are far from arbitrary, and are the simplest thing that one could come up with, and (ii) you could have come up with them if you wanted to, without knowing any calculus!

In order to obtain the power rule, we made the reasonable assumption (called an **axiom**) that a nonempty collection of natural numbers must have a smallest natural number. This assumption, called the **well-ordering principle**, together with the (also very reasonable) assumption that each nonzero natural number n has a "predecessor" $n-1$, can be used to prove many results in mathematics, both in calculus and elsewhere. The two combinations are also widely used (in an equivalent form) outside of mathematics, for example to prove the correctness of many algorithms.

Challenge 4 Use the well-ordering principle to show that if we have n functions $f_1, f_2, \ldots, f_n$, for some positive natural number n, then $(f_1 + f_2 + \cdots + f_n)' = f_1' + f_2' + \cdots + f_n'$. This is also called the **sum rule** for derivatives.

Polynomials

Combining the sum rule and the power rule allows us to find the derivatives of a large class of functions. For example, it is straightforward to calculate the derivative of $f : n \mapsto 3600n^5 + 70000n^4 + 42n + 9$ and $g : k \mapsto k^2 + k$. Such functions are examples of *polynomials*.

A **polynomial of degree** m **(on the variable** $\square$**)** is an expression of the form

$$c_m\square^m + c_{m-1}\square^{m-1} + c_{m-2}\square^{m-2} + \cdots + c_2\square^2 + c_1\square + c_0,$$

where the **coefficients** $c_m, c_{m-1}, \ldots, c_1, c_0$ are allowed to be any number, including 0, with the exception of c_m, which must be nonzero. A polynomial of degree **at most** m includes all polynomials of degree less than or equal to m.

[10] This number will be greater than 3 because we checked the formula up until the number 3.

The expression $3600n^5 + 70000n^4 + 42n + 9$ is a polynomial of degree 5 (on the variable n), and the expression $k^2 + k$ is a polynomial of degree 2 (on the variable k).

It will be convenient to introduce the following notation, called the **summation notation**. For the natural numbers p and q with $p \leq q$, the expression $\sum_{\diamond=p}^{q} h(\diamond)$ means $h(p) + h(p+1) + \cdots + h(q)$. In particular, $\sum_{\diamond=p}^{p} h(\diamond) := h(p)$. Using this notation, a polynomial of degree m on the variable $\square$ may be written compactly as $\sum_{\diamond=0}^{m} c_\diamond \square^\diamond$, or equivalently as $\sum_{\diamond=0}^{m} c_{m-\diamond} \square^{m-\diamond}$. The latter respects the ordering of each term in our definition, while the former reverses it from back to front.

If f is a polynomial of degree m on the variable t, then by the sum rule,

$$f' = \left(\sum_{i=0}^{m} c_i t^i\right)' = \sum_{i=0}^{m} (c_i t^i)'.$$

Applying the product rule on the constant function $c_i : t \mapsto c_i$ and the function $t^i : t \mapsto t^i$ gives $(c_i t^i)' = c_i (t^i)'$. By the power rule $(t^i)' = i t^{i-1}$, and so

$$f' \left(\sum_{i=0}^{m} c_i t^i\right)' = \sum_{i=0}^{m} (c_i i) t^{i-1}. \tag{2.4}$$

This is a fairly symbol heavy way to write down what we already knew. For example, $(3600n^5 + 70000n^4 + 42n + 9)' = 3600 \cdot 5n^4 + 70000 \cdot 4n^3 + 42$ and $(k^2 + k)' = 2k + 1$. The key idea is that we can take any polynomial, calculate its derivative term by term, then add them up to get the derivative of the polynomial. That is all that Formula 2.4 is saying.

Challenge 5

(a) Write the expression $1^3 + 2^3 + 3^3 + \cdots + k^3$ using the summation notation.
(b) Using $(1 + X)^2 := (1 + X)(1 + X) = 1 + X + X + X^2 = 1 + 2X + X^2$, expand $(1 + X)^3$.
(c) Check that $\sum_{k=1}^{n} k = n(n+1)/2$ holds when $n = 1$, $n = 2$, and $n = 3$.
(d) Use the well-ordering principle to show that the equation $\sum_{k=1}^{n} k = n(n+1)/2$ holds for each positive natural number n.

2.2 What is a Velocity?

The definition

We have worked out the arithmetic of derivatives, so now it is time to figure out what a derivative is. Recall that the notion of a derivative generalizes the idea of a velocity. Why do we care about velocity? We usually care about our velocity when we are in a car, so let us start from there. Why is there a speedometer in every car? I suppose it can help us avoid getting speeding tickets. But what if we didn't have to worry about tickets? Speedometers are there so that we can gauge when we will get to our destination. If our speedometer says 70 km/hr (or mi/hr if you wish), then we know that if we go for an hour at that speed, then we will be able to cover a distance of 70 km.

Let us denote our current time by t, our position function by f, our current velocity of 70 km/hr by v, and the time interval we wish to look into the future (an hour) by α. If we manage to travel at

exactly 70 km/hr for the next hour without any change in our velocity, then we can calculate our future position an hour later using our current position with the following formula.

$$\underbrace{f(t+\alpha)}_{\text{future pos.}} = \underbrace{f(t)}_{\text{current pos.}} + \underbrace{v \cdot \alpha}_{\text{travel dist.}}$$

In reality, it is impossible to stick to an exact constant velocity for an hour. Because our velocity will deviate during the hour, the correct formula will be given by

$$\underbrace{f(t+\alpha)}_{\text{future pos.}} = \underbrace{f(t)}_{\text{current pos.}} + \underbrace{v \cdot \alpha + X}_{\text{projected travel dist.}} \tag{2.5}$$

where X is the error in our projection caused by our velocity deviations during the next hour.

What can we say about our velocity deviations? That there will be deviations happening constantly, and so it makes no sense to try and track them all down! So instead, let's simplify and try to summarize our velocity deviations in a sinlge number. We cannot keep track of all the velocity deviations, but we know that their cumulative effect is given by the distance error X. We also know that the longer into the future we try to predict (3 hours for example), the greater the error. Conversely, the shorter we look into the future (3 minutes for example), the lesser the error. Hence the length of the time interval α will is correlated with how much velocity deviations occur. X is a Length and α is a time, and so X/α is a speed, which is what we are looking for to summarize our velocity deviation. We will define the **rogue velocity** to be X/α, a quantity we will use to summarize the amount of velocity deviations we experience during time interval α.

What can we say about our rogue velocity X/α? If we choose smaller values of α, then it becomes smaller. How small can we choose α? Any positive number α is fair game because then Equation 2.5 can be used to make a projection into the future, which is the whole point of wanting to know velocity. If α is negative, then we are no longer making a projection, we are looking into the past, so that's no good. Similarly, if α is zero, then we are no longer making a projection, we are looking into the present, so that's no good either. So as long as α is positive, we can make it as large or as small as we wish. Except, we don't want α to be large, because our projections will be garbage, so we want $\alpha > 0$ to be small.

Now, suppose we call a friend and ask what they are doing. The question we ask is "what are you doing right now?". But what we mean is not the same as the words we say. "What are you doing **right now**?" is short for, "what were you doing **before** you picked up the phone?" Otherwise, our question will always be answered with: "I'm on the phone" or "I'm talking to you". Duh, we meant *before* that!

We will do the exact same thing. We know that rogue velocity decreases as we drop α. But drop to what? There is no smallest positive number to drop to.[11] To get around the issue of having no smallest positive number to drop to, we will say that we "drop α to zero" (just as we say "what are doing **right now**" to mean "what were you doing **before** picking up the phone?"). Using this language, we will say: "the rogue velocity drops to zero as we drop α to zero", with the understanding that we are not actually taking anything to zero. In symbols we will write: as $\alpha \to 0$, $X/\alpha \to 0$, which we read as "as α drops to 0, (rogue velocity) X/α drops to 0.

[11] If $a > 0$ is a candidate for the smallest positive number, then $a/2$ is an even smaller positive number.

We will need to write this so often that an even simpler notation will be very helpful. We will write $\square = o_\alpha(1)$ to mean that the quantity $\square$ has the property that as $\alpha \to 0$, $\square \to 0$. Hence rogue velocity $X/\alpha = o_\alpha(1)$, because it has the property that as $\alpha \to 0$, $X/\alpha \to 0$. Multiplying both sides of the equation by α, we see that the distance error $X = \alpha \cdot o_\alpha(1)$.

We now take our projection Formula 2.5 and replace error X by $\alpha \cdot o_\alpha(1)$, because $X = \alpha \cdot o_\alpha(1)$.

Definition 1. A function f is **differentiable** at input t if there is a number v such that the following equation holds.

$$f(t + \alpha) = f(t) + v \cdot \alpha + \alpha \cdot o_\alpha(1)$$

If such a number v exists, then v is called the **derivative** of f at t. We will also denote the derivative of f at t using the notation $f'(t)$. If function f is differentiable at every input, then f is said to be a **differentiable function**, and the derivative of f is denoted by the symbol f'.[12]

To recap, the whole point of wanting to know our velocity is to predict our future position at α (minutes, say) into the future. Our current velocity v times the time interval α tells us how much we expect to have moved, but we recognize that there will be an error X caused by our velocity deviations away from v during our travel. To quantify the velocity deviations, we define a rogue velocity X/α, which has the property of droping to 0 as we drop the time interval α to 0. Hence $X/\alpha = o_\alpha(1)$, where the symbol $o_\alpha(1)$ denotes a quantity that drops to 0 as we drop α to 0.

Time

When we are observing objects traveling across a line (like a straight path/road), there is a notion of what is located on the right and what is located on the left. The notion of orientation, what direction is left and what direction is right, is not unique. For example, if we are having a face to face conversation, your right is my left and my right is your left.

When we say *time*, we will be using it in the exact same manner as *position*. Just as we can measure lengths and distances, we can measure time differences. Just like we can travel left to right or right to left, we can go from a smaller time value to a larger time value, but also from a larger time value to a smaller time value. Just as the orientation of what is left versus right is not unique and is a matter of convention, the flow of time is not unique and is a matter of convention.

Therefore, our discussion of predicting position “in the future" must work for folks whose flow of time is the opposite of ours, and are thus (in our view) calculating position in the past. In our view, they will be taking negative α values then “upping" it to 0, but everything will work in the same manner. Since dropping α to 0 and “upping" α to 0 are the same action, just in different time flow conventions, we will denote both by the symbol $\alpha \to 0$.

To really drive the point home that α can be taken to be either positive or negative, we will write the defining equation of a derivative at an input t as

$$f(t + \alpha) = f(t) + f'(t)\alpha + |\alpha| o_\alpha(1). \tag{2.6}$$

This modification does not change the equation and its interpretation. The symbol $|a|$ is used to denote the **absolute value** of a number a, and is defined to be a if a is positive or zero, and $-a$ if a

[12] Since $t + \alpha$ is an input of f, α must have the dimension of an input of f. Furthermore, for $f(t) + v \cdot \alpha$ to make sense, $v \cdot \alpha$ must have dimension $f(t)$. Therefore, a derivative has a dimension of f divided by its input, as expected.

is less than zero. For example, the absolute value of -2, written $|-2|$ is 2, while the absolute value of 2, written $|2|$ is still 2.[13] An **absolute value function** $|x|$ is the function $x \mapsto |x|$, which switches the sign of negative inputs. Because the value $|a|$ changes only when a is negative, writing the definition of the derivative as Equation 2.6 reminds us that α can be negative.

The absolute value of a velocity is the speed. For example, if an object has the velocity of -5 m/s, then it is moving to the left at a speed of 5 m/s. Similarly, an object with the velocity of 5 m/s is moving to the right at a speed of 5 m/s.

Little oh of one

Since we have a new object $o_\alpha(1)$ we best describe how to do arithmetic with it. It is going to be so simple and magical: if we have $o_\alpha(1)$ and add/subtract/multiply another $o_\alpha(1)$ to it, it does nothing! Even better, if we multiply a constant to $o_\alpha(1)$, it stays the same. How could such a thing be possible? Let's try and build some intuition about the behavior of $o_\alpha(1)$.

From now on, we will simplify the notation even further by omitting the subscript α and writing $o(1)$. For example, we will write the definition of a derivative at an input t as

$$f(t+\alpha) = f(t) + f'(t)\alpha + |\alpha| o(1). \tag{2.7}$$

This is because we will always be taking $\alpha \to 0$, and so the subscript α is redundant. Rest assured, if there is a potential for confusion, then I will write down the necessary subscripts.

Recall that if something is denoted by the symbol $o(1)$, then it drops to 0. Imagine a sleigh on a snowy hill headed towards the ground level, which we take to be 0 (meters). We will denote the position of our sleigh over time by the symbol X, which satisfies $X = o(1)$. If we moved our sleigh and placed it on top of a hill twice as high or twice as small, the sleigh will still drop towards the ground. That is, $2X = o(1)$ and $(1/2)X = o(1)$. In fact, the number 2 is not special, for we could have picked any positive number. Hence, if c is a positive constant, then $cX = o(1)$. We will thus write that for each positive constant c, we have $c \cdot o(1) = o(1)$.

If the multiplying factor is *not* a constant, this may no longer be true! Indeed, $(1/X) \cdot X = 1 \neq o(1)$ because a nonzero constant (like 1) will never drop to 0, it's a nonzero *constant*!

Suppose $Y \leq X$ and $X = o(1)$. Can we conclude that $Y = o(1)$? Suppose Y denotes the location of a spectator, moving about underground (thus $Y < 0$) and never approaching the ground level, hence $Y \neq o(1)$. Then $Y \leq X$, but $Y \neq o(1)$. What if we look at the absolute value $|Y|$ instead and pretend that the spectator is on top of the hill? Now we can see that the spectator's location is not dropping to 0, because the spectator is not on a sleigh rolling down, So one way to check if $Y = o(1)$ is to see if $|Y| \leq o(1)$.

One of the advantages of using $o(1)$ notation is that absolute values are built in. To see this, we use the fact that the orientation of direction is not unique. We will flip the convention of up and down and denote everything higher than the base of the hill with a negative sign. Thus a hill of height 5 meters is now of height -5 meters. This means that the position of our sleigh is now $-X$, yet this will not change the fact our sleigh will still drop down towards the ground. Hence $-X = o(1)$, and so $-o(1) = o(1)$.[14] Since $c \cdot o(1) = o(1)$ for positive c, we see that $-co(1) = -o(1) = o(1)$. Therefore, $c \cdot o(1) = o(1)$ for each constant c (whether negative, positive or zero).

[13] Thus the absolute value of a nonzero number is always positive (the absolute value of 0 is 0).

[14] As a consequence $|\alpha| o(1)$ and $\alpha \cdot o(1)$ are interchangeable, regardless of the sign of α.

Suppose we have two sleighs that arrive at the bottom of the hill at the same time, whose positions we denote by X and Y respectively. We know that $X = o(1)$ and $Y = o(1)$. Sum their position functions to define $Z := X + Y$. If $Z \neq o(1)$, that means Z does not drop to 0; suppose Z never drops below $k > 0$. But X and Y both stay below height $k/2$ after some time has passed because they drop to the ground, meaning that their sum will drop below k. Hence $Z = X+Y = o(1)$ and so $o(1) + o(1) = o(1)$. This is also consistent with the fact that $2 \cdot o(1) = o(1)$.

Using $-o(1) = o(1)$ and $o(1) + o(1) = o(1)$, we have $o(1) - o(1) = o(1) + o(1) = o(1)$. Therefore, $o(1) - o(1) \neq 0$. This makes sense because $X = o(1)$ and $0 = o(1)$ (0 drops to 0 for sure!), but $X - 0 = X \neq 0$. Here we find a peculiarity: $0 = o(1)$, yet $o(1) \neq 0$. Confusing? Not really, because $0 = o(1)$ means 0 falls to 0, which is true. But saying anything is equal to 0, as in $o(1) \neq 0$, is false unless that thing is itself 0. Similarly, $X = o(1)$ means X drops to 0, but $o(1) \neq X$ because the quantity represented by the notation $o(1)$ is not necessarily X.

Finally, the product satisfies $o(1)o(1) = o(1)$. To check this, set our origin for the time axis to when our sleigh reaches the ground. Then the sleigh reaches the ground at $t = 0$, and is dropping down during negative time. Suppose our sleigh always remains below the height of 3 meters above ground after $t = -5$ seconds. In other words, from $t = -5$ and onwards, $X < 3$. Now take $\alpha = -5$ and then up it to 0. Ignoring everything that happened before time $t = -5$, we have $Xo(1) \leq 3 \cdot o(1) = o(1)$, and since $X = o(1)$, we have $o(1)o(1) = o(1)$.

Our findings, summarized below, will simplify calculations greatly.

(a) $o(1) \square o(1) = o(1)$, where $\square$ can be $+$, $-$, or $\times$. If c is a constant, then $c \cdot o(1) = o(1)$.
(b) To check if $f = o(1)$, put it on the slope! If $|f| \leq o(1)$, then $f = o(1)$.

Basic properties

We will now check that our definition of the derivative satisfies the arithmetic rules we deduced at the beginning. It will require more work than dimensional analysis, but everything is still just arithmetic: adding, subtracting, multiplying, and dividing. The twist is that we will be using arithmetic with $o(1)$, but that makes things simpler! Remember, if we multiply $o(1)$ with itself or a constant (a derivative of a function at a point is a constant), then the result is still $o(1)$. But if we multiply $o(1)$ with a variable (like α, which we want to drop to 0), then we cannot simplify further.

Uniqueness of derivatives

When we speak of a velocity of an object, we are speaking about *the* velocity of an object. That is to say, there should be one unique velocity of an object. Suppose a function f has a derivative of a at t. This means that the equation $f(t + \alpha) = f(t) + a\alpha + |\alpha|o(1)$ holds. Is it possible that there is a different number that satisfies the above equation? What if there is a number b, with $a \neq b$ such that the following holds?

$$f(t + \alpha) = f(t) + b\alpha + |\alpha|o(1)$$

This would be a big problem, because we will be unable to agree on exactly which derivative $f'(t)$ we are talking about: do we mean the number a or the number b? Is our definition too weak to rule such cases out?

Let us check and see. Suppose a function f is differentiable at t with derivative a and b. Here, f is a function, while t, a, and b are all numbers. To show that a derivative is unique, it is sufficient

to show that $a - b = 0$. The definition of a derivative gives us the two equations

$$f(t+\alpha) = f(t) + a\alpha + |\alpha|o(1),$$
$$f(t+\alpha) = f(t) + b\alpha + |\alpha|o(1).$$

Equate these two to get $f(t) + a\alpha + |\alpha|o(1) = f(t) + b\alpha + |\alpha|o(1)$. Subtract the terms $f(t)$ and $b\alpha$ from both sides and we have

$$a\alpha - b\alpha + |\alpha|o(1) = |\alpha|o(1).$$

Recall that $-o(1)$ and $o(1)$ are the same. We divide both sides by the nonzero term α to get

$$a - b + o(1) = o(1).$$

Denote the left side of the equation above by A and the right side by B. Now take $\alpha \to 0$ and observe that $A \to (a - b)$ and $B \to 0$. Since $A = B$, we see that $a - b = 0$. Whenever derivatives exist, we know that they must be unique!

Constant rule

Let $f : x \mapsto c$ be a constant function. Since $0 = o(1)$, we have $0 = |\alpha|o(1)$. Then for each t,

$$f(t+\alpha) = c = c + 0 + 0 = f(t) + 0 \cdot \alpha + 0 = f(t) + 0 \cdot \alpha + |\alpha|o(1).$$

This is true for any input t. Therefore constant functions are differentiable, and the zero function is its derivative, as we expected.

Sum rule

Suppose functions f and g are differentiable at t. By the definition of the derivative,

$$f(t+\alpha) = f(t) + f'(t)\alpha + |\alpha|o(1), \qquad g(t+\alpha) = g(t) + g'(t)\alpha + |\alpha|o(1).$$

Taking the sum gives

$$f(t+\alpha) + g(t+\alpha) = \big(f(t) + f'(t)\alpha + |\alpha|o(1)\big) + \big(g(t) + g'(t)\alpha + |\alpha|o(1)\big).$$

Since $|\alpha|o(1) + |\alpha|o(1) = |\alpha|\big(o(1) + o(1)\big) = |\alpha|o(1)$, we have

$$(f+g)(t+\alpha) := f(t+\alpha) + g(t+\alpha) = \big(f(t) + g(t)\big) + \big[f'(t) + g'(t)\big]\,\alpha + |\alpha|o(1).$$

Therefore, the sum function $(f+g) : t \mapsto [f(t)+g(t)]$ is differentiable at t, with derivative $(f'+g')(t)$.

Product rule

Suppose functions f and g are differentiable at t. By the definition of the derivative,

$$f(t+\alpha)g(t+\alpha) = \big(f(t) + f'(t)\alpha + |\alpha|o(1)\big) \cdot \big(g(t) + g'(t)\alpha + |\alpha|o(1)\big).$$

The product is simple, but will look much more complicated than it is! The product multiplies out to:

$$f(t+\alpha)g(t+\alpha) = f(t)g(t) + \big[f'(t)g(t) + f(t)g'(t)\big]\,\alpha + f'(t)g'(t)\alpha^2 + |\alpha|Ao(1) \tag{2.8}$$

where $A := f(t) + g(t) + f'(t)\alpha + g'(t)\alpha + |\alpha|o(1)$. Here is a friendly reminder: $f(t), g(t), f'(t), g'(t)$ are all constants. Because we multiply A to $o(1)$, all the constants vanish, and we get $A = \alpha + \alpha o(1)$. Furthermore, the term $\alpha^2 = \alpha o(1)$ because if we divide it by α and take $\alpha \to 0$, then what's left (just α) drops to zero. Equation 2.8 is thus

$$\begin{aligned} f(t+\alpha)g(t+\alpha) &= f(t)g(t) + \left[f'(t)g(t) + f(t)g'(t)\right]\alpha + f'(t)g'(t)\alpha o(1) + |\alpha|[\alpha + \alpha o(1)]o(1) \\ &= f(t)g(t) + \left[f'(t)g(t) + f(t)g'(t)\right]\alpha + \alpha o(1) + \alpha^2 o(1) + \alpha^2 o(1)o(1) \\ &= f(t)g(t) + \left[f'(t)g(t) + f(t)g'(t)\right]\alpha + \alpha o(1) + \alpha o(1)o(1) + \alpha o(1)o(1)o(1). \end{aligned}$$

Since $o(1)o(1) = o(1)$ and $\alpha o(1) = |\alpha|o(1)$, we have

$$f(t+\alpha)g(t+\alpha) = f(t)g(t) + \left[f'(t)g(t) + f(t)g'(t)\right]\alpha + |\alpha|o(1)$$

and so the product function is differentiable at t with derivative $f'(t)g(t) + f(t)g'(t)$. Whew!

Quotient Rule

Just as we saw before, the rule for division follows from the product rule. Suppose functions f and g are differentiable at a, and $g(a)$ is nonzero. Furthermore, assume that the function $(f/g) : x \mapsto f(x)/g(x)$ is differentiable at a. Then we can apply the product rule to $f = (f/g) \cdot g$ to obtain

$$f'(a) = (f/g)'(a) \cdot g(a) + (f/g)(a) \cdot g'(a).$$

This gives us the **quotient rule** $(f/g)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{[g(a)]^2}$.[15] Taking $f : x \mapsto 1$ gives us the **reciprocal rule**: if $1/g$ is differentiable at a, then $(1/g)'(a) = -g'(a)/[g(a)]^2$.

2.3 The Chain Rule

Dual numbers

Let us take a second look at our definition of the derivative of function f at t:

$$f(t+\alpha) = f(t) + f'(t)\alpha + |\alpha|o(1).$$

The term α is not exactly a number—it is, until we drop it to zero.[16] So really, we are using the term α to mean the same thing as $a \cdot o(1)$, where a is the starting value of α.[17] So let us substitute the term α with $a \cdot o(1)$ into the definition of the derivative:

$$f(t + a \cdot o(1)) = f(t) + af'(t)o(1) + |a|o(1).$$

Since $f'(t)$ is a constant, $af'(t)o(1) = o(1)$, and thus the term $af'(t)o(1)$ can be absorbed into the final term $|a|o(1)$. But that's not what we want! We need the term $af'(t)o(1)$ to stay, because we are

[15] It is much easier to remember the rule as $(f/g)' = (f'g - fg')/(g^2)$. Or to derive it yourself!

[16] This is why we are using a Greek letter to denote it. It is not like the other numbers.

[17] Indeed $a \cdot o(1) = o(1)$, but let us leave the constant for now.

defining $f'(t)$ to satisfy the equation above. Instead, we will let the term $af'(t)o(1)$ absorb the term $|a|o(1)$. This gives us a simpler equation:

$$f(t + a \cdot o(1)) = f(t) + af'(t)o(1).$$

That's better, but notice that we have no α quantity for us to drop. Since the notation $o(1)$ makes little sense, we will replace the notation $o(1)$ with the Greek letter ϵ to write:

$$f(t + a\epsilon) = f(t) + af'(t)\epsilon. \tag{2.9}$$

Since we are no longer taking $\alpha \to 0$, the term ϵ is no longer $o(1)$. But the number ϵ should still preserve the key characteristics of the object $o(1)$. The three properties of $o(1)$ that we have needed in our derivations so far were: $c \cdot o(1) = o(1)$ for each constant c, $o(1) + o(1) = o(1)$, and $o(1)o(1) = o(1)$. We do not want to preserve the first property, because if we do, then Equation 2.9 becomes $f(t + a\epsilon) = f(t) + af'(t)\epsilon = f(t) + \epsilon$, and we have lost the crucial $f'(t)$ term. In addition, we do not want to preserve the second property: if $\epsilon + \epsilon = 2\epsilon = \epsilon$, then we lose the uniqueness property of derivatives, for if v is a derivative satisfying Equation 2.9, then so does $2v$ (the 2 is absorbed by ϵ). The only requirement left is $\epsilon^2 = 0$. This is also a problem, because the only number that squares to a zero is zero. Hence $\epsilon = 0$, in which case Equation 2.9 becomes $f(t) = f(t)$, useless!

Is this approach doomed to fail? Let us backtrack a bit. We know that we cannot bring over the properties $c \cdot o(1) = o(1)$ and $o(1) + o(1) = o(1)$. However, the only objection with bringing over $o(1)o(1) = o(1)$ is that there is no nonzero number that squares to zero. From the very beginning, we have tried to be more lax on what we mean by a number—indeed, a unit is not a number, but doing arithmetic with it as if it were turned out to be very useful! We will take the same approach and agree that ϵ is no ordinary number. We will define ϵ to be a nonzero quantity such that $\epsilon^2 = 0$.

A **dual number** is a number $a + b\epsilon$ for ordinary numbers a and b and a symbol ϵ such that $\epsilon \neq 0$ but $\epsilon^2 = 0$.Using dual numbers, the derivative of a function is defined by Equation 2.9 from before. A function f is **differentiable** at t if the following equation holds for nonzero a:

$$f(t + a\epsilon) = f(t) + af'(t)\epsilon$$

and the number $f'(t)$ is called the **derivative** of f at t.

Is this definition any good? Is it even correct? Let us check and see if this new definition obeys the same rules as before. Uniqueness is easy to check: if $\clubsuit$ and $\spadesuit$ are derivatives of f at t, then $f(t) + a\clubsuit\epsilon = f(t) + a\spadesuit\epsilon$. Subtract $f(t)$ from both sides and divide by a and ϵ which are both nonzero to get $\clubsuit = \spadesuit$.

The constant rule is also easy to check: if f is a constant function, then

$$f(t + a\epsilon) = f(t) + a \cdot 0 \cdot \epsilon$$

and so f has the zero derivative everywhere.

Our new definition really starts to shine when verifying the sum rule and the product rule. The sum rule is verified as follows.

$$(f + g)(t + a\epsilon) = \left[f(t) + af'(t)\epsilon\right] + \left[g(t) + ag'(t)\epsilon\right] = \left[f(t) + g(t)\right] + a\left[f'(t) + g'(t)\right]\epsilon$$

The product rule, a monstrosity using our previous definition, is now quite manageable:

$$\begin{aligned}(fg)(t+a\epsilon) &= \left[f(t)+af'(t)\epsilon\right]\left[g(t)+ag'(t)\epsilon\right] \\ &= f(t)g(t)+a\left[f'(t)g(t)+f(t)g'(t)\right]\epsilon + a^2f'(t)g'(t)\epsilon^2 \\ &= \left[f(t)g(t)\right]+a\left[f'(t)g(t)+f(t)g'(t)\right]\epsilon.\end{aligned}$$

So it seems like our new definition is all good to go! Let us go one step further. There is one important operation that we cannot do with units, but we can do with functions. This is the *chaining* operation: we can use one function as an input to another function. Suppose we chain the outputs of a function g into another function f. We write this using the notation $f \circ g$. Let us assume that function g is differentiable at t and that function f is differentiable at $g(t)$. A natural question to ask is whether the chained function $f \circ g$ is differentiable at t, and if so, what is the derivative? Let us check and see!

As we have done before, we consider the expression $(f \circ g)(t+a\epsilon)$, which (by differentiability of g at t) is the same thing as $f\left(g(t)+ag'(t)\epsilon\right)$. Let us denote $ag'(t)$ by the letter $\bar{a}$ and $g(t)$ by $\bar{t}$. Since function f is differentiable at $g(t)$, by the new definition of a derivative, $f\left(\bar{t}+\bar{a}\epsilon\right) = f\left(\bar{t}\right)+\bar{a}f'\left(\bar{t}\right)\epsilon = f\left(g(t)\right)+ag'(t)f'\left(g(t)\right)\epsilon$. We reorganize what we have found in the following line.

$$(f\circ g)(t+a\epsilon) = f\left(g(t)+ag'(t)\epsilon\right) = f\left(g(t)\right)+ag'(t)f'\left(g(t)\right)\epsilon = (f\circ g)(t)+a(f'\circ g)(t)\cdot g'(t)\epsilon$$

So we see that if g is differentiable at t and f is differentiable at $g(t)$, then the chained function $(f \circ g)$ is differentiable at t, with $(f\circ g)'(t) = (f'\circ g)(t)\cdot g'(t)$. This is called the **chain rule**.

Dual numbers are incredibly useful because they simplify calculations of derivatives enormously. Nevertheless, they utilize a suspect object ϵ which is nonzero while squaring to zero. Until we have the means to understand exactly what such an object is, we will stick to our previous definition of the derivative using $o(1)$.

Absolute values

We are going to have to verify our new result (the chain rule) independently using our definition of the derivative. This will require some preparation. As a first step, let us return to the absolute value function.

The absolute value function, which takes in a number and outputs the number's absolute value, has two important properties called the *triangle inequality* and *homogeneity*.

The **triangle inequality** states that for two numbers a and b, the inequality $|a+b| \le |a|+|b|$ holds. Notice that if a and b are both positive or both negative, or at least one of them is zero, then $|a+b| = |a|+|b|$. The inequality holds because the inequality becomes an equality.

The only remaining possibility is when exactly one of the numbers a, b is positive and the other is negative. For definiteness, let $a > 0$ and $b < 0$. There are two possibilities: either $a+b \ge 0$ or $a+b < 0$. In the former case,

$$|a+b| = a+b < a+(-b) = |a|+|b|$$

while in the latter case,

$$|a+b| = -(a+b) = -a-b = -a+(-b) = -a+|b| < |a|+|b|.$$

This completes our verification of the triangle inequality.

Homogeneity of the absolute value function states that for two numbers a and b, we have $|ab| = |a||b|$. If at least one of the numbers is zero, then the equation holds. The full result is verified by trying out all three cases: (i) $a \geq 0$ and $b \geq 0$, (ii) $a \leq 0$ and $b \leq 0$ (iii) exactly one of the numbers is positive, while the other is not.

Challenge 6

(a) By exhausting the cases (as in the proof of the triangle inequality), show that $|ab| = |a||b|$.
(b) If b is nonzero, use part (a) to show that $|1/b| \cdot |b| = 1$ and conclude that $|1/b| = 1/|b|$.
(c) If b is nonzero, use parts (a) and (b) to show that $|a/b| = |a|/|b|$.
(d) Show that $|c| - |d| \leq |c - d|$. [*Hint:* The triangle inequality says that $|a + b| - |b| \leq |a|$.]

The chain rule

Recall that if we have two functions f and g, and use the output of g as the input to f, then the chained function is written $f \circ g$. Thus $f \circ g : x \mapsto f(g(x))$ and the output of the chained function for input x is denoted by $(f \circ g)(x)$ or by $f(g(x))$.

Let us bring in differentiation once again. Suppose function g is differentiable at input t and function f is differentiable at input $g(t)$. Is the chained function $f \circ g$ differentiable at a? An equivalent question is: is there some number $\clubsuit$ that satisfies the equation below?

$$(f \circ g)(t + \alpha) = (f \circ g)(t) + \clubsuit \cdot \alpha + |\alpha| o_\alpha(1) \tag{2.10}$$

Let us begin with what we know. Differentiability of the function g at t and differentiability of function f at $s := g(a)$ gives

$$g(t + \alpha) = g(t) + g'(t)\alpha + |\alpha| o_\alpha(1), \tag{2.11}$$
$$f(s + \beta) = f(s) + f'(s)\beta + |\beta| o_\beta(1) \tag{2.12}$$

where $o_\alpha(1) \to 0$ as $\alpha \to 0$ and likewise, $o_\beta(1) \to 0$ as $\beta \to 0$. The subscripts are back because there are now two variables at play (α and β), and as a result, the notation $o(1)$ is ambiguous.

We need to chain these expressions together, where the former is the input to the latter. In particular, we are looking for an expression for $(f \circ g)(t + \alpha)$. Since the input of the "outer" function f is the value $g(t + \alpha)$, this is the chain in our link. Define $\beta := g(t + \alpha) - g(t)$ so that $g(t + \alpha) = g(t) + \beta = s + \beta$, which is exactly what we need to connect the two functions in Equation 2.11 and Equation 2.12.

By Equation 2.11, $\beta := g(t + \alpha) - g(t) = g'(t)\alpha + |\alpha| o_\alpha(1)$, where $g'(t)$ is some constant. Since $g'(t)\alpha = o_\alpha(1)$, when we take $\alpha \to 0$, then $\beta \to 0$ too. Therefore, $o_\beta(1) = o_\alpha(1)$.

Now let us consider the chained function $f \circ g$ together with our link. We have

$$\begin{aligned}
(f \circ g)(t + \alpha) &= f(s + \beta) = f(s) + f'(s)\beta + |\beta| o_\beta(1) \\
&= (f \circ g)(t) + f'(s) \left[g'(t)\alpha + |\alpha| o_\alpha(1) \right] + |\beta| o_\beta(1) \\
&= (f \circ g)(t) + \left[(f' \circ g)(t) \cdot g'(t) \right] \alpha + f'(s)|\alpha| o_\alpha(1) + |\beta| o_\beta(1) \\
&= (f \circ g)(t) + \left[(f' \circ g)(t) \cdot g'(t) \right] \alpha + |\alpha| o_\alpha(1) + |\beta| o_\beta(1)
\end{aligned}$$

where we have used the fact that $f'(s)$ is a constant to obtain the final equality. In order to obtain Equation 2.10, all we need to do is to check that $|\alpha|o_\alpha(1) + |\beta|o_\beta(1) = |\alpha|o_\alpha(1)$.

It suffices to show that $|\beta|o_\beta(1) = |\alpha|o_\alpha(1)$, because then $|\alpha|o_\alpha(1)+|\beta|o_\beta(1) = |\alpha|o_\alpha(1)+|\alpha|o_\alpha(1) = |\alpha|o_\alpha(1)$. We will use the two properties of an absolute value function from earlier. Using the triangle inequality $|a + b| \leq |a| + |b|$ and homogeneity $|ca| = |c||a|$, we have

$$|\beta| = \left|g'(t)\alpha + |\alpha|o_\alpha(1)\right| \leq |g'(t)||\alpha| + |\alpha||o_\alpha(1)| = |g'(t)||\alpha| + |\alpha|o_\alpha(1).$$

Since $o_\beta(1) = o_\alpha(1)$, we have

$$\frac{|\beta|o_\beta(1)}{|\alpha|} \leq \left(|g'(t)||\alpha| + |\alpha|o_\alpha(1)\right)\frac{o_\beta(1)}{|\alpha|} = o_\beta(1) + o_\alpha(1)o_\beta(1) = o_\alpha(1) + o_\alpha(1)o_\alpha(1) = o_\alpha(1).$$

Recall that if $|f| \leq o(1)$, then $f = o(1)$. Since $|\beta o_\beta(1)/\alpha| \leq o_\alpha(1)$, we have $|\beta|o_\beta(1)/|\alpha| = o_\alpha(1)$, as desired. Equation 2.10 is satisfied and we are done!

Theorem 2 (The Chain Rule)**.** If g is differentiable at t and f is differentiable at $g(t)$, then $f \circ g$ is differentiable at t with

$$(f \circ g)'(t) = (f' \circ g)(t) \cdot g'(t).$$

What if we want to chain more than two functions? Suppose f, g, and h are differentiable functions and we want to find the derivative of the function that is the chain of all three function. First, we must resolve some ambiguity: is $f \circ (g \circ h)$ the same as $(f \circ g) \circ h$? If not, then we might have two different derivatives for the composition of three functions, which is a problem! Luckily, when chaining functions (also called **function composition**), we are guaranteed that $f \circ (g \circ h) = (f \circ g) \circ h$. This guarantee is called **associativity**, and so function composition is said to be **associative**. To see this, pick some input x. Then $\left[f \circ (g \circ h)\right](x) = f\left((g \circ h)(x)\right) = f\left(g\left(h(x)\right)\right)$. But this is the same as $\left[(f \circ g) \circ h\right](x) = (f \circ g)\left(h(x)\right) = f\left(g\left(h(x)\right)\right)$.

Therefore, to take the derivative of a composition of three functions $f \circ g \circ h$, we may use the chain rule to get $\left([f \circ g] \circ h\right)' = ([f \circ g]' \circ h) \cdot h'$ or equivalently $\left(f \circ [g \circ h]\right)' = \left(f' \circ [g \circ h]\right) \cdot [g \circ h]'$. Both will give the same answers, so we choose whichever one is more convenient. Just as the chaining of three functions can be reduced to the case of two functions, the case of any finite number of function composition can also be handled by the chain rule.

2.4 Higher Derivatives

We can continue to take derivatives repeatedly on a function, as long as the derivative exists at each step. The interpretation is that if we wish to know an object's **acceleration**, we need to calculate the rate of change of the object's *velocity*. The *second* derivative of a function is a generalization of the concept of acceleration. The second derivative of a function f is denoted by the symbol f'' and it is the derivative of the function f'. The derivative of f'', if it exists, is written $f^{(3)}$. The expression $f^{(k)}(a)$ for a positive integer k is the k-th derivative of function f at point a. If $k = 1$, then $f^{(1)}(a)$ is the number $f'(a)$, while $f^{(0)}(a)$ is just the number $f(a)$.

Challenge 7 For $0 \leq k \leq n$, the **binomial coefficient** $\binom{n}{k}$ (read "n choose k") is the number of ways we can choose an unordered selection of k items from n distinct items. For example, there are 10 ways to choose 2 items from 5 elements (first we have 5 choices for the first item, then

there are 4 choices for the second item, but since the order we drew which item does not matter, we are double counting, which we account for by dividing by 2) and so $\binom{5}{2} = 10$. In general, $\binom{n}{k} = \frac{n\times(n-1)\times\cdots\times(n-k+1)}{k\times(k-1)\times\cdots\times1}$. In factorial notation, where $k! := k \times (k-1) \times \cdots 2 \times 1$ and $0! := 1$, we have $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Observe that $\binom{n}{0} = \binom{n}{n} = 1$.

(a) Let f be a polynomial of degree n. Convince yourself of the following equation.[18]

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

[*Hint:* if $f : x \mapsto 7x^5 + 2x^3 + 5$, what is the expression above saying?]

(b) Let $f : x \mapsto (x + b)^n$. Calculate $f^{(0)}(x)$, $f^{(1)}(x)$ and $f^{(2)}(x)$ and convince yourself that the following holds.

$$f^{(k)}(x) = k!\binom{n}{k}(x + b)^{n-k}$$

(c) Apply the result of part (b) to part (a) to conclude that

$$(x + b)^n = \sum_{k=0}^{n} \binom{n}{k} x^k b^{n-k}$$

and substitute the symbol x with the symbol a to obtain the **binomial formula**.

The binomial formula has a nice combinatorial interpretation when a and b are both natural numbers. If we have a pool of k distinct items, from which we were to draw n items sequentially with replacement, there are k^n possibilities (we make n draws, where at each stage there are k choices). Similarly, if we have a pool of k_1 distinct items and a pool of k_2 distinct items, from which we were to select n items sequentially, there are $(k_1 + k_2)^n$ possibilities. This is because we could pool each pile together each into one pile of $k_1 + k_2$ distinct items.

An alternative way to count the number of possibilities is to do the actual selection algorithmically, case by case. We could pick n objects from k_1, or $n-1$ objects from k_1 and 1 object from k_2. or pick $n-2$ objects from k_1 and 2 objects from k_2, ..., or pick 0 objects from k_1 and pick n objects from k_2. Adding all of these separate cases is exactly what the expression $\sum_{i=0}^{n} \binom{n}{i} k_1^{n-i} k_2^i$ means. Since either way of counting must give the same results, we conclude that $(k_1 + k_2)^n = \sum_{i=0}^{n} \binom{n}{i} k_1^{n-i} k_2^i$.

2.5 Nonexamples

Nonexample 1: the absolute value function

Now that we have discussed some examples of derivatives, let us examine some nonexamples. Consider the absolute value function $f : x \mapsto |x|$. At the origin, for $\alpha > 0$, we have $f(0 + \alpha) = |0 + \alpha| = |0| + 1|\alpha| = f(0) + 1\alpha$, and so it seems like we can conclude that the absolute value function is differentiable at the origin, with $f'(0) = 1$. However, what if we take $\alpha < 0$? We cannot stop anyone from taking $\alpha < 0$ because one person's preferred orientation of the x axis can be the opposite of the other (see Figure 2.13).

[18] Since $\sum_{\square=p}^{q} h(\square)$ means $h(p) + h(p+1) + \cdots + h(q)$, we have $\sum_{j=0}^{1} \frac{f^{(j)}(0)}{j!} x^j := \frac{f^{(0)}(0)}{0!} x^0 + \frac{f^{(1)}(0)}{1!} x^1$.

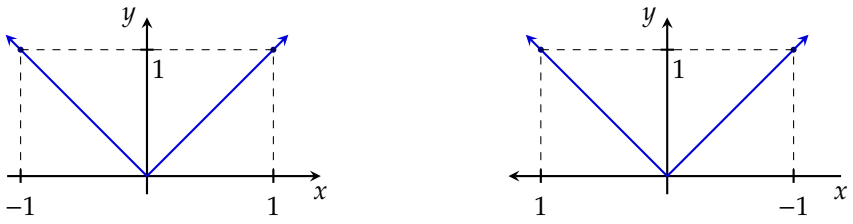


Figure 2.13: Changing the orientation of the x-axis changes nothing.

Then for $\alpha < 0$, we have $f(0+\alpha) = |0+\alpha| = -\alpha = |0| - \alpha = f(0) - 1\alpha$. Thus there is a disagreement on exactly what the value of $f'(0)$ is, and derivatives are known to be unique. We therefore conclude that the absolute value function is not differentiable at the origin.

Challenge 8 Recall that to use the chain rule $(f \circ g)'(t) = (f' \circ g)(t) \cdot g'(t)$, we need both $(f' \circ g)(t)$ and $g'(t)$ to exist. We investigate whether derivatives of chained functions can exist even if one of the component function is *not* differentiable.

(a) Let $f : x \mapsto x^2$ and $g : x \mapsto |x|$. We saw that g is *not* differentiable at 0 and so $g'(0)$ does not exist. Nevertheless, show that $(f \circ g)'(0)$ and $(g \circ f)'(0)$ both exist.

(b) The **relu** function (rectified linear unit) is defined by $\text{relu} : x \mapsto \max(0, x)$. Sketch the relu function. Show that the derivative of relu is zero for negative inputs, while for positive x, $\text{relu}'(x) = 1$. Furthermore, show that the relu function is not differentiable at 0.

(c) Let $n > 1$ be a natural number and let $f : x \mapsto x^n$. Show that even though the relu function is not differentiable at 0, both $(f \circ \text{relu})'(0)$ and $(\text{relu} \circ f)'(0)$ exist.

(d) Let $f : y \mapsto y - \text{relu}(y)$ and $g : x \mapsto \frac{1}{2}x + \frac{1}{2}\text{relu}(x)$. Show that although neither $f'(0)$ nor $g'(0)$ exist, both $(f \circ g)'(0)$ and $(g \circ f)'(0)$ exist.

Nonexample 2: a step function

What is going on with the function graphed in Figure 2.14? It logs the position (denoted by the symbol x and measured in meters from some origin) of an object over time (denoted by the symbol t and measured in seconds). The graph suggests that our object is perfectly still at all times, yet has managed to teleport from one location to another instantaneously.

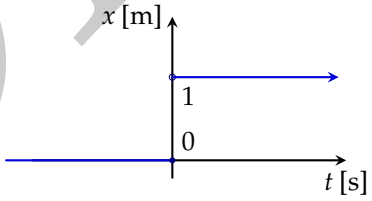


Figure 2.14: Graph of a function defined by $t \mapsto 1$ for $x > 0$ and $t \mapsto 0$ for $t \leq 0$.

We cannot allow such a behavior. A function cannot have zero derivative (no velocity) and be a non-constant function (display motion). In this case, the problem is that our function is not

continuous at time $t = 0$ due to an instantaneous teleportation.

How do we know if a function is continuous? Like a differentiable function, a function is **continuous** if it is continuous at each point that the function is defined. How do we know if a function f is continuous at an input t? As with a derivative, first take some nonzero step α, which is allowed to be either positive or negative. The difference between $f(t+\alpha)$ and $f(t)$ should then drop to zero as we dial down α, that is, drop $\alpha \to 0$.

Definition 3. A function f is **continuous** at an input t if $f(t+\alpha) - f(t) = o(1)$.

For example, let us denote the function graphed in Figure 2.14 with the symbol g. Then for $\alpha < 0$, $g(0+\alpha) - g(0) = 0 - 0 = 0 = o(1)$, *but* for $\alpha > 0$, we have $g(0+\alpha) - g(0) = 1 - 0 = 1 \neq o(1)$. Therefore, function g is not continuous at 0.

Continuity is not sufficient to guarantee differentiability, as the absolute value function demonstrates. However, differentiable functions are always continuous. Indeed, if f is differentiable at t, then $f(t+\alpha) - f(t) = f'(t)\alpha + |\alpha|o(1)$. Drop $\alpha \to 0$ and observe that $f'(t)\alpha \to 0$ and $|\alpha|o(1) \to 0$, and so $f(t+\alpha) - f(t) = o(1) + o(1) = o(1)$.

Proposition 4. If a function is differentiable at an input t, then it is continuous at t.

Challenge 9 Consider a mystery function h that satisfies the following: for each t, we have $h(t+\alpha) - h(t-\alpha) = o(1)$. Can we conclude that h is continuous? If not, come up with a counter example of a function that satisfies the given property, but is not continuous.

Nonexample 3: holes

Perhaps the simplest way to manufacture functions that are not continuous is by taking one that is continuous, and puncturing a hole in it. Consider a function h defined by $t \mapsto 1$ if $t < 0$ and $t \mapsto 1$ if $t > 0$. That is to say, h is *almost* a constant function, but the function is not defined at $t = 1$, and so the value $h(1)$ is undefined. Because we have introduced a hole, the function h is not continuous. In particular function h is not continuous at $t = 1$.

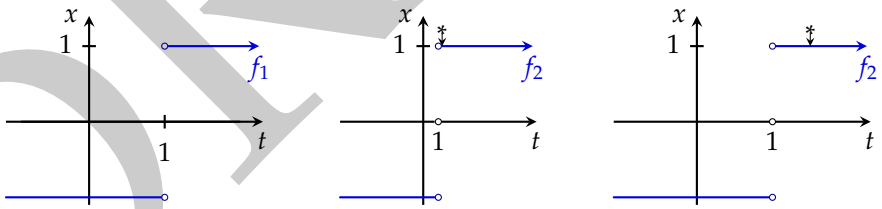


Figure 2.15: A step function f_1 and a step function f_2 defined on a defective time axis.

Let us examine the step function once more. The leftmost graph in Figure 2.15 is a depiction of the function f_1, defined by $t \mapsto 1$ for $t > 1$ and $t \mapsto -1$ for $t < 1$. In particular, the function is *not* defined on the point 1, and so the value $f_1(1)$ does not exist. As we discussed with the almost constant function h previously, the function f_1 is not continuous at the point $t = 1$.

However, there is a different type of hole we can introduce. Consider the function f_2 depicted in the middle of Figure 2.15. Just like the function f_1, the function f_2 is defined by the rules $t \mapsto 1$

for $t > 1$ and $t \mapsto -1$ for $t < 1$. The difference here is that we pretend that the point $t = 1$ does not exist by introducing a hole in the time axis. Thus the function f_2 cannot be defined at $t = 1$ even if we wanted to. This was not the case with our previous step function f_1.

If the function f_2 continuous? Surprisingly, yes! Even though we have introduced a hole, the function turns out to be continuous.

To check that a function is continuous, we have to make sure that the function is continuous at each point it is defined. There are two cases to check: points greater than 1 and points less than 1. The latter case is essentially the same as the previous case, so let us consider the case of $t > 1$. It is visually simple to check that for points far away from $t = 1$, the function f_2 is continuous. So let us pick a point very close to $t = 1$, the point $*$ shown in the middle graph of Figure 2.15. But notice that the units we use to measure time is completely arbitrary. So we may "zoom in" by introducing a smaller unit of time so that the distance between $t = 1$ and $*$ is much more pronounced.[19] Now there is no problem in seeing that function f_2 is also continuous at point $*$.

Isn't the function f_2 not continuous at $t = 1$? That is an invalid question, because time $t = 1$ does not exist. The function f_2 is continuous everywhere it can be defined on.

Completeness

At this point, we have broken calculus. We can have continuous functions describing the positions of objects teleporting at will. In such a setting, trying to sensibly ascribe velocity becomes impossible.

In order to prevent this from happening and to keep calculus intact, we must insist that the number axis we are dealing with has no holes. To fix this problem, let us return back to the step function f_2. We observed the fact that no matter how "close" we got to $t = 1$, by a suitable choice of units, we discovered that we were in fact "not close" to $t = 1$. Among the numbers $t > 1$, there is no smallest number which is objectively "close" to the number 1. All of them can be made "not close" to $t = 1$.

We have previously encountered the concept of a *smallest* number. In our proof of the power rule, we used the fact that if we have a nonempty collection of natural numbers, then there must be a smallest element. This is the well-ordering principle. This principle doesn't hold here, because there is no smallest number among $t > 1$.

Actually, it is even easier to break the well-ordering principle. Recall that the integers are the collection of natural numbers and its negative counterparts. The integers are thus the numbers $-1, -2, -3, \ldots$, as well as the usual $0, 1, 2, 3, \ldots$ from the natural numbers. The integers do not obey the well-ordering principle because if we consider a collection of negative integers $-1, -2, -3, \ldots$, this collection has no smallest element.

Nevertheless, this situation can be fixed. If we consider nonempty collections of integers whose members are all above a certain lower limit, then there has to be a smallest integer. The well-ordering principle of the natural numbers is itself a special case of this, for it states that anytime we have a collection of integers that are not negative, and thus greater than -1, there will always be a smallest element. We say that -1 is a **lower bound** of the natural numbers, or equivalently, that the natural numbers is **bounded from below** by -1. In fact, any negative integer is a lower bound

[19] You might object that if we use a different unit of time, then the meaning of 1 has changed. If this bothers you, replace the hole in our time axis by 0. This choice was not made to enhance legibility.

of natural numbers. This allows us to apply the well-ordering principle to nonempty collections of integers bounded from below.

We now import this fix. Going back to our step function f_2 and our defective time axis, the collection of numbers $t > 1$ was bounded from below (by numbers smaller than 1), but there was no smallest number. We have already visually seen that there cannot be a smallest number $t > 1$ by zooming in our graph,[20] so let us approach this from the other side with numbers $t < 1$. These numbers form the lower bounds of the numbers $t > 1$. Is there a largest?

By a **set**, we mean a collection of objects. If we can list out the objects of a set (called **elements** of a set), then we list out the objects between the braces { and }. For example, if S is a set of natural numbers less than 5, then we can write $S = \{0, 1, 2, 3, 4\}$ (the order of the elements is immaterial). A set with no elements is called an **empty set** and is denoted by the symbol $\emptyset$. Thus $\emptyset = \{\}$.

A number system is **complete** (in the sense of not having holes) if each set of numbers (from the number system) that is bounded from below, a *greatest* among all the lower bounds always exists. The greatest among all the lower bounds of the set is called the **greatest lower bound** or the **infimum** of the set. As a shorthand, the infimum of a set S is denoted by "$\inf S$".

We saw from function f_2 that calculus requires a *complete* number system. The number system we use, represented as an axis on a graph, is called the **real numbers**. The symbol for denoting the set of real numbers is $\mathbb{R}$. If T is the set of real numbers greater than 1, then $\inf T = 1$. Our problem with function f_2, or rather, our defective number system, was that $\inf T$ was not a part of the number system. Real numbers do not have this problem with holes because the number 1, and indeed any value of length or value of time we can think of, can be depicted on a line (as we have done so far), and are real numbers. In symbols, we write $1 \in \mathbb{R}$ to mean that 1 is a member of (or is an **element of**) the set of real numbers $\mathbb{R}$. More generally, we write $a \in S$ to mean that a is an element of the set S and we write $b \notin S$ to mean that b is *not* an element of set S.

As emphasized many times before, the orientation of an axis is completely arbitrary. Just as a direction of left to one is a direction of right to another, a negative number is a positive number to another. Thus the set of real numbers $\mathbb{R}$ also has the equivalent property that: each set of real numbers that is **bounded from above** has a *smallest* among all the **upper bounds**. The smallest among upper bounds is called the **least upper bound** or the **supremum** of the set. If a set S has a least upper bound, we denote the supremum using the symbol "$\sup S$". A set is **bounded** if it is both bounded from above and bounded from below.

If we take a positive α and then drop $\alpha \to 0$, the symbol zero denotes the infimum of the set of positive real numbers. Similarly, if we take a negative α and then up $\alpha \to 0$, the symbol zero denotes the supremum of the set of negative real numbers.

The natural numbers (denoted by the symbol $\mathbb{N}$) and the integers (denoted by the symbol $\mathbb{Z}$) are not complete and are thus not sufficient for calculus. For example, we cannot describe lengths or times with decimal points using integers. Surprisingly, fractions are not sufficient either. The **rational numbers** (denoted by the symbol $\mathbb{Q}$) are the numbers of the form $\frac{a}{b}$, where a is an integer and b is a nonzero natural number. For example, $0.1 = \frac{1}{10}$ and so 0.1 is a rational number.

The classic counterexample is that the diagonal length of a square with side length 1 cannot be expressed as a fraction. We will not pursue such theoretical issues further, as we wish to return to calculus. We will simply be content that there is a complete number system that has no holes

[20] An alternative way to see this is that if someone claims that $*$ is the smallest among the numbers $t > 1$, then $(* - 1)/2$ is even smaller! In fact, it is halfway between 1 and $*$.

called the real numbers in which we can do calculus in, and that we no longer have problems with instantaneous teleportation and the like. In particular, motion cannot happen in the absence of velocity. Or in calculus language, if $f' = 0$, then f is a constant function.

Intervals

Of course, if velocity of an object is zero during a time interval, then we know its position must be constant. Thus the notion of an *interval* will be useful as we discuss displacement arising from motion starting at one time and ending at another time. In order to make calculus work, we will be working with *subsets* of the set of real numbers $\mathbb{R}$. If S is a set, then A is a **subset** of S (written $A \subset S$) if each element of A is an element of S. The following are some convenient notation to specify *connected* subsets of $\mathbb{R}$. For real numbers a and b with $a < b$:

(a) the symbol (a, b) denotes the set of real numbers x such that $a < x < b$,
(b) the symbol $[a, b]$ denotes the set of real numbers x such that $a \leq x \leq b$,
(c) the symbol $[a, b)$ denotes the set of real numbers x such that $a \leq x < b$,
(d) the symbol $(a, b]$ denotes the set of real numbers x such that $a < x \leq b$.

Sometimes, the symbols ∞ and $-\infty$ are used in a similar context. For each real number a:

(a) the symbol $(-\infty, a)$ denotes the set of real numbers x such that $x < a$,
(b) the symbol (a, ∞) denotes the set of real numbers x such that $a < x$,
(c) the symbol $(-\infty, a]$ denotes the set of real numbers x such that $x \leq a$,
(d) the symbol $[a, \infty)$ denotes the set of real numbers x such that $a \leq x$,
(e) the symbol $(-\infty, \infty)$ is another way to denote the set of real numbers $\mathbb{R}$.

The nine sets defined above are called **intervals**. To distinguish the intervals involving the symbols ∞ or $-\infty$ from those that do not, intervals defined by real numbers only (the first four kinds) are called **finite intervals**. There is a potential source of confusion since the ordered pair of real numbers (a, b) may mean an interval or a coordinate on the x-y plane.[21] Whether the ordered pair means an interval or a coordinate will be clear from the context.

An **open interval** is an interval that does not contain its endpoints. Thus the intervals (a, b), $(-\infty, a)$, (a, ∞), $(-\infty, \infty)$ are open intervals. On the other hand, an interval is a **closed interval** if it is an interval that contains all *real* numbers between the endpoints, inclusive. Thus $[a, b]$ is a closed interval, but so are $[a, \infty)$ and $(-\infty, a]$ because these intervals also contain all real numbers between the endpoints, inclusive (the symbols ∞ and $-\infty$ are not real numbers).[22]

We can combine two intervals into one by taking their common elements. If S and T are sets then $S \cap T$ (read "S **intersection** T") is defined to be the set of elements in S that are also elements of T. Thus $\{1, 2\} \cap \{2, 3\} = \{2\}$, $\{1, 2\} \cap \{3, 4\} = \emptyset$, $\emptyset \cap \{1\} = \emptyset$, and $(0, 2) \cap (1, 3) = (1, 2)$.

We can also combine two intervals into one by "joining" them together. If S and T are sets then $S \cup T$ (read "S **union** T") is defined to be the set of elements in S that are also elements in T. Hence $\{1, 2\} \cup \{2, 3\} = \{1, 2, 3\}$, $\emptyset \cup \{1\} = \{1\}$, and $(0, 1] \cup [1, 2) = (0, 2)$.

[21] For example, an object at position 5 m to the right of the origin at time 8 s can be denoted $(5, 8)$ which is *not* an interval.
[22] When we speak of an interval (a, b) or an interval $[a, b]$, we assume that $a < b$.

3

Integration

The following optional Challenge is designed to get us into the mood for discussing displacements. In particular, all the symbols mean something *physical*. Our discussion in the first section of this chapter will be especially simple to understand and easy to remember if we stay grounded in the physical world.

Challenge 10 Consider an object constrained to motion along a line. Let t be the time since we started to keep track of the object and denote the object's initial position by the constant x_i and the object's initial velocity by the constant v_i. The object's position is denoted by $x(t)$ and its velocity is denoted by $v(t)$. To simplify matters, assume the object is under constant acceleration a (this constant could be positive, negative or zero).

(a) Our object's position x may be calculated using the initial position x_i, initial velocity v_i, current velocity v, and time t. Use dimensional analysis and apply a simple case (or common sense) to find the formula for x. What does the formula say?

(b) Repeat part (a), but this time use acceleration a in place of velocity v.

(c) Our objects's velocity v may be calculated from the initial position v_i, acceleration a and time t. Use dimensional analysis and apply some simple cases to find the formula for v.

(d) The squared velocity v^2 can be calculated from the initial velocity v_i, acceleration a, and displacement $x - x_i$. Use dimensional analysis and apply some simple cases to find the formula for v^2.[1]

(e) Use the derivative rules to show that your answer from part (b) gives the correct velocity and acceleration for our object. Verify your formula from part (d) by taking the time derivative of the formula from part (b), solving for time t, and then plugging the formula for t back into the formula from part (b).

The formulas are also called the big four kinematics equations.

$$x = x_i + \frac{v_i + v}{2}t \qquad x = x_i + v_i t + \frac{a}{2}t^2$$
$$v = v_i + at \qquad v^2 = v_i^2 + 2a(x - x_i)$$

[1] *Hint:* Although time t does not make an appearance in this formula, to check cases, nothing is stopping us from for example, taking $t = 1$ to simplify values of $x - x_i$ and v.

3.1 The Fundamental Theorems

Displacements

We studied velocity in the previous chapter, in particular, velocity functions and arithmetic with velocity functions. In this chapter, we will examine displacements and displacement functions. For simplicity, we will only consider objects in motion along a line moving back and forth.

Suppose we have some velocity function f at hand. As we saw in the previous section, there are headaches with functions that are not continuous, so we will always assume f is continuous. Furthermore, it is tricky to talk about displacements with unbounded velocity. We will assume our velocity function f is bounded, at the very least within the time interval we are considering (a function is **bounded** if the set of its outputs are bounded from above and below by real numbers). In order to calculate the displacement of an object between an initial time t_i and final time t_f, we could follow these basic steps.

First, we divide the time interval into smaller chunks. Second, for each smaller time interval we pick some representative value of f. The velocity function f is assumed to be bounded, so we may take the supremum or infimum of the values of f in that time interval. The third and final step: for each time interval $[t_i, t_j]$, we calculate an estimate of displacement during that time with $(t_j - t_i) \times \inf f$ or $(t_j - t_i) \times \sup f$, depending on our choice made in step two, then add all the estimates up.

These steps are simply a more detailed version of what we could imagine how a car's odometer calculates distance travelled using information from its speedometer: (i) given some time interval, (ii) pick a representative speed during that time interval, and (iii) multiply the representative speed with the time interval and accumulate to the previous estimate of distance travelled.[2]

Alternatively, we could view these steps as describing properties of displacement.

(a) We 'break down" a time interval into smaller chunks because our displacement during the day from 9AM (t_0) to 9PM (t_2) is the same as accumulating our displacement from 9AM (t_0) to noon (t_1) with our displacement from noon (t_1) to 9PM (t_2).

(b) Our estimate for displacement can change depending on our representative velocity chosen, because for an object traveling in one direction, a faster velocity leads to greater displacement.

(c) We estimate our displacement during a time interval as if we are moving at a constant velocity at that time interval with the representative velocity. With this assumption, our displacement is given by the product of our representative velocity with the length of the time interval.

It will be convenient to introduce a notation due to Gottfried Leibniz and Joseph Fourier. If f is a velocity function, then the displacement from time ♠ to time ♡ is denoted by the symbol

$$\int_{\spadesuit}^{\heartsuit} f(\square)\, d\square$$

where the two boxes $\square$ may be replaced by your choice of exactly one symbol, with the exception of the symbols used to represent the time endpoints—in this case ♠ and ♡. For example, $\int_{\spadesuit}^{\heartsuit} f(x)\, dx$ and $\int_{\spadesuit}^{\heartsuit} f(t)\, dt$ will both be equally acceptable. The reason we need the box is that the velocity

[2] A car's odometer measures distance travelled by looking at speed, which is the absolute value of velocity. We are looking to measure displacement by looking at velocity. What's the difference? We can reverse our car and create negative velocity, reducing displacement, but we cannot reduce an odometer reading by driving our car in reverse.

function f may have several symbols and we will need to distinguish the constants from the variables. For example, suppose we have a velocity function $f : t \mapsto at^2 + bt + c$, for some constants a, b, and c. Then we will denote the displacement from time t_i to time t_f by $\int_{t_i}^{t_f} (at^2 + bt + c)\, dt$. As another example, suppose we have a different velocity function $g : x \mapsto \alpha x + \beta$, for some constants α and β. Then we will denote the displacement from time a to time b by $\int_a^b (\alpha x + \beta)\, dx$.

We recast our properties using this new notation for some continuous and bounded function f.

(P1) Displacement from time t_0 to time t_2 is the same as the displacement from time t_0 to time t_1 added to the displacement from time t_1 to t_2.

$$\int_{t_0}^{t_2} f(t)\, dt = \int_{t_0}^{t_1} f(t)\, dt + \int_{t_1}^{t_2} f(t)\, dt$$

(P2) If w is some continuous and bounded function with $v(t) \leq w(t)$ for each $t \in (t_i, t_f)$, then

$$\int_{t_i}^{t_f} v(t)\, dt \leq \int_{t_i}^{t_f} w(t)\, dt.$$

Consistently faster objects exhibit greater displacement.

(P3) If v is a constant function $v : t \mapsto c$ for some constant c, over a time interval t_i to t_f, then

$$\int_{t_i}^{t_f} v(t)\, dt := c(t_f - t_i).$$

Objects traveling at a constant velocity have a simple formula for calculating displacement.

In property 1, that is (P1), there is no restriction that time t_1 be between t_0 and t_2. To see how this works, imagine watching a marathon from start to finish. If we rewind the marathon footage, we will see marathoners running −42.195 km. The marathoners will need to run 42.195 km to return to the finish line. So there is no problem calculating the displacement of an object from 9PM to 9AM of the same day, as long as we put on a minus sign at the end. In symbols, our convention is

$$\int_{t_i}^{t_f} f(t)\, dt = -\int_{t_f}^{t_i} f(t)\, dt. \tag{3.1}$$

While we are on the subject of technicalities, recall that even if we are standing perfectly still, because the earth is moving, so are we. Thus we get a boost exceeding 1600 km/hr (exact figure depends on our location), even if we are staying perfectly still. To account for such differences, we can take our original velocity function f and subtract some predetermined constant v, where v could be 1600 km/hr. In such a case, the displacement between time t_i and time t_f could be denoted by $\int_{t_i}^{t_f} \big(f(t) - v\big)\, dt$,[3] where we have subtracted the velocity due to earth's motion. Alternatively, we could continue to measure displacements as before, and only when we need to conform to other conventions, make up for the difference. This is done using property (P3) to calculate $\int_{t_i}^{t_f} f(t)\, dt - v(t_f - t_i)$. This establishes the equality:

$$\int_{t_i}^{t_f} \big(f(t) - v\big)\, dt = \int_{t_i}^{t_f} f(t)\, dt - v(t_f - t_i). \tag{3.2}$$

[3]Here, v is being used as a constant function. Thus $\int_{t_i}^{t_f} f(t) - v(t)\, dt$ would mean the same thing.

First fundamental theorem of calculus

In our study of differentiation in Chapter 2, we were interested in velocity *functions*, rather than velocity itself. Likewise, we will turn our attention to displacement *functions*. From some initial time t_i, we define a displacement function F associated to a velocity function f as the function

$$F : t \mapsto \int_{t_i}^{t} f(x)\,dx.$$

The first thing we should note is that the rate of change of a displacement should be its velocity. That is, we expect

$$F' = f. \tag{3.3}$$

We can also write Equation 3.3 with the symbol $\frac{\mathrm{d}}{\mathrm{d}t}$, which means differentiate with respect to t

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{t_i}^{t} f(x)\,dx = f(t).$$

Let us verify that this is indeed the case. We take a bounded function f that is continuous at t. Define the displacement function $F : t \mapsto \int_{t_i}^{t} f(x)\,dx$, measured from some initial time t_i. What we want to show is that

$$F(t+\alpha) = F(t) + f(t)\alpha + |\alpha| o(1).$$

To achieve this, it is sufficient to show that

$$\left|F(t+\alpha) - F(t) - f(t)\alpha\right| \leq |\alpha| o(1).$$

Analogous to the steps for estimating displacements, we first consider a time slice of nonzero length α:

$$F(t+\alpha) - F(t).$$

By property (P1), we have

$$F(t+\alpha) - F(t) = \int_{t_i}^{t+\alpha} f(x)\,dx - \int_{t_i}^{t} f(x)\,dx = \int_{t}^{t+\alpha} f(x)\,dx.$$

We will define $f(t)$ to be our base level for velocity (just as we established the speed of the earth to be the base level for velocity by subtracting the speed of the earth in Equation 3.2). To match units, we multiply $f(t)$ by α and subtract to get

$$F(t+\alpha) - F(t) - f(t)\alpha = \int_{t}^{t+\alpha} f(x)\,dx - f(t)\alpha.$$

Using Equation 3.2, and applying absolute values everywhere to suppress questions about the sign of α, we have

$$\left|F(t+\alpha) - F(t) - f(t)\alpha\right| = \left|\int_{t}^{t+\alpha} f(x) - f(t)\,dx\right|.$$

Next, we move on to the second step of the estimation of displacements: we will pick a representative velocity for the time interval from time t to $t+\alpha$. In particular, we know f is bounded,[4] so there is a least upper bound for the values of the function f during the time interval $[t, t+\alpha]$. In fact, since $f(t)$ is a constant, there will be a least upper bound for the values of the function $|f - f(t)|$ during the time interval $[t, t+\alpha]$, which we will denote by $\sup_{x\in[t,t+\alpha]} |f(x) - f(t)|$. We take this as the representative and use property (P2) to obtain the following.

$$\left|F(t+\alpha) - F(t) - f(t)\alpha\right| = \left|\int_t^{t+\alpha} f(x) - f(t)\,dx\right| \leq \left|\int_t^{t+\alpha} \sup_{x\in[t,t+\alpha]} |f(x) - f(t)|\,dx\right|$$

Finally, we move on to the final step of the estimation of displacements. By property (P3),

$$\left|F(t+\alpha) - F(t) - f(t)\alpha\right| \leq \left|\int_t^{t+\alpha} \sup_{x\in[t,t+\alpha]} |f(x) - f(t)|\,dx\right| = |\alpha| \sup_{x\in[t,t+\alpha]} |f(x) - f(t)|.$$

Now let us imagine reducing the time interval by taking $\alpha \to 0$. Then each point x in the time interval $[t, t+\alpha]$ drops to t. Hence as $\alpha \to 0$, x drops to t, and by continuity of f at time t: $f(x) - f(t) = o_\alpha(1)$. Thus $\sup_{x\in[t,t+\alpha]} |f(x) - f(t)| = o_\alpha(1)$. Therefore,

$$F(t+\alpha) - F(t) - f(t)\alpha = |\alpha| o_\alpha(1)$$

and we conclude that F is differentiable at t with $F'(t) = f(t)$.[5]

Theorem 5 (First Fundamental Theorem of Calculus)**.** Suppose f is a bounded function defined on a closed interval $[t_i, t_f]$ that is continuous on $t \in [t_i, t_f]$ and we take

$$F : t \mapsto \int_{t_i}^{t} f(x)\,dx.$$

Then F is differentiable at t with $F'(t) = f(t)$.

This important result shows that we can generalize our intuitive idea that the rate of change of a displacement is its velocity, and apply them to functions beyond velocities and displacements. Just like we generalized the concept of a velocity into the notion of a derivative, we now generalize the notion of a displacement. The objects with the symbol $\int$ that obey properties (P1), (P2), (P3), and Equation 3.2 are called **integrals**. The calculation of integrals is called **integration**.

There are several types of integrals. Suppose we have an integral $I := \int_a^b f(x)\,dx$. If b is a constant, then I is a real number (generalizing "displacement"). If b is a variable, then the integral I is a function $I : b \mapsto \int_a^b f(x)\,dx$ (generalizing a "displacement function"). The latter is bad style, for we expect b to be a symbol for a constant, not a variable. It would be better in this case, as an example, to define $I : t \mapsto \int_a^t f(x)\,dx$ or $I : x \mapsto \int_a^x f(t)\,dt$.

[4] Recall that if something is bounded, it has both an upper bound and a lower bound

[5] Notice that it was necessary to take absolute values, for if $F(t+\epsilon) - F(t) - f(t)\epsilon \leq |\epsilon| o(1)$, we cannot conclude that $F(t+\epsilon) - F(t) - f(t)\epsilon = |\epsilon| o(1)$. Negative functions are smaller than $o(1)$, but are not necessarily $o(1)$. As we discussed in Section 2.2, if $|F(t+\epsilon) - F(t) - f(t)\epsilon| \leq |\epsilon| o(1)$, then we know that $F(t+\epsilon) - F(t) - f(t)\epsilon = |\epsilon| o(1)$

There is yet another type of integral. Compare the expression $\left(\frac{x^2}{2}\right)' = x$ with the expression $F' = f$ from the Fundamental Theorem of Calculus (Equation 3.3). We see that $F := \frac{x^2}{2}$ has an interpretation of a "displacement function" for the "velocity function" $f : x \mapsto x$. However, $\frac{x^2}{2}$ is not unique in this regard. For example $(\frac{x^2}{2} + 1)' = x$, $(\frac{x^2}{2} + 2)' = x$, and so on. This makes sense, for there are infinitely many conventions to measure displacements: one convention where the measurement starts from the King's palace, one convention where the measurement starts from the library, etc. It is natural to classify all such functions into one group.

The **antiderivative** or the **indefinite integral** of a function f, written $\int f(\square)\, d\square$ is the set of functions whose derivative is f.[6] For example, $\int x\, dx = \frac{x^2}{2} + c$, where c denotes the arbitrary constant representing the degree of freedom in choosing where we can set the origin for measuring "displacements". In contrast to the indefinite integral, integrals of the form $\int_{\spadesuit}^{\heartsuit} f(\square)\, d\square$ are called **definite integrals**. For example, $\int_a^b x\, dx$ is a *definite* integral.

Second fundamental theorem of calculus

Now that we have discussed a fair bit about displacement functions, we now turn to the natural question: how to do we calculate displacements? For example, what is the real number corresponding to the definite integral $\int_0^1 x\, dx$?

To make this concrete, let us imagine that we are walking up a very long stairwell and we wish to measure how much height we have traversed. One way would be count the number of steps per second say, and add them all up.

An easier way would be to use an altimeter, any one that works, and then (i) measure our altitude at the beginning of the journey and (ii) measure our altitude at the end of our climb, then (iii) calculate: final altitude − beginning altitude.

Notice how the altitude the altimeter is calibrated to makes no difference to the result: whether the altitude begins at sea level, or the peak of Mount Everest at a certain year, they are both ok. However, it is crucial that we stick to the same altimeter. If we swap out one for another in the middle, then this method is no good.

Let us use this thinking to calculate the definite integral $\int_0^1 x\, dx$. We know from our discussion before that $\int x\, dx = \frac{x^2}{2} + c$. Pick an "altimeter"—we'll pick $\frac{x^2}{2} + 3.141592$. At the start time of $t_i = 0$, we have an altimeter reading of $\frac{0^2}{2} + 3.14192$. At the end time of $t_f = 1$, we have an altimeter reading of $\frac{1^2}{2} + 3.141592$. Subtract the former reading from the latter and we see that $\int_0^1 x\, dx = \frac{1}{2}$.

Let us see how our method could fail. Well, if we are allowed to move about with sudden jumps, or move with zero velocity (Examples 2 and 3 in Section 2.5), then our method will not work. So we will only be able to apply this method to continuous functions, and we will have to disallow instantaneous teleportations (motion without velocity). By working with real numbers (Section 2.5), we do not have to worry about the latter, for a function with zero derivative (no velocity) will be a constant function (no motion).

[6] This is analogous to the expression $o(1)$, since $o_\alpha(1)$ is actually a collection of functions that drop to zero as $\alpha \to 0$. Even though antiderivatives and $o(1)$ are sets of functions, we treat them like functions.

Now let us verify that our method works. Consider a continuous function f, and let F be an antiderivative of f. The "manual way" of calculating "altitude" can be expressed by the symbol $\int_a^x f(t)\,dt$. By the Fundamental Theorem of Calculus,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\,dt = \frac{\mathrm{d}}{\mathrm{d}x}F(x).$$

The subtraction rule for derivatives tells us that $g' = h'$ is equivalent to $(g-h)' = 0$, and so

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\int_a^x f(t)\,dt - F(x)\right) = 0.$$

Since the derivative is zero and instantaneous teleportations are not permitted, the function inside the brackets must be a constant function:

$$\int_a^x f(x)\,dx - F(x) = c.$$

To find the value of the constant, we will evaluate the function at the starting time a and use the third property of an integral (P3) to obtain the following.

$$c = \int_a^a f(x)\,dx - F(a) = 0 - F(a)$$

Theorem 6 (Second Fundamental Theorem of Calculus)**.** If f is bounded and continuous on a closed interval $[t_i, t_f]$ with antiderivative F, then

$$\int_{t_i}^{t_f} f(x)\,dx = F(t_f) - F(t_i).$$

Sometimes we will find it convenient to use the shorthand

$$F(x)\Big|_{x=t_i}^{t_f} := F(t_f) - F(t_i).$$

For example,

$$\int_0^1 x\,dx = \frac{x^2}{2}\Big|_{x=0}^{1} := \frac{1^2}{2} - \frac{0^2}{2} = \frac{1}{2}.$$

Here is a comment on the theorem. The right hand side is *not* a definition of the definite integral on the left. The theorem simply says that *if* an antiderivative is available, then there is a shortcut to computing the definite integral. A particular altimeter from one manufacturer is not the definition of the elevation of a location, but it we have one available, why not use it?

3.2 Arithmetic of Displacements

We now port some of the essential differentiation rules we obtained in Chapter 2 for use with integrals.

Linearity of Integrals

Recall that $(f+g)' = f' + g'$ and $(cf)' = cf'$ for real c. Suppose f and g are continuous and thus have antiderivatives F and G, respectively. Ignoring the arbitrary constants (which are subsumed), we have

$$\int (f+g) = (F+G) = \int f + \int g \qquad\qquad \int (cf) = cF = c\int f.$$

Similarly,

$$\begin{aligned}\int_a^b (f+g)(x)\,dx &= (F+G)(b) - (F+G)(a) = [F(b)-F(a)] + [G(b)-G(a)] \\ &= \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.\end{aligned}$$

and

$$\int_a^b (cf)(x)\,dx = (cF)(b) - (cF)(a) = c[F(b)-F(a)] = c\int_a^b f(x)\,dx.$$

Integration by parts

Is there an analogue of the product rule $(fg)' = (f'g) + (fg')$ for integration? Taking the antiderivative of both sides gives

$$\int (fg)' = \int (f'g) + \int (fg').$$

For the antiderivative of $(fg)'$, we pick fg (with an arbitrary constant of zero), and we have **integration by parts**. Repeating the derivation for the definite integral gives an analogous result.

Theorem 7 (Integration by Parts)**.** If f and g are differentiable and f' and g' are continuous, then

$$\int fg' = fg - \int f'g \qquad\qquad \int_a^b f(x)g'(x)\,dx = fg\Big|_a^b - \int_a^b f'(x)g(x)\,dx.$$

Substitution rule

We seek an analogue of the chain rule for integration. Recall the chain rule states that

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

The right term is fairly complex, but the left term admits a simple application of the Second Fundamental Theorem of Calculus:

$$\int_a^b (f\circ g)'(x)\,dx = (f\circ g)(b) - (f\circ g)(a).$$

We have an opportunity to apply the Second Fundamental Theorem of Calculus once more:

$$(f\circ g)(b) - (f\circ g)(a) = \int_{g(a)}^{g(b)} f'(u)\,du.$$

Therefore, the following holds (the second equality is an application of the chain rule).

$$\int_{g(a)}^{g(b)} f'(u)\,du = \int_a^b (f\circ g)'(x)\,dx = \int_a^b (f'\circ g)(x)\cdot g'(x)\,dx$$

This is the **substitution rule**. We will make a minor cosmetic change, replacing each symbol "f'" in the above with the symbol "f".

Theorem 8 (Substitution Rule)**.** If f is continuous, g is differentiable, and g' is continuous, then

$$\int_{g(a)}^{g(b)} f(u)\,du = \int_a^b (f\circ g)(x)\cdot g'(x)\,dx. \tag{3.4}$$

Taylor's theorem (optional)

The position x of an object traveling at speed v starting from an initial position x_0 is given by

$$x = x_0 + vt + |t|o(1).$$

By Challenge 10 part (b), if the acceleration of our object is a constant, then we know the $|t|o(1)$ term exactly, with

$$x = x_0 + v_0 t + \frac{1}{2}at^2 \tag{3.5}$$

where v_0 is the initial velocity of our object. In particular, if our object has acceleration $a = 0$, then $x = x_0 + vt$ (notice the velocity must be a constant). What if our object undergoes variable acceleration over time?

In order to make the dependencies more transparent, we will make the change of notation: f instead of x, f' instead of v, and f'' instead of a. We will also denote the input by x rather than t.

If the object's acceleration is variable, so is the object's velocity. Thus integration is in order and by the Fundamental Theorem of Calculus,

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\,dt.$$

What next? There is not a whole lot to try, the only thing that is applicable from what we have done so far is integration by parts.

The relevant product for integration by parts is $t\cdot f'$ so that $\int f'(t)\,dt = \int t'\cdot f'(t)\,dt$. Applying integration by parts gives

$$f(x) = f(x_0) + t\cdot f'(t)\Big|_{t=x_0}^{x} - \int_{x_0}^{x} t'\cdot f''(t)\,dt = f(x_0) + \big(xf'(x) - x_0 f'(x_0)\big) - \int_{x_0}^{x} t\cdot f''(t)\,dt.$$

We are looking for something of the form of Equation 3.5 above. This means that we need to change the $f'(x)$ in the second term on the right side into $f'(x_0)$. This can be done with the Fundamental Theorem of Calculus: $f'(x) = f'(x_0) + \int_{x_0}^{x} f''(t)\,dt$. Applying this gives

$$f(x) = f(x_0) + x\left(f'(x_0) + \int_{x_0}^{x} f''(t)\,dt\right) - x_0 f''(x_0) - \int_{x_0}^{x} t\cdot f''(t)\,dt.$$

We can tidy up using the integration rule $\int a + \int b = \int (a + b)$ to get

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \int_{x_0}^{x} (x - t)f''(t)\,dt.$$

This is the analogue of Equation 3.5 for variable acceleration f''.

As long as function f has the continuous derivatives, we can continue this procedure. Apply integration by parts on the integral $\int (x - t)f''(t)\,dt$ with the product $-\frac{(x-t)^2}{2}f''$ to get

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(x_0) + \int_{x_0}^{x} \frac{(x - t)^2}{2}f'''(t)\,dt.$$

Once more! Integration by parts on the integral $\int \frac{(x-t)^2}{2}f'''(t)\,dt$ with the product $-\frac{(x-t)^3}{3!}f'''$ gives

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(x_0) + \frac{(x - t)^3}{3!}f'''(x_0) + \int_{x_0}^{x} \frac{(x - t)^3}{3!}f^{(4)}(t)\,dt.$$

Our calculations suggest the following result.

Theorem 9 (Taylor's Theorem)**.** If function f is $(k + 1)$-times differentiable, then

$$f(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!}(x - x_0)^k\,dt. \tag{3.6}$$

The polynomial $\sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$ is called the **k-th order Taylor polynomial of f at x_0**.

Proof. We use the well-ordering principle. Suppose $k+1$ is the smallest integer for which the result does not hold. By assumption, Equation 3.6 holds. As before, we integrate by parts. The product is $-\frac{(x-x_0)^{k+1}}{(k+1)!}f^{(k+1)}$. The result of the integration by parts is

$$\int_{x_0}^{x} \frac{f^{(k+1)}(t)}{(k + 1)!}(x - x_0)^k\,dt = \frac{(x - x_0)^{k+1}}{(k + 1)!}f^{(k+1)}(x_0) + \int_{x_0}^{x} \frac{(x - x_0)^{k+1}}{(k + 1)!}f^{(k+2)}(t)\,dt. \tag{3.7}$$

Therefore,

$$f(x) = \sum_{n=0}^{k+1} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \int_{x_0}^{x} \frac{f^{(k+2)}(t)}{(k + 1)!}(x - x_0)^{k+1}\,dt.$$

But this is simply Equation 3.6 with k replaced by $(k+1)$. Therefore, Equation 3.6 always holds. □

3.3 Area Under a Curve

Consider the three diagrams in Figure 3.8. Each curve may be interpreted as telling us the velocity of an object from the time t_i to time t_f. The first is the simplest, our object is moving at a

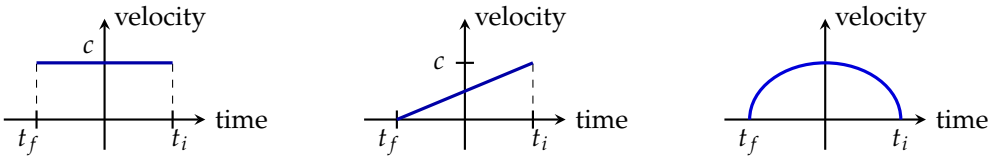


Figure 3.8: Velocities of objects from time t_i to t_f. The displacement is the area under the curve.

constant velocity. Using a definite integral and the Fundamental Theorem of Calculus, we know that our object is subject to the displacements d_1, whose value is

$$d_1 = \int_{t_i}^{t_f} c\,dt = ct\Big|_{t_i}^{t_f} = c(t_f - t_i).$$

The displacement takes a simple form, as it should: it says that the displacement d_1 is the velocity times the duration of travel. But we can interpret this as the area of a square whose height is c and base length is $t_f - t_i$. Such a square is literally drawn in our diagram: it is the shape that is enclosed inside the curve, the x-axis, and the equations $t = t_i$ and $t = t_f$.

We thus have another interpretation of integration: as an area under a curve. Our method of calculating the displacement of an object by accumulating its velocity is the same as that for calculating area under a curve!

Let us try this out for the function graphed in the second diagram in Figure 3.8. The area under the curve is a **right triangle**: a triangle where one of the angle measures 90°. A triangle with height h and base length l occupies precisely half the area of a square with height h and base length l. Therefore, a triangle with height h and base length l has area $hl/2$. Applying this to our curve, we see that the displacement d_2 of our object is: $c(t_f - t_i)/2$. Let us repeat this calculation with integration. The formula for a line is given by $t \mapsto wt + b$ where the constant w is called the **slope**, or **weight**, of the line and the constant b is called the **bias**. The slope measures the rate of change of the line. In this case, the rate of change is $\frac{c-0}{t_f - t_i}$ since it steadily increased from 0 to c over the time t_i to t_f. To find the bias, pick any point on the line. Any point suffices, but the point $(t_i, 0)$ is a particularly simple one. We then apply the x-coordinate of the point to our formula and correct for the difference with the y-coordinate. The formula is $\frac{c}{t_f - t_i}t + b$, so plugging in the input t_i into the variable t gives $\frac{ct_i}{t_f - t_i} + b$ which must equal the y-coordinate: 0. Therefore, the bias b is given by $b := -\frac{ct_i}{t_f - t_i}$ and our formula for the line is

$$\frac{c}{t_f - t_i}t - \frac{ct_i}{t_f - t_i}.$$

The definite integral for the function above from t_i to t_f is

$$\begin{aligned} d_2 &= \int_{t_i}^{t_f} \left(\frac{c}{t_f - t_i}t - \frac{ct_i}{t_f - t_i}\right) dt = \frac{c}{t_f - t_i}\int_{t_i}^{t_f} t\,dt - \frac{ct_i}{t_f - t_i}\int_{t_i}^{t_f} 1\,dt \\ &= \frac{ct^2}{2(t_f - t_i)}\Big|_{t_i}^{t_f} - \frac{ct_i t}{t_f - t_i}\Big|_{t_i}^{t_f} = \frac{c(t_f^2 - t_i^2)}{2(t_f - t_i)} - \frac{ct_i(t_f - t_i)}{(t_f - t_i)} = \frac{c(t_f^2 - 2t_f t_i + t_i^2)}{2(t_f - t_i)} = \frac{c(t_f - t_i)^2}{2(t_f - t_i)}. \end{aligned}$$

Since $t_f - t_i$ is nonzero, we may cancel out the common factors in the fraction to get $d_2 = c(t_f - t_i)/2$. A whole lot more work to get the obvious answer!

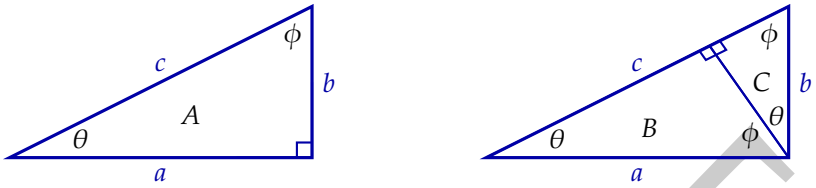


Figure 3.9: A single line is sufficient to prove the Pythagorean theorem.

The horrendous calculation for the area of a triangle shows us the advantage of having multiple perspectives. A difficult problem in one perspective might turn out to be far simpler in another. So how about we try out another perspective on triangles?

Have a look at the right triangle depicted on the left diagram of Figure 3.9. In either of our calculations, we never had to use the length c of the **hypotenuse** (the longest side of a right triangle). How about we try to calculate the area of a right triangle without using the lengths a and b? The other pieces of information we have available are the length of the hypotenuse c and two angles θ and ϕ (labeled in the left diagram of Figure 3.9). Since the angles of a triangle add up to 180°, we have $\theta + \phi = 90°$.

An angle is the ratio of two lengths and is therefore dimensionless (see footnote: angle ψ is the length of the red arc divided by the circumference of the blue circle; this ratio is independent of the radius involved).[7] The only **dimensionful** quantity (quantity with a dimension) is the length c. By dimensional analysis, the area of the triangle A will then be given by

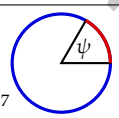
$$A = f(\theta, \phi)c^2$$

where f is some dimensionless function of our angles θ and ϕ. Draw a line from the right angle to the hypotenuse such that two new right angles are formed (diagram on the right in Figure 3.9). There are now three right triangles in one diagram. Denote the area of the larger of the new triangle by B and the area of the smaller of the new triangle by C. All three right triangles have the angles θ and ϕ. The area B is given by $f(\theta, \phi)a^2$ and the area C is given by $f(\theta, \phi)b^2$. Since $B + C = A$, we have $f(\theta, \phi)a^2 + f(\theta, \phi)b^2 = f(\theta, \phi)c^2$. Since $f(\theta, \phi)$ must be nonzero, we can divide both sides by $f(\theta, \phi)$ to obtain the **Pythagorean theorem**:

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

Circles and ellipses

The function $f(\theta, \phi)$ and our attempt to calculate a triangle's area with it is an example of a *MacGuffin*. True to a MacGuffin's purpose we immediately return to the plot: we want to use



7

integrals to calculate areas under curves. The curves corresponding to constant velocity (square) and constant acceleration (triangle) were simple. How about an arc as shown in the third diagram of Figure 3.8? The arc corresponds to the top half of an ellipse. Before discussing ellipses, it would be better to talk about circles, which are simpler. Notice that an ellipse former needs two real numbers (width and height) to describe, while a circle is described by a single number (radius).

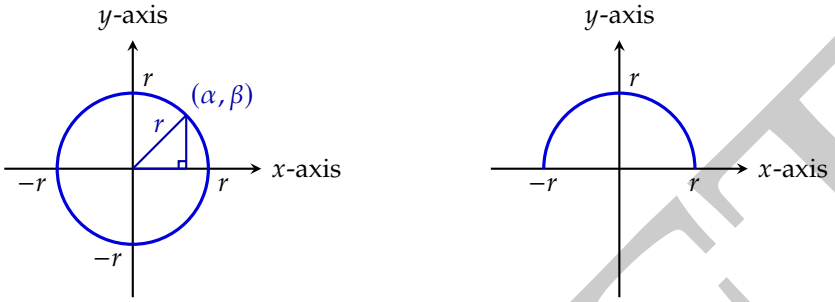


Figure 3.10: A circle of radius r at the origin is given by the equation $x^2 + y^2 = r^2$ (left). The top semicircle is given by the equation $y = \sqrt{r^2 - x^2}$ (right).

To apply integration, we need a function that describes a curve. What is the equation of a circle of radius r centered at the origin? Take any point in a circle that is not on an axis, and label the x-coordinate by α and the y-coordinate by β (see the diagram on the left in Figure 3.10). We may create a right triangle whose base is on the x-axis. By the Pythagorean theorem, $\alpha^2 + \beta^2 = r^2$. Thus points on the circle that are not located in the axis are described by the equation $x^2 + y^2 = r^2$. But the points located in the axis also satisfy the equation $x^2 + y^2 = r^2$ because one of the term in the left side is r^2 and the other is zero. Therefore, the equation of a circle of radius r centered at the origin is given by $x^2 + y^2 = r^2$. To find the equation of the top half of a circle, called the top **semicircle**, which we may interpret as describing the velocity of an object from time $-r$ to time r, we subtract x^2 from both sides of the equation and use the fact that $y > 0$ on the top half to take the square root. This gives the equation $y = \sqrt{r^2 - x^2}$.

Suppose we knew nothing about area formulas for circles and ellipses. By dimensional considerations, we guess that the area of a circle of radius r should be cr^2 for some dimensionless constant c. What is the constant c? Normally we would plug in the value $r = 1$ to find the value of c, but we are starting from scratch so there is no other information to help us. We have no choice but to define the constant. A **unit circle** is a circle of radius 1. The constant π is defined to be the value of the area of a unit circle.

A unit circle may be depicted on a plane. If we position the x-axis and the y-axis to be the origin at the center of the unit circle as shown in the left of Figure 3.11, the graph of the unit circle is given by the equation: $x^2 + y^2 = 1$. In particular, the equation for the top semicircle, shown on the right of Figure 3.11 is given by $y = \sqrt{1 - x^2}$. To see this, subtract x^2 from both sides of the equation to get $y^2 = 1 - x^2$ then take square roots on both sides (which is ok to do since $y > 0$ on this side of the circle).

If we think of the equation $y = \sqrt{1 - x^2}$ as describing the velocity y of a car at time x from time -1 to time 1, Then the integral of the function $\sqrt{1 - x^2}$ from -1 to 1 is the accumulated velocity

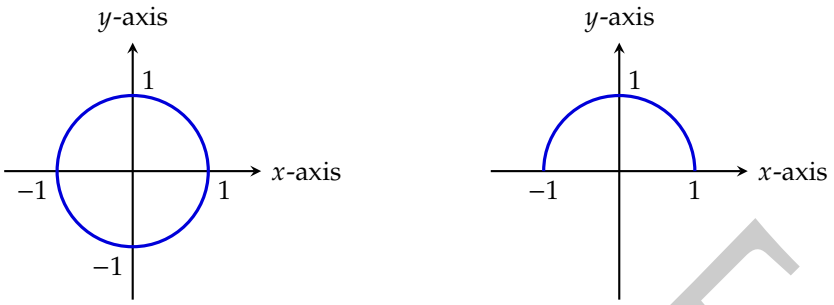


Figure 3.11: The graphs of equation $x^2 + y^2 = 1$ (left) and equation $y = \sqrt{1 - x^2}$ (right).

during this time. Geometrically, this corresponds to the area enclosed between the semicircle and the x-axis. Since the area of a circle is twice that of the area of a semicircle of the same radius,

$$\pi := 2 \int_{-1}^{1} \sqrt{1 - x^2}\, dx.$$

Now that we have defined π such that it is (the value of) the area of the unit circle, let us use this information to check our guess that the area of a circle with radius r is πr^2. Since we already know the area of a unit circle (defined to be π), the most straightforward path will be to reduce the circle of radius r into a unit circle. The function f describing the top semicircle of radius r is given by $\sqrt{r^2 - x^2}$. We want the accumulated "velocity" from time $-r$ to r. The integral we wish to calculate is thus $\int_{-r}^{r} \sqrt{r^2 - x^2}\, dx$, where radius r is a positive constant. The function $f : x \mapsto \sqrt{r^2 - x^2}$ can be made to resemble the function describing the top semicircle of a unit circle by pulling out the r term:

$$\sqrt{r^2 - x^2} = \sqrt{r^2(1 - x^2/r^2)} = r\sqrt{1 - (x/r)^2}.$$

To fully reduce the term $\sqrt{1 - (x/r)^2}$ into $\sqrt{1 - x^2}$, we will make the substitution $g : x \mapsto x/r$. This calls for the substitution rule with $u := g(x)$

$$\int_a^b (f \circ g)(x) \cdot g'(x)\, dx = \int_{g(a)}^{g(b)} f(u)\, du.$$

Now $g'(x) = 1/r$, which is a problem because we have a factor of r instead in $r\sqrt{1 - (x/r)^2}$. We remedy this by multiplying and dividing by r:

$$f(x) = \sqrt{r^2 - x^2} = r\sqrt{1 - (x/r)^2} = r^2(1/r)\sqrt{1 - (x/r)^2}.$$

All the preparation is done and we just have to apply the substitution rule. The area $A(r)$ of a circle of radius r is given by

$$A(r) = 2 \int_{-r}^{r} \sqrt{r^2 - x^2}\, dx = 2r^2 \int_{-r}^{r} \sqrt{1 - (x/r)^2} \cdot \frac{1}{r}\, dx \tag{3.12}$$

$$= \left(2 \int_{g(-r)}^{g(r)} \sqrt{1 - g(x)^2} \cdot g'(x)\, dx \right) r^2 = \left(2 \int_{-1}^{1} \sqrt{1 - u^2}\, du \right) r^2 = \pi r^2. \tag{3.13}$$

The integral $2\int_{-1}^{1}\sqrt{1-u^2}\,du$ is defined to be π, and so $A(r)=\pi r^2$, just as we guessed.

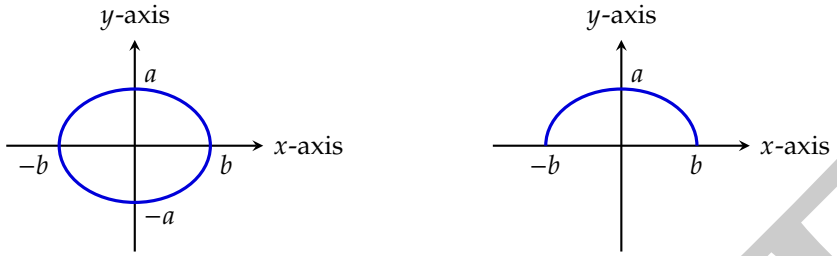


Figure 3.14: The graphs of equation $\frac{x^2}{b^2}+\frac{y^2}{a^2}=1$ (left) and equation $y=\sqrt{a^2-(ax/b)^2}$ (right).

We now turn to the ellipse. What is an equation that describes an ellipse of height $2a$ and width $2b$ centered at the origin? Since the answer is not obvious at all, let us try to reduce this problem into a simpler one. If we measure the x axis in units of b so that $b=1$, then our ellipse will have width of 2. Similarly, if we measure the y axis in units of a so that $a=1$, then our ellipse will have a height of 2. In other words, with our new choice of units, our ellipse becomes a unit circle! The unit circle's equation is given by $x^2+y^2=1$. We see that with the substitution $x\mapsto x/b$ (this makes $b=1$) and $y\mapsto y/a$ (this makes $a=1$) we obtain the equation of a unit circle. Therefore, the equation of an ellipse is $\frac{x^2}{b^2}+\frac{y^2}{a^2}=1$. To get the top half of an ellipse, which allows us to interpret the area under a curve as a displacement (and thus an integral), we subtract both sides of the equation of an ellipse by $\frac{x^2}{b^2}$ and then multiply both sides by a^2 to isolate the y^2 term. Since $y>0$ on the top half of an ellipse, we can take a square root of both sides to get the equation $y=\sqrt{a^2-(ax/b)^2}$.

Let us use the interpretation of area under the curve as an integral to find a formula for the area of an ellipse. We will reduce our problem into one we have already solved: the formula for the area of a circle. The equation for the top half of an ellipse $y=\sqrt{a^2-(ax/b)^2}$ can be transformed into the equation for the top semicircle of radius a given by $\sqrt{a^2-x^2}$ using the substitution $g:x\mapsto ax/b$. Since $g'(x)=a/b$, the substitution rule with $u:=g(x)$ gives the area $A(a,b)$ of an ellipse as

$$A(a,b)=2\int_{-b}^{b}\sqrt{a^2-(ax/b)^2}\,dx=2\int_{-b}^{b}\frac{b}{a}\cdot\frac{a}{b}\sqrt{a^2-(ax/b)^2}\,dx=\frac{2b}{a}\int_{-b}^{b}\sqrt{a^2-(ax/b)^2}\cdot\frac{a}{b}\,dx$$
$$=\frac{b}{a}\left(2\int_{g(-b)}^{g(b)}\sqrt{a^2-g(x)^2}\cdot g'(x)\,dx\right)=\frac{b}{a}\left(2\int_{-a}^{a}\sqrt{a^2-u^2}\,du\right)=\frac{b}{a}\left(\pi a^2\right)=\pi ab$$

where we have used the fact that $2\int_{-a}^{a}\sqrt{a^2-u^2}\,du$ is the area of a circle of radius a (Equation 3.12). The answer $A(a,b)=\pi ab$ confirms our guess from dimensional analysis at the beginning of Chapter 2.

Observe that in both of our calculations for the area of a circle and an ellipse, the only substitution we needed was a rescaling of the variable x. In the former case it was $g:x\mapsto x/r$, while in the latter case it was $g:x\mapsto ax/b$. Since r,a, and b are all positive constants, these substitutions are simply a change of units. For example, in the former case our substitution simply rescales our x axis such that the number r becomes our unit of measurement. It has the effect of setting $r=1$ (if

a meter is our unit of measurement, then the length of a meter becomes 1) and turning our circle into a unit circle. The case of the ellipse is similar where we are setting our unit of measurement such that $\frac{a}{b} = 1$, in other words: $a = b$, which turns our ellipse into a circle!

This demonstrates the special case of the substitution rule: if we measure the x-axis in units of a nonzero constant c so that $c = 1$, then $\int_{-c}^{c} f(x)\,dx = c\int_{-1}^{1} f(x)\,dx$.

Solids of revolution

We were able to calculate areas by interpreting area under a curve as the displacement of an object, moving with velocity described by the curve. Can we measure volume in a similar way?

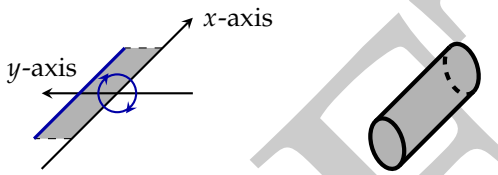


Figure 3.15: Rotating the area underneath the constant function defined on a finite interval sweeps out a cylinder.

Suppose we have a function f defined on an interval $[a, b]$ with $f \geq 0$ on each point it is defined on (we will call such functions **positive functions**).[8] If we rotate the area enclosed by the function f, the x-axis, and the lines $x = a$ and $x = b$, then we sweep out a geometrical solid. From Figure 3.15 we see that rotating an area of a square sweeps out a cylinder. Rotating an angled line (with positive function values only) sweeps out a cone, rotating a semicircle sweeps out a sphere. Much like we can accumulate velocity to obtain displacement, we should be able to accumulate area to obtain volume. This will enable us to use the machinery of integrals to calculate the formulas of volumes for a large class of geometrical objects. Let us guess the formula of the volume of a solid obtained by sweeping a positive function defined in a finite interval $[a, b]$.

Recall that a derivative of g has dimension of g divided by the dimension of its input. Since integration is an inverse operation of differentiation, due to the Fundamental Theorem of Calculus ($\int_{t_i}^{x} g(t)\,dt)' = g(x)$, an integral of g has dimension of g multiplied by the dimension of its input. For example, for velocity g with input time t, the integral of g (displacement) has dimension velocity (Length/Time) multiplied by Time, which is Length.

Under the interpretation of an integral as an area, a function f and its input x both have the dimension of Length, allowing $\int_a^b f(x)\,dx$ to represent an area with dimension Length^2. We see that the expression $\int_a^b [f(x)]^2\,dx$ is the simplest one that has the desired dimension Length^3 of a volume. The only thing missing is our ignorance about dimensionless constants. We therefore guess that the volume V is given by $\int_a^b c[f(x)]^2\,dx$ for some dimensionless constant c.[9]

[8] Thus a zero function is also a "positive" function. It rolls off the tongue better than "nonnegative functions".

[9] Technically speaking the expression $c\left(\int_a^b f(x)\,dx\right) f$ also has the dimension of a volume, but it cannot be a volume because it is not a number but a function.

To find the constant c, we consider the simple case of a cylinder. The cylinder's volume can be found by taking the product of the base circle of radius r (whose area is πr^2) and its height h. Therefore, a cylinder of height h and radius r has volume $\pi r^2 h$. Let us set up our integral. The constant function $f : x \mapsto r$ defined on the interval $[0, h]$ will give us a rectangle with the desired shape. Our guess will thus give

$$\pi r^2 h = \int_0^h c[f(x)]^2\,dx = \int_0^h c \cdot r^2\,dx = cr^2 \int_0^h 1\,dx = cr^2 x\Big|_{x=0}^{x=h} = cr^2 h$$

and we see that the dimensionless constant is π.

The volume of a **solid of revolution** obtained by rotating a positive function f defined on an interval $[a, b]$ around the x-axis is given by

$$\int_a^b \pi \left[f(x)\right]^2\,dx.$$

Challenge 11

(a) Let r and h be positive real numbers and let $f : x \mapsto rx/h$ be defined on the interval $[0, h]$. Use the method of solid of revolution on the function f to verify that the volume of a cone of height h and circular base of radius r is given by the formula $\frac{\pi r^2 h}{3}$.

(b) Apply the solid of revolution to the equation for a semi circle of radius r to verify that a sphere of radius r has volume $\frac{4}{3}\pi r^3$.

3.4 Exponentiation Revisited

The logarithm function

Question: What is an antiderivative of the function $1/x$? Answer: Easy! Assign the dimension of Length to variable x so that $1/x$ has dimension Length^{-1}. Its antiderivative must then have dimension Length $\times$ Length^{-1}, in other words, it must be dimensionless. So we guess that the antiderivative of $1/x$ is an arbitrary constant c.

This is completely wrong! By the constant rule, we know that $c' = 0$, which is definitely not $1/x$. Now, we could ignore this problem and pretend that everything is ok. However, $1/x$ is such a simple yet important function that describes division by a variable. We will have to resolve this.

As we saw, if $1/x$ has an antiderivative, it must be dimensionless, so we have no clue to help us our. Just as we calculated areas of circles by defining the (value of the) area of a unit circle to be π, our solution will be to define a function that differentiates to $1/x$.

Definition 10. For each $x \in (0, \infty)$, the (**natural**) **logarithm** function is defined as

$$\log : x \mapsto \int_1^x \frac{1}{u}\,du.$$

If u is a positive constant, then $\frac{1}{u+\alpha} - \frac{1}{u} = -\frac{\alpha}{u^2+u\alpha}$. Drop $\alpha \to 0$ and we have $\frac{1}{u+\alpha} - \frac{1}{u} \to 0$. Therefore, $1/u$ is continuous and the Fundamental Theorem of Calculus gives $\log'(x) = 1/x$ for each $x \in (0, \infty)$.

By construction, the logarithm function is **strictly increasing**: if $0 < a < b$ then $\log(a) < \log(b)$. Property (P3) of integrals gives $\log 1 = \int_1^1 1/u \, du = 0$. The derivative of the logarithm function never vanishes (that is, the derivative is never zero), and since differentiability implies continuity, we conclude that the logarithm function is continuous.

The following properties of the logarithm function are the guarantors of the function's utility.

Proposition 11. Let x and y be positive real numbers; then

(a) $\log(xy) = \log x + \log y$,
(b) $\log \frac{x}{y} = \log x - \log y$,
(c) if p is a real number, then $\log x^p = p \log x$.

Proof. (a) We have two variables x and y. To make things more manageable, we first fix y and consider a function of x only. We will ignore $\log y$ and move $\log x$ to the left side by defining the following function: for positive y, let $g : x \mapsto \log(xy) - \log x$. We apply the differentiation rules, treating y as a constant to get

$$g'(x) = \frac{1}{xy} \frac{\mathrm{d}}{\mathrm{d}x}(xy) - \frac{1}{x} = \frac{1}{xy} \cdot y - \frac{1}{x} = 0.$$

The symbol $\frac{\mathrm{d}}{\mathrm{d}x}$ means, take the derivative with respect to x. This was necessary because the expression $(xy)'$ in a vacuum might be ambiguous, whereas $\frac{\mathrm{d}}{\mathrm{d}x}(xy)$ and $\frac{\mathrm{d}}{\mathrm{d}y}(xy)$ are both unambiguous.[10]

Our calculation shows that g is a function of x, whose rate of change with respect to x is zero. Thus g is actually a constant. To calculate g, observe that $g(1) = \log y - \log 1 = \log y - 0 = \log y$. This gives $\log(xy) - \log x = g(y) = \log y$, as desired.

(b) We multiply by 1 and use property (a) to get

$$\log x = \log\left(\frac{x}{y} \cdot y\right) = \log \frac{x}{y} + \log y.$$

Rearranging, we have $\log x - \log y = \log \frac{x}{y}$.

(c) We will return to this later.

□

We note one further property of the real numbers. We formalize the idea that a ruler, no matter how small, may be used to measure *any* length in finitely many steps, no matter how long. The proof will be reminiscent of our previous encounters with the well-ordering principle.

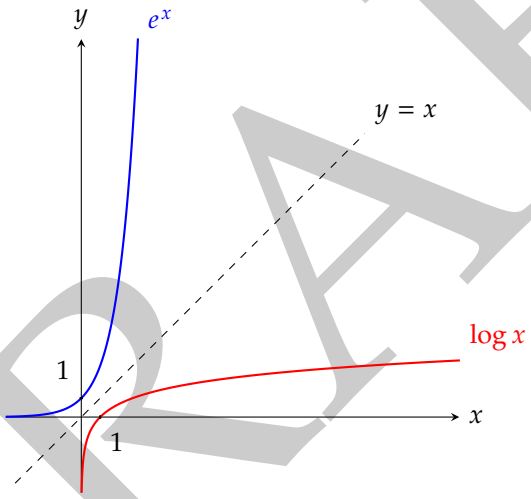
Theorem 12 (Archimedean property of $\mathbb{R}$)**.** If x is a positive real number and y is a real number, then there is some natural number n such that $nx > y$.

[10] Contrast this with the unambiguous expression $(cx^k)'$. Our convention is that c is a constant.

Proof. To obtain a contradiction, suppose there is no natural number n such that $nx > y$. Let S be the set of numbers nx for each natural number n. Since the set S is a nonempty set of real numbers containing 0, bounded from above by y, there is a least upper bound $u := \sup S$. Since x is positive we know that $u - x < u$, and so $u - x$ is not an upper bound of S. Since $u - x$ is *not* an upper bound of S, there must be a natural number m satisfying $mx > u - x$. Then $(m+1)x > u - x + x = u$, where $m + 1$ is a natural number. This contradicts our assumption that u is an upper bound of S. □

The exponential function

Recall that the logarithm function is strictly increasing with $\log 1 = 0$. Thus for any $\alpha > 1$, we have $\log \alpha > 0$. By Proposition 11 part (a), $\log(\alpha^n) = n \log \alpha$ for each natural number n. We conclude using the Archimedean property that the logarithm function is not bounded from above. Similarly, if $\alpha \in (0, 1)$, then $\log \alpha < 0$ and Proposition 11 part (a) tells us that the logarithm function is not bounded from below. Therefore, each positive real input x to the logarithm function is unambiguously associated with a unique real number $\log x$. We flip this relation and associate to each real number $\log x$, a unique positive real number x.



To formalize this, we define an *inverse* function of the logarithm function on $\mathbb{R}$.

Definition 13. The **exponential** function exp is defined on $\mathbb{R}$ such that $\exp \circ \log$ and $\log \circ \exp$ are the identity maps $x \mapsto x$.[11] More commonly, we write the exponential function as e^x, where $e^{\log x} = x$ for $x \in (0, \infty)$ and $\log(e^x) = x$ for $x \in \mathbb{R}$. The constant e (called **Euler's number**) is defined to be $\exp(1)$.

Since $\log \circ \exp$ is the constant map $x \mapsto x$, we have $(\log \circ \exp)' = 1$. Assuming exp is differentiable (to be justified in Chapter 4), applying the chain rule gives $(\log \circ \exp)' = \frac{1}{\exp} \cdot \exp'$. Hence $\exp'/\exp = 1$, and so $\exp' = \exp$. We restate this important property.

[11] Observe that $\exp \circ \log$ and $\log \circ \exp$ are different maps, because although exp is defined on $\mathbb{R}$, the logarithm function log is only defined on the positive real numbers.

Proposition 14. The derivative of the exponential function is itself. That is $(e^x)' = e^x$.

The following property is the exponential function's analogue of Proposition 11 part (a).

Theorem 15. If x and y are real numbers, then $e^x \cdot e^y = e^{x+y}$.

Challenge 12

(a) Show that an input to the exponential function must be dimensionless. As the logarithm function is an inverse function of the exponential function, it follows that outputs of the logarithm function are dimensionless (flip the graph). Verify this using $\frac{\mathrm{d}\log x}{\mathrm{d}x} = x^{-1}$.

(b) Use the definition of the logarithm function to show that an input to the logarithm function must be dimensionless. Later, we will see functions cos and sin satisfying $\frac{\mathrm{d}^2\cos x}{\mathrm{d}x^2} = -\cos x$ and $\frac{\mathrm{d}^2\sin x}{\mathrm{d}x^2} = -\sin x$, respectively.[12] Show that inputs to the cos and sin functions must be dimensionless.

(c) Show that $e^x \cdot e^y = e^{x+y}$ and deduce that $e^0 = 1$. [*Hint:* start with $\log(e^x e^y)$]

Hyperbolic functions

Definition 16. A function f is **even** (an **even function**) if $f(x) = f(-x)$ for each input x. A function g an **odd** (an **odd function**) if $g(x) = -g(-x)$ for each input x.

For example, the absolute value function is even, while the identity function $x \mapsto x$ is odd.

Challenge 13

(a) Suppose we have a function f. Let $f_e : x \mapsto \frac{f(x)+f(-x)}{2}$ and $f_o : x \mapsto \frac{f(x)-f(-x)}{2}$. Show that f_e is even and f_o is odd. Conclude that a function can be written as the sum of an even and an odd function.

(b) Let function f be written as the sum $f(x) = f_1(x) + f_2(x)$, where f_1 is even and f_2 is odd. By part (a), such a *decomposition* is always possible. We show that a function's decomposition into odd and even functions is unique. Find an expression for $f(-x)$, then solve for f_1 and f_2. Use the decomposition to show that $f_1 = f_e$ and $f_2 = f_o$, as defined in part (a).

Definition 17 (Hyperbolic functions)**.** Let x be a real number. Define the functions $\sinh x$ (read *sinch*), $\cosh x$ (read *cosh*), and $\tanh x$ (read *tanch*) by the following.[13]

$$\sinh x := \frac{e^x - e^{-x}}{2} \quad \cosh x := \frac{e^x + e^{-x}}{2} \quad \tanh x := \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Challenge 14 Use the definitions to show the following.[14]

(a) $\cosh^2 x - \sinh^2 x = 1$,

(b) $\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$,

(c) $\sinh x = \frac{\tanh x}{\sqrt{1-\tanh^2 x}}$,

[12] The notation $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$ means the second derivative of f.

[13] Notice that $\cosh x$ is the even function of $\exp(x)$, while $\sinh x$ is the odd function of $\exp(x)$.

[14] Here $\square^2 x$ means $(\square x)^2$. Hence $\cosh^2 x := (\cosh x)^2$, $\sinh^2 x := (\sinh x)^2$, and $\tanh^2 x := (\tanh x)^2$.

(d) $\cosh x = \frac{1}{\sqrt{1-\tanh^2 x}}$,

(e) $\sinh' x = \cosh x$, $\cosh' x = \sinh x$, and $\tanh' x = 1/(\cosh^2 x)$.

Exponentiation

We now come full circle and obtain the exponentiation rules we encountered in Chapter 1, but in far greater generality.

Definition 18. Let a be a positive real number. For each real number x, the expression a^x is defined by $a^x := e^{x \log a}$.

Proposition 19. Let a and b be positive real numbers. Let x and y be real numbers. Then

(a) $a^0 = 1$,
(b) $a^{-x} = \frac{1}{a^x}$,
(c) $a^x \cdot a^y = a^{x+y}$,
(d) $(a \cdot b)^x = a^x \cdot b^x$,
(e) $(a^x)^y = a^{xy}$.

Proof. These all follow from the properties of the exponential function and the logarithm function. The only property that is tricky is the final one. By definition, $(a^x)^y = e^{y \log a^x}$. Once again, by definition, $a^x = e^{x \log a}$. Using the fact that $\log \circ \exp$ is an identity map, we have

$$(a^x)^y = e^{y \log a^x} = e^{y \log\left(e^{x \log a}\right)} = e^{y(x \log a)} = e^{(yx) \log a} = e^{(xy) \log a} = a^{xy}. \qquad \square$$

We now obtain Proposition 11 part (c): for real p, the equality $\log x^p = p \log x$ holds.

Proof. We use the exponentiation rule $(a^\alpha)^\beta = a^{\alpha\beta}$.[15] Let $y := \log x$ so that (i) $yp = p \log x$. By our definition of y, we know that $x = e^y$. From the exponentiation rules, $x^p = (e^y)^p = e^{yp}$. By the definition of the logarithm function, $x^p = e^{yp}$ can be written (ii) $\log(x^p) = yp$. As (i) = (ii), we are done. $\square$

Challenge 15

(a) Prove Proposition 19.

(b) Let a be a positive real number. Show that the function a^x is differentiable and find $(a^x)'$. Conclude that $\int a^x \, dx = \frac{a^x}{\log a} + c$.

Challenge 16 (The Power Rule) Let a be a real number and let $f : x \mapsto x^a$ for $x \in (0, \infty)$. Show that f is differentiable and find f'. Deduce that for $a \neq -1$, $\int x^a \, dx = \frac{x^{a+1}}{a+1} + c$.

The antiderivative of x^a for $a = -1$ is given by the logarithm function: $\int x^{-1} \, dx = \log|x| + c$. Indeed, if $x < 0$, then the chain rule gives $(\log|x|)' = \big(\log(-x)\big)' = \frac{1}{-x} \cdot (-1) = x^{-1}$. The case of $x = 0$ is undefined because $1/0$ is undefined, while the positive case follows from the definition of the logarithm function.[16]

[15] This is Proposition 19 part (e).

[16] Recall that the logarithm function is only defined on $(0, \infty)$. By chaining the function with the absolute value function, we can define the function on negative real numbers.

Challenge 17

(a) Differentiable functions are a nice class of continuous functions. This does *not* mean differentiable functions can necessarily be integrated! Define $f : x \mapsto 1/x$ on the interval $(0, 1)$. By the power rule, f is differentiable. Calculate the integral of f on the interval $(\alpha, 1)$, for each α satisfying $0 < \alpha < 1$. Argue that f cannot be integrated on the interval $(0, 1)$.

(b) We find the integral $\int_{-1}^{1} f(x)\,dx$ for the function $f : x \mapsto 1/x^2$. By the power rule, $(-x^{-1})' = 1/x^2$ and so the Fundamental Theorem of Calculus gives $\int_{-1}^{1} f(x)\,dx = (-x^{-1})\big|_{x=-1}^{1} = -2$. Even though f is a positive function, its integral is negative! What did we do wrong?

What about negative numbers? Since the logarithm function is undefined for negative numbers (and also 0), we do not have a way to define x^a for all real x. Indeed, can we make sense of the expression x^a if $x = -1$ and $a = 1/2$? This question essentially asks: is there a number squared that equals -1? Right now, the answer is a no, for we cannot square any real number to get a negative number.

4

Limits

We have been able to develop quite a bit of calculus. Nevertheless, there is an Achilles heel. Suppose we presented our findings, beginning with what a derivative is. The question we are going to get is: what is this object $o(1)$?

I hope that after working with it for quite some time, to both of us the object $o(1)$ makes intuitive sense. However, perhaps it is time we really think about what exactly $o(1)$ is.

Let us revisit the definition of a derivative. If a function f is differentiable at t, then there is a number $f'(t)$ such that the following equation holds.[1]

$$f(t+\alpha) = f(t) + f'(t)\alpha + \alpha o(1)$$

Subtract the number $f(t)$ from both sides of the equation and divide both sides by α to get

$$\frac{f(t+\alpha) - f(t)}{\alpha} = f'(t) + o(1).$$

What the equation above means is that if we drop $\alpha \to 0$, then $f'(t)$ is given by the quotient on the left side. Let us spell out the fact that we take $\alpha \to 0$ by using the notation "lim" as follows.

$$f'(t) = \lim_{\alpha \to 0} \frac{f(t+\alpha) - f(t)}{\alpha}$$

Recall that a function f is continuous at t if $f(t+\alpha) = f(t) + o(1)$. We can also state the fact that we drop α to zero explicitly by using the notation "lim" as follows.

$$\lim_{\alpha \to 0} f(t+\alpha) = f(t)$$

We say that a derivative is a *limit*, and that continuity of a function is defined by a *limit*.[2]

We see that when we were using $o(1)$ and when we were working with derivatives and continuous functions, we have been secretly working with limits. So what then is a limit?

[1] Since $-o(1)$ and $o(1)$ are the same thing, we have removed the absolute value on α.

[2] Integrals can also be thought of being the result of some limiting process.

4.1 What is a Limit?

Continuity

In the definition of the derivative, the limit is used to show us that the quotient $\frac{f(t+\alpha)-f(t)}{\alpha}$ gets closer to the number $f'(t)$ as α drops to smaller values. Similarly, in the definition of continuity, the limit is used to signify that $f(t+\alpha)$ gets *closer* to the number $f(t)$ as α drops to smaller values, that is: as α gets *closer* to 0. We will have to quantify what we mean by "close" in both instances.

To be concrete, suppose we have a function f that takes as input time, and outputs distances. For example, we can imagine that function f describes the location of an object over time.

In order to quantify closeness in distances, we will need to pick a unit of measurement. But here is a question: is 5 meters close? It is incredibly close in galactic scales, but quite far for an ant. A micrometer will satisfy an ant, but is huge in atomic scales. Because of this, it is actually impossible to satisfy everyone on what closeness means. So we will accept the fact that not everyone will be in agreement, only that some will be in agreement. Then, we will consider all possible choice of units of a distance so that at the end of the day, everyone will be satisfied.

So let us denote one possible choice of unit of distances u_0. Units of measurement must be positive, so $u_0 > 0$. Once again, some will be disappointed at our choice of unit, but they will get their turn because we will exhaust all possible units. We are simply beginning with u_0. For this turn, we will agree that the values $f(x)$ and $f(t)$ are close if their difference is within one unit, u_0. The naive expression $f(x) - f(t) < u_0$ will hold if the left side is negative, regardless of whether $f(x)$ and $f(t)$ are close or not. Therefore, we will need to use absolute values, and we will say that the values $f(x)$ and $f(t)$ are close if $|f(x) - f(t)| < u_0$.

All done? Well not quite. Where are the inputs to function f coming from? The inputs are time, and we want distances $f(x)$ and $f(t)$ to be close whenever times x and t are close. To measure closeness in time, once again, we will need to choose a unit of time. This choice of unit will depend on the proportions of u_0. For example, if u_0 is of galactic scales, tens of thousands of years could be sufficient, but in the scale of ants something much smaller will be required. But in any case, once there is some unit of time $u(u_0)$ which provides a closeness measure in time, we can proceed to check that each time x within that closeness measure of t will allow $f(x)$ to be close to $f(t)$. If such a unit $u(u_0)$ exists, then we have satisfied some people that f is continuous (nearby time maps to nearby distance). We then choose another unit of distance and repeat the process.

To summarize: for each unit of distance $u_0 > 0$, if there is some unit of time $u(u_0) > 0$ such that $|f(x)-f(t)| < u_0$ for each time x satisfying $|x-t| < u(u_0)$, then we can conclude that f is continuous at t. We write this compactly as $\lim_{x\to t} f(x) = f(t)$. Notice that redoing our previous discussion, but replacing the input x with $t+\alpha$ gives the analogous statement for $\lim_{\alpha\to 0} f(t+\alpha) = f(t)$.

A limit

Now let us turn to the definition of a derivative of a function f at t. We will know that f is differentiable at t with derivative $f'(t)$ once we verify that: for each unit of distance $u_0 > 0$, there is some unit of time $u(u_0) > 0$ such that

$$\left|\frac{f(t+\alpha)-f(t)}{\alpha} - f'(t)\right| < u_0$$

for each time $t+\alpha$ satisfying $|(t+\alpha)-t| < u(u_0)$.

There is a problem here. The time $t+\alpha$ for $\alpha = 0$ satisfies $|(t+\alpha)-t| < u(u_0)$ because $0 < u(u_0)$. But if $\alpha = 0$, then the quotient $\frac{f(t+\alpha)-f(t)}{\alpha}$ is undefined as it is a division by zero! We will have to fix this by insisting that we ignore the time t. Instead of looking at points $t+\alpha$ that are close to t by the unit u, we will look at points x that are close to t by the unit u, but not equal to t.

We summarize our finding. We will know that f is differentiable at t with derivative $f'(t)$ once we verify that: for each unit of distance $u_0 > 0$, there is some unit of time $u(u_0) > 0$ such that

$$\left|\frac{f(x)-f(t)}{x-t} - f'(t)\right| < u_0$$

for each time input $x_{\neq t}$ satisfying $|x-t| < u(u_0)$.[3] We write this compactly as $\lim_{x\to t} \frac{f(x)-f(t)}{x-t} = f'(t)$. Analogously, by replacing x with $t+\alpha$: f is differentiable at t with derivative $f'(t)$ once we verify that: for each unit of distance $u_0 > 0$, there is some unit of time $u(u_0) > 0$ such that

$$\left|\frac{f(t+\alpha)-f(t)}{\alpha} - f'(t)\right| < u_0$$

for each time input $t+\alpha$ with $0 < |\alpha| < u(u_0)$. This is written $\lim_{\alpha\to 0} \frac{f(t+\alpha)-f(t)}{\alpha} = f'(t)$.

We have essentially obtained the definition of a *limit*. Tradition dictates that we denote the unit of measurement for the output u_0 by the Greek letter ϵ and the unit of measurement for the input $u(u_0)$ by the Greek letter $\delta(\epsilon)$.

Definition 20. A function f has a **limit** l at input t, written $\lim_{x\to t} f(x) = l$, if for each $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that whenever $x_{\neq t}$ satisfies $|x-t| < \delta(\epsilon)$, we have $\left|f(x)-l\right| < \epsilon$.

From the definition, it is sufficient to exhibit a strictly positive function δ with the property that for each input $\epsilon > 0$, whenever $x_{\neq t}$ satisfies $|x-t| < \delta(\epsilon)$, we have $\left|f(x)-l\right| < \epsilon$.[4] This simply formalizes the idea that we have a rule δ associating each unit of output ϵ to a unit of input $\delta(\epsilon)$.

As an example, let us show that a constant function $f : x \mapsto c$ for some constant c satisfies $\lim_{x\to t} f(x) = c$ for each t. For $\epsilon > 0$ let $\delta(\epsilon) := \epsilon$. Then for each $x_{\neq t}$ such that $|x-t| < \delta(\epsilon) = \epsilon$, we have $|f(x)-f(t)| = |c-c| = 0 < \epsilon$, as desired. The proof that $g : x \mapsto x$ satisfies $\lim_{x\to t} g(x) = t$ is essentially the same, with the only difference being the last part: $|g(x)-g(t)| = |x-t| < \delta(\epsilon) = \epsilon$.

The definition of a limit looks very complicated. But it is complicated mainly because we have several things to keep track of, necessitating the employment of many different symbols. The definition itself is as natural and as simple as it could be: for each unit of measurement ϵ for outputs, there will be a unit of measurement $\delta(\epsilon)$ for inputs such that each input that is close by $\delta(\epsilon)$ to t (but not close by zero) will map to outputs that are close to l by ϵ. This modern definition of a limit is due to Karl Weierstrass from the mid 19th Century (building upon the work of many predecessors like Bernard Bolzano and Augustin Cauchy), almost two Centuries after the invention of calculus!

With the definition of a limit settled, the definition of continuity is simple.

Definition 21. A function f is **continuous at** t if $\lim_{x\to t} f(x) = f(t)$.

[3] The notation $x_{\neq t}$ means: "the number x, which is not equal to t".

[4] A function f is **strictly positive** if its values are greater than 0, wherever it is defined.

Let us check that the square root function $f : x \mapsto \sqrt{x}$ is continuous on the interval $(0, \infty)$. We will show that if $t \in (0, \infty)$, then $\lim_{x \to t} f(x) = \sqrt{t}$. Suppose δ is some strictly positive function. For each x satisfying $|x - t| < \delta(\epsilon)$, we use a multiplication by 1 trick and homogeneity to get

$$\left|\sqrt{x} - \sqrt{t}\right| = \left|(\sqrt{x} - \sqrt{t})\frac{\sqrt{x} + \sqrt{t}}{\sqrt{x} + \sqrt{t}}\right| = \left|\frac{x - t}{\sqrt{x} + \sqrt{t}}\right| = \frac{|x - t|}{|\sqrt{x} + \sqrt{t}|} < \frac{\delta(\epsilon)}{|\sqrt{x} + \sqrt{t}|}.$$

These expressions only make sense if $x \geq 0$ because a square root of a negative number is undefined. We thus have a clue that we require $\delta(\epsilon) \leq t$. Next, we observe that the absolute value function is an increasing function, and so if $|x - t| < t$,[5] then $|\sqrt{x} + \sqrt{t}| \geq |\sqrt{t}|$. Therefore, $\frac{1}{|\sqrt{x}+\sqrt{t}|} \leq \frac{1}{|\sqrt{t}|}$. Now define $\delta : \epsilon \mapsto \min\left(\epsilon\sqrt{t}, t\right)$.[6] Since $\delta(\epsilon) \leq t$, we know that $\sqrt{x}$ is defined. Furthermore, since $\delta(\epsilon) \leq \epsilon\sqrt{t}$,

$$\left|\sqrt{x} - \sqrt{t}\right| < \frac{\delta(\epsilon)}{|\sqrt{x} + \sqrt{t}|} \leq \frac{\delta(\epsilon)}{|\sqrt{t}|} \leq \frac{\epsilon\sqrt{t}}{|\sqrt{t}|} = \epsilon.$$

We conclude that the square root function is continuous on the interval $(0, \infty)$.

4.2 Arithmetic of Limits

Uniqueness

With the definition of a limit at hand, we proceed as we did for derivatives and see what kind of arithmetic rules they permit.[7] But first, we need to check that a limit of a function at a point is unique, otherwise we will be in trouble!

Proposition 22 (Limits are unique)**.** If $\lim_{x \to t} f(x) = l_1$ and $\lim_{x \to t} f(x) = l_2$, then $l_1 = l_2$.

Proof. Let $\epsilon > 0$. Since $\lim_{x \to t} f(x) = l_1$, by the definition of a limit, there is a $\delta_1(\epsilon) > 0$ such that $|f(x) - l_1| < \epsilon$ for each input $x_{\neq t}$ with $|x - t| < \delta_1(\epsilon)$. Similarly, since $\lim_{x \to t} f(x) = l_2$, there is a $\delta_2(\epsilon) > 0$ such that $|f(x) - l_2| < \epsilon$ for each input $x_{\neq t}$ with $|x - t| < \delta_2(\epsilon)$.

For one unit of measurement for the output there are two units of measurement for the input. Two units of measurement for the input is one too many. We will err on the side of caution and pick the smaller of the two by setting $\delta(\epsilon) := \min(\, \delta_1(\epsilon), \delta_2(\epsilon)\,)$. The reasoning is this: we want to measure closeness of inputs, and by being more stringent and picking a smaller unit of measurement, we will offend no one. On the other hand, if we picked the larger unit of measurement, then some will no longer agree that the inputs are close.

Our choice of unit $\delta(\epsilon)$ means that for each $x_{\neq t}$ with $|x - t| < \delta(\epsilon)$, we satisfy both $|x - t| < \delta_1(\epsilon)$ and $|x - t| < \delta_2(\epsilon)$. Therefore, $|f(x) - l_1| < \epsilon$ and $|f(x) - l_2| < \epsilon$ are both true whenever $x_{\neq t}$ is within $\delta(\epsilon)$ of t.

All that is left is to check that $|l_1 - l_2| = 0$. By the triangle inequality,

$$|l_1 - l_2| = |l_1 - f(x) + f(x) - l_2| \leq |l_1 - f(x)| + |f(x) - l_2|.$$

[5] This is simply there to make sure x is positive and thus $\sqrt{x}$ is defined.

[6] The function "min" takes two inputs and outputs whichever is smaller. For example, $\min(-10, 2) = -10$.

[7] The material in the rest of the chapter, although important, is not used in any essential way except for Section 6.4 and can be skipped or read (very slowly) in parallel with Chapter 5 and Chapter 6.

By homogeneity, $|l_1 - f(x)| = |-1||f(x) - l_1| = |f(x) - l_1|$. Therefore,

$$|l_1 - l_2| \leq |f(x) - l_1| + |f(x) - l_2| < \epsilon + \epsilon = 2\epsilon.$$

But this must be true for any unit of measurement ϵ, no matter how small. Thus the real number $|l_1 - l_2|$ is a lower bound on the set of positive real numbers, and must be zero or smaller. By the definition of the absolute value function, $|l_1 - l_2| \geq 0$, and so $|l_1 - l_2| = 0$. □

Sum rule

As we have done previously, the first arithmetic operation we will discuss is the summation of limits. The *sum rule* for limits states that the sum of limits behaves just as expected.

Proposition 23 (Sum Rule)**.** If $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} g(x) = l_2$, then

$$\lim_{x\to t}(f + g)(x) = l_1 + l_2.$$

Proof. Let $\epsilon > 0$. Our goal is to find a unit of measurement $\delta(\epsilon) > 0$ that makes each $x_{\neq t}$ close to t map within unit ϵ of $l_1 + l_2$.

Since $\lim_{x\to t} f(x) = l_1$, by the definition of a limit, there is some $\delta_1(\epsilon) > 0$ such that $|f(x) - l_1| < \epsilon$ for each $x_{\neq t}$ with $|x - t| < \delta_1(\epsilon)$. Similarly, since $\lim_{x\to t} f(x) = l_2$, by the definition of a limit, there is some $\delta_2(\epsilon) > 0$ such that $|f(x) - l_2| < \epsilon$ for each $x_{\neq t}$ with $|x - t| < \delta_2(\epsilon)$.

Once again, there are two units of measurement for the input. We set $\delta(\epsilon) := \min(\delta_1(\epsilon), \delta_2(\epsilon))$ so that each input $x_{\neq t}$ with $|x - t| < \delta(\epsilon)$ will satisfy both $|f(x) - l_1| < \epsilon$ and $|f(x) - l_2| < \epsilon$.

By the triangle inequality,

$$\big|(f + g)(x) - (l_1 + l_2)\big| = \big|f(x) - l_1 + g(x) - l_2\big| = \big|f(x) - l_1\big| + \big|g(x) - l_2\big| < \epsilon + \epsilon = 2\epsilon.$$

The definition requires that in order to conclude $\lim_{x\to t}(f + g)(x) = l_1 + l_2$, we need $\big|(f + g)(x) - (l_1 + l_2)\big| < \epsilon$. But this can be achieved by changing the first statement of the proof to "Let $\epsilon/2 > 0$." and then substituting all instances of ϵ by $\epsilon/2$. So we are done! □

Product rule

Proposition 24 (Product Rule)**.** If $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} g(x) = l_2$, then

$$\lim_{x\to t}(fg)(x) = l_1 \cdot l_2.$$

The product rule for limits states that the products of limits behaves just as expected. However, the proof will be quite hairy because it will not be sufficient to take the unit of input to be $\delta(\epsilon) := \min(\delta_1(\epsilon), \delta_2(\epsilon))$. To see this, let us see what our end goal of the proof is. Ultimately, we want to show that each output $(fg)(x)$ is close to $l_1 l_2$. That is, there is a suitable unit of measurement for inputs such that inputs $x_{\neq t}$ close to t will guarantee $\big|(fg)(x) - l_1 l_2\big| < c \cdot \epsilon$ for some positive constant c.[8] By applying the triangle inequality and homogeneity on a sneaky addition

[8] As in the proof of the sum rule, we can always scale ϵ by a positive constant c.

and subtraction of the term $f(x)l_2$, the following holds.

$$|(fg)(x) - l_1l_2| = |f(x)g(x) - f(x)l_2 + f(x)l_2 - l_1l_2| \leq \underbrace{|f(x)|}_{<???} \underbrace{|g(x) - l_2|}_{<\epsilon} + \underbrace{|f(x) - l_1|}_{<\epsilon} |l_2|$$

Since $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} f(x) = l_2$, the terms $|f(x) - l_1|$ and $|g(x) - l_2|$ are each less than ϵ. The term $|l_2|$ is a constant, so that's ok, but the term $|f(x)|$ is *not* a constant, and that is a problem. Our solution will be to choose a unit of measurement for the input such that inputs $x_{\neq t}$ close to t will satisfy $|f(x)| < |l_1| + 1$. Then, we will have

$$\big|(fg)(x) - l_1l_2\big| \leq |f(x)||g(x) - l_2| + |f(x) - l_1||l_2| < (|l_1| + 1)\epsilon + \epsilon|l_2| = (|l_1| + |l_2| + 1)\epsilon,$$

where $(|l_1| + |l_2| + 1)$ is simply a positive constant.

So how can ensure that $|f(x)| < |l_1| + 1$? We need a slight variation on the triangle inequality: $|a| - |b| \leq |a - b|$. Since $|f(x)| < |l_1| + 1$ is the same as $|f(x)| - |l_1| < 1$, the modified triangle inequality shows us that it is sufficient to choose a unit of inputs $\delta(\epsilon)$ such that x maps to values satisfying $|f(x) - l_1| < 1$. What does this mean? Well, if the unit of outputs ϵ satisfies $\epsilon \leq 1$, then we know that $|f(x) - l_1| < \epsilon \leq 1$ is true for an appropriate choice of unit $\delta(\epsilon)$, and all is well. The problem is when the unit of outputs ϵ is greater than one, because now we can have situations where $|f(x) - l_1| < \epsilon$ but $|f(x) - l_1| \geq 1$, and the value $|f(x)|$ may stray too far from l_1.

But there is an easy fix! Whenever we have to make our choice of unit $\delta(\epsilon)$ and we are faced with $\epsilon > 1$, we pretend that $\epsilon = 1$. For example, if ϵ is the distance from the sun to the earth, when it comes time to pick our unit of inputs, we will be pessimistic and pick $\delta(\epsilon)$ as if ϵ is the distance from the earth to the moon. This way, the values of $|f(x)|$ will be even closer to l_1 than usual, and we can guarantee that $|f(x)| < |l_1| + 1$, say.

We now proceed to the proof of the product rule, which states that if $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} g(x) = l_2$, then

$$\lim_{x\to t}(fg)(x) = l_1 \cdot l_2.$$

Proof. We wish to show that there is some strictly positive function δ such that whenever $x_{\neq t}$ is within the distance of $\delta(\epsilon)$ to t, then $\big|(fg)(x) - l_1 \cdot l_2\big| < c \cdot \epsilon$ for some positive constant c.[9] From $\lim_{x\to t} f(x) = l_1$, we know there is a strictly positive function δ_1 such that whenever $x_{\neq t}$ is within $\delta_1(\epsilon)$ of t, then $\big|f(x) - l_1\big| < \epsilon$. Similarly, from $\lim_{x\to t} g(x) = l_2$, we know there is a strictly positive function δ_2 such that whenever $x_{\neq t}$ is within $\delta_2(\epsilon)$ of t, then $\big|g(x) - l_2\big| < \epsilon$. The triangle inequality and homogeneity gives the following inequality, as we discussed before.

$$|(fg)(x) - l_1l_2| \leq |f(x)||g(x) - l_2| + |f(x) - l_1||l_2|$$

Let δ be the function defined by $\delta : \epsilon \mapsto \min\ (\ \delta_1(\min(1, \epsilon)), \delta_2(\epsilon)\ )$. The definition of δ is difficult to parse, so below is the same in pseudocode. It is quite simple, we want to take the minimum of $\delta_1(\epsilon)$ and $\delta_2(\epsilon)$, but before we do so, in order to make $|f(x)|$ closer to $|l_1|$ than usual,

[9] The nonzero constant c has to stay the same, regardless of the value of ϵ. Remember, the idea is that it is possible to finish the proof, go back and do the substitution $\epsilon \mapsto \epsilon/c$. If c changes with ϵ, a substitution is no longer possible.

we pretend $\epsilon = 1$ for $\delta_1(\epsilon)$ whenever $\epsilon > 1$.

```
def δ(ε):
    if ε ≤ 1:
        d_1 ← δ_1(ε)
    else:
        d_1 ← δ_1(1)
    return min(d_1, δ_2(ε))
```

There are two possibilities: either $\epsilon \leq 1$ or $\epsilon > 1$. In the former case, for each input $x_{\neq t}$ within $\delta(\epsilon)$ of t, we have

$$\left|(fg)(x) - l_1 l_2\right| \leq \underbrace{|f(x)|}_{(*)} \underbrace{|g(x) - l_2|}_{(**)} + \underbrace{|f(x) - l_1|}_{(***)} |l_2| < \underbrace{(|l_1| + \epsilon)}_{(*)} \underbrace{\epsilon}_{(**)} + \underbrace{\epsilon}_{(***)} |l_2|$$

The inequality $|f(x)| < |l_1| + \epsilon$ holds because $|f(x)| - |l_1| \leq |f(x) - l_1| < \epsilon$. Since $\epsilon \leq 1$,

$$\left|(fg)(x) - l_1 l_2\right| < (|l_1| + \epsilon)\epsilon + \epsilon \cdot |l_2| \leq (|l_1| + 1)\epsilon + \epsilon \cdot |l_2| = (|l_1| + |l_2| + 1)\epsilon.$$

How about the case when $\epsilon > 1$? We treat ϵ as if it is 1, which was covered in the previous case, so we are done! □

The proof of the product rule for limits is difficult. I still remember first seeing a proof of this result and being absolutely terrified! The digestion of this proof is not necessary to understand and practice calculus, which is why we are diving into these matters *after* seeing calculus in action.

Now that we are done with the proof, let us note that only two new ideas were needed. First was the sneaky manipulation

$$|(fg)(x) - l_1 l_2| = |f(x)g(x) - f(x)l_2 + f(x)l_2 - l_1 l_2| \leq |f(x)||g(x) - l_2| + |f(x) - l_1||l_2|. \tag{4.1}$$

Second was the realization that defining $\delta : \epsilon \mapsto \min\left(\delta_1(\epsilon), \delta_2(\epsilon)\right)$ is not enough. By Equation 4.1 above, we need to make sure that $|f(x)|$ is small. We accomplished this by ensuring $|f(x)|$ is within distance 1 of $|l_1|$, regardless of the value of ϵ.

Challenge 18 Suppose we have a function f such that $\lim_{x \to t} f(x) = l$ holds. By definition, there is a function δ that takes as input a positive real number ϵ and outputs a positive real number such that each $x_{\neq t}$ that is within $\delta(\epsilon)$ distance of t satisfies $|f(x) - l| < \epsilon$.

(a) Let $\delta' : \epsilon \mapsto \delta(\epsilon)/2$. For each $x_{\neq t}$ within $\delta'(\epsilon)$ of t, can we guarantee that $|f(x) - l| < \epsilon$ holds? Repeat for $\delta'' : \epsilon \mapsto \delta(\epsilon)/c$, where $c > 1$ is a constant.

(b) Instead of dividing, suppose we define $\delta' : \epsilon \mapsto 2 \cdot \delta(\epsilon)$. If $x_{\neq t}$ is within $\delta'(\epsilon)$ of t, can we guarantee that $|f(x) - l| < \epsilon$? Repeat for $\delta'' : \epsilon \mapsto c \cdot \delta(\epsilon)$, where $c > 1$ is a constant.

(c) Let $c > 1$ be a constant. For each of the following definitions of δ_i, identify the ones that guarantee that each $x_{\neq t}$ within $\delta_i(\epsilon)$ of t satisfies $|f(x) - l| < \epsilon$.

$$\delta_1 : \epsilon \mapsto \delta(\epsilon/2), \quad \delta_2 : \epsilon \mapsto \delta(2\epsilon), \quad \delta_3 : \epsilon \mapsto \delta(\epsilon/c), \quad \delta_4 : \epsilon \mapsto \delta(c\epsilon)$$

(d) Now suppose we had two functions δ^* and δ^{**} such that each $x_{\neq t}$ within $\delta^*(\epsilon)$ satisfies $|f(x) - l| < \epsilon$, and each $x_{\neq t}$ within $\delta^{**}(\epsilon)$ satisfies $|f(x) - l| < \epsilon$. Does the function $\tilde{\delta}$ defined below ensure that each $x_{\neq t}$ within $\delta^*(\epsilon)$ satisfy $|f(x) - l| < \epsilon$?

$$\tilde{\delta} : \epsilon \mapsto \min\left(\delta^*(\epsilon), \delta^{**}(\epsilon) \right)$$

(e) Repeat part (d), but this time assume that δ^{**} is some mystery function that takes in a positive real number and outputs some random positive real number. The function δ^* is the same as before.

The definition of a limit is often written concisely using the symbol $\forall$, which reads: "for each", the symbol $\exists$, which reads: "there is" or "there exists", and the symbol $\implies$, which reads "implies". Using these symbols the expression $\lim_{h\to y} f(h) = l$ means

$$(\forall \epsilon > 0)(\exists \delta(\epsilon) > 0)(\forall x \in \mathbb{R}) \left(0 < |x - t| < \delta(\epsilon) \implies |f(x) - l| < \epsilon\right).$$

Challenge 19

(a) Suppose someone told you that: a function f has a limit l at a point t, if for each $\delta > 0$, there is some $\epsilon > 0$ such that for each $x_{\neq t}$ satisfying $|x - t| < \delta$, we have $\left|f(x) - l\right| < \epsilon$. Write this "definition" down using the symbols $\forall$, $\exists$, and $\implies$. Give some intuition as to why this "definition" is incorrect.[10] Our habit of writing $\delta(\epsilon)$ should tip you off immediately!

(b) Suppose we were trying to prove that $\lim_{x\to t} f(x) = l$, but when working with a specific value of ϵ, we failed to find a $\delta(\epsilon) > 0$ that guarantees $0 < |x - t| < \delta(\epsilon) \Rightarrow |f(x) - l| < \epsilon$. We are forced to conclude that $\lim_{x\to t} f(x) \neq l$. Write the definition of $\lim_{x\to t} f(x) \neq l$ using the symbols $\forall$, $\exists$, and $\Rightarrow$.[11]

(c) Show that the following proposed "definition" of $\lim_{x\to t} f(x) = l$ is incorrect by using it prove that if $f : x \mapsto c$ for some constant c, then $\lim_{x\to 0} f(x) \neq c$.[12]

$$(\forall \epsilon > 0)(\exists \delta(\epsilon) > 0)(\forall x \in \mathbb{R}) \left(|f(x) - l| < \epsilon \implies 0 < |x - t| < \delta(\epsilon)\right)$$

Quotient rule

We have one more arithmetic rule left: division.[13]

Proposition 25 (Quotient Rule)**.** Let $\lim_{x\to t} f(x) = l_1$ and let g be a nonzero function with $\lim_{x\to t} g(x) = l_2 \neq 0$. Then

$$\lim_{x\to t}(f/g)(x) = l_1/l_2.$$

This is also tricky, but no more difficult than the product rule. We will first show that $\lim_{x\to t}(1/g)(x) = 1/l_2$, and then apply the product rule. As usual, we wish to show that

[10] Here is one answer. A function f is continuous at t if $\lim_{x\to t} f(x) = f(t)$. So the "definition" tells us that if we zoom into the graph of the function by decreasing the unit of measurement of the input, we will see whether the function is continuous. Consider a constant $\hbar$ whose numerical value is about 10^{-34}. Try zooming in on the x-axis of the step function $f : x \mapsto \hbar$ if $x < 0$ and $f : x \mapsto 0$ if $x \geq 0$, and it won't help at all! The graph will continue to look like a constant function with no change. What we need to do is zoom into the graph by decreasing the unit of measurement in the y-axis so that we can make out the step from zero to $\hbar$ around $x = 0$.

[11] Answer: $(\exists \epsilon > 0)(\forall \delta(\epsilon) > 0)(\exists x \in \mathbb{R}) \left(0 < |x - t| < \delta(\epsilon) \Rightarrow |f(x) - l| \geq \epsilon\right)$.

[12] *Hint:* we will first need to repeat part (b) for this new "definition".

[13] Subtraction is verified in the same way as the subtraction rule for derivatives.

$|(1/g)(x) - 1/l_2| < c \cdot \epsilon$ for some constant c. Using homogeneity and the identity $\frac{1}{a} - \frac{1}{b} = \frac{b-a}{ab}$ gives

$$\left|\frac{1}{g(x)} - \frac{1}{l_2}\right| = \frac{|l_2 - g(x)|}{|g(x)l_2|} = \underbrace{\frac{|g(x) - l_2|}{|l_2|}}_{<\epsilon/|l_2|} \cdot \underbrace{\frac{1}{|g(x)|}}_{<???}.$$

Like before, we need to pick some unit of measurement for the input such that $1/|g(x)|$ is bounded.

Let us try to bound $1/|g(x)|$ by some positive constant. Observe that if $|g(x) - l_2| < \min(\epsilon, X)$, then the inequality $|a| - |b| \leq |a - b|$ gives

$$|l_2| - |g(x)| \leq |l_2 - g(x)| < X.$$

The idea is to repeat what we did in the product rule: if ϵ is too big (for the product rule, whenever $\epsilon > 1$), then we pretend ϵ is smaller. In this case, if ϵ is too big, then we pretend as if $\epsilon = X$. Assuming that $|l_2| - X$ is positive, rearranging the inequality above gives

$$|l_2| - X < |g(x)| \implies \frac{1}{|g(x)|} < \frac{1}{|l_2| - X}.$$

For example, if $X := |l_2|/2$, then $|l_2| - X = X$ and we have $1/|g(x)| < 2/|l_2|$.

Proof. First, we show that if g is nonzero with $\lim_{x\to t} g(x) = l_2 \neq 0$, then $\lim_{x\to t}(1/g)(x) = 1/l_2$. By assumption, there is some positive function δ' such that each $x_{\neq t}$ within $\delta'(\epsilon)$ satisfies $|g(x) - l_2| < \epsilon$. Recall from our preliminary discussion that homogeneity gives

$$\left|\frac{1}{g(x)} - \frac{1}{l_2}\right| = \frac{|g(x) - l_2|}{|l_2|}\frac{1}{|g(x)|}.$$

Observe that if $|g(x) - l_2| < \min(\epsilon, |l_2|/2)$, then the inequality $|a| - |b| \leq |a - b|$ gives

$$|l_2| - |g(x)| \leq |l_2 - g(x)| < \frac{|l_2|}{2} \implies |l_2| - \frac{|l_2|}{2} < |g(x)| \implies \frac{1}{|g(x)|} < \frac{2}{|l_2|}.$$

Let $\delta : \epsilon \mapsto \delta'(\min(\epsilon, |l_2|/2))$. Then each $x_{\neq t}$ within distance $\delta(\epsilon)$ of t satisfies

$$\left|\frac{1}{g(x)} - \frac{1}{l_2}\right| = \underbrace{\frac{|g(x) - l_2|}{|l_2|}}_{(*)}\underbrace{\frac{1}{|g(x)|}}_{(**)} < \underbrace{\frac{\epsilon}{|l_2|}}_{(*)}\underbrace{\frac{2}{|l_2|}}_{(**)} = \frac{2}{|l_2|^2}\epsilon.$$

We have found a strictly positive function δ such that no matter what $\epsilon > 0$ we may need to work with, our function δ ensures that each $x_{\neq t}$ within $\delta(\epsilon)$ of t satisfies

$$\left|(1/g)(x) - 1/l_2\right| < c \cdot \epsilon$$

for the constant $c := 2/|l_2|^2$. Hence $\lim_{x\to t}(1/g)(x) = 1/l_2$. The final result is obtained by applying the product rule to $f \cdot (1/g)$ for a function f satisfying $\lim_{x\to t} f(x) = l_1$. □

The quotient rule is incredibly useful. For one thing, a derivative is a limit of a quotient! As an example, let us calculate the derivative of the square root function $f : x \mapsto \sqrt{x}$ for $x > 0$ from scratch. There are now many definitions to choose from. How about we use $f'(x) := \lim_{\alpha \to 0} \frac{f(x+\alpha)-f(x)}{\alpha}$. First, we need to do some algebraic manipulations. We use the trick of multiplying by 1 and simplifying to get the following.

$$\begin{aligned}\frac{f(x+\alpha)-f(x)}{\alpha} &= \frac{\sqrt{x+\alpha}-\sqrt{x}}{\alpha} = \frac{\sqrt{x+\alpha}-\sqrt{x}}{\alpha}\left(\frac{\sqrt{x+\alpha}+\sqrt{x}}{\sqrt{x+\alpha}+\sqrt{x}}\right) \\ &= \frac{\alpha}{\alpha(\sqrt{x+\alpha}+\sqrt{x})} = \frac{1}{\sqrt{x+\alpha}+\sqrt{x}}\end{aligned}$$

Earlier on, we showed that the square root function is continuous. Thus $\lim_{\alpha \to 0} \sqrt{x+\alpha} = \sqrt{x}$ and we have

$$f'(x) = \lim_{\alpha \to 0} \frac{1}{\sqrt{x+\alpha}+\sqrt{x}} = \frac{\lim_{\alpha \to 0} 1}{\lim_{\alpha \to 0}(\sqrt{x+\alpha}+\sqrt{x})} = \frac{1}{\sqrt{x}+\sqrt{x}} = \frac{1}{2\sqrt{x}} = (1/2)(x^{-1/2}).$$

As we expected from the power rule: $(x^{1/2})' = (1/2)(x^{-1/2})$.

The differentiation rule that corresponds to the quotient rule for limits is the quotient rule for derivatives. Let us rederive the quotient rule for derivatives, albeit in slightly greater generality than we have done before. We will first obtain the reciprocal rule $(1/f)' = -f'/f^2$.

Proposition 26 (Reciprocal Rule)**.** Suppose f is differentiable at t and $f(t)$ is nonzero. Then

$$(1/f)'(t) = -\frac{f'(t)}{\big[f(t)\big]^2}.$$

Proof. We start with a definition of the derivative and go from there.

$$(1/f)'(t) = \lim_{\alpha \to 0} \frac{\frac{1}{f(t+\alpha)} - \frac{1}{f(t)}}{\alpha} = \lim_{\alpha \to 0} \frac{f(t)-f(t+\alpha)}{\alpha f(t+\alpha)f(t)} = \lim_{\alpha \to 0}\left(\frac{f(t)-f(t+\alpha)}{\alpha} \cdot \frac{1}{f(t+\alpha)f(t)}\right)$$

We are in a position to apply the product rule. Since f is differentiable at t, it is continuous at t. Hence $\lim_{\alpha \to 0} f(t+\alpha) = f(t)$ and we have

$$\left(\frac{1}{f}\right)'(t) = \lim_{\alpha \to 0} \frac{-[f(t+\alpha)-f(t)]}{\alpha} \lim_{\alpha \to 0} \frac{1}{f(t+\alpha)f(t)} = -f'(t)\frac{\lim_{\alpha \to 0} 1}{f(t)\lim_{\alpha \to 0} f(t+\alpha)} = -\frac{f'(t)}{f(t)^2}.$$

□

The reciprocal can be generalized into the quotient rule. This time we no longer need the assumption that (f/g) is differentiable. That (f/g) is differentiable is a consequence of the quotient rule.

Proposition 27 (Quotient Rule)**.** Suppose f and g are differentiable at t, with $g(t) \neq 0$. Then (f/g) is differentiable at t with $(f/g)'(t) = \frac{f'(t)g(t)-f(t)g'(t)}{[g(t)]^2}$.

Proof. By the reciprocal rule, the function $(1/g)$ is differentiable at t and so we may apply the product rule to obtain $(f \cdot 1/g)'(t) = (f'/g)(t) + [f(1/g)'](t)$. By the reciprocal rule, $(1/g)'(t) = -g'(t)/[g(t)^2]$ and so

$$\left(\frac{f}{g}\right)'(t) = \frac{f'(t)}{g(t)} - \frac{f(t)g'(t)}{g(t)^2} = \frac{f'(t)g(t) - f(t)g'(t)}{g(t)^2}.$$

□

This rule is rather difficult to memorize *correctly* (but it will be memorized after you apply this rule many times over). Initially, it might be easier to obtain the correct formula from scratch by obtaining the product rule from dimensional analysis and then applying it to the product $(f/g) \cdot g$ to find the quotient rule. The reciprocal rule then also comes for free and there is no need to worry about whether you got the minus sign in the correct place.

The following is an application of the quotient rule for limits.

Theorem 28 (L'Hospital's Rule)**.** Suppose f and g are differentiable at t with $g'(t) \neq 0$. Furthermore, assume that $f(t) = g(t) = 0$. Then

$$\lim_{x \to t} \frac{f(x)}{g(x)} = \frac{f'(t)}{g'(t)}.$$

In addition if f' and g' are continuous at t, then

$$\lim_{x \to t} \frac{f(x)}{g(x)} = \lim_{x \to t} \frac{f'(x)}{g'(x)}.$$

Proof. Since f and g are differentiable at t,

$$\frac{f'(t)}{g'(t)} = \frac{\lim_{x \to t} \frac{f(x)-f(t)}{x-t}}{\lim_{x \to t} \frac{g(x)-g(t)}{x-t}}.$$

By the quotient rule, the limit can be pulled out. Since $f(t) = g(t) = 0$, and $x - t$ is nonzero (by the definition of a limit), we have

$$\frac{f'(t)}{g'(t)} = \lim_{x \to t} \frac{\frac{f(x)-f(t)}{x-t}}{\frac{g(x)-g(t)}{x-t}} = \lim_{x \to t} \frac{\frac{f(x)}{x-t}}{\frac{g(x)}{x-t}} = \lim_{x \to t} \frac{f(x)}{g(x)}.$$

If f' and g' are continuous at t, then certainly $\frac{f'}{g'}(t) = \lim_{x \to t} \frac{f'}{g'}(x)$. □

Since continuity is defined by limits, the sum, difference, products, and quotients of continuous functions are also continuous.

4.3 Further Notions

Little oh

The object $o(1)$ denotes the set of functions f with the property $\lim_{x \to 0} f(x) = 0$ and $f(0) = 0$. Thus the expression $f = o(1)$ means that f is an element of $o(1)$. This means that the expression

$o(1) = f$ is incorrect because the set $o(1)$ cannot be an element of a function. On the other hand, the expression $c \cdot o(1) = o(1)$ means that the objects on both sides of the equation are the same objects.

There is a more general concept: if g is a nonzero function, then $f = o(g)$ means that

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = 0 \quad \text{with} \quad (f/g)(0) = 0.$$

If $g : x \mapsto 1$, then we recover $f = o(1)$. Indeed, from the definition, for each positive $c > 0$, $o(c) = o(1)$. Once again, the object $o(g)$ is the set of functions with the above property, and $f = o(g)$ means that f is an element of the set $o(g)$. Using this notation, the definition of the derivative may be written using $o(\alpha)$ in place of $|\alpha|o(1)$. Thus f is differentiable at t if there is a number $f'(t)$ such that the following holds.

$$f(t + \alpha) = f(t) + f'(t)\alpha + o(\alpha)$$

Challenge 20

(a) Use the well-ordering principle to show that $x^n = o(e^x)$ for each positive integer n.

(b) For a nonzero constant C and integer k, check that $\lim_{|x-a| \to 0} \left[\int_a^x C(x-a)^k \, dt \right] / (x-a)^k = 0$. Hence $\int_a^x C(x-a)^k \, dt = o(|x-a|^k)$.

(c) By definition $\alpha \cdot o(1)$ and $o(\alpha)$ are the same objects (α is *not* a constant!). Check that $o(\alpha) + o(\alpha) = o(\alpha)$, that for each constant c, $c \cdot o(\alpha) = o(\alpha)$, and that $o(\alpha)o(\alpha) = o(\alpha)$.

Taylor Series

Recall Taylor's Theorem (Theorem 9) that if function f is $(k+1)$-times differentiable then

$$f(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \int_{x_0}^{x} \frac{f^{(k+1)}(t)}{k!}(x - x_0)^k \, dt.$$

If $f^{(k+1)}$ is unbounded, we are in trouble, so we may assume that it is bounded by some constant C on the interval with endpoints x and x_0. By Challenge 20, Taylor's Theorem takes the form

$$f(x) = \sum_{n=0}^{k} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + o(|x - x_0|^k).$$

Taking $k = 1$ allows us to recover the definition of the derivative at x.

A **critical point** is a point x such that $f'(x) = 0$ or f is not differentiable at x. If f is three times differentiable we can use Taylor's theorem for tiny perturbations $\alpha := x - x_0$ to check whether $f(x)$ is smaller/larger than points nearby. If $f'(x) = 0$, then for tiny α, we can ignore the remainder term $o(|\alpha|^k)$ to get

$$f(x + \alpha) - f(x) = f'(x)\alpha + f''(x)\alpha^2/2 = f''(x)\alpha^2/2.$$

If $f''(x) > 0$, then $f(x + \alpha) \geq f(x)$ for each tiny α and so $f(x)$ is a **local minimum**. On the other hand, if $f''(x) < 0$, then $f(x + \alpha) \leq f(x)$ for each tiny α and so $f(x)$ is a **local maximum**. This method for identifying local minima/maxima is called the **second derivative test**.

If function f is infinitely differentiable, that is differentiable arbitrarily many times (like e^x or any polynomial), then if x and x_0 are "close", we can ignore the little oh term and write

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \frac{f'''(x_0)}{3!}(x - x_0)^3 + \cdots$$

where the dot dot dot means we can add as many terms to the Taylor polynomial as we wish. The above is called a **Taylor series of f at x_0** or a **Taylor expansion of function f about x_0**.

Challenge 21 If f is four times differentiable with $f'(x) = f''(x) = 0$ and $f^{(3)}(x) \neq 0$, argue that $f(x)$ cannot be a local maximum nor a local minimum. What can we say about the critical point of the function $f : x \mapsto x^3$? Let g be five times differentiable with $g'(x) = g''(x) = g^{(3)}(x) = 0$. Obtain the *fourth derivative test* and apply it to the critical point of function $g : x \mapsto x^4$.

Challenge 22 Check that the following Taylor series at 0 hold.

$$\begin{aligned}
(1 + x)^n &= 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots \\
\sqrt{1 + x} &= 1 + \frac{x}{2} - \frac{x^2}{8} + \cdots \\
1/\sqrt{1 + x} &= 1 - \frac{x}{2} + \frac{3x^2}{8} + \cdots \\
e^x &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \\
\log(1 + x) &= x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots
\end{aligned}$$

One sided limits

Outside this book, if you see $f = o(g)$ in the wild, it will mean the following limit is satisfied.

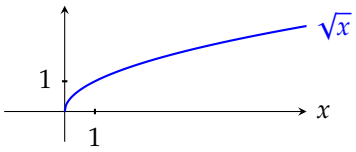
$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

The meaning is the same: f is *negligible* compared to g, or in the case of $f = o(1)$, that f is *negligible*.

The symbol $\lim_{x \to \infty} f(x) = l$ means: as x is allowed to grow, $f(x)$ approaches the number l. To capture the idea that the input is allowed to grow, we pick some height level $n > 0$ and then check that for each input x exceeding that height, $|f(x) - l| < \epsilon$. But one height that is large for one entity will be microscopic to another. So we need to consider all possible height levels of the input. The formal definition of the expression $\lim_{x \to \infty} f(x) = l$ is then: for each $\epsilon > 0$, there is some positive integer n such that for each $x > n$, we have $|f(x) - l| < \epsilon$.

Challenge 23 Show that if $\lim_{x \to \infty} f(x) = l_1$ and $\lim_{x \to \infty} g(x) = l_2$ then $\lim_{x \to \infty}(f + g)(x) = l_1 + l_2$. Also show that if c is a real number then $\lim_{x \to \infty} c \cdot f(x) = c \cdot l_1$.

In the case of $\lim_{x \to \infty}$, there is a distinguished direction in which we take the limit: from smaller values of x to larger values (to our right). On the other hand, the square root function $f : x \mapsto \sqrt{x}$ graphed below is undefined for negative real numbers, so there is no way to take a limit from smaller x to larger x at the origin, because x will be negative.



Nevertheless, we wish to speak of a limit of the square root function at the origin, as it clearly should take the limit value of 0. We formalize this with a **one sided** limit.

Function f has a **limit** l **from above** at input t, if for each $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ such that each $x \in (t, t + \delta(\epsilon))$ satisfies $f(x) \in (l - \epsilon, l + \epsilon)$. We write this using the notation $\lim_{x \to t+} f(x) = l$.

Notice that the statement $f(x) \in (l - \epsilon, l + \epsilon)$ is equivalent to the statement $|f(x) - l| < \epsilon$. Both statements mean the same thing: $f(x)$ is within distance ϵ of point l. Hence if $|f(x) - l| < \epsilon$ then $l - \epsilon < f(x) < l + \epsilon$.

Challenge 24

(a) Verify that $\lim_{x \to 0+} \sqrt{x} = 0$.

(b) Formulate a definition for a function f to have a **limit** l **from below** at point t. The symbol used in such a case is $\lim_{x \to t-} f(x) = l$.

(c) Suppose we have a function f such that $\lim_{x \to t} f(x) = l$. Show that not only do both $\lim_{x \to t+} f(x)$ and $\lim_{x \to t-} f(x)$ exist, but they are equal.

(d) Suppose we have a function f with the property that $\lim_{x \to t+} f(x) = \lim_{x \to t-} f(x)$. Show that $\lim_{x \to t} f(x)$ exists. Combining parts (c) and (d), we say that $\lim_{x \to t} f(x)$ exists **if and only if** $\lim_{x \to t+} f(x) = \lim_{x \to t-} f(x)$. The term "if and only if" indicates equivalence.

Although a derivative is a limit, we cannot conclude that the derivative at an input exists because both its limits from above and below at the input exist and are equal. For example, if $g : x \mapsto x(x + 1)/x$ (defined for nonzero real numbers), then $g' : x \mapsto 1$ for nonzero x. Therefore $\lim_{x \to 0+} g'(x) = \lim_{x \to 0-} g'(x) = 1$, however $g'(0)$ does not exist (cf. Challenge 9). The problem of course is that g is not continuous at the origin, as it is undefined there. If in addition, we define $g(0) := 1$, then there is no such problem.

Limits and inequalities

Whenever we use properties like if $|f| \leq o(1)$, then $f = o(1)$, we are using limits with inequalities. We now check that limits work as expected with inequalities.

Proposition 29. If $f \leq g$ with $\lim_{x \to t} f(x) = l_1$ and $\lim_{x \to t} g(x) = l_2$, then $l_1 \leq l_2$.

Proof. We need to show that $l := l_2 - l_1 \geq 0$. In order to derive a contradiction, suppose $l < 0$.

By assumption, the function $h : x \mapsto g(x) - f(x)$ is a positive function. By the sum rule, $\lim_{x \to t} h(x) = l_2 - l_1 = l < 0$, and so there is some positive function δ such that for each $x_{\neq t}$ within $\delta(\epsilon)$ of t, we have $|h(x) - l| < \epsilon$. In particular, $|l|/2$ is the positive real number $-l/2$ and so each $x_{\neq t}$ within $\delta(-l/2)$ of t satisfies $|h(x) - l| < -l/2$. Since $h(x)$ is within $-l/2$ of l, we have $h(x) < l - l/2 = l/2 < 0$, contradicting the fact that h is a positive function. □

How about if we bound a function from above and below and then *squeeze*?

Theorem 30 (Squeeze Theorem)**.** Suppose we have functions f, g and h with $f(x) \leq h(x) \leq g(x)$. If $\lim_{x\to t} f(x) = \lim_{x\to t} g(x) = l$, then $\lim_{x\to t} h(x) = l$.

Challenge 25

(a) If f is differentiable on (a, b) with $f(x)$ a local maximum/minimum, show that $f'(x) = 0$.

(b) Prove the Squeeze Theorem.[14]

And that is it, you have successfully tackled the most difficult topic in calculus! As I mentioned before, the idea of a limit is the culmination of nearly two centuries of investigation. It is truly a difficult concept, but know you know what a limit is, and why we need it.

4.4 Continuous Functions

Continuity and intervals

The idea of continuous functions is that close inputs map to close outputs. We were able to formalize this idea with the definition of a limit. Yet, something is very off: if inputs are close by $\delta(\epsilon)$, then we check if their outputs are close by ϵ. But this is not what we have done: we are satisfied as long as the outputs are close by $c \cdot \epsilon$ for some positive constant c. As an example, in Challenge 25, you will have likely concluded that

$$|h(x) - l| \leq |h(x) - f(x)| + |f(x) - l| \leq |g(x) - l| + |l - f(x)| + |f(x) - l| < \epsilon + \epsilon + \epsilon = 3\epsilon.$$

This is comparatively mild, for in the proof of product rule, we obtained a constant $|l_1| + |l_2| + 1$. Imagine if $|l_1| = 10^{100}$! At this point, can we really say that the outputs are nearby?

Well yes! Even a gigantic number like 10^{100} can be scaled to a small number like 1 with some choice of unit. Once again, something being large is a relative statement, rather than an absolute one. This is why we need to consider all possible values of ϵ.

But this raises a conundrum, for "closeness" is also a relative statement, not an absolute one. This suggests that chasing after "closeness" is perhaps not right.

Here is our current definition of continuity: a function f is continuous at t if for each $\epsilon > 0$, there is some $\delta(\epsilon) > 0$ such that for each x satisfying $|x - t| < \delta(\epsilon)$, we have $|f(x) - f(t)| < \epsilon$. It is the same as the definition of a limit, except that since $\lim_{x\to t} f(x) = f(t)$, we replaced the letter l with $f(t)$, and we allow x to take the value t by removing the restriction $x \neq t$.

Challenge 26 Suppose function g is continuous at t and function f is continuous at $g(t)$. Show that their composition $f \circ g$ is continuous at t.

Recall from Challenge 24 that $\lim_{x\to t} f(x) = l$ if and only if $\lim_{x\to t+} f(x) = \lim_{x\to t-} f(x) = l$. Since both are equivalent, we may take the statement $\lim_{x\to t+} f(x) = \lim_{x\to t-} f(x) = f(t)$ as the definition of continuity. There are actually two statements: $\lim_{x\to t+} f(x) = f(t)$ and $\lim_{x\to t-} f(x) = f(t)$. Combining the two statements into one, we have: a function f is continuous at t if for each $\epsilon_1 > 0$ and for each $\epsilon_2 > 0$, there is some $\delta_1(\epsilon_1) > 0$ and some $\delta_2(\epsilon_2) > 0$ such that for each $x \in \big(t - \delta_1(\epsilon_1), t + \delta_2(\epsilon_2)\big)$, we have $f(x) \in \big(f(t) - \epsilon_1, f(t) + \epsilon_1\big)$ and $f(x) \in \big(f(t) - \epsilon_2, f(t) + \epsilon_2\big)$.

Observe that this unfamiliar definition of continuity no longer appears as a statement about closeness. It is a statement about open intervals: the first part about $\epsilon_1 > 0$ and $\epsilon_2 > 0$ specifies an interval I_o in the output axis (y-axis) while the second part about $\delta_1(\epsilon_1) > 0$ and some $\delta_2(\epsilon_2) > 0$

[14] *Hint:* $|h(x) - l| \leq |h(x) - f(x)| + |f(x) - l| < |g(x) - f(x)| + \epsilon \leq |g(x) - l + l - f(x)| + \epsilon$.

specifies a corresponding interval I_i in the input axis (x-axis). In particular, each input in the interval I_i must map to the interval I_o. As a shorthand, we will write $f(I_i)$ to denote the set of outputs $f(x)$ for $x \in I_i$. With this notation, we wish to show that $f(I_i) \subset I_o$.

Theorem 31. Suppose a function f satisfies the following: for each open interval I_o containing $f(t)$, there is a corresponding open interval I_i containing t such that $f(I_i) \subset I_o$. Then function f is continuous at t.

Proof. We want to show that a function f satisfying the condition outlined is indeed continuous. Let $\epsilon > 0$. Then $\big(f(t)-\epsilon, f(t)+\epsilon\big)$ is an interval containing $f(t)$. Then there will be a corresponding open interval $I_i := (t-a, t+b)$ containing t, where a and b are positive constants. Take $\delta(\epsilon) = \min(a, b)$ and observe that $I := \big(t-\delta(\epsilon), t+\delta(\epsilon)\big) \subset I_i$. Therefore, $f(I) \subset I_o$, in particular, for each $|x-t| < \delta(\epsilon)$, we have $|f(x) - f(t)| < \epsilon$. We see that function f is indeed continuous. □

Challenge 27 Let f be continuous at t. Show that for each open interval I_o containing $f(t)$, there is a corresponding open interval I_i containing t such that $f(I_i) \subset I_o$.

Theorem 31 and Challenge 27 allows us to take the definition of continuity as follows.

Definition 32. A function f is **continuous** at t if each open interval I_o containing $f(t)$ has an open interval I_i containing t such that $f(I_i) \subset I_o$.

Recall from Section 2.5 that completeness of the real numbers was necessary to ensure that continuous functions would not skip over a point. This was because a hole could render a step function to be continuous. We now verify that completeness does indeed make sure that continuous functions do not skip over points.

Theorem 33 (Intermediate Value Theorem)**.** If f is a continuous function and a, b are real numbers such that $f(a) \leq f(b)$, then for each $y \in \big[f(a), f(b)\big]$ there is some $x \in [a, b]$ such that $f(x) = y$.

Proof. If $f(a) = f(b)$ then the statement is satisfied, so we may assume that $f(a) < f(b)$. The idea is as follows: we take the collection of points p such that points to the left of p take on values $f(x) > y$. Then the rightmost point x of such collection should satisfy $f(x) = y$ because any point to the right of x will take on values $f(x) > y$.

Let S be the set of $\alpha \in [a, b]$ such that $f(\alpha) \leq y$. Observe that S contains the endpoint a and is thus a nonempty set of real numbers. Furthermore, because $y \leq f(b)$ we see that the set S is bounded from above by b. By completeness, there is a least upper bound $x := \sup S$.

All that remains is to check that $f(x) = y$. We will do this by showing that neither $f(x) < y$ and $f(x) > y$ are possible. First, suppose that $f(x) < y$ and consider the open interval $I_o := (-\infty, y)$. Since f is continuous at x, there is an open interval I_i containing x such that $f(I_i) \subset I_o$. Let $\beta \in I_i$ such that $x \leq \beta \leq b$. Since $f(I_i) \subset f(I_o)$, we see that $f(\beta) < y$ and so $\beta \in S$. This contradicts the fact that x is an upper bound of S and so $f(x) \not< y$.

Next, suppose that $f(x) > y$. Once again, by continuity of function f at x, there is an open interval I_i containing x such that $f(I_i) \subset I_o$ where $I_o := (y, \infty)$. Let $\alpha \in I_i$ such that $\alpha < x$. Since $[\alpha, x] \subset I_o$ we see that $f([\alpha, x]) \subset f(I_o)$. This means that α is an upper bound of S, contradicting the fact that $x = \sup S$. Therefore $f(x) \not> y$ and the only possibility is that $f(x) = y$, as desired. □

As continuous functions cannot skip over points, a continuous function must map an interval into another.

Corollary 34. If f is a continuous function defined on an interval I, then $f(I)$ is also an interval.

Proof. There is nothing to show if $f(I)$ consists of a single element, so we will assume that $f(I)$ contains at least two points. How would we know if $f(I)$ is *not* an interval? We would know that $f(I)$ is not an interval if it has a hole. That is to say, $f(I)$ is not an interval if there some point p such that for $u, v \in f(I)$ with $u < p < v$, we see that $p \notin f(I)$. This is equivalent to checking that there exists $u, v \in f(I)$ with $u < v$ such that $[u, v] \not\subset f(I)$. The negation of this statement (see Challenge 19) is that $f(I)$ is an interval if each $u, v \in f(I)$ with $u < v$ satisfies $[u, v] \subset f(I)$. But this is guaranteed from the Intermediate Value Theorem and so we are done. □

Once we begin to talk about functions defined on intervals that are not $(-\infty, \infty)$, the definition of continuity must be modified very slightly to take into account the fact that the function is undefined outside the interval.

Definition 35. A function f defined on an interval I is **continuous** at t if each open interval I_o containing $f(t)$ has an open interval I_i containing t such that $f(I_i \cap I) \subset I_o$.

Challenge 28 (Continuing Section 2.5) Define function f on $(-\infty, 0) \cup (0, \infty)$ such that $f : x \mapsto 0$ if $x < 0$ and $f : x \mapsto 1$ if $x > 0$. Show that f is continuous at each $x \in (-\infty, 0) \cup (0, \infty)$. Define function g by $g : x \mapsto 0$ for $x \in (-\infty, 0]$ and $g : x \mapsto 1$ for $x \in (0, \infty)$. Show that g is *not* continuous.

Returning to Corollary 34, a natural question to ask is whether a continuous function f defined on a finite interval can be unbounded, that is, have $f(I)$ that is not a finite interval. If we define $f : x \mapsto 1/x$ on the open interval $(0, 1)$, then f is differentiable with $f((0, 1)) = (1, \infty)$, which is unbounded. Can we define a continuous function f on a closed interval $[a, b]$ such that $f(I)$ is unbounded? This seems implausible since the function would need to: start from $f(a)$ then go infinitely high or infinitely low and then hit $f(b)$ without skipping any points along the way.

So let us suppose that f is a continuous function defined on a closed interval $[a, b]$. As in the proof of the Intermediate Value Theorem, let S be the set of $\alpha \in [a, b]$ such that $f([a, \alpha])$ is bounded. Observe that S is a nonempty set (containing a) that is bounded from above (by b) and so there is some $x := \sup S$. Let us first check to see if $f([a, x])$ is bounded. Since function f is continuous, there is some open interval I_i containing x such that $f(I_i) \subset I_o$ where we will put $I_o := (f(x) - 1, f(x) + 1)$ for definiteness. Let $\beta < x$ be an element of interval I_i so that $f([a, \beta])$ is bounded. But $f([\beta, x]) \subset I_o$ and so the union $f([a, x])$ is also bounded.

We check whether it is possible that $x < b$. If $x < b$, there is some $\gamma \in I_i$ such that $x < \gamma < b$. But then $f([x, \gamma]) \subset I_o$ and so $f([a, \gamma])$ is bounded. This contradicts the fact that x is an upper bound of set S. Therefore, $x = b$ and we conclude that $f([a, b])$ is bounded from above. Switching the orientation of the y-axis shows that $f([a, b])$ is bounded from below. We restate our result below.

Theorem 36 (Boundedness Theorem)**.** If f is a continuous function defined on a closed interval $[a, b]$ then $f([a, b])$ is bounded.

If $f([a, b])$ is bounded, then there is some $y := \sup f([a, b])$. A constant function will always attain this least upper bound y, but does this hold in general? Our intuition suggest yes, for if $f([a, b])$ is only able to attain some value at most $y_0 < y$, then y would not be able to be the supremum of $f([a, b])$. We now verify that this is indeed the case.

Theorem 37 (Extreme Value Theorem)**.** Let f be a continuous function defined on a closed interval $I := [a, b]$ and (by the Boundedness Theorem, we may) put $m := \inf f(I)$ and $M := \sup f(I)$. Then there is some $\alpha, \beta \in I$ such that $f(\alpha) = m$ and $f(\beta) = M$.

Proof. We first show that there is some $\beta \in I$ such that $f(\beta) = M$. If not, then the function $1/[M - f(x)]$ is continuous on the interval I. As M is the least upper bound of $f(I)$, we see that for each $\epsilon > 0$, there is some $x \in I$ such that $M - f(x) < \epsilon$. This means that $1/[M - f(x)] > 1/\epsilon$. But the function $1/\epsilon$ defined on $(0, \infty)$ is not bounded from above. Hence the function $1/[M - f(x)]$ is also not bounded from above, contradicting the Boundedness Theorem. Changing the orientation of the y-axis shows that there is some $\alpha \in I$ such that $f(\alpha) = m$. □

Thus a function defined on $[a, b]$ is guaranteed to attain its extremum. But as we saw with the function $x \mapsto 1/x$ on the interval $(0, 1]$, this is not the case once we lose at least one of the endpoints. As we may change units by any constant multiplicative factor, working on an open interval (a, b) is like working with the entire set of real numbers $(-\infty, \infty)$. This freedom allows us to define continuous functions with values as large as we want. However, by adding in the end points, the interval ironically becomes "smaller". There is something curious at work. Pursuing these threads leads to the field of *topology*.

Inverse functions

As continuous functions map intervals into another, we can consider the especially nice case where each point on an interval is mapped uniquely into a point in another interval.

Let f be a function defined on an interval I. Function g is an **inverse function** of f if each $x \in I$ satifies $(g \circ f)(x) = x$ and each $y \in f(I)$ satifies $(f \circ g)(y) = y$. We previously saw in Section 3.4 that the exponential function is an inverse function of the logarithm function.

Challenge 29 Let function f be defined on an interval I with an inverse function g.

(a) Check that if $f(a) = f(b)$, then $a = b$. We say that function f is **injective**. Conclude that each $y \in f(I)$ has a *unique* $x \in I$ such that $f(x) = y$.

(b) Suppose h is also an inverse function of f. Show that $g = h$ by checking that $g(y) = h(y)$ for each $y \in f(I)$. Conclude that an inverse function is unique.

(c) Assume in addition that f is a strictly increasing function, that is $f(a) < f(b)$ for each $a < b$ in the interval I. Show that g is a strictly increasing function on $f(I)$. Deduce that the exponential function is a strictly increasing function.

(d) State and confirm the analogue of part (c) for strictly *decreasing* functions.

We check that the exponential function is continuous from the definition. We want to show that $\lim_{t \to y} \exp(t) = \exp\big(y\big)$ for each real number y. Let $\epsilon > 0$; we wish to find some positive $\delta(\epsilon)$ value such that for each t satisfying $|t - y| < \delta$, we have $|\exp(t) - \exp\big(y\big)| < \epsilon$. In other words, $\exp\big(y\big) - \epsilon < \exp(t) < \exp\big(y\big) + \epsilon$ whenever $y - \delta(\epsilon) < t < y + \delta(\epsilon)$.

To minimize confusion, let us define the intermediate variable $\log x := y$. The logarithm function is a strictly increasing function and so $\log(x - \epsilon) < \log x < \log(x + \epsilon)$.[15] Take $\delta(\epsilon) := \min\big(\log(x + \epsilon) - \log(x), \log(x) - \log(x - \epsilon)\big)$. Then whenever $y - \delta(\epsilon) < t < y + \delta(\epsilon)$, we have the

[15] What if $x - \epsilon \leq 0$ and so $\log(x - \epsilon)$ is undefined? No problem! We can always pretend that ϵ is small enough (Challenge 18) and make the substitution $\epsilon \mapsto \min(\epsilon, x/2)$. The value x here is fixed because $\log x := y$, where y is fixed.

following inequalities.

$$\log(x-\epsilon) = y - \big(\log x - \log(x-\epsilon)\big) \le y - \delta(\epsilon) < t$$
$$t < y + \delta(\epsilon) \le y + \big(\log(x+\epsilon) - \log x\big) = \log(x+\epsilon)$$

By Challenge 29, the exponential function is strictly increasing and so using $x = \exp\big(y\big)$ gives

$$\exp\big(y\big) - \epsilon = (\exp\circ\log)(x-\epsilon) < \exp(t) < (\exp\circ\log)(x+\epsilon) = \exp\big(y\big) + \epsilon$$

as desired.

We can also check that the exponential function is differentiable (simply assumed in Section 3.4). Recall that the logarithm function is differentiable at each $x \in I := (0,\infty)$ with $\log'(x) > 0$. Put $y = \log x$ and let α be nonzero. Since the exponential function is the inverse function of the logarithm function, there is some nonzero $\bar{\alpha}$ such that $y + \alpha = \log(x+\bar{\alpha})$.

By construction, we have the following.

$$\frac{\exp\big(y+\alpha\big) - \exp\big(y\big)}{\alpha} = \frac{x+\bar{\alpha}-x}{\log(x+\bar{\alpha}) - \log x} = \frac{1}{\big[\log(x+\bar{\alpha}) - \log(x)\big]/\bar{\alpha}}$$

Because the exponential function is continuous, if $|(y+\alpha)-y| \to 0$, then $|\exp\big(y+\alpha\big) - \exp\big(y\big)| \to 0$. Therefore if $\alpha \to 0$, then $\bar{\alpha} \to 0$. We may apply the quotient rule of limits to get

$$\lim_{\alpha\to 0} \frac{\exp\big(y+\alpha\big) - \exp\big(y\big)}{\alpha} = \lim_{\bar{\alpha}\to 0} \frac{1}{[\log(x+\bar{\alpha}) - \log(x)]/\bar{\alpha}} = \frac{\lim_{\bar{\alpha}\to 0} 1}{\lim_{\bar{\alpha}\to 0}[\log(x+\bar{\alpha}) - \log(x)]/\bar{\alpha}}$$
$$= \frac{1}{\log'(x)} = x = \exp y.$$

We see that the exponential function is differentiable at each real number y, with

$$\exp'(y) = \exp\big(y\big).$$

Challenge 30 Let f be a strictly increasing continuous function defined on an open interval I with the inverse function g.

(a) Show that the inverse function g must also be continuous by repeating the argument for the exponential function.

(b) Show that if f is differentiable at $x \in I$ with $f'(x) \neq 0$, then g is differentiable at $f'(x)$ with

$$g'\left(f(x)\right) = 1/f'(x).$$

(c) Show that parts (a) and (b) continue to hold if continuous f is assumed to be strictly *decreasing*.

The Mean Value Theorem

In Section 2.5 we saw that calculus breaks down if we allow motion without velocity. In other words, if $f'(x) = 0$ on an interval, then we absolutely need f to be a constant on that interval. Let us now verify that this is indeed the case.

Now it is very difficult to visualize how a differentiable function f with zero velocity could be non constant. Indeed, the only counterexample seems to be step functions, but we know from the Intermediate Value Theorem that step functions cannot be continuous. So instead of trying to think about functions with zero velocity, let us study the behavior of continuous functions whose behavior in the end point are fixed as a constant, but some motion is permitted in the middle.

Here is the simplest thing we could check. As long as f has velocity between the endpoints, can we say for certain that a point x exist on the interval such that $f'(x) = 0$? If not, we are at a dead end, so this is something that a calculus skeptic might be interested in.

Suppose we have a continuous function f defined on a closed interval $[a, b]$ such that the values of f are equal at the endpoints. Let us further assume that f is differentiable between a and b. If f is a constant, then we are done, so let us assume f is not a constant. We obtained a great result called the Extreme Value Theorem that tells us that there is some $\alpha \in [a, b]$ and $\beta \in [a, b]$ such that $f(\alpha) = \inf f([a, b])$ and $f(\beta) = \sup f([a, b])$. Since f is not a constant, at least one $f(\alpha)$ and $f(\beta)$ are distinct from $f(a)$. Without loss of generality, suppose $f(\alpha) > f(a)$. Then $f(\alpha)$ is a maximum and by Challenge 25, $f'(\alpha) = 0$.

Let us go through the argument once again. Since $f'(\alpha)$ is differentiable, the following limits from above and below exist and are equal.

$$\lim_{x \to \alpha+} \frac{f(\alpha) - f(x)}{\alpha - x} = \lim_{y \to \alpha-} \frac{f(\alpha) - f(y)}{\alpha - y} \tag{4.2}$$

Since $f(\alpha)$ is a maximum, the numerators of the above satisfies $f(\alpha) - f(x) > 0$ and $f(\alpha) - f(y) > 0$. But $\alpha - x < 0$ and $\alpha - y > 0$ and since limits preserve inequalities, we have

$$\lim_{x \to \alpha+} \frac{f(\alpha) - f(x)}{\alpha - x} \leq 0 \qquad \lim_{y \to \alpha-} \frac{f(\alpha) - f(y)}{\alpha - y} \geq 0.$$

Equation 4.2 is satisfied only when

$$\lim_{x \to \alpha+} \frac{f(\alpha) - f(x)}{\alpha - x} = \lim_{y \to \alpha-} \frac{f(\alpha) - f(y)}{\alpha - y} = 0. \tag{4.3}$$

Since f is differentiable at α, we conclude that $f'(\alpha) = 0$.

This result, which we restate below, was first obtained (for polynomials) by Michel Rolle during his years as a calculus skeptic.

Theorem 38 (Rolle's Theorem)**.** Let f be a continuous function defined on the closed interval $[a, b]$ such that $f(a) = f(b)$. Furthermore, suppose f is differentiable on (a, b). Then there is some point $\alpha \in (a, b)$ such that $f'(\alpha) = 0$.

Very good. The next step towards showing that $f' = 0 \implies f = c$ is that each value of f must take a certain value (hopefully the endpoints). In fact, a general statement about continuous functions taking a certain value can be made with no reference to derivatives.

Proposition 39. If f is a continuous function defined on $[0, 1]$ with $f([0, 1]) = [0, 1]$, then there is some $\alpha \in [0, 1]$ such that $f(\alpha) = \alpha$. Such a point α is called a **fixed point**.

Proof. The proof is very sneaky: just like we wish to obtain a result about constant functions by not looking at constant functions, we will consider a different function g. Let $g : x \mapsto x - f(x)$; we will show that there is some α such that $g(\alpha) = 0$. If g is zero at the endpoints, we are done. Suppose $g(0) \neq 0$ and $g(1) \neq 0$. Then $f(0) > 0$ and $f(1) < 1$. Therefore, $g(0) = 0 - f(0) < 0$ and $g(1) = 1 - f(1) > 0$. By the Intermediate Value Theorem, there is some α such that $g(\alpha) = 0$. □

The *average* velocity of an object during some time interval t is the displacement of the object divided by the time interval t. For example, a car that moved 100 km to the right in 1 hour, then the car has an average velocity of 100km/hr. On the other hand, if the car had zero displacement, then no matter the motion, the car has an average velocity of 0km/hr. So this is it! If we can show that the velocity of the car always attains an average velocity of 0 km/hr, then it must have had zero displacement all the time and the car must have been stationary.

We will first show that the average velocity is guaranteed to be attained.

Theorem 40 (Mean Value Theorem)**.** Let f be a continuous function defined on the closed interval $[a, b]$ that is differentiable on (a, b). Then there is some point $\alpha \in (a, b)$ such that

$$f'(\alpha) = \frac{f(b) - f(a)}{b - a}.$$

Proof. As before, we consider a different function h; the derivative of h will measure the difference between $f'(x)$ and $[f(b) - f(a)]/(b - a)$:

$$h'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

The simplest candidate is

$$h(x) := f(x) - \frac{f(b) - f(a)}{b - a}x. \tag{4.4}$$

Let's check the behavior of h at the endpoints: how do they differ?

$$\begin{aligned} h(b) - h(a) &= \left[f(b) - \frac{f(b) - f(a)}{b - a}b\right] - \left[f(a) - \frac{f(b) - f(a)}{b - a}a\right] \\ &= f(b) - f(a) - \frac{f(b) - f(a)}{b - a}(b - a) = 0 \end{aligned}$$

So h is a differentiable function fixed at the endpoints. But from Rolle's Theorem, we already know that such a function must have vanishing derivative at some point $\alpha \in (a, b)$. Therefore

$$0 = g'(\alpha) = f'(\alpha) - \frac{f(b) - f(a)}{b - a}.$$

□

Corollary 41. If f' is zero on (a, b), then f is a constant function on (a, b).

Proof. For each (time) interval $[t_i, t_f] \subset (a, b)$, the Mean Value Theorem guarantees that there is some $t \in [t_i, t_f]$ such that

$$0 = f'(t) = \frac{f(t_f) - f(t_i)}{t_f - t_i}.$$

So $f(t_f) = f(t_i)$. Since our choice of final time t_f and initial time t_i were arbitrary, f is a constant. □

The following are a few results we can obtain easily from our work.

Corollary 42. If $f' = g'$ on (a, b), then $f - g$ is a constant.

Proof. Put $h : x \mapsto f(x) - g(x)$ and observe that $h' : x \mapsto 0$ on (a, b). By Corollary 41, h is a constant function on (a, b). □

Corollary 43. Each antiderivative of f differs by a constant.

Proof. By definition, antiderivatives of a function must have the same derivatives, so they can only differ by a constant. □

Challenge 31 Let f and g be continuous functions on $[a, b]$ and differentiable on (a, b).

(a) If $f'(x) > 0$ for each $x \in (a, b)$, show that f is strictly increasing.
(b) If $f'(x) \geq 0$ for each $x \in (a, b)$, show that f is increasing.
(c) If $f'(x) < 0$ for each $x \in (a, b)$, show that f is strictly decreasing.
(d) If $f'(x) \leq 0$ for each $x \in (a, b)$, show that f is decreasing.
(e) If $f'(x) \leq g'(x)$ on $[a, b]$, use part (b) to show that $f(x) - f(a) \leq g(x) - g(a)$ for each $x \in [a, b]$.
(f) If $\alpha \leq f'(x) \leq \beta$ on $[a, b]$, use part (e) to show that $\alpha(x - a) \leq f(x) - f(a) \leq \beta(x - a)$ for each $x \in [a, b]$. Obtain the **mean value inequality**: $\alpha \leq \frac{f(x)-f(a)}{x-a} \leq \beta$.

Let us take a second look at equality of the Mean Value Theorem:

$$\frac{f'(\alpha)}{1} = \frac{f(b) - f(a)}{b - a}.$$

This is really a statement about two functions, for if we let $g : x \mapsto x$, because $g'(x) = 1$ we have

$$\frac{f'(\alpha)}{g'(\alpha)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

Will such an equality hold in general?

Suppose f and g are continuous functions defined on the closed interval $[a, b]$ that are differentiable on (a, b) with $g'(x) \neq 0$ on (a, b). Simply take the definition of the difference function h from Equation 4.4 and make the change variables $a \mapsto g(a)$, $b \mapsto g(b)$, and $x \mapsto g(x)$ to obtain the function $\tilde{h}$:

$$h(x) := f(x) - \frac{f(b) - f(a)}{b - a}x \implies \tilde{h}(x) := f(x) - \frac{f(b) - f(a)}{g(b) - g(a)}g(x).$$

The function $\tilde{h}$ is undefined if $g'(x) = 0$ anywhere on (a, b) or if $g(b) = g(a)$. The former is ruled out by assumption and the second is impossible by the Mean Value Theorem.[16] By the sum, product,

[16] If $g(b) - g(a) = 0$, the Mean Value Theorem guarantees there is some $\alpha \in (a, b)$ such that $g'(\alpha) = 0$.

quotient rule for limits and derivatives, function $\tilde{h}$ is continuous on $[a,b]$ and differentiable on (a,b). Furthermore, the value of function $\tilde{h}$ equal at the endpoints:

$$\begin{aligned}\tilde{h}(a) - \tilde{h}(b) &= \left[f(a) - \frac{f(b)-f(a)}{g(b)-g(a)}g(a)\right] - \left[f(b) - \frac{f(b)-f(a)}{g(b)-g(a)}g(b)\right] \\ &= f(a) - f(b) - \frac{f(b)-f(a)}{g(b)-g(a)}g(a) + \frac{f(b)-f(a)}{g(b)-g(a)}g(b) \\ &= f(a) - f(b) + \frac{f(b)-f(a)}{g(b)-g(a)}[g(b)-g(a)] = f(a) - f(b) + f(b) - f(a) = 0.\end{aligned}$$

By Rolle's Theorem, there is some $\alpha \in (a,b)$ such that $f'(\alpha) - \frac{f(b)-f(a)}{g(b)-g(a)}g'(\alpha) = 0$; that is

$$\frac{f'(\alpha)}{g'(\alpha)} = \frac{f(b)-f(a)}{g(b)-g(a)}.$$

This result is called the **Generalized Mean Value Theorem**.

Theorem 44 (Generalized Mean Value Theorem)**.** Let f and g be continuous function defined on the interval $[a,b]$ that are differentiable on (a,b). If $g'(x) \neq 0$ for each $x \in (a,b)$, then there is some $\alpha \in (a,b)$ such that

$$\frac{f'(\alpha)}{g'(\alpha)} = \frac{f(b)-f(a)}{g(b)-g(a)}.$$

The Generalized Mean Value Theorem is a result on a quotient of differentiable functions. We know of another result on a quotient of differentiable functions: L'Hospital's rule (Theorem 28). Let us use the Generalized Mean Value Theorem to obtain a slight variation on L'Hospital's Rule. First, here is a variation on the symbol $\lim_{x\to\infty} f(x) = l$ for when f grows boundlessly.

Definition 45. The symbol $\lim_{x\to\infty} f(x) = \infty$ means: for each positive integer m (analogue of ϵ) there is some positive integer n (analogue of δ) such that for each $x > n$ we have $f(x) > m$.

Theorem 46 (L'Hospital's Rule)**.** Let f and g be differentiable on (a,∞) with $g(x) \neq 0$ and $g'(x) \neq 0$ for each $x > a$. If $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = \infty$ and the limit $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$ exists then

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{f'(x)}{g'(x)}.$$

Proof. Let $\epsilon > 0$ and let $l := \lim_{x\to\infty} \frac{f'(x)}{g'(x)}$. By definition there is some positive integer n such that for each $x > n$ we have

$$\left|\frac{f'(x)}{g'(x)} - l\right| < \epsilon. \tag{4.5}$$

Observe that $g(x) \neq g(n)$ because if $g(x) = g(n)$ then Rolle's Theorem guarantees some $\alpha \in (n,x)$ such that $g'(\alpha) = 0$, which is forbidden by assumption. Applying the Generalized Mean Value Theorem to the interval $[n,x]$ tells us that there is some $\beta \in (n,x)$ such that

$$\frac{f(x)-f(n)}{g(x)-g(n)} = \frac{f'(\beta)}{g'(\beta)}.$$

Inequality 4.5 holds for each $x > n$ and in particular $x = \beta$. Therefore

$$\left|\frac{f(x)-f(n)}{g(x)-g(n)} - l\right| = \left|\frac{f'(\beta)}{g'(\beta)} - l\right| < \epsilon. \tag{4.6}$$

It is sufficient to check that

$$\left|\frac{f(x)-f(n)}{g(x)-g(n)} - \frac{f(x)}{g(x)}\right| < \epsilon$$

because then we can use the triangle inequality and Inequality 4.6 to conclude that

$$\left|\frac{f(x)}{g(x)} - l\right| \leq \left|\frac{f(x)}{g(x)} + \frac{f(x)-f(n)}{g(x)-g(n)}\right| + \left|\frac{f(x)-f(n)}{g(x)-g(n)} - l\right| < \epsilon + \epsilon = 2\epsilon.$$

Multiplying and dividing the term $\frac{f(x)-f(n)}{g(x)-g(n)}$ by $g(x)$ gives

$$\frac{f(x)-f(n)}{g(x)-g(n)} = \left[\frac{f(x)}{g(x)} - \frac{f(n)}{g(x)}\right] \Big/ \left[1 - \frac{g(n)}{g(x)}\right].$$

Since $\lim_{x\to\infty} f(x) = \infty$ and $\lim_{x\to\infty} g(x) = \infty$ the values of $f(x)$ and $g(x)$ grow higher without bounds. This means that the quotients $f(n)/g(x)$ and $g(n)/g(x)$ can be made as close to zero as we wish by taking sufficiently large values of x. Thus for sufficiently large x the following must always hold, as desired.

$$\left|\frac{f(x)-f(n)}{g(x)-g(n)} - \frac{f(x)}{g(x)}\right| < \epsilon$$

□

As a result of L'Hospital's rule we may conclude that e^x grows faster than x^k for each positive integer k. Indeed $\lim_{x\to\infty} k!/e^x = k! \lim_{x\to\infty} 1/e^x = 0$ and by L'Hospital's rule

$$0 = \lim_{x\to\infty} \frac{[k\cdot(k-1)\cdots(2)]x}{e^x} = \lim_{x\to\infty} \frac{[k\cdot(k-1)\cdots(3)]x^2}{e^x} = \cdots = \lim_{x\to\infty} \frac{kx^{k-1}}{e^x} = \lim_{x\to\infty} \frac{x^k}{e^x}.$$

In fact, we see that e^x grows faster than x^a for each real a because if we take k to be any positive integer greater than a, then $\lim_{x\to\infty} |x^a/e^x| \leq \lim_{x\to\infty} x^k/e^x = 0$. Using the sum rule for limits (adapted to $\lim_{x\to\infty}$) shows that if P is a polynomial on the variable x, then $\lim_{x\to\infty} \frac{P}{e^x} = 0$.

As a slight variation, we can introduce $\phi_i := x^{-i}$ for each positive integer i to obtain the following.

$$\lim_{x\to 0} \frac{e^{-1/x^i}}{x^k} = \lim_{\phi_i\to\infty} \frac{\phi_i^{n/i}}{e^{\phi_i}} = 0 \tag{4.7}$$

Challenge 32 Define the function f by

$$f : x \mapsto \begin{cases} 0 & x = 0, \\ e^{-1/x^2} & x \neq 0. \end{cases}$$

(a) Use Equation 4.7 to check that f is differentiable at 0 with $f'(0) = 0$.

(b) By the chain rule, f is differentiable whenever $x \neq 0$. Calculate $f'(x)$ for nonzero x.

(c) Use the product rule to calculate $f''(x)$ for nonzero x.

(d) Use the well-ordering principle to show that for each natural number j, if $x \neq 0$ then $f^{(j)}(x) = P_j e^{-1/x^2}$ where P_j is some polynomial on the variable $y := 1/x$. By parts (b) and (c), $P_1 = 2y^3$ and $P_2 = -6y^4 + 2y^3$.

(e) If $x < 0$ then $-1/x > -1/x^2$. Check that if $x \in (0, 1)$ then $-1/x > -1/x^2$. Conclude that if $x < 1$ is nonzero, then $\left|e^{-1/x}/x^n\right| > \left|e^{-1/x^2}/x^n\right|$.

(f) Use the well-ordering principle and the fact that e^y grows faster than any polynomial in y to show that for each natural number j, the derivative $f^{(j)}(0)$ exists with $f^{(j)}(0) = 0$. Observe that part (a) is the case when $j = 1$.

Function f defined in Challenge 32 can be differentiated as many times as we wish (and is thus called **infinitely differentiable**) and so we can create Taylor polynomials of f for any degree k. However, each Taylor polynomials of f at 0 will vanish, no matter how high the degree. Hence the Taylor series of f at 0 also vanishes.

5

Dynamics

5.1 Forces and Energy

Force

The two key concepts in calculus are that of derivatives and integrals. We were able to obtain each concept by examining the ideas behind that of velocities and displacements, respectively. So far, we have taken the motion of objects, their velocities and displacements as a given. We will now go further and examine what *causes* objects to have their velocities and displacements.

To turn a stationary object into one in motion or vice versa, we will need to apply some sort of *force*. Anyone who has gone up a ski piste using a ski lift knows that force is not proportional to velocity, but acceleration. The resistance to acceleration given a force is known as **mass**, and so $F = ma$, where F is the (total) force acting on our object of study, m is the mass of the object, and a is acceleration of the object. This is **Newton's second law**, and it is not to be taken as the definition of force, but rather as a succinct summarization of observations and experiences. This law is in fact incorrect, but a very good approximation in our ordinary lives to a more fundamental law called Schrödinger's equation from quantum mechanics. Notice that force has the dimension Mass $\times$ Length $\times$ Time^{-2}. Force is measured in **newtons** N, where 1 N is defined to be 1 kg·m·s^{-2}.

Since acceleration is a second derivative of position, Newton's second law is an example of a **differential equation**, an equation containing derivative(s) of unknown function(s). Many physical systems are modeled using differential equations. In the context of classical mechanics, we solve differential equations for the unknown function which models the *dynamics* of the system, that is, how the system changes over time.

Work and energy

Two natural question arise after studying velocities and displacements. (1) How much effort must we exert in order to get an object to attain a certain velocity? (2) How much effort must we exert in order to displace an object by a certain distance? For both questions, we will need the object to have moved, for we cannot calculate nonzero velocities or displacements without any movement. Then, to calculate the total effort exerted, we accumulate the force we applied at each location the object was in, until our desired velocity or displacement was attained. The **work done**

W by a force F from position x_i to x_f on a line is the definite integral $W := \int_{x_i}^{x_f} F(x)\,dx$. The "total effort we exert" is quantified by work done, and as we might expect, involves an integral. The dimension of work done is Force × Length, that is, Mass × Length^2 × Time^{-2}.

Let us tackle the first question: how much work must we do to get an object of mass m to some velocity v? Immediately by dimensional analysis, we see that the answer must take the form $c \cdot mv^2$ for some constant c. We will assign the constant c by considering the simplest case. The simplest case we can imagine is applying a constant force F to our object. Then the total work done is $F \cdot (x_f - x_i)$, where x_i is the initial position of our object and x_f is the final position of our object. We cannot assume that the object's velocity during our hard work is constant, because we *want* the velocity to change. However, the next simplest thing to assume is that the object goes from a velocity of 0 and steadily increases to the velocity v. In other words, the object has constant acceleration a. Then the displacement is the average velocity of the object $\frac{v-0}{2}$ multiplied by the total time t we worked on the object. The constant acceleration a is given by the total gain in velocity divided by the time it took to reach that velocity: v/t. Applying Newton's second law gives

$$\text{work needed } = F \cdot (x_f - x_i) = ma \cdot \left(\frac{v-0}{2} \cdot t\right) = m\frac{v}{t} \cdot \left(\frac{v}{2} \cdot t\right) = \frac{1}{2}mv^2$$

and so the dimensionless constant we wanted was 1/2. The quantity $\frac{1}{2}mv^2$ is called the **kinetic energy** of an object, and is denoted by the symbol K. The kinetic energy is often written slightly differently using *momentum*. The **(linear) momentum** p of an object is given by mv. It tells us how quickly we should get away from the object's path. Using this notation, $K = \frac{p^2}{2m}$.

Next we turn to the second question: how much work must we do to get an object from point o to point r? Let us consider an example: suppose we want to lift a box from the floor straight up. Then we must work against the force of gravity. The effort we need to exert will be easier on the moon compared to the earth, so our answer will have to depend on the force we are working *against* to lift up our box. Therefore,

$$\text{work needed } = \int_o^r -F(x)\,dx,$$

where we have a minus sign because we must apply force to *counteract* external forces. This quantity is called the **potential energy** of an object moving along a line. The location o is called the **reference point** or reference position. That the potential energy of an object depends on its reference point might be unsettling, but it is really not. If we want to determine an elevation of a location, we need to establish a reference point: say the ground level, or the sea level, etc, but this does not worry us, as long as we are in agreement on what the reference point is. The potential energy of an object is denoted by the symbol V.

There is however, one subtlety. The potential energy is the work we need to do to *displace* an object from point o to point r. There are actually an infinite number of ways to do this. The normal way to lift a box from the ground to a height r is straight up. However, if we lifted the box up halfway to a height of $r/2$, then returned the box to the ground, then brought the box up to height r, the final *displacement* of the box is still r. A force F is **conservative** if a potential energy can be defined unambiguously no matter how weird we decide to move the object. More precisely, force F is conservative if the integral $\int_o^r -F(x)\,dx$ is defined unambiguously. All forces we will encounter in this book are conservative. An example of a force that is nonconservative is frictional force. If

we are moving an object against frictional force, then the work we need to do will increase with the number of backtracks we take.

If force F is conservative so that $V(x) := \int_o^x -F(\alpha)\,d\alpha$ is unambiguous, then by the Fundamental Theorem of Calculus, $\frac{dV}{dx} = -F(x) + F(o)$. Since we are free to choose our reference point, we choose our reference point such that $F(o) = 0$. Once again, this is analogous to talking about an elevation of a location. Technically we need to specify what our reference elevation is, but a natural reference point is always implicitly used, and so we may talk about an elevation without ambiguity. Hence, even though *the* elevation of a location technically does not make sense, we have no problem ignoring this problem in practice. Similarly, even though it may not make sense to talk about *the* potential energy of an object, we can do so in practice. By defining a point of reference o with $F(o) = 0$ for a conservative force, we have

$$F(x) = -\frac{dV}{dx}.$$

The **mechanical energy**, or *total energy*, of an object is defined to the sum of the object's kinetic and potential energy $K + V$. As long as we are only dealing with conservative forces, the total energy of an object remains the same, and we say that energy is conserved.

Theorem 47 (Conservation of Energy)**.** Suppose we have an object of mass m confined to move along a line. If only conservative forces are at play, then mechanical energy is conserved.

Proof. The notation $\dot{\square}$ for a function $\square$ means $\frac{d}{dt}\square$. The symbol $\dot{x}$ denotes the velocity of our object (the rate of change of the position x of our object with respect to time). Multiplying $\dot{x}$ into both sides of Newton's second law gives $F(x)\dot{x} = m\ddot{x}\dot{x}$. Since F is conservative, $F(x) = -\frac{dV}{dx}$ and so by the chain rule,

$$0 = m\ddot{x}\dot{x} + \frac{dV(x)}{dx}\dot{x} = \frac{d}{dt}\left(m\frac{\dot{x}^2}{2} + V(x)\right) = \frac{d}{dt}(K + V).$$

Notice that $\frac{d}{dt}V(x) = \frac{dV}{dx}\dot{x}$ and *not* $\frac{dV}{dt}$, because x here is used to denote the position function. □

Simple harmonic oscillator

The simplest *nontrivial* force we can imagine is a force $F(x) := cx$ for some constant c. We can (approximately) realize such a force in a spring and mass system, as shown in Figure 5.1, where the longer we pull on the mass, the spring exerts a force proportional to the displacement of pull, which wants to restore the mass *back* to the resting point. We choose the origin of the x-axis to be the resting point of our mass. The assumptions are that we are not pulling too much to damage the spring, and that no other forces (such as gravity, friction, air resistance, etc) are working on our system. Such a system is called the **simple harmonic oscillator**, where the only force F is defined by $F : x \mapsto -kx$, in which k is the spring constant. This force is called **Hooke's law**. The spring constant k is a property of the spring which dictates the strength of its pull. Notice the minus sign: the spring is working against us, not for us, as we displace the attached mass.

We could apply Newton's second law $F = ma$ to get the equation $ma = -kx$ and solve for x to find out the oscillator's motion. Instead, let us examine the system's energy. Since $F(0) = 0$, we set the reference point at the origin. At a displacement of r, the potential energy V of our system is

$$V = \int_o^r -F(x)\,dx = \int_0^r kx\,dx = \frac{1}{2}kr^2.$$

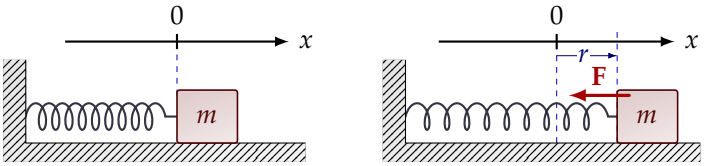


Figure 5.1: A mass and spring system at rest and displaced by r. (By Izaak Neutelings at `tikz.net`)

Therefore, the total energy E of our mass and spring system is given by

$$E := K + V = \frac{p^2}{2m} + \frac{1}{2}kx^2.$$

By conservation of energy, the quantity E is conserved for all time and is thus a constant.

A squared term (with some constant) plus a squared term (with other constant) equals another constant. Where have we seen something like that before? To make this more explicit, let us divide both sides by the nonzero constant E to get $\frac{x^2}{2E/k} + \frac{p^2}{2mE} = 1$.[1] Setting $a^2 := 2E/k$ and $b^2 := 2mE$ gives us an equation of an ellipse!

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

Figure 5.2 shows our ellipse with the x-axis representing position and the y-axis representing momentum. When we represent a system in terms of its position and momentum, as we are doing now, we are working in **phase space**.

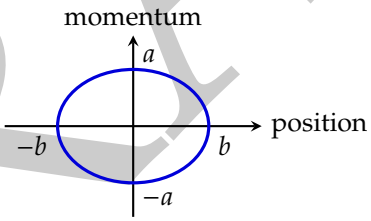


Figure 5.2: Our simple harmonic oscillator has an elliptical trajectory in phase space.

Phase space

Let us try to get some intuition about phase space. Consider Figure 5.3, where an object has been displaced from position x_i to x_f. We say that our object has undergone a **(spatial) translation**. What caused the the object to undergo spatial translation? Momentum p was applied to our object. We say that momentum *generates* translation.

Now let us add time. The time axis will work just the same as it did in calculus: there is no preferred sense of direction (just like left or right are equally valid), we can "move" through it just

[1] We have used the fact that $\frac{b}{1/a} = ab$.

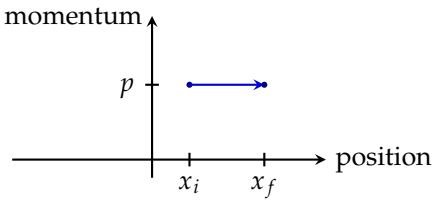


Figure 5.3: Momentum p is applied to an object, displacing it from x_i to x_f.

like any spatial dimension. Suppose we added energy to our oscillator and we graphed the system in phase space, as in the right of Figure 5.4. What is going on there?

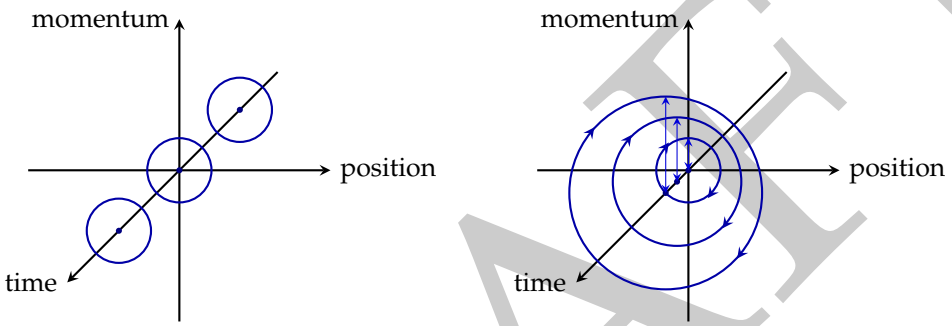


Figure 5.4: The phase space of an oscillator and another with energy being added.

Let us review the simpler case of momentum being added to the object (Figure 5.3). Initially, the object is stationary, with no motion, at position x_i. Since position x_i is simply a label, we are free to assign any numerical value: it could be the origin, or not. Now we apply momentum to the object, and the object is translated to position x_f. Once again, the position x_f is simply a label, we are free to assign any numerical value. Suppose we define the origin to be position x_f. Then $x_f = 0$. However, x_i can no longer be the origin. It can be some positive number or negative number depending on our choice of left/right, but it cannot be zero. We therefore conclude that spatial translation has occurred, and it was caused by momentum. We say that momentum generates translation.

Now we return to the case of adding energy to our oscillator (Figure 5.4). Because time is simply a coordinate (like position), we can assign any time to be the origin. In particular, by energy conservation, as long as no energy is added or removed, the diagram in our phase space will continue to be the same. So we could assign the time of zero to any of them and no one will be the wiser (left diagram in Figure 5.4). Now suppose we add in some energy. Time is an axis like any other and we can define the origin to be anywhere we want. So suppose we define the time to be 0 right after energy is added to the oscillator. Now the oscillator in the previous state with less energy has a different diagram from the new one and cannot be said to be at time zero: it could be positive or negative depending on our choice of direction, but it is not zero. We see that **time translation** has occurred, and it was caused by energy. We say that energy *generates* time

translation.

In the context of phase space, we call the total energy of a system the **Hamiltonian** and denote it by the symbol H. Thus $H(x,p) := \frac{p^2}{2m} + V(x)$ for the potential energy V of the system.

5.2 Vectors and Matrices

Vectors

We now return to the *very* beginning and ask ourselves again, what is 1 + 1? The answer is still 2, and once again, we will insist that these numbers mean something. To each number 1, we attach the meaning of 1 apple and 1 orange, respectively. We now ask again, what is 1 apple plus 1 orange?

Our position at the beginning of the book was that this question involves quantities that cannot be matched, and therefore, it is a sum which cannot be resolved. This led to the idea of dimensional analysis. Yet there is another answer that is just as reasonable. 1 apple plus 1 orange is 2 fruits! Let us see where this takes us.

We begin with most important question: why? Why are we trying to sum different fruits? An obvious application is to make a fruit salad or a fruit juice or a platter of fruits. Let us suppose we want to create a fruit salad.

Although it is convenient to clump things together under a bigger label (in this case, "fruits"), we have the additional complexity of having to keeping track of things. For example, 1 apple plus 1 orange is 2 fruits, but so is 1 tomato and 1 olive. It will be necessary to distinguish between the pile of 1 apple and 1 orange versus the pile of 1 tomato and 1 olive, because they are *not* interchangeable when making a fruit salad.

For simplicity, let us assume there are only three different types of fruits: apples, oranges, and tomatoes. We can express the sum 1 apple + 1 orange by the list $\begin{pmatrix}1\\1\\0\end{pmatrix}$, where we have established the convention that the first (or top) of our list is the number of apples, the second (or middle) of our list is the number of oranges, and the third (or bottom) of our list is the number of tomatoes. To create a recipe, we need some standardized form of units: grams, pounds, cups, etc. So depending on the unit we choose, the list will look different. But even if the numbers look different with different units, the fruit salad that we have in mind will still be the same. We will call lists of numbers by **vectors**, and the **dimension** of the vector is the length of the list. In our case, we are dealing with vectors of dimension 3, because we only consider three types of fruits. Notice that the object (in this case, a fruit salad) is *represented* as a vector, but the representation is not unique because we can always change the units. To emphasize that objects are unbound to a particular vector, we will write them using a special symbol. For example, a fruit salad named A made with 100 grams of apples and 20 grams of oranges could be represented as the vector $\begin{pmatrix}100\\20\\0\end{pmatrix}$. But we will refer to the fruit salad itself by the symbol $|A\rangle$.

The next thing to do is to try and furnish an arithmetic. Now, it is straightforward to add and subtract using vectors. For example, if we have a fruit salad $|A\rangle$ and another fruit salad $|B\rangle$, then

we may add them together to get a bigger fruit salad by representing each as a vector in the *same* units, and adding each up. For example, if $|A\rangle$ has 100 grams of apples and $|B\rangle$ has 20 grams of oranges, then

$$A + B = \begin{pmatrix} 100 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 20 \\ 0 \end{pmatrix} = \begin{pmatrix} 100 \\ 20 \\ 0 \end{pmatrix}.$$

Subtraction works the same way, but the plus sign becomes a minus sign. How about multiplication and division? Well, it makes little sense to multiply two fruit salads, or to divide a fruit salad by another, so we will not attempt to define a multiplication of vectors.[2] However, it makes perfect sense to double a portion of fruit salad or halve a portion of fruit salad. The **scalar multiplication** of a scalar c on a vector $D = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$, written cD is defined by

$$cD := \begin{pmatrix} c \cdot d_1 \\ c \cdot d_2 \\ c \cdot d_3 \end{pmatrix}.$$

In fact, scalar multiplication could mean a change in portion, but also a change of units. For example, to convert a vector whose unit in each entry is a gram (for standardization, we insist that all entries in a vector share the same unit), then to convert it into a vector whose unit in each entry is a kilogram, we do a scalar multiplication by 1/1000.

Linearity

Much like we can calculate the displacement of an object from time t_i to t_f by applying integration to a function, or calculate the velocity of an object by applying differentiation to a function, we can change the units of the entries in a vector from one to another by applying a change of units. Let us denote the last operation by the symbol ^{o} (much like a derivatives are indicated by $'$). With respect to the two arithmetic operations we know, vector addition and scalar multiplication by scalar c (a real number will sometimes be referred to as a **scalar**), the following holds:

$$(A + B)^o = A^o + B^o \text{ and } (cA)^o = cA^o. \tag{5.5}$$

The first simply reflects the fact that combining fruit salads and then changing units of measurement gives us the same result as changing units after adding two fruit salads. The second comes from the fact that halving a portion and then changing units is the same as changing units and then halving a portion.

We have seen this before. For a real number c and differentiable functions f and g, we have

$$(f + g)' = f' + g', \text{ and } (cf)' = cf'.$$

Ditto for integration of bounded continuous functions f and g defined on an interval $[t_i, t_f]$:

$$\int_{t_i}^{t_f} \big[f(x) + g(x)\big]\, dx = \int_{t_i}^{t_f} f(x)\, dx + \int_{t_i}^{t_f} g(x)\, dx, \text{ and } \int_{t_i}^{t_f} cf(x)\, dx = c\int_{t_i}^{t_f} f(x)\, dx.$$

[2] We will revisit this matter later.

Since this pattern has already occurred three times with respect to the most important operations, we will single this out and call this property **linearity**. Thus an operation that satisfies the two conditions in Equations 5.5 is said to be **linear**.

Matrices

We began with fruits and salads, yet unexpectedly returned to calculus. How about we consider an example from calculus? One of the simplest nontrivial thing we can do with calculus is to calculate derivatives of polynomials. Let us try representing polynomials using vectors. The catch is that, like capping the number of fruits, we will need to cap the degree of the polynomials we are considering. Let us fix the maximum degree at 2 and consider polynomials of the form ax^2+bx+c. We may write such a polynomial using vector notation as $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$. The derivative is the vector $\begin{pmatrix} 0 \\ 2a \\ b \end{pmatrix}$, as you can verify using the differentiation rules.

But which rules in which order? By linearity of the derivative operation,

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}' = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix}' + \begin{pmatrix} 0 \\ b \\ 0 \end{pmatrix}' + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}' = a\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}' + b\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' + c\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}'. \tag{5.6}$$

Observe that is precisely an application of the sum rule and the product rule. By the power rule,

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}' = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} \qquad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.7}$$

Plugging these values back into Equation 5.6, we have

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}' = a\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}' + b\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' + c\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' = a\begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix} + b\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + c\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 2a \\ b \end{pmatrix}.$$

That was a long roundabout way of doing something we knew from the very beginning. Or is it? Notice how once we have the values of derivatives at each entry calculated upfront, as in Equations 5.7, all that is needed is just scalar multiplication and vector addition. No additional calculus needed!

Since our strategy from the beginning was to do as little calculus as possible and replace it with as much arithmetic as possible, this is very good news! Let us systemize this procedure.

Let $\mathbb{R}^n$, called the n dimensional *real vector space*, denote the set of vectors of dimension n with real entries, equipped with the vector addition and scalar multiplication operations. The (ordered) **standard basis** of the n dimensional real vector space are the n vectors $e_1, e_2, \ldots, e_n$ defined by the following, and listed in that order.

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix} \qquad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix} \qquad e_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix} \qquad \cdots \qquad e_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

To calculate the derivative of a polynomial of degree $n-1$, all we need to do is cache the result of the power rule applied to each polynomial represented by the standard basis. After that, all we need to do is scalar multiplication and vector addition. By the power rule,

$$e_1' = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix} \qquad e_2' = \begin{pmatrix} 0 \\ 0 \\ n-2 \\ \vdots \end{pmatrix} \qquad \cdots \qquad e_{n-1}' = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \qquad e_n' = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

There is no reason we need to keep track of n vectors separately. How about we squash them all together into one object, as shown below? We will call this **concatenation** of vectors.

$$D := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ n-1 & 0 & \cdots & 0 \\ 0 & n-2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{5.8}$$

Our new procedure for taking the derivative of a polynomial of degree at most $n-1$ is as follows. Suppose we have vectors $v_1, v_2, \ldots, v_n$ and scalars $c_1, c_2, \ldots, c_n$. A **linear combination** of vectors $v_1, \ldots, v_n$ with coefficients $c_1, c_2, \ldots c_n$ is the expression

$$c_1 v_1 + c_2 v_2 + \cdots + c_n v_n.$$

For each polynomial $p = a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_n$, we turn its vector representation v into a linear combination

$$v = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = a_1 e_1 + a_2 e_2 + \cdots a_n e_n.$$

The derivative can be found by looking at each column of the matrix D defined in Equation 5.8:

$$v' = a_1 e_1' + a_2 e_2' + \cdots a_n e_n' = a_1 \text{ column 1 of } D + a_2 \text{ column 2 of } D + \cdots + a_n \text{ column } n \text{ of } D. \tag{5.9}$$

Because D contains all the information about derivatives of polynomials we need, it is our familiar derivative operator, but represented as a *matrix*. A **matrix** is a rectangular table of numbers. A matrix with m rows and n columns is said to have **dimension** $m \times n$. Notice that unlike a vector which records static information about an object, a matrix is *dynamic*, taking a vector and transforming it into another. In this case, the matrix D takes in a polynomial of degree at most $n-1$ and transform it to another polynomial (its derivative).

Let us try and apply what we have discovered to fruits. The key operator here is the change of units. Suppose we were measuring fruits in grams and we wished to measure instead in kilograms. Using the notation ^{o} to signify the change of units, we have

$$e_1^o = \begin{pmatrix} 1/1000 \\ 0 \\ 0 \end{pmatrix} \qquad e_2^o = \begin{pmatrix} 0 \\ 1/1000 \\ 0 \end{pmatrix} \qquad e_3^o = \begin{pmatrix} 0 \\ 0 \\ 1/1000 \end{pmatrix}.$$

Then the change of unit can be represented as a matrix

$$C := \begin{pmatrix} 1/1000 & 0 & 0 \\ 0 & 1/1000 & 0 \\ 0 & 0 & 1/1000 \end{pmatrix}.$$

To do a change of units for a fruit salad recipe $|r\rangle$, we take its vector representation $r = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$, turn it into a linear combination of standard basis vectors, then apply the rule

$$r^o = r_1 \text{ column 1 of } C + r_2 \text{ column 2 of } C + r_3 \text{ column 3 of } C. \tag{5.10}$$

Since it gets rather tedious to right out the expressions in Equations 5.9 and Equations 5.10, we will use the shorthand Dv and Cr, respectively.

Thus if we have a matrix A and a vector v defined by

$$A := \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \qquad \text{and} \qquad v := \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

then the expression Av is the vector defined by the sum

$$Av := v_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{m1} \end{pmatrix} + v_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{m2} \end{pmatrix} + \cdots + v_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{mn} \end{pmatrix}.$$

Computing the vector additions above, we have

$$Av := \begin{pmatrix} A_{11}v_1 + A_{12}v_2 + \cdots + A_{1n}v_n \\ A_{21}v_1 + A_{22}v_2 + \cdots + A_{2n}v_n \\ \vdots \\ A_{m1}v_1 + A_{m2}v_2 + \cdots A_{mn}v_n \end{pmatrix}. \tag{5.11}$$

This looks a little scary, but do not fear, for we are simply restating what we have been doing with derivatives of polynomials and changing units of fruits.

Partial derivatives

We interrupt this program to bring you some calculus! Suppose we have a function f that takes as inputs vectors of dimension n and outputs a real number. We can think of function f as taking n inputs, and a natural question to ask is what its rate of change is with respect to one of the n inputs.

In order to do this for the kth input, we start from $f(t)$ and vary t by αe_k. If there is a number $\partial_k f(t)$ such that the following equation holds:

$$f(t + \alpha e_k) = f(t) + \partial_k f(t)\alpha + o(\alpha)$$

then we know that the rate of change of function f at t is given by $\partial_k f(t)$. The number $\partial_k f(t)$ is called the **partial derivative** of f at t with respect to the kth variable.

Back in Section 3.4 when we were showing that $\log(xy) = \log x + \log y$, we defined a function $g : x \mapsto \log(xy) - \log x$. The function g is a function of a single variable because the quantity y was treated as a constant. Taking the derivative of g gave us

$$g'(x) = \frac{1}{xy}\frac{\mathrm{d}}{\mathrm{d}x}(xy) - \frac{1}{x} = \frac{y}{xy} - \frac{1}{x} = 0.$$

We could have achieved the same thing by defining f as a function of two variables x and y defined by $f(x, y) := \log(xy) - \log x$ and then taking the partial derivative with respect to x to get:[3]

$$\partial_x f(x, y) = \frac{1}{xy}\partial_x(xy) - \frac{1}{x} = \frac{y}{xy} - \frac{1}{x} = 0.$$

The reason is the same as why $(cx)' = cx'$ whenever c is a fixed number. In the definition of a partial derivative, the only thing we vary is the kth variable by adding αe_k, while all other inputs are fixed numbers. For example, if $h(x, y) = 3xy + y^2$ and we want to know $\partial_1 h(2, 5)$, then y is no longer a variable: it is the constant 5. This means that partial derivatives obey the same differentiation rules as our ordinary derivatives.[4]

A popular notation that we will use is that if we have a function f which has inputs denoted by the variables ♣, ♠, then we will write $\frac{\partial f}{\partial \clubsuit}$ and $\frac{\partial f}{\partial \spadesuit}$ to denote the partial derivatives with respect to ♣ and ♠, respectively. If the function f is twice partial differentiable with respect to the ♣ variable, then we write $\frac{\partial^2 f}{\partial \clubsuit^2}$.

If f is a function of n variables, then the **gradient** of f at t, denoted by $\nabla f(t)$, is defined to be

$$\nabla f(t) := \begin{pmatrix} \partial_1 f(t) \\ \vdots \\ \partial_n f(t) \end{pmatrix}.$$

There is also a **Laplacian** operator, denoted by the symbol ∇^2 or Δ, defined by

$$\nabla^2 f := \sum_{i=1}^{n} \partial_i^2 f.$$

[3]The notation ∂_x is used in place of ∂_1 because we know that x is the first variable.

[4]Feel free to check this. It amounts to defining functions of one variable and then applying the usual differentiation rules. The process is the same as finding $\partial_1 h(2, 5)$ for $h(x, y) = 3xy + y^2$ by defining $\tilde{h}(x) = 3x \cdot 5 + 5^2$ to obtain a function of one variable, then taking the derivative $\tilde{h}'$ and plugging in 2 for x to get $\partial_1 h(2, 5) = h'(2) = 15$.

Matrix multiplication

Although units do not support chaining, functions do. Vectors do not support chaining, but we can chain matrices together, as in $(A \circ B)v := A(Bv)$. We will use the shorthand ABv to mean the same thing. Looking at the ghastly expression of Equation 5.11, it may seem like we are asking for trouble. But once again, matrices are nothing scary. All they do is tell us how to transform vectors.

We got our first matrix D from Equation 5.8 by concatenating (squashing) vectors together.

$$e_1' = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix} \quad e_2' = \begin{pmatrix} 0 \\ 0 \\ n-2 \\ \vdots \end{pmatrix} \quad \cdots \quad e_n' = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \quad \Longrightarrow \quad D := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ n-1 & 0 & \cdots & 0 \\ 0 & n-2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

Recalling this fact, we can study the chain AB independently of the input vector v just as we can study $f \circ g$ independently from its input t. Since a matrix exists to transform vectors, the matrix A in the chain AB is looking for a vector. But matrix B is simply a concatenation of vectors, just like our matrix D was a concatenation of vectors. In particular, the jth column of a matrix B, denoted by the notation B_j, is a vector, which is exactly what A is looking for! The following is an example from our derivative matrix D.

$$D_1 = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix} := e_1' \quad D_2 = \begin{pmatrix} 0 \\ 0 \\ n-2 \\ \vdots \end{pmatrix} := e_2' \quad \cdots \quad D_{n-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} := e_{n-1}' \quad D_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} := e_n'$$

So we simply repeat what we have done and apply B to each of the columns of A, then concatenate them together.

Let us see this in action for our polynomial derivative matrix D. We are allowed to take derivatives as many times as we wish with polynomials. So if we want to take a derivative a second time, then we simply apply the matrix D to each of $D_1, D_2, \ldots, D_n$ using the fact that $D(D_i) := D(e_i') = (e_i)''$. To obtain our matrix for taking derivatives twice, which we will refer to as D^2, we simply do what we did before: concatenate the vectors $D(D_i)$ together. To recap, the operation D^2 to take derivatives twice is given by the matrix whose kth column is given by the vector $D(D_k)$, or in matrix form:

$$D^2 := DD = \begin{pmatrix} | & | & \cdots & | \\ D(D_1) & D(D_2) & \cdots & D(D_n) \\ | & | & \cdots & | \end{pmatrix}.$$

Challenge 33

(a) Find the matrix D for differentiating a polynomial of degree at most four (polynomials of the form $ax^3 + bx^2 + cx + d$). Check your answer agrees with our derivative matrix given above.

(b) Find the matrix D^2 two different ways. First, by calculating $(ax^3 + bx^2 + cx + d)''$ and caching the rule for transforming standard basis vectors e_1, e_2, e_3, and e_4 as one object through concatenation. Second, calculate the vectors $D(D_1)$, $D(D_2)$, $D(D_3)$, and $D(D_4)$, then concatenate the four vectors as one object. Verify that your results from both methods are equal.

Let us review this prescription for a general matrix. Suppose B is a matrix that takes in a vector v of dimension n and produces a vector Bv of dimension m, where n and m are positive integers. We want to feed this result into a second matrix A, which takes in a vector of dimension m and produces a vector of positive integer dimension l. Then their product matrix $C := AB$ takes in a vector of dimension n and returns a vector of dimension l. This means that to figure out the product matrix C, we need l piece of information: how C transforms each of the vectors of the standard basis $e_1, e_2, \ldots, e_l$. Once we have those information, we can combine them together as one object:

$$C := AB = \begin{pmatrix} | & | & \cdots & | \\ Ce_1 & Ce_2 & \cdots & Ce_n \\ | & | & \cdots & | \end{pmatrix}.$$

Using the fact that $C := AB$ and that $B_i := Be_i$, we have the following.

$$C = \begin{pmatrix} | & | & \cdots & | \\ Ce_1 & Ce_2 & \cdots & Ce_n \\ | & | & \cdots & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ ABe_1 & ABe_2 & \cdots & ABe_n \\ | & | & \cdots & | \end{pmatrix} = \begin{pmatrix} | & | & \cdots & | \\ AB_1 & AB_2 & \cdots & AB_n \\ | & | & \cdots & | \end{pmatrix}$$

We restate what we have found.

Definition 48. Let A be a matrix of dimension $l \times m$ and B be a matrix of dimension $m \times n$. The **matrix multiplication** AB of the two matrices A and B results in a matrix C of dimension $l \times n$ defined by $C_k := AB_k$.

Notice that a matrix cannot transform just any old vector, and so there is some restriction in our ability to do matrix multiplication. For example, if the vector v has dimension 5, but matrix B has dimension 1×1, then the matrix B cannot transform the vector v. In order for the chain ABv to work, matrix A must have dimension $\clubsuit \times m$, where $\clubsuit$ is any positive integer and m is the dimension of Bv.[5] To recap, we can chain a matrix A of dimension $\clubsuit \times m$ with a matrix B of dimension $m \times \spadesuit$ to get the chain AB, but we may *not* form the chain BA unless $\clubsuit = \spadesuit$.

What about chaining three or more matrices? Consider the chain ABC for matrices A, B, and C (with the appropriate dimensions). There is potentially some ambiguity, for ABC could mean the matrix multiplication $(AB)C$ or $A(BC)$. Well, a matrix is nothing more than a way to cache the rules for transforming vectors. Hence matrices are simply a concrete way of writing down a particular class of functions (linear functions). Recall that function composition is associative. Thus the results $(f \circ g) \circ h(x)$ and $f \circ (g \circ h)(x)$ are the same. If we represent a linear function f by the matrix A, a linear function g by the matrix B, and a linear function h by the matrix C, then for each vector v with the appropriate dimension, $(AB)Cv$ and $A(BCv)$ will give the same result. Therefore, the expression ABC is unambiguous, and matrix multiplication is associative: $(AB)C = A(BC)$.

There is a distinguished matrix I, defined by $I_k := e_k$. That is,

$$I := \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

[5] Recall that a matrix has dimension $\clubsuit \times n$ if the matrix has $\clubsuit$ rows and n columns.

The rule that matrix I uses to transform vectors is: transform a standard basis vector e_i into e_i. Hence for each vector v, the vector $Iv = v$. By the definition of matrix multiplication, for each matrix M, the matrix multiplications $IM = MI = M$. Because the matrix I does nothing, it is called the **identity matrix**. We will also denote it by 1, because the number 1 is the distinguished real number such that for each number c, $1 \cdot c = c \cdot 1 = c$.

There is also a rather silly matrix called the **zero matrix**, which we will denote 0, defined as the matrix with zero everywhere:

$$0 := \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

For each matrix, M, we have $M0 = 0M = 0$. Once again, this is in analogy to the real number 0, with the property $0 \cdot c = c \cdot 0 = 0$ for each real number c.

We can multiply a real number by another real number. We can also multiply a real number to a matrix. For a real number c and matrix M, the matrix cM is the matrix whose entries have each been multiplied by c. For example, in the context of matrices, -1 denotes the identity matrix 1 multiplied by the real number -1:

$$-1 := \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Challenge 34 Denoting a matrix I by the number 1 and the zero matrix by the number 0 is our first step in accepting matrices as numbers we can do arithmetic with (just like we did for units and functions). However, matrices and matrix multiplication exhibit some odd behavior that we have not seen with real numbers. This makes matrices more exciting!

(a) As a warm up, use the definition of matrix multiplication and the matrix transformation rule given in Equation 5.11 to show that if $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B := \begin{pmatrix} w & x \\ y & z \end{pmatrix}$ then

$$AB = \begin{pmatrix} aw + by & ax + bz \\ cw + dy & cx + dz \end{pmatrix}.$$

(b) Let $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. Show that $AB \neq BA$. We say that matrix multiplication is not *commutative*, because changing the order of multiplication may change the result.

(c) Let $\epsilon := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Show that $\epsilon^2 := \epsilon\epsilon = 0$, even though $\epsilon \neq 0$. This justifies the mysterious dual numbers. It is perfectly possible to have non-zero things that square to a zero.

(d) Let $A := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $B := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Show that $A^2 := AA = -1$ and $B^2 := BB = -1$.

The German theoretical physicist Werner Heisenberg was one of the founders of the field of quantum theory. He published his Nobel Prize winning paper on matrix mechanics at the age of 24, laying the foundation of quantum mechanics. It which would make obsolete "old quantum

theory", which were heuristics used to attempt to explain quantum mechanical phenomena. He would later receive the Nobel Prize in Physics at the age of 31 “for the creation of quantum mechanics".[6] You now know more about matrices than Heisenberg did when he was creating his matrix formulation of quantum mechanics! Matrix theory was considered at the time to be abstract mathematics (matrix multiplication was only first written down in the 19th Century). Congratulations on making it this far!

5.3 The Complex Field

An algebra

We have so far made zero attempt to multiply two vectors, even though vector multiplication seems like a natural thing to try and figure out. Rather than engaging in excessive generality, we will explore a nontrivial yet simple setting: multiplying two vectors, each with two real entries.[7] So here is the plan: we have two vectors v and w, and we want to create their product u. If u and v were basis vectors, it would make sense for us to write this out as a linear combination:

$$u = v \cdot w = av + bw$$

where a, b are constants. So let us consider the multiplication of basis vectors.

Everything is happening in the simple setting of two entries and so two basis vectors is sufficient to describe all our vectors involved, including products. Since all our product vectors can be expressed as a linear combination of two basis vectors, let us try to identify a good candidate for these two basis vectors, which we will call α and β. First, let us assign a vector to α, so we have something to work with. The simplest thing would be to consider the vector α as the number zero, but with our intuition from real numbers, we would expect everything multiplied to a zero to become zero. This is far too trivial: for each vector v, we have $\alpha \cdot v = \alpha$.[8] The next simplest is to consider the vector α as the number 1, so that $\alpha \cdot v = v \cdot \alpha = v$. Now what about the vector β? Let us write down what we have figured out so far:

$$\alpha \cdot \alpha = \alpha,^{9} \quad \alpha \cdot \beta = \beta \cdot \alpha = \beta,^{10} \quad \beta \cdot \beta = a\alpha + b\beta$$

where a and b are scalar constants (and *not* vectors).

What do we do next? There is no more information to go by. The only knob we have at our disposal is our freedom in how we define β. So let's see what happens if we pull apart the vector α out from β. Define $\bar{\beta} := \beta - c\alpha$, where c is a real number.[11] Then

$$\bar{\beta} \cdot \bar{\beta} = (\beta - c\alpha) \cdot (\beta - c\alpha) = \beta \cdot \beta - 2c\alpha \cdot \beta + c^2 \alpha \cdot \alpha.$$

Using our known information $\beta \cdot \beta := a\alpha + b\beta$, $\alpha \cdot \alpha = \alpha$, $\alpha \cdot \beta = \beta$ and simplifying, we have

$$\bar{\beta} \cdot \bar{\beta} = (a + c^2)\alpha + (b - 2c)\beta.$$

[6] Obviously no one *created* quantum phenomena, but someone had to work out the *theory* of quantum mechanics.

[7] We already know how to multiply two vectors, each with one real entry. Two entries is the next simplest.

[8] It turns out that such a simple structure is the foundation of some very applicable mathematics, but we will not deal with this in this book.

[9] This is the analogue of $1 \cdot 1 = 1$.

[10] This is the analogue of $1 \cdot \beta = \beta \cdot 1 = \beta$.

[11] Since we do not know a priori how much α we need to pick out of β, we will quantify our ignorance with this new constant c.

Now check this out! If $2c := b$, then $\bar{\beta} \cdot \bar{\beta} = d\alpha$, for some constant d. The choice of basis vector $\bar{\beta} := \beta - (b/2)\alpha$ is superior, so let us forget that β existed by replacing it with $\bar{\beta}$ to get

$$\bar{\beta} \cdot \bar{\beta} = (a + [b/2]^2)\alpha + (b - 2c)[\beta - (b/2)\alpha] = (a + b^2/4)\alpha + (b - 2c)\bar{\beta} = (a + b^2/4)\alpha.$$

What we have successfully done is to turn our abstract problem of trying to multiply two vectors into something that's like multiplying two real numbers: α corresponds to the real number 1 and $\bar{\beta}^2$ corresponds to the real number $d := a + b^2/4$. There are three possibilities for the real number d: (i) $d = 0$ or (ii) $d > 0$ or (iii) $d < 0$. In actuality, because we can choose any units to scale things as we wish, there are really only three unique values we need to contemplate. Either d is 0, 1, or -1. In other words, we have $\bar{\beta}^2 = 0$ or $\bar{\beta}^2 = 1$ or $\bar{\beta}^2 = -1$.

Now this is very interesting, we have already seen the case of $\bar{\beta}^2 = 0$ in our encounter with dual numbers. The case of $\bar{\beta}^2 = 1$ is not super interesting because $\alpha^2 = 1$ as well. But what's this? The case of $\bar{\beta}^2 = -1$, now that's something! Where have we seen this before? We have seen such a behavior in Challenge 34 with the matrices $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, which both square to the matrix -1. Since we have to make a choice, we will take

$$\alpha := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \bar{\beta} := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and we will call the matrix $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ the *conjugate* of $\bar{\beta}$.

So how did we go from starting with an attempt to multiply two vectors and end up with matrices? Indeed, the "vectors" α and $\bar{\beta}$ look like matrices, each with four entries, and they do not look like "vectors". To see that these are also vectors with two entries, but in a different notation, consider the linear combination of the basis vectors:

$$x\alpha + y\bar{\beta} = x \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + y \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}.$$

The "matrices" we are dealing with only have two knobs to change about, and so they can be described by vectors of dimension two. In fact, how about we make this explicit right now?

Complex numbers

We will now treat the matrices α and $\bar{\beta}$ as numbers. There is no problem thinking of α as the real number 1, as we have done so before, but the catch is that we have to remember that $\bar{\beta}$ squares to -1. Because of this curious property, we call $\bar{\beta}$ the **imaginary number** and denote it with the symbol i. Of course, there is nothing more imaginary about i compared to the real numbers, but this is the nomenclature.

Thus the linear combination $x\alpha + y\bar{\beta}$ for real numbers x and y will now be written as the number $x + yi$, and we call the set of such numbers the **complex numbers**. The set of complex numbers is denoted by the symbol $\mathbb{C}$. We have a new number system, so let us explore its arithmetic.

We may think of a complex number $x + yi$ as a vector of dimension two (or perhaps as a fruit salad where we accept only two different types of fruits). Thus to add two complex numbers

$z_1 := a + bi$ and $z_2 := c + di$, where a, b, c, and d are real numbers, we use vector addition:

$$\begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} a + c \\ b + d \end{pmatrix}.$$

Hence the sum of our two complex numbers z_1, z_2 is given by the complex number $z_1 + z_2 := (a + c) + (b + d)i$. In fact, since, a complex number can also be represented as a matrix, it should be possible to write the above as

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} + \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} a + c & -[b + d] \\ b + d & a + c \end{pmatrix}.$$

To ensure this, **matrix addition** should be defined for matrix with matching dimensions as follows.

$$\begin{pmatrix} A_{11} & \cdots & A_{1n} \\ A_{21} & \cdots & A_{2n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix} + \begin{pmatrix} B_{11} & \cdots & B_{1n} \\ B_{21} & \cdots & B_{2n} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mn} \end{pmatrix} = \begin{pmatrix} A_{11} + B_{11} & \cdots & A_{1n} + B_{1n} \\ A_{21} + B_{21} & \cdots & A_{2n} + B_{2n} \\ \vdots & \ddots & \vdots \\ A_{m1} + B_{m1} & \cdots & A_{mn} + B_{mn} \end{pmatrix}$$

Ok, so we know how to add complex numbers. Subtracting a complex number from a complex number is just as simple: $z_1 - z_2 := (a - c) + (b - d)i$. How about multiplying two complex numbers? Here it will be useful to recall the definition of matrix multiplication. We will use the shortcut from Challenge 34: if $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B := \begin{pmatrix} w & x \\ y & z \end{pmatrix}$, then

$$AB = \begin{pmatrix} aw + by & ax + bz \\ cw + dy & cx + dz \end{pmatrix}.$$

Therefore, if $z_1 := a + bi$ and $z_2 := c + di$ are complex numbers, then their product $z_1 z_2$ can be represented in matrix form by

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} c & -d \\ d & c \end{pmatrix} = \begin{pmatrix} ac - bd & -bd - ad \\ ad + bc & -bd + ac \end{pmatrix} = \begin{pmatrix} ac - bd & -(ad + bc) \\ ad + bc & ac - bd \end{pmatrix}.$$

Therefore, the product of two complex numbers z_1 and z_2 is given by the complex number

$$z_1 z_2 := (ac - bd) + (ad + bc)i.$$

Addition of complex numbers was quite simple, but multiplication looks complicated. Yet there is some method to the madness. Let us turn off the pesky i term by setting $b = d = 0$ in our complex numbers $z_1 := a + bi$ and $z_2 := c + di$. Then addition of two complex numbers is $z_1 + z_2 = (a + b) + 0i$ and multiplication of two complex numbers is $z_1 z_2 = ac + 0i$. We have been able to recover the familiar addition and multiplication of real numbers! To amplify the fact that something familiar is still with us, we use the following definition.

If $z := x + yi$ is a complex number for real numbers x and y, then $\operatorname{Re} z := x$ is called the **real part** of z and $\operatorname{Im} z := y$ is called the **imaginary part** of z.

We can divide real numbers. Can we divide a complex number by another complex number?

Challenge 35 If A is a matrix with dimension $n \times n$, then matrix A is said to be **invertible** if there is a matrix B such that $AB = BA = 1$.[12] The matrix B is called the **inverse matrix** of A, and is denoted by the symbol A^{-1}.

(a) As a warmup, show that if B is a matrix inverse of A, then A is a matrix inverse of B. Use the fact that matrix multiplication is associative and $B = 1B = B1$ to show that matrix inverses are unique by supposing B and C are matrix inverses of A and concluding that $B = C$.[13]

(b) Let $A := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B := \begin{pmatrix} w & x \\ y & z \end{pmatrix}$ so that $AB = \begin{pmatrix} aw + by & ax + bz \\ cw + dy & cx + dz \end{pmatrix}$. For B to be an inverse of A, it is necessary (but not sufficient) that $AB = 1$, in particular:

$$aw + by = 1, \qquad ax + bz = 0, \qquad cw + dy = 0, \qquad cx + dz = 1.$$

The values of a, b, c, and d are constants and we wish to find the values of the real numbers w, x, y, and z so that the above holds. Find the values w, x, y, z.[14]

(c) Show that your answer from part (b) can be written as $w = d/(ad - bc)$, $x = -b/(ad - bc)$, $y = -c/(ad - bc)$, and $z = a/(ad - bc)$.

(d) Show that if $ad - bc \neq 0$, then $BA = 1$, where the entries of matrix B are as you found in part (b) or part (c). Conclude that the matrix A with dimension 2×2 has an inverse when $ad - bc \neq 0$ with

$$A^{-1} := \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

where $\det A := ad - bc$ is the **determinant** of a matrix A of dimension 2×2. If a determinant is nonzero, the matrix A is invertible. If $B = \begin{pmatrix} b \end{pmatrix}$ is a 1×1 matrix, then matrix B is invertible if it is not the zero matrix, and so $\det B := b$ and its inverse matrix is given by $B^{-1} := \begin{pmatrix} \frac{1}{b} \end{pmatrix}$.

(e) Show that if A and B are two 2×2 matrices then $\det(AB) = \det A \det B$.

(f) Let $z := x + yi$ be a complex number where x and y are real numbers. Define $1/z$ (the **multiplicative inverse** of z) to be the complex number such that $z \cdot (1/z) = (1/z) \cdot z = 1$. By part (a), this number is unique. Find a formula for $1/z$. When does a complex number z *not* have a multiplicative inverse?

(g) Verify that your answer from part (f) matches our intuition from real numbers by setting $y := 0$ and checking that it is the same as that of real numbers.

(h) Define the division of a complex number $z_1 := a + bi$ by another complex number $z_2 := c + di$ by the product $z_1 \cdot (1/z_2)$, whenever $(1/z_2)$ exists. What is $\mathrm{Re}(z_1/z_2)$ and $\mathrm{Im}(z_1/z_2)$? Check that it matches our intuition from real numbers by setting $b := 0$ and $d := 0$.

Challenge 35 shows that we can divide complex numbers by other nonzero complex numbers,[15] just like real numbers! In fact, we see that real numbers are a special case of complex numbers where the imaginary part is 0. A number system where we can do all the arithmetic operations (addition, subtraction, multiplication, division by nonzero numbers) as with real numbers, is called

[12] For the two matrix multiplications to work, we see that if B exists, it must have dimension $n \times n$. A matrix is called a **square matrix** it has the same number of rows and columns. We see that non square matrices do not have matrix inverses (there are however, *pseudo* inverses).

[13] *Hint:* $B = 1B = (CA)B$.

[14] *Hint:* the second equation tells us $x = -bz/a$. Plugging this into the fourth equation gives us a formula for z in terms of the constants a, b, c, and d. Then you also know the formula for x, and are halfway done!

[15] A nonzero complex number is a complex number with at least one nonzero real part or imaginary part.

a **field**. Because we can do all the arithmetic operations with complex numbers just as we do with the real numbers, the complex numbers with its arithmetic operations form a field called the **complex field**. The numbers in a field are called **scalars**, and since we upgrade our number system from real numbers to the complex numbers, by a scalar, we mean a complex number.

Further concepts

The sum of the squares of the real part and imaginary part of a complex number appeared numerous times in Challenge 35, and so it is useful to isolate this concept. For a complex number $z := x + yi$, the **absolute value** of z, written $|z|$, is defined as the number $\sqrt{x^2 + y^2}$.

Observe that because x and y are real numbers, the absolute value of a complex number is always a real number. Furthermore, if $y = 0$, then this matches our definition of an absolute value of a real number. In fact, the only complex number with absolute value 0 is the real number 0.

There is an alternative way of calculating the absolute value of a complex number z. The **complex conjugate** of a complex number $z := x + yi$, denoted by the symbol z^*, is the complex number $x - yi$. That is, the complex conjugate of a complex number z is the same number, with $\operatorname{Im} z$ switching signs. Using the formula for the products of complex numbers, we obtain:

$$\sqrt{zz^*} = \sqrt{(z^*)z} = |z|.$$

A real number x has no imaginary part, and so $x^* = x$.

By the definition of an absolute value for a complex number z, we have $|z|^2 = (\operatorname{Re} z)^2 + (\operatorname{Im} z)^2$. In particular, if $|z| = 1$, then we have the equation $(\operatorname{Re} z)^2 + (\operatorname{Im} z)^2 = 1$. This is an equation we have seen several times already! It is the equation of a unit circle.

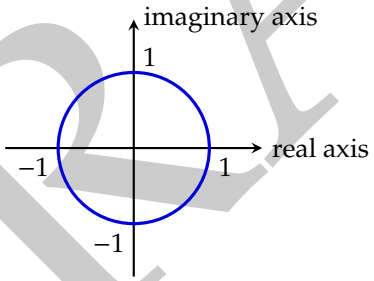


Figure 5.12: The set of complex numbers z with $|z| = 1$ form a unit circle (the blue circle).

A diagram of the plane, where the x-axis represents the values of the real part of a complex number, and the y-axis represents the values of the imaginary part of a complex number, is called an **Argand diagram**.[16] Figure 5.12 is an example of an Argand diagram.

From Figure 5.13, we see that geometrically the conjugation operation on a complex number is a reflection across the real axis. Recall that when we were creating the complex numbers, there were two matrices that squared to -1, the matrices $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. We thus had to make

[16]The identification of complex numbers as geometric objects (points on a plane) was apparently done first in 1799 by the mathematician Caspar Wessel.

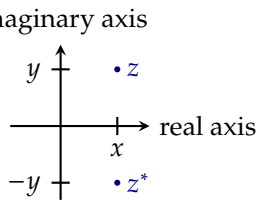


Figure 5.13: A complex number $z := x + yi$ and its complex conjugate $z^* = x - yi$.

a choice on which matrix to assign to the imaginary number i. Geometrically, the choice was on deciding which side of the imaginary axis is up and which is down. To see this, if we had chosen the second matrix as the imaginary number i, all our conventions would have the opposite sign in the imaginary axis of the Argand diagram. As we are comfortable with making such choices from calculus, we see that there was no loss in generality by making one choice over the other.

Challenge 36 Let u and w be complex numbers.

(a) Show that $(u^*)^* = u$, $(u + w)^* = u^* + w^*$, and $u^*w^* = (uw)^*$.
(b) Show that $\operatorname{Re} u = (u + u^*)/2$ and $\operatorname{Im} u = (u - u^*)/(2i)$.
(c) Show that $|u|^2 = uu^*$, $|u^*| = |u|$, $|uw| = |u||w|$, and $|\operatorname{Re} u| \leq |u|$ and $|\operatorname{Im} u| \leq |u|$.
(d) If w is nonzero, then show that $u/w = (uw^*)/|w|^2$.

Theorem 49 (Triangle Inequality)**.** If u and w are complex numbers, then $|u + w| \leq |u| + |w|$.

Proof. Since $2\operatorname{Re}(uw^*) = uw^* + (uw^*)^* = uw^* + u^*w$, and $2\operatorname{Re}(uw^*) \leq 2|uw^*|$, we have

$$|u + w|^2 = (u + w)(u^* + w^*) = uu^* + uw^* + wu^* + ww^* = |u|^2 + 2\operatorname{Re}(uw^*) + |w|^2$$
$$\leq |u|^2 + 2|uw^*| + |w|^2 = |u|^2 + 2|u||w| + |w|^2 = (|u| + |w|)^2 .$$

Since $|u+w|^2, (|u| + |w|)^2$ are nonnegative real numbers, we may take square roots on both sides. □

Challenge 37

(a) Identify the complex number $w := 3 - 4i$ on an Argand diagram and calculate $|w|$.
(b) Verify that the complex number $u := \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i$ satisfies $|u| = 1$.
(c) Calculate the product uw and use the fact that $\sqrt{2}$ is approximately 1.41 to place the complex number uw on the Argand diagram from part (a).[17] What is $|uw|$?
(d) Let $z_1 := a + bi$ and $z_2 := c + di$ be complex numbers, with $|z_1| = 1$. Show that $z_3 := z_1z_2$ satisfies $|z_3| = |z_2|$. Conclude that geometrically, multiplying a complex number z_2 by a complex number z_1 in a unit circle amounts to rotating the number z_2 on an Argand diagram.
(e) We need two numbers to identify a complex number z unambiguously: $\operatorname{Re} z$ and $\operatorname{Im} z$. Part (d) suggests an alternative way. Start with the real number $|z|$, and then rotate it to where z belongs. Show that for each complex number z, there is some complex number u_z with

[17] Here is a way to figure out that $\sqrt{2}$ is approximately 1.41. The number $\sqrt{2}$ is the length of the diagonal of a unit square, so it is a real number greater than 0. Define a real number $\epsilon > 0$, for example $\epsilon := 0.01$ and put $\alpha = 0$. Continue to increment the value of α by ϵ while $\alpha^2 < 2$. At some point, $\alpha^2 \geq 2$ and we will know that $\alpha - \epsilon < \sqrt{2} \leq \alpha$.

$|u_z| = 1$ such that $z = |z|u_z$. If $z = 0$, then the complex number u_z is not unique. For nonzero z, convince yourself that u_z is unique (this should be obvious geometrically).

We pause for a word on calculus. A function is said to be **complex valued** if its outputs are complex numbers. If a complex valued function's outputs are always real numbers, then the function is said to be **real valued**. If we have a complex valued function f that takes complex numbers as inputs, then f is differentiable at z if there is a number $f'(z)$ such that

$$f(z + \alpha) = f(z) + f'(z)\alpha + |\alpha| o(1)$$

where $|\alpha|$ is the absolute value of the complex number α that we drop to 0. How about integration? If we write a complex valued function f as the sum of its real and imaginary parts, then for real inputs, $f : t \mapsto \operatorname{Re} f(t) + i \operatorname{Im} f(t)$. Thus for a function f that maps real numbers in the interval $[a, b]$ to complex numbers, we have

$$\int_a^b f(t)\,dt := \int_a^b \operatorname{Re} f(t)\,dt + i \int_a^b \operatorname{Im} f(t)\,dt.$$

Complex matrices

Now that we know about complex numbers, we need no longer restrict ourselves to matrices whose entries are real numbers. A **complex matrix** is a matrix whose entries are complex numbers. A matrix whose entries are all real numbers may still be considered as a complex matrix, but we will refer to it as a **real matrix**. A complex number is an example of a real matrix.

We know how to add, subtract, and multiply two real or complex matrices (with compatible dimensions). Using matrix inverses, as discussed in Challenge 35, we could even speak of "dividing" a matrix by another. If a matrix B is invertible, then AB^{-1} is the analogue of dividing a matrix A by matrix B. In fact, we defined the division operator for complex numbers in this manner.

How about the complex conjugation operation? Let $z := x + yi$ be a complex number. It matrix representation is $\begin{pmatrix} x & -y \\ y & x \end{pmatrix}$. The complex conjugate of z is $z^* := x - yi$, whose matrix representation is $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$. The complex conjugate of z^* is z, with the matrix representation given by the first matrix. How can we transform the first matrix into the second matrix, and vice versa? It appears that we need to "flip" the matrix entries over its diagonal. We formalize this below.

Let A be an $m \times n$ matrix as shown below.

$$A := \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \tag{5.14}$$

The **transpose** of matrix A, denoted $A^\intercal$, is the matrix of dimension $n \times m$ defined by

$$A^\intercal := \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{m1} \\ A_{12} & A_{22} & \cdots & A_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2m} & \cdots & A_{mn} \end{pmatrix}.$$

For example, for the real matrix $z := \begin{pmatrix} x & -y \\ y & x \end{pmatrix}$, its transpose matrix is

$$z^\top = \begin{pmatrix} x & y \\ -y & x \end{pmatrix}.$$

So is the analogue of a complex conjugation for matrices the transpose of a matrix? Well, we have only been working with *real* matrices so far. We want to talk about the more general class of *complex* matrices. A complex matrix is a matrix where each entry is a real matrix of dimension 2×2. We need to take the transpose of each entry (complex conjugation) in addition to transposing the matrix. This is the *conjugate transpose* operation.

If A is a complex matrix with entries as given in Equation 5.14 above, then its **conjugate transpose** is the matrix $A^\dagger$ defined by

$$A^\dagger := \begin{pmatrix} A^*_{11} & A^*_{21} & \cdots & A^*_{m1} \\ A^*_{12} & A^*_{22} & \cdots & A^*_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ A^*_{1n} & A^*_{2m} & \cdots & A^*_{mn} \end{pmatrix}.$$

Recall that a complex number z is real if $z^* = z$. A complex matrix H is **Hermitian** if $H^\dagger = H$.[18] A Hermitian matrix is thus the complex matrix analogue of a real number.

The complex numbers located geometrically in the unit circle in an Argand diagram provided the role of rotation (Challenge 37). What is the complex matrix analogue? Recall that a complex number z is located geometrically in the unit circle if $zz^* = 1$. A complex matrix U is **unitary** if $UU^\dagger = U^\dagger U = 1$.[19] A unitary matrix is thus the complex matrix analogue of a complex number in the unit circle, and it rotates complex vectors (with compatible dimensions).

We can also take the exponential and logarithm of matrices as with real numbers. Recall from Chapter 4 that the Taylor series of e^x is given by $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots$ (This is a variant of the Taylor polynomial). We can use the Taylor series to define the **matrix exponential** of an $n \times n$ matrix X as the following.

$$e^X := 1 + X + \frac{1}{2}X^2 + \frac{1}{6}X^3 + \cdots$$

Similarly, we can use the Taylor series for $\log(1 + x)$ to define the **matrix logarithm** of an $n \times n$ matrix X as below.

$$\log X = \log(1 + [X - 1]) = (X - 1) - \frac{1}{2}(X - 1)^2 + \frac{1}{3}(X - 1)^3 - \frac{1}{4}(X - 1)^4 + \cdots.$$

5.4 Quantum Dynamics

The Schrödinger equation

We return to the topic of dynamics that we began this chapter with and see if we can gain new insights with what we have developed. The only mechanical system we know of is the

[18] We can deduce that H must be a square matrix.

[19] Observe that $U^\dagger$ is the matrix inverse of U. Hence a unitary matrix is always a square matrix with an inverse.

simple harmonic oscillator (which we examined at the beginning of this chapter). It is in some sense a system that is the perfect setting for the machinery we have developed, for we saw that an oscillator's motion in phase space is an ellipse. An ellipse with a suitable choice of units is a unit circle, so we will discard any unnecessary complexity and simplify even more to consider an oscillator whose motion in phase space is a unit circle.

The location of an oscillator exists at a point in time, regardless of whether we choose to catalogue it with some choice of units. Hence the "state" our oscillator is in is like a fruit salad $\left|f\right\rangle$, which exists in the physical world without us writing down its contents as a vector in some choice of units to quantify its ingredients. If someone pressed us for the ingredients, then we could present a representation of $\left|f\right\rangle$ as a vector f under some choice of units. We will denote the **abstract state** of our oscillator at time t by $|\Psi(t)\rangle$. If someone insists that we represent the location of our oscillator with some unit of measurement, we will represent $|\Psi\rangle$ as a complex vector Ψ. Notice we are using a *complex* vector. A real vector is an example of a complex vector, and once our eyes are open to the existence of complex numbers, complex vectors and complex matrices, there is little reason for us to insist on real numbers.

We want to create a mathematical model for the motion of our oscillator through time. A complex matrix is built to do just that, since a complex matrix exists to turn a complex vector into another complex vector. Since a unitary matrix is the analogue of a complex number in the unit circle, we will say that the state of pendulum at time t "evolves" into the state at time $t + \alpha$ with the following rule for some unitary matrix $U(\alpha)$.

$$|\Psi(t+\alpha)\rangle = U(\alpha)\,|\Psi(t)\rangle \tag{5.15}$$

At our initial state at time $t = 0$, we have

$$|\Psi(0)\rangle = U(0)\,|\Psi(0)\rangle$$

since there is no time evolution. Thus $U(0) = 1$. Now U is a representation of a function that takes in complex vectors and outputs complex vectors. One thing we want for the *motion* of our oscillator is that the motion should be continuous. So we will assume that U is continuous. By continuity, $U(\alpha) = U(0) + o(1)$.

Since we are dealing with a physical object, we may eventually want to do things like measure the oscillator's displacement away from the origin, and so on. Lengths are represented by real numbers, or complex numbers z such that $z^* = z$. We saw that the complex matrix analogue of this is a Hermitian matrix. Let us introduce a Hermitian matrix to the mix.

$$U(\alpha) = U(0) + o(1) = U(0) - \alpha H + o(\alpha).^{20}$$

Our decision to put a minus sign in front of H is by convention, and could easily be accounted for (or removed) by replacing H with $-H$.[21]

But there is a problem here, do you see it? We are thinking of $U(\alpha)$ as a complex number. $U(0) = 1$ and so it corresponds to a real number 1, and the Hermitian matrix H also corresponds to a real number. What we are saying is that the complex number $U(\alpha)$ is a sum of two real numbers

[20] The symbol $o(\alpha)$ is the same as $|\alpha|o(1)$ (Section 4.3).

[21] Notice that $-H$ is the matrix $(-1)H$, where -1 is a real number.

(plus a term negligible with respect to α which we can ignore). This assumption is unnecessarily restrictive. It would be much better to put an i in:

$$U(\alpha) = 1 - i\alpha H + o(\alpha).$$

We are simply doing the obvious: a complex number is being taken apart into a real part 1 and an imaginary part $-\alpha H$, where the minus sign is as convention dictates and can be removed if you wish by relabeling H with $-H$.

We plug this back into Equation 5.15 and use linearity to get

$$\begin{aligned} |\Psi(t+\alpha)\rangle = U(\alpha)|\Psi(t)\rangle &= (1 - i\alpha H + o(\alpha))|\Psi(t)\rangle \\ &= |\Psi(t)\rangle - i\alpha H|\Psi(t)\rangle + o(\alpha). \end{aligned}$$

Where have we seen this kind of expression before? In the definition of a derivative! Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}|\Psi(t)\rangle = -iH|\Psi(t)\rangle$$

What is our equation is telling us? It is telling us that the time evolution of the state of our oscillator is generated by applying a Hermitian matrix H. That's good, because H corresponds to a real number! But before we celebrate our victory, let us recall our earlier discussion from the beginning of this chapter that time translation is generated by energy. Since H is generating time evolution of our oscillator, H is actually the total energy of our oscillator with dimension of energy. Since we are working in phase space, we will call the Hermitian matrix H the **Hamiltonian**. Observe that our equations

$$U(\alpha) = 1 - i\alpha H + o(\alpha) \quad \text{and} \quad |\Psi(t+\alpha)\rangle = |\Psi(t)\rangle - i\alpha H|\Psi(t)\rangle + o(\alpha)$$

do not make sense dimensionally. To fix this we introduce a new dimensionful constant. Since α has dimension Time and H has dimension Energy, we will cancel them out by introducing a new constant $\hbar$, called the **reduced Planck constant**, with dimension Energy × Time. We go back and write $U(\alpha) = 1 - \frac{i}{\hbar}\alpha H + o(\alpha)$, from which we deduce $|\Psi(t+\alpha)\rangle = |\Psi(t)\rangle - \frac{iH}{\hbar}\alpha|\Psi(t)\rangle + o(\alpha)$, giving us the equation $\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = -\frac{i}{\hbar}H|\Psi\rangle$. This is the **Schrödinger equation**, usually written as the following.

$$\boxed{i\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = H|\Psi\rangle}$$

Time-independent Hamiltonian

We began with the relation $|\Psi(t)\rangle = U(t)|\Psi(0)\rangle$. If we plug in this relation to the Schrödinger equation $\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = -\frac{i}{\hbar}H|\Psi\rangle$, we have $\frac{\mathrm{d}}{\mathrm{d}t}U(t)|\Psi(0)\rangle = -\frac{i}{\hbar}HU(t)|\Psi(0)\rangle$. Now, we are free to choose the initial state $|\Psi(0)\rangle$, so let's put $|\Psi(0)\rangle = 1$ to get the following.

$$\frac{\mathrm{d}}{\mathrm{d}t}U(t) = -\frac{i}{\hbar}HU(t) \tag{5.16}$$

Recall that by conservation of energy, our harmonic oscillator had a constant total energy E, even though the kinetic and potential energies fluctuated. Similarly, we will assume that the Hamiltonian H is not a function of time, that is, H is a **time-independent Hamiltonian**.

In Equation 5.16 we see that $U'(t) = -\frac{i}{\hbar}HU(t)$, where H is a constant. Thus Equation 5.16 says that the derivative of U is itself times a constant. Where have we seen such a derivative before? We have seen such a derivative with the exponential function, where $(e^{cx})' = ce^x$ for constant c. We therefore see that if H is a constant, then $U(t) = e^{-iHt/\hbar}$.

The equation in 1 dimension

We have been dealing with abstract states $|\Psi\rangle$ thus far. How about if we wish to talk about the state's represention in a complex vector Ψ? This is the analogue of a "position" of our oscillator, and we insist that position functions in the physical world are differentiable so that we can calculate velocities. Thus we will assume that Ψ is differentiable to get

$$\Psi(x-\alpha) = \Psi(x) - \alpha\frac{\mathrm{d}}{\mathrm{d}x}\Psi(x) + o(\alpha) = \left(1 - \alpha\frac{\mathrm{d}}{\mathrm{d}x}\right)\Psi(x) + o(\alpha)$$

where the minus sign is once again simply conforming to our convention from before, and the second equality is due to linearity. What does this equation tell us? That to spatially translate our oscillator from location x to $x-\alpha$, we are applying the operation $\left(1-\alpha\frac{\mathrm{d}}{\mathrm{d}x}\right)$, modulo some terms negligible compared to α.[22] Thus the **translation operator** $T(\alpha)$ is given by

$$T(\alpha) = 1 - \alpha\frac{\mathrm{d}}{\mathrm{d}x}.$$

I don't know about you, but this doesn't have enough i's and $\hbar$'s for my taste. We know that the time evolution operator $U(\alpha)$ is given by $U(\alpha) = 1 - \frac{i}{\hbar}\alpha H$. To maintain consistency with the time evolution operator, we write

$$T(\alpha) = 1 - \frac{i}{\hbar}\alpha\left(-i\hbar\frac{\mathrm{d}}{\mathrm{d}x}\right).$$

Recall that translation is generated by momentum, and so just like H was an energy term, the term in the brackets is a momentum term. We call $P := -i\hbar\frac{\mathrm{d}}{\mathrm{d}x}$ the **momentum operator**, and in fact, as you should verify, it has the correct dimension of momentum![23]

Recall that the mechanical energy of a system is the sum of the kinetic energy $\frac{p^2}{2m}$ and potential energy V. Since $P^2 f = PPf = \left(-i\hbar\frac{\mathrm{d}}{\mathrm{d}x}\right)\left(-i\hbar\frac{\mathrm{d}}{\mathrm{d}x}\right)f = -\hbar^2\frac{\mathrm{d}^2}{\mathrm{d}x^2}f$, we have $H = -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + V$. Plugging this into the Schrödinger equation for the complex vector Ψ, we obtain the **one-dimensional Schrödinger equation** for a particle confined to a line with mass m shown below.

$$\boxed{i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V\Psi}$$

Since Ψ is a function of not only time t but also of space x, the derivative $\frac{\mathrm{d}}{\mathrm{d}t}$ has been replaced with a partial derivative $\frac{\partial}{\partial t}$. We also have a second partial derivative $\frac{\partial^2}{\partial x^2}$ from the momentum operator of the kinetic energy.

[22] We are doing a translation $x \mapsto x - \alpha$ because we are actually doing a *passive transformation*, translating our origin while keeping our oscillator fixed.

[23] All of this is in one dimension. In three dimensions, we will have $P_x := -i\hbar\partial_x$, $P_y := -i\hbar\partial_y$, and $P_z := -i\hbar\partial_z$.

Challenge 38 Consider a particle of mass m moving on a line located at position x_0 at initial time $t = 0$. Put

$$U(x,t) := \sqrt{\frac{m}{2\pi i\hbar t}} e^{im(x-x_0)^2/(2\hbar t)}.$$

(a) Show that

$$\frac{\partial}{\partial x}U = \frac{im(x-x_0)}{\hbar t}U.$$

(b) Show that

$$\frac{\partial^2}{\partial x^2}U = \frac{im}{\hbar t}U - \frac{m^2(x-x_0)^2}{\hbar^2 t^2}U.$$

(c) Show that

$$\frac{\partial}{\partial t}U = -\frac{1}{2t}U - \frac{im(x-x_0)^2}{2\hbar t^2}U.$$

(d) Conclude that

$$-i\hbar\frac{\partial}{\partial t}U = -\frac{\hbar}{2m}\frac{\partial^2}{\partial x^2}U$$

and thus U is a solution to the one-dimensional Schrödinger equation with $V = 0$.

The circle

The dynamics of an oscillator in phase space is that of an ellipse. Nevertheless, its motion in space is a mass simply moving back and forth. It seems proper that we look into an object whose motion is that of a circle. Now, things are made of atoms, and the most basic of those is that of a **hydrogen atom**, which consists of an electron orbiting a proton (see Figure 5.17). Not only is the hydrogen atom the simplest, it is also by far the most abundant type of an atom.

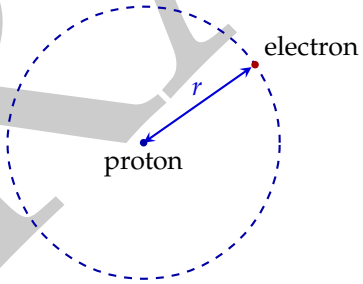


Figure 5.17: A hydrogen atom (diagram not to scale).

Both the electron and the proton are *charged particles* with a charge of $-1e$ and e, respectively, where e is the *elementary charge*. The electric force acting on each other due to the charge is described by *Coulomb's law*. Let us consider two particles with charge q_1 and q_2 that are distance r away of each other. First of all, particles of opposite charges attract and particles of like charges repel, with their attraction or repulsion proportional to the product of their charges: q_1q_2.

The force of attraction/repulsion falls off with distance, and the strength of the force is felt equally for all charged particle of the same charge on the same distance away from the source

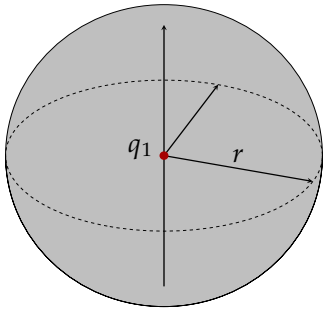


Figure 5.18: All charged particles of equal charge on the boundary of a sphere of radius r centered at a point charge q_1 feels the same electric force.

charge.[24] Thus all charged particles with charge q_2 in the boundary of a sphere of radius r (see Figure 5.18) are affected equally from the particle q_1 in the the origin. To calculate the drop off in strength as we increase the distance r, let us imagine the electric force from charge q_1 as it tries to reach infinitely far away. As the reach of the force r increases, the force must apply equally to all charges of the same charge that are equidistance from the source charge q_1. Thus as the force reaches distance r, the force is sweeping out a volume of a sphere of radius r. The drop off in force over distance r is the rate of change of the volume of the sphere: in other words, the derivative of the volume with respect to r.[25]

From Challenge 11, we know that the volume of a sphere of radius r is given by $\frac{4}{3}\pi r^3$. The rate of change of the volume of a sphere is then $\left(\frac{4}{3}\pi r^3\right)' = 4\pi r^2$ (this is actually the surface area of a sphere of radius r; due to the uniform rate of change of an area of a circle in all directions, the method gives the circumference of a circle of radius r as $(\pi r^2)' = 2\pi r$). Therefore, the force law is

$$F = \frac{q_1 q_2}{4\pi r^2 \epsilon_0} \tag{5.19}$$

where ϵ_0 is a dimensionful constant that allows us match the units in both sides of the equation. Equation 5.19 is called **Coulomb's law**.

Let us calculate the potential energy for the hydrogen atom, where $q_1 q_2 = -e \cdot e = -e^2$. We take the reference point to be infinitely far away from our proton, where the force due to our proton is zero. The potential energy V of the work needed to bring in an electron from infinitely far away to within distance r of a proton is:

$$V = \int_o^r -\left(\frac{-e^2}{4\pi\epsilon_0}\frac{1}{x^2}\right) dx = \frac{e^2}{4\pi\epsilon_0}\int_o^r \frac{1}{x^2}\, dx = -\frac{e^2}{4\pi\epsilon_0}\frac{1}{r} + 0 = -\frac{e^2}{4\pi\epsilon_0}\frac{1}{r}.$$

Now let us bring in some quantum mechanics. The one-dimensional Schrödinger equation

[24] It would be weird if there was some distinguished axis where the force was stronger or weaker.

[25] This argument works because we assume all particles of charge q_2 of the same distance r away from the source charge q_1 are affected equally. This argument will not work if the force felt depends on the direction of the charge away from q_1 because the rate of change will not be uniform.

contains a twice spatial derivative for the x-axis, $\frac{\partial^2}{\partial x^2}$:

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi + V\Psi.$$

The Schrödinger equation in three dimensions is given by the following, where the Laplacian of Ψ $\nabla^2\Psi = \frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2}$ takes the place of $\frac{\partial^2\Psi}{\partial x^2}$.[26]

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi$$

For the hydrogen atom, the mass term is now the mass of the electron m_e, and the potential energy $V = -\frac{e^2}{4\pi\epsilon_0}\frac{1}{r}$. Therefore, the Schrödinger equation for the hydrogen atom is

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m_e}\nabla^2\Psi - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r}\Psi.$$

The table below contain the dimensionful quantities of the equation, with the (most probable) radius of the hydrogen atom denoted by the symbol a_0. The letter L stands for the dimension Length, the letter M stands for the dimension Mass, and the letter T stands for the dimension Time.

	Variable	Dimension	Rough Value
Radius of hydrogen atom	a_0	L	10^{-10} m
Reduced Planck constant	$\hbar$	ML^2/T	??? J·s
Electron mass	m_e	M	10^{-30} kg
Coulomb term	$\frac{e^2}{4\pi\epsilon_0}$	ML^3/T^2	10^{-28} N·m^2

Challenge 39

(a) Show that it is not possible to find a combination of integers p and q such that the equation

$$a_0 = \beta\cdot(m_e)^p\left(\frac{e^2}{4\pi\epsilon_0}\right)^q$$

holds, where β is some dimensionless constant. Hence the constant $\hbar$ is indispensable.

(b) The simplest formula for expressing $\hbar$ using the other three variables is

$$\hbar = \gamma\cdot a_0^x\cdot(m_e)^y\cdot\left(\frac{e^2}{4\pi\epsilon_0}\right)^z \tag{5.20}$$

where γ is some dimensionless constant. Find integers x, y, and z that satisfy the formula.

(c) Very rough values of a_0, m_e, and the Coulomb term $\frac{e^2}{4\pi\epsilon_0}$ in the nearest powers of 10 are given in the table above.[27] For example, 5.29×10^{-11} m would be rounded up to 10^{-10} m. Ignoring the dimensionless constant γ for now, use the values to give a rough estimate of $\hbar$ in powers

[26] Since energy is a scalar quantity, we cannot replace $\frac{\partial^2\Psi}{\partial x^2}$ with the vector $\left(\frac{\partial^2\Psi}{\partial x^2} \quad \frac{\partial^2\Psi}{\partial y^2} \quad \frac{\partial^2\Psi}{\partial z^2}\right)^t$.

[27] IRecall that 1 N (newton) is defined as 1 kg·m·s^{-2} and 1 J (joule) is defined as 1 kg·m^2·s^{-2}.

of 10.[28] The value of $\hbar$ is so tiny for us that we can ignore the dimensionless constant γ. The actual value of $\hbar$ is about 1.05457×10^{-34} J·s. Max Planck first proposed the constant $h := 2\pi\hbar$ and calculated its value in 1901.

Since $\hbar$ is so tiny, the Schrödinger equation appears to describe a world that is imperceptible. Can this equation have any relevance to us?

Waves and superposition

Around the time of the invention of calculus, there was a controversy over the nature of light. Christiaan Huygens argued that light was a wave, while Newton argued that light must be a particle. Although Newton initially had the upper hand, Thomas Young's experiments in 1801 seemed to settle the question in favor of Huygens. Subsequently, there was a great deal of effort to try and bridge the wave nature of light with that of ordinary particle dynamics. A key result of such investigations is one of the crowning jewels of mathematical physics of the 19th Century: the **Hamilton–Jacobi equation**

$$-\frac{\partial S}{\partial t} = H\left(x, p := \frac{\partial S}{\partial x}, t\right). \tag{5.21}$$

The function H is the Hamiltonian of the system, with momentum defined by $p := \frac{\partial S}{\partial x}$. It is equivalent to Newton's second law, but derived using the machinery of infinite dimensional calculus.

Can our own investigations lead to any illumination on this issue? Let us first investigate what we can about wave phenomena. As with most physical phenomena, we will need a differential equation to describe waves. This equation, which we will call the *wave equation*, will model how a wave changes over time.

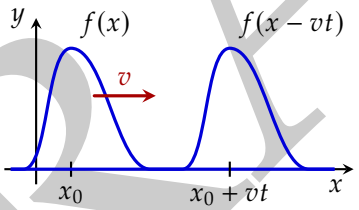


Figure 5.22: A wave traveling at speed v to the right.

Imagine a wave which we represent by a function f that is traveling to the right at some speed v (see Figure 5.22). To simplify matters, we will assume an idealized situation in which the wave does not widen or drop over time. We could imagine a water wave, and the number $f(x_0, t_0)$ will tell us how much water is elevated in the x-coordinate x_0 at time t_0. Let us denote the initial wave at time $t = 0$ by the function g, that is, $g : x \mapsto f(x, 0)$. After time t, the wave will have travelled to the right by distance vt. Thus all the numbers $g(x)$ will have shifted to the right by vt. Suppose an object is standing still, but we have shifted all the x-coordinates to the left. Then the object will have shifted to the right! Similarly, we can shift the number $g(x)$ to the right by a *substitution* shifting all the x-coordinates to the left: $x \mapsto x - vt$. Therefore, after time t,

$$f(x, t) = g(x - vt).$$

[28] We will return to this this later in Challenge 46.

For a wave moving to the *left*, the same reasoning gives $f(x,t) = g(x+vt)$. Now, if we throw a pebble to a pool of water and take a cross section, waves are traveling not only to the left, but also to the right at the same time. So our wave equation must satisfy both cases. In fact, if we imagine the pebble thrown into a pool of water and take a cross section, we not only see two waves dispersing away, but there are multiple of different sizes at the same time! Therefore, our wave equation must allow not just the sum of the functions $g(x-vt)$ and $g(x+vt)$, but each of the linear combination:

$$a \cdot g(x-vt) + b \cdot g(x+vt).$$

This looks like a tall order, can we do it? First, because we want a differential equation that describes the dynamics of the wave over time, the equation will involve some time derivative of f. This causes a problem, because the chain rule tells us that the time derivative of $f(x,t) := g(x-vt)$ and the time derivative of $f(x,t) := g(x+vt)$ will differ by a minus sign. But we need both cases to be solutions! Thus one time derivative will not be sufficient: in order to make both functions work as solutions to our wave equation, we must take two time derivatives of f.

Let us crank out the time derivatives. Since we know that the twice time derivatives of $g(x-vt)$ and $g(x+vt)$ will equal, we will only do it for the former. By the chain rule,

$$\frac{\partial f}{\partial t} = -vg'(x-vt), \qquad \frac{\partial^2 f}{\partial t^2} = v^2 g''(x-vt).$$

We see that there is a twice spatial derivative involved. Now,

$$\frac{\partial f}{\partial x} = g'(x-vt), \qquad \frac{\partial^2 f}{\partial x^2} = g''(x-vt).$$

Therefore, the one-dimensional **wave equation** for a wave with speed v is given by the following.

$$\boxed{\frac{\partial^2 f}{\partial t^2} = v^2 \frac{\partial^2 f}{\partial x^2}} \tag{5.23}$$

This is a **linear differential equation** because linear combinations of solutions to the wave equation are also solutions.[29]

Now let us examine the one-dimensional Schrödinger equation. To simplify, let us consider a **free particle**, which is a particle with no forces acting on it. Then $V = 0$ and so the equation is simply

$$-\frac{i}{\hbar}\frac{\partial \Psi}{\partial t} = \frac{1}{2m}\frac{\partial^2 \Psi}{\partial x^2}.$$[30]

This doesn't really look like a wave equation because we are missing a derivative with respect to time. But check this out, remember the Schrödinger equation for abstract states $|\Psi\rangle$? It was

$$\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = -\frac{i}{\hbar}H|\Psi\rangle. \tag{5.24}$$

[29] I encourage you to verify this by using the derivative rules to check that $h : t \mapsto ag(x-vt) + bg(x+vt)$ satisfies the wave equation $\frac{\partial^2 h}{\partial t^2} = v^2 \frac{\partial^2 h}{\partial x^2}$. This should be simple, for (partial) derivatives are linear!

[30] We have divided both sides by the nonzero constant $\hbar^2$ and multiplied both sides by -1.

We see that the term $-\frac{i}{\hbar}$ is like a time derivative! So we could consider the Schrödinger equation for the free particle to be a wave equation. Because of this connection, the object Ψ is called a **wavefunction**. In fact, just like the wave equation, the general Schrödinger equation 5.24 is a linear differential equation. Indeed, derivatives are linear and (Hermitian) matrices are linear, so for two states $|\Psi_1\rangle$, $|\Psi_2\rangle$, and scalars a, b:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(a\,|\Psi_1\rangle + b\,|\Psi_2\rangle\Big) = a\frac{\mathrm{d}}{\mathrm{d}t}|\Psi_1\rangle + b\frac{\mathrm{d}}{\mathrm{d}t}|\Psi_2\rangle = -a\frac{i}{\hbar}H|\Psi_1\rangle - b\frac{i}{\hbar}H|\Psi_2\rangle = -\frac{i}{\hbar}H\Big(a\,|\Psi_1\rangle + b\,|\Psi_2\rangle\Big)$$

which verifies that linear combinations of solutions to the Schrödinger equation are also solutions. Linear combinations are also called **superpositions**. Linear equations like the wave equation and the Schrödinger equation are said to obey the **superposition principle**.

We started this section by trying to upgrade the mathematical apparatus for describing a particle (a simple oscillator) and got an equation that has so much in common with waves! The distinction between particles and waves are so blurred, it is no wonder that scientists were debating about whether light was a wave or a particle.

Challenge 40 Recall that the wavefunction Ψ is a complex vector. Let us take the special case where Ψ is a complex valued function of position x and time t (like a wave, but complex). Put $\Psi := \rho e^{i\omega/\hbar}$ where ρ is a real valued function of x and t that determines the scaling and ω is some real valued function of x and t that determines the rotation.

(a) Use the product rule to show that

$$\frac{\partial\Psi}{\partial t} = \left(\dot\rho + i\frac{\dot\omega}{\hbar}\rho\right)e^{i\omega/\hbar}, \qquad \frac{\partial\Psi}{\partial x} = \left(\rho' + i\frac{\omega'}{\hbar}\rho\right)e^{i\omega/\hbar}.$$

(b) Apply the product rule on $\frac{\partial\Psi}{\partial x}$ once more to show that

$$\frac{\partial^2\Psi}{\partial x^2} = \left(\rho'' + 2i\frac{\rho'\omega'}{\hbar} + i\frac{\omega''}{\hbar}\rho - \frac{(\omega')^2}{\hbar^2}\rho\right)e^{i\omega(x,t)/\hbar}.$$

(c) The one-dimensional Schrödinger equation states that $i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} + V\Psi$. Plug in your answers from part (a) and part (b) into the one-dimensional Schrödinger equation, divide both sides by $e^{i\omega/\hbar}$ and do all the multiplication by $i\hbar$ (on the left side) and multiplication by $\frac{\hbar^2}{2m}$ (on the right side) to obtain the equation

$$i\hbar\dot\rho - \rho\dot\omega = -\frac{\hbar^2}{2m}\rho'' - i\frac{\hbar}{m}\rho'\omega' - i\frac{\hbar}{2m}\omega''\rho + \frac{1}{2m}(\omega')^2\rho + V\rho. \tag{5.25}$$

(d) Equation 5.25 from part (c) is far too complicated to reason with and it looks nothing like the Schrödinger equation it is supposed to be! But notice how all the $\hbar$'s in the bottom of the fractions (denominators) have magically disappeared. Since $\hbar$ is so tiny, we might as well as drop it. Take $\hbar \to 0$ to obtain a much simpler equation and then divide through by R to get the following.[31]

$$-\frac{\partial\omega}{\partial t} = \frac{1}{2m}\left(\frac{\partial\omega}{\partial x}\right)^2 + V(x) \tag{5.26}$$

[31] What does it mean to drop a constant to 0? Suppose we were measuring length by the height of a building h and we took $h \to 0$. That means we are scaling up everything much larger than the building, while taking the length of our building and everything of roughly the same size or smaller to be negligible. Thus by taking $\hbar \to 0$, we are taking the constant $\hbar$ to be negligible.

(e) The right side of Equation 5.26 from part (d) is a Hamiltonian (total energy) with momentum $p := \frac{\partial \omega}{\partial x}$. Conclude that

$$-\frac{\partial \omega}{\partial t} = H\left(x, p := \frac{\partial \omega}{\partial x}, t\right). \tag{5.27}$$

Have we seen Equation 5.27 before? It is simply the Hamilton–Jacobi equation (Equation 5.21)! We see that classical mechanics is a special case of this new theory in the limit $\hbar \to 0$. Thus in situations of scale $\hbar$, we need to use Schrödinger's equation, but in situations involving scales where $\hbar$ is negligible, then we can use classical mechanics. Recall that the value of $\hbar$ is about 1.05457×10^{-34} J·s, a negligible amount indeed! Such a value can be considered practically zero in our dally lives.

This is a theory of an extremely tiny world, a world where our classical intuition in trying to distinguish between waves and particles are doomed to a failure. This is the realm of **quantum mechanics**. Nevertheless, this theory of tiny particles is used everywhere. Everyone carries around in their hands or their pockets a proof that the Schrödinger equation works.

5.5 Trigonometry

In the previous section, we saw that the dynamics of a quantum particle was governed by a unitary matrix U, which rotates complex vectors. We now consider the simplest case of a time-independent hamiltonian with $H := \hbar$ so that $U = e^{-it}$. We take $x := -t$ and examine the function e^{ix} which rotates complex numbers.

Radians

From Challenge 37, we know that we can obtain any complex number in the unit circle by rotating the complex number 1. We begin by rotating the complex number 1 into the number e^{ix} on the unit circle, with the rotation starting from the complex number 1 counterclockwise.

Recall that the circumference of a circle has length 2π. Each angle of rotation x corresponds uniquely to a point on the unit circle, which in turn corresponds uniquely to a length in the interval $[0, 2\pi)$ (see Figure 5.28). It is therefore natural to measure angles of rotation as lengths such that a full 360 degree rotation is defined to be 2π. This way of measuring angles is called **radians**. A right angle in radians is thus $\pi/2$ as we need four right angles to cover the circumference of a circle. Similarly, the angle corresponding to a semicircle is π since we need two of them to cover the circumference of a circle. Using radians to measure rotation, $e^{2\pi i}$ is the complex number 1 (a rotation of 360 degrees), $e^{\pi i}$ is the complex number -1 (a rotation of 180 degrees), and $e^{i\pi/2}$ is the complex number i.

By convention a rotation by a negative angle $-x$ is a rotation by angle x clockwise from the complex number 1 (see Figure 5.29). As this is geometrically a reflection across the real axis, it is algebraically equivalent to the complex conjugation operation. Therefore, $(e^{ix})^* = e^{-ix}$.

Trigonometric functions

We saw that the complex number e^{ix} on the unit circle is fully specified by the angle x (measured in radians). We also know that each complex number z is fully specified by the two real numbers $\operatorname{Re} z$ and $\operatorname{Im} z$. (Shown in Figure 5.30.)

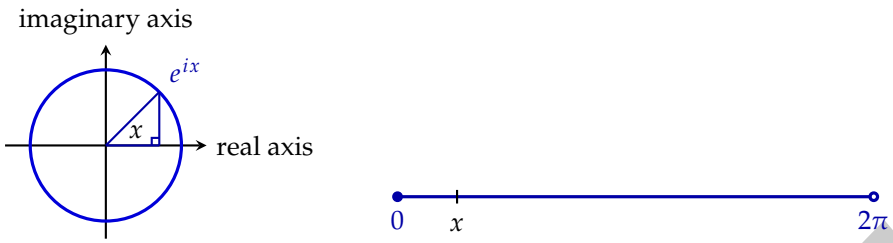


Figure 5.28: An angle of rotation x uniquely corresponds to a length on the interval $[0, 2\pi)$.

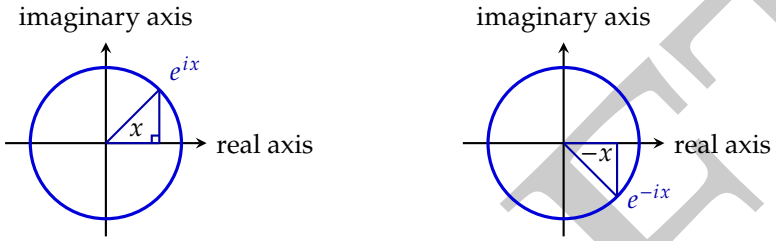


Figure 5.29: Argand diagram of e^{ix} (left) and e^{-ix} (right).

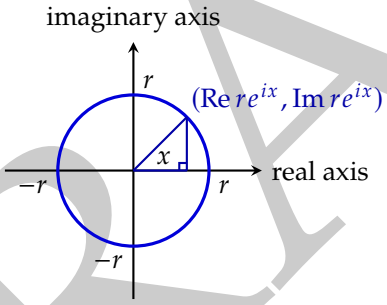


Figure 5.30: A complex number re^{ix} can be specified by $\operatorname{Re} re^{ix}$ and $\operatorname{Im} re^{ix}$ or angle x.

By Challenge 36, $\operatorname{Re} z = \frac{1}{2}(z + z^*)$ and $\operatorname{Im} z = \frac{1}{2i}(z - z^*)$ and so

$$\operatorname{Re} e^{ix} = \frac{e^{ix} + (e^{ix})^*}{2} \qquad \operatorname{Im} e^{ix} = \frac{e^{ix} - (e^{ix})^*}{2i}.$$

As the number e^{ix} is also identifiable by the angle x, it will be convenient to define the above using just x. Since $(e^{ix})^* = e^{-ix}$, we arrive at the following definition.

Definition 50. The **cosine** function cos and the **sine** function sin are defined by

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} \qquad \sin x = \frac{e^{ix} - e^{-ix}}{2i}.$$

The **tangent** function tan is defined by $\tan : x \mapsto \sin x / \cos x$.

Challenge 41

(a) Verify that cosine is an even function and sine is an odd function.

(b) Calculate the values of the cosine, sine, and tangent functions for the angles $0, \pi/6, \pi/4, \pi/3$, and $\pi/2$. To find the values for the angles $\pi/6$ and $\pi/3$, apply the Pythagorean theorem to the equilateral triangle in Figure 5.31 (an **equilateral triangle** is a triangle whose sides are all equal; each of its three angles are $\pi/3$).

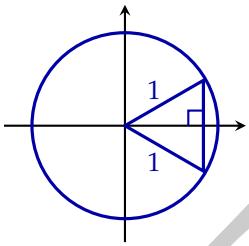


Figure 5.31: A bisected equilateral triangle of side length 1 inscribed in a unit circle.

By definition the following equation, called **Euler's formula**, holds.

$$e^{ix} = \cos x + i \sin x$$

Rotating the number 1 by angle π gives the number -1. This is **Euler's identity**

$$e^{i\pi} = -1.$$

The Pythagorean theorem tells us that $(\operatorname{Re} e^{ix})^2 + (\operatorname{Im} e^{ix})^2 = 1$, which is equivalent to

$$\cos^2 x + \sin^2 x = 1$$

where $\cos^2 x := (\cos x)^2$ and $\sin^2 x := (\sin x)^2$.

Challenge 42

(a) Check that the following trigonometric derivatives hold.

$$(\cos x)' = -\sin x \qquad (\sin x)' = \cos x \qquad (\tan x)' = \frac{1}{\cos^2 x}$$

(b) Use the Taylor series for e^{ix} to obtain the following Taylor series for cos and sin.[32]

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

Challenge 43 Use Euler's formula to check that following identities hold.

$$\cos\left(x + \frac{\pi}{2}\right) = -\sin x \qquad \sin\left(x + \frac{\pi}{2}\right) = \cos x$$
$$\cos\left(x - \frac{\pi}{2}\right) = \sin x \qquad \sin\left(x - \frac{\pi}{2}\right) = -\cos x$$
$$\cos\left(\frac{\pi}{2} - x\right) = \sin x \qquad \sin\left(\frac{\pi}{2} - x\right) = \cos x$$

[32] This part uses material from Chapter 4.

[*Hint*: since $e^{i\pi/2} = i$, we know that $e^{i(x+\pi/2)} = i\cos x - \sin x$.]

A rotation of a nonzero complex number z by angle $(x_1 + x_2)$ is achieved by taking the product $z \cdot e^{i(x_1+x_2)}$. The same rotation can be achieved by rotating it first by angle x_1 with $z \cdot e^{ix_1}$ and then rotating the result by angle x_2 by taking a second product $(z \cdot e^{ix_1}) \cdot e^{ix_2}$. Therefore,

$$e^{i(x_1+x_2)} = e^{ix_1}e^{ix_2}. \tag{5.32}$$

By Euler's formula,

$$\begin{aligned} e^{i\theta_1}e^{i\theta_2} &= (\cos\theta_1 + i\sin\theta_1)(\cos\theta_2 + i\sin\theta_2) \\ &= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2 + i(\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2). \end{aligned}$$

Since the above must equal $e^{i(x_1+x_2)}$, we see that

$$\cos(x_1 + x_2) = \cos x_1\cos x_2 - \sin x_1\sin x_2 \qquad \sin(x_1 + x_2) = \cos x_1\sin x_2 + \sin x_1\cos x_2.$$

These are the **trigonometric addition formulas**.

Challenge 44

(a) Obtain the trigonometric addition formula for cosines by taking the derivative of both sides of the addition formula for sines with respect to x_1. Obtain the addition formula for sines from the addition formula for cosines.

(b) Use the fact that the cosine function is even and the sine function is odd to find the formulas for $\cos(x_1 - x_2)$ and $\sin(x_1 - x_2)$. Each formula can be combined with the corresponding addition formula and written as follows.

$$\cos(x_1 \pm x_2) = \cos x_1\cos x_2 \mp \sin x_1\sin x_2 \qquad \sin(x_1 \pm x_2) = \cos x_1\sin x_2 \pm \sin x_1\cos x_2$$

(c) Deduce the **double angle formulas**:

$$\cos(2x) = \cos^2 x - \sin^2 x \qquad \sin(2x) = 2\sin x\cos x. \tag{5.33}$$

(d) Use the double angle formula for cosine with the Pythagorean theorem to show that

$$\cos^2 x = \frac{1 + \cos(2x)}{2} \qquad \sin^2 x = \frac{1 - \cos(2x)}{2}.$$

(e) Use part (b) to obtain the following formulas:

$$\begin{aligned} \cos x_1\cos x_2 &= \frac{1}{2}\left[\cos(x_1 + x_2) + \cos(x_1 - x_2)\right], \\ \sin x_1\sin x_2 &= \frac{1}{2}\left[-\cos(x_1 + x_2) + \cos(x_1 - x_2)\right], \\ \cos x_1\sin x_2 &= \frac{1}{2}\left[\sin(x_1 + x_2) + \sin(x_1 - x_2)\right]. \end{aligned}$$

(f) Applying the well-ordering principle to Equation 5.32 gives us the fact that for each natural number n, $(e^{ix})^n = e^{inx}$. Use it to obtain **de Moivre's formula**:

$$(\cos x + i\sin x)^n = \cos(nx) + i\sin(nx).$$

We obtain a useful relation between an angle of a triangle with the three side lengths. Suppose we have a triangle whose side lengths are a, b, and c. Furthermore, suppose we know the angle x (in radians) between the sides of length a and b (see left diagram in Figure 5.34). Place the axis such that the side of length a form the x-axis with the vertex with angle x as the origin and such tht the entirety of the triangle lies above the x-axis (see middle diagram in Figure 5.34). By construction, one of the vertices will be the complex number a and the other will be the complex number be^{ix}.

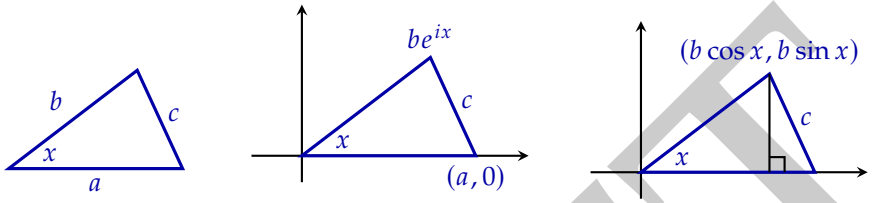


Figure 5.34: Cosine Law

The Pythagorean theorem gives the following.

$$\begin{aligned} c^2 &= (a - b\cos x)^2 + (b\sin x)^2 = a^2 - 2ab\cos x + b^2\cos^2 x + b^2\sin^2 x \\ &= a^2 - 2ab\cos x + b^2(\cos^2 x + \sin^2 x) = a^2 + b^2 - 2ab\cos x \end{aligned}$$

The formula $c^2 = a^2 + b^2 - 2ab\cos x$ is called the **cosine law**. Observe that if $x = \pi/2$ (a right angle), then we recover the Pythagorean theorem.

To wrap off our discussion of trigonometric functions, we calculate some useful integrals. First, we calculate the integral $\int x\cos(ax)\,dx$. This is an integral of a product, so we use integration by parts: $\int fg' = fg - \int f'g$ with $f(x) = x$ and $g'(x) = \cos(ax)$. Then $g(x) = \frac{1}{a}\sin(ax)$ and we have

$$\begin{aligned} \int x\cos(ax)\,dx &= x\left(\frac{1}{a}\sin(ax)\right) - \int \frac{1}{a}\sin(ax)\,dx = \frac{1}{a}\int -\sin(ax)\,dx + \frac{x}{a}\sin(ax) \\ &= \frac{1}{a}\left(\frac{1}{a}\cos(ax)\right) + \frac{x}{a}\sin(ax) + c = \frac{1}{a^2}\cos(ax) + \frac{x}{a}\sin(ax) + c. \end{aligned}$$

Similarly, we can calculate the integral $\int x\sin(ax)\,dx$ using integration by parts with $f(x) = x$ and $g'(x) = \sin(ax)$ to get

$$\begin{aligned} \int x\sin(ax)\,dx &= x\left(-\frac{1}{a}\cos(ax)\right) - \int -\frac{1}{a}\cos(ax)\,dx = \frac{1}{a}\int \cos(ax)\,dx - \frac{x}{a}\cos(ax) \\ &= \frac{1}{a}\left(\frac{1}{a}\sin(ax)\right) - \frac{x}{a}\cos(ax) + c = \frac{1}{a^2}\sin(ax) - \frac{x}{a}\cos(ax) + c. \end{aligned}$$

Challenge 45 Let a and b be real numbers such that $a^2 \neq b^2$. Show that

(a)
$$\int \cos(ax)\cos(bx)\,dx = \frac{\sin([a+b]x)}{2(a+b)} + \frac{\sin([a-b]x)}{2(a-b)} + c,$$

[*Hint:* $\cos(ax) = (e^{iax} + e^{-iax})/2$.]

(b)
$$\int \sin(ax)\sin(bx)\,dx = -\frac{\sin([a+b]x)}{2(a+b)} + \frac{\sin([a-b]x)}{2(a-b)} + c,$$

(c)
$$\int \cos(ax)\sin(bx)\,dx = -\frac{\cos([a+b]x)}{2(a+b)} - \frac{\cos([a-b]x)}{2(a-b)} + c.$$

(d) Conclude that the following hold for distinct positive integers m and n:

$$\int_{-\pi}^{\pi} \cos(mx)\cos(nx)\,dx = 0 \quad \int_{-\pi}^{\pi} \sin(mx)\sin(nx)\,dx = 0 \quad \int_{-\pi}^{\pi} \cos(mx)\sin(nx)\,dx = 0.$$

[*Hint:* the cosine function is an even function and the sine function is odd.]

(e) Justify each step of the following calculation.

$$\int \cos(ax)\sin(ax)\,dx = \int \frac{\sin(2ax)}{2}\,dx = -\frac{\cos(2ax)}{4a} + c = \frac{2\sin^2(ax)-1}{4a} + c = \frac{\sin^2(ax)}{2a} + c$$

Show that

$$\int \cos^2(ax)\,dx = \frac{x}{2} + \frac{\sin(ax)\cos(ax)}{2a} + c$$
$$\int \sin^2(ax)\,dx = \frac{x}{2} - \frac{\sin(ax)\cos(ax)}{2a} + c$$

and conclude that the following hold for each positive integer n.

$$\int_{-\pi}^{\pi} \cos^2(nx)\,dx = \pi \qquad \int_{-\pi}^{\pi} \sin^2(nx)\,dx = \pi \qquad \int_{-\pi}^{\pi} \cos(nx)\sin(nx)\,dx = 0$$

(f) Let m and n be integers. Show that if $m = -n$ then $\int_{-\pi}^{\pi} e^{imx}e^{inx}\,dx = 2\pi$. Furthermore, show that if $m \neq -n$ then $\int_{-\pi}^{\pi} e^{imx}e^{inx}\,dx = 0$.

Challenge 46 Let us imagine the hydrogen atom as an electron rotating counterclockwise in a *constant* speed around a proton (see Figure 5.35).

(a) Let r be the radius of the electron's circular motion and let T be the time needed for the electron to do one full rotation. Define the **angular velocity** of the electron by $\omega := 2\pi/T$. What does the constant ω represent?

(b) Let us assume that at time $t = 0$ the electron was located at the real number r on the Argand diagram. The position function p of the electron over time t is then $p : t \mapsto re^{i\omega t}$. Calculate the (complex-valued) velocity function $v := p'$ and the (complex-valued) acceleration function $a := v'$. The *speed* of the electron is then $|v|$. Verify that $|v| = r\omega$ and $|a| = |v|^2/r$. Notice that even though the electron has constant speed, $|a|$ is nonzero!

(c) Use Coulomb's law and Newton's second law $|F| = m|a|$ to conclude that $\frac{e^2}{4\pi\epsilon_0 r^2} = m\frac{|v|^2}{r}$.

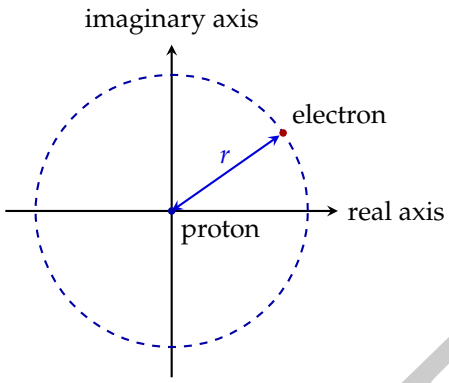


Figure 5.35: A hydrogen atom on an Argand diagram with the origin fixed at the proton's position.

(d) The *angular momentum* from the origin of a point particle rotating in a circle of radius r at constant speed $|v|$ is given by $L = m|v|r$.[33] Notice that the constant $\hbar$ has dimensions of angular momentum. We have not yet given a definition of $\hbar$. By dimensional considerations, the constant can represent an angular momentum of something. In the quantum context, the simplest possible something would be an electron rotating around a proton. We thus define $\hbar$ as the angular momentum of the electron in our model of the hydrogen atom. Conclude that $r = \frac{4\pi\epsilon_0\hbar^2}{e^2 m}$. If we switch the variable names with $r \mapsto a_0$ and $m \mapsto m_e$ we obtain our guess from Challenge 39. In particular, the dimensionless constant is 1.

Fourier series

Recall that the one-dimensional Schrödinger equation for a free particle is

$$-i\frac{2m}{\hbar}\frac{\partial\Psi}{\partial t} = \frac{\partial^2\Psi}{\partial x^2}.$$

Let us solve a simpler version of this equation. Replace Ψ with a real valued function u that takes as input position $x \in [0,1]$ and time $t \geq 0$. Ignoring all the constants, the imaginary number i, and the negative sign, the partial differential equation takes the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

which is a *heat equation*,[34] where u is the temperature at position x at time t. We can imagine this equation describes the temperature of a thin metal rod.

Let us assume that $u(0,t) = u(1,t) = 0$ and solve the heat equation for u. We begin by finding *one* simple solution. Suppose u is a product of two function X and T, which are functions of position x and position t, respectively. That is, we will try a solution of the form $u(x,t) = X(x)T(t)$.[35]

[33] We will derive this quantity later (Challenge **??**).
[34] We will obtain the heat equation at the end Section 6.6.
[35] This technique is called **separation of variables**.

Plugging this guess into the heat equation gives the equation $X(x)T'(t) = X''(x)T(t)$, which is equivalent to the equation

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)}.$$

Observe that the right side of the equation is a function of position x only, while the left side is a function of t only. Since both sides must equal, we see that neither side can be a function of x or t. This means that both sides must equal a constant which we will call c.

We will first solve the equation $X''(x)/X(x) = c$, which is the same as the equation $X''(x) = cX(x)$. We will assume $c \neq 0$.[36] What is a function whose twice derivative is equal to itself times a constant? We know of the exponential function, sines and cosines. Since the former is the sum of the latter two, a general solution to the equation has the form $X(x) = e^{\lambda x}$ for some constant λ. Then $X''(x) = \lambda^2 X(x)$ and so $\lambda = \pm\sqrt{c}$. Recall that the linear combination of solutions of a linear differential equation is itself a solution. Since $X''(x) = cX(x)$ is a linear differential equation, we see that

$$X(x) = Ae^{\sqrt{c}x} + Be^{-\sqrt{c}x}$$

for some constants A and B solves the equation $X''(x) = cX(x)$. We have two conditions that our solution must satisfy: (i) $X(0) = 0$ and (ii) $X(1) = 0$. The first condition gives $0 = X(0) = A + B$ and so $B = -A$. The second condition gives $0 = X(1) = A\left(e^{\sqrt{c}} - e^{-\sqrt{c}}\right)$. Since $e^{\sqrt{c}} = e^{-\sqrt{c}}$, we see that $e^{2\sqrt{c}} = 1$. The equation $e^{2\sqrt{c}} = 1$ is satisfied when $2\sqrt{c} = 2n\pi i$ for each positive integer n. Hence $c = -n^2\pi^2$. In our search for one solution X, we have found infinitely many solutions, each of which we will label X_n, with $X_n(x) = A_n\left(e^{n\pi ix} - e^{-n\pi ix}\right)$. Notice that the term inside the brackets is a sine term. We may thus write

$$X_n(x) = 2iA_n \sin(n\pi x).$$

We now turn to the equation $T'(t)/T(t) = c = -n^2\pi^2$, which is equivalent to the equation $T'(t) = -n^2\pi^2 T(t)$. The solution to the equation is $T(t) = Ce^{-n^2\pi^2 t}$ for some constant C. Once again, we have infinitely many solutions, and we will label each by the symbol T_n:

$$T_n(t) = C_n e^{-n^2\pi^2 t}.$$

Therefore, a solution to our heat equation is

$$u(x,t) = X_n(t)T_n(t) = a_n \sin(n\pi x)e^{-n^2\pi^2 t}$$

where $a_n := 2iA_nC_n$. Since the heat equation is a linear equation, sums of the solutions is also a solution to the heat equation. Therefore,

$$a_1 \sin(\pi x)e^{-1^2\pi^2 t} + a_2 \sin(2\pi x)e^{-2^2\pi^2 t} + a_3 \sin(3\pi x)e^{-3^2\pi^2 t} + a_4 \sin(4\pi x)e^{-4^2\pi^2 t} + \cdots$$

is also a solution to our heat equation. We will write this solution as

$$u(x,t) = \sum_{n=1}^{\infty} a_n \sin(n\pi x)e^{-n^2\pi^2 t}.$$

[36] If $c = 0$, then integrating twice gives $X(x) = Ax + B$ for some constants A and B. In such a case the conditions $X(0) = X(1) = 0$ forces both X and u to be the zero function, which is not a useful solution.

We still need to find the constants a_n for each positive integer n. An easy calculation (the same as Challenge 45) shows that for each positive integer m,

$$\int_0^1 \sin(n\pi x)\sin(m\pi x)\,dx = \frac{1}{2}\delta_{nm}$$

where $\delta_{nm} := 1$ when $n = m$ and $\delta_{nm} := 0$ if $n \neq m$ (this function is called the **Kronecker delta**). Then

$$\int_0^1 u(x,0)\sin(m\pi x)\,dx = \int_0^1 \left[\sum_{n=1}^{\infty} a_n \sin(n\pi x)\right]\sin(m\pi x)\,dx.$$

What next? How about we interchange the sum and the integral (this trick was first used by Fourier) to get

$$\int_0^1 u(x,0)\sin(m\pi x)\,dx = \sum_{n=1}^{\infty} a_n \int_0^1 \sin(n\pi x)\sin(m\pi x)\,dx.$$

With this trick, the inner integral is trivial. Only one term in the sum survives:

$$\int_0^1 u(x,0)\sin(m\pi x)\,dx = \sum_{n=1}^{\infty} a_n \frac{1}{2}\delta_{nm} = \frac{a_m}{2}.$$

Therefore, each constant a_n is given by $a_n = 2\int_0^1 u(x,0)\sin(n\pi x)\,dx$ and so

$$u(x,t) = \sum_{n=1}^{\infty}\left[2\int_0^1 u(x,0)\sin(n\pi x)\,dx\right]\sin(n\pi x)e^{-n^2\pi^2 t}$$

solves our heat equation. Our solution agrees with our intuition that the heat distribution at time t depends on the initial heat distribution at time $t = 0$ and that as $t \to \infty$, $u \to 0$ (since the terms $e^{-n^2\pi^2 t} \to 0$ for each positive integer n).

Fourier's work on heat initiated an explosion of innovations in mathematics, sciences, and engineering. We mention just two examples from mathematics. Georg Cantor's initiation of set theory, including his establishment of different kinds of infinities began here. The Riemann integral and Lebesgue's integration theory also sprung out from here (under what circumstances does Fourier's trick of interchanging integrals and summation work?).

5.6 Introduction to Groups

Let us return to the rotation of complex numbers in a unit circle. We can visualize this as a wheel we are free to rotate about. Suppose we wanted to implement this wheel in software, what would we need?

(A1) First, we will need to make sure that the rotation operation is implement properly such that if a user rotates the wheel, it still remains a wheel (if we rotate a complex number on the unit circle, we get a complex number on the unit circle.)

(A2) Second, a user can choose to not rotate the wheel, and leave it as is (we can leave a complex number on the unit circle as is by rotating the complex number by 0, that is multiplying it by $e^{0i} = 1$).

(A3) Third, the user should be able to undo any rotation (rotation of a complex number on the unit circle by x is undone by rotation by $-x$).
(A4) Fourth, our program should be able to resolve rotations unambiguously when the user specifies multiple rotation inputs (if z is a complex number on the unit circle, then $ze^{ix}e^{iy} := (ze^{ix})e^{iy}$ must equal $z(e^{ix}e^{iy})$).

Observe that a complex number on the unit circle z that we are rotating is itself describing a rotation! Indeed, each complex number on the unit circle takes the form e^{ix} for some angle x. The set of rotations along with the rotation operation form what we call a *group*.

Definition 51. A **group** is a set G with an operation $\star$ such that the following hold.

(A1) (Closure) The $\star$ operation takes two elements in G and outputs another element in G. That is, for each pair $a, b \in G$, we guarantee that $a \star b$ is an element of G.
(A2) (Identity) There is an element $e \in G$ such that for each $a \in G$, we have $e \star a = a \star e = a$. An element that satisfies this is called an **identity** of G.
(A3) (Inverse) For each $a \in G$, there is a $b \in G$ such that $a \star b = b \star a = e$. We call b an **inverse** of a.
(A4) (Associativity) For each $a, b, c \in G$, we have $(a \star b) \star c = a \star (b \star c)$.

If G is a group with operation $\star$, we will refer to it by $(G, \star)$ or simply by G. The following is a basic consequence of the conditions for a group.

Proposition 52. Group $(G, \star)$ has a unique identity and each $a \in G$ has a unique inverse.

Proof. Let e and e' be elements of G such that for each $a \in G$ we have $e \star a = a \star e = a$ and for each $b \in G$ we have $e' \star b = b \star e' = b$. Put $b := e$ and $a := e'$ to get

$$e = e \star e' = e'.$$

Let $a \in G$ and let $b, c \in G$ satisfy $a \star b = b \star a = e$ and $a \star c = c \star a = e$. Then

$$b = b \star e = b \star (a \star c) = (b \star a) \star c = e \star c = c.$$

□

Since the identity of a group is unique, there is no problem in denoting the identity of G by the symbol e or by e_G. Similarly, as the inverse of an element of a group is unique, there is no problem in denoting the inverse of element $\square$ by the symbol $\square^{-1}$.

Challenge 47 For each element a, b in group $(G, \star)$ show that $(a^{-1})^{-1} = a$ and $(a \star b)^{-1} = b^{-1} \star a^{-1}$.

In general, the order in which we rotate an object (in 3D say) matters. However, the order in which we rotate a complex number on the unit circle does not matter (that is, $ze^{ix} = e^{ix}z$). Such special groups are said to be *abelian*.

Definition 53. Group $(G, \star)$ is **abelian** if for each $a, b \in G$ we have $a \star b = b \star a$.

The real numbers under addition represent spatial translation of an object in a line (an *active transformation*), or equivalently a rearrangement of the origin (a *passive transformation*) and so $\mathbb{R}$ with the addition operation is a group. In fact, it is an abelian group as $a + b = b + a$ for each real numbers a and b.

The set of complex numbers on the unit circle under the (complex) multiplication operation is called the $U(1)$ group. But we know that each complex number e^{ix} on the unit circle can be represented as a matrix:

$$e^{ix} = \cos x + i \sin x = \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix}.$$

The set of matrices of the form above under the matrix multiplication operation is called the $SO(2)$ group. For all intents and purposes, the groups $U(1)$ and $SO(2)$ are the "same".

What do we mean that the groups $U(1)$ and $SO(2)$ are the "same"? It means that for each complex number in $U(1)$, there is a unique corresponding matrix in $SO(2)$ such that the operations in each group are respected.

Definition 54. A **domain** of a function f is the set of inputs to f. The **codomain** of a function f is the set that the outputs of f belong. Thus complex valued functions have the set $\mathbb{C}$ as codomain while real valued functions have the set $\mathbb{R}$ as codomain. If function f has domain A and codomain B, then we write $f : A \to B$.

Definition 55. A function f is **injective** if each $x \neq y$ implies $f(x) \neq f(y)$. A function $f : A \to B$ is **surjective** if for each $y \in B$ there is a $x \in A$ such that $f(x) = y$. A function f is **bijective** if it is both injective and surjective.

For example, the exponential function $\exp : \mathbb{R} \to \mathbb{R}$ is injective because it is strictly increasing, but it is not surjective as there are no inputs which map to negative real numbers (but $\exp : \mathbb{R} \to (0, \infty)$ is bijective). On the other hand, the logarithm function $\log : (0, \infty) \to \mathbb{R}$ is bijective. This is what allowed us to consider an *inverse* function of the logarithm function (the exponential function).

Definition 56. Groups $(G, \star)$ and $(H, *)$ are **isomorphic** if there is a map $f : G \to H$ that satisfies the following two conditions.

(a) Function f is bijective.
(b) For each $a, b \in G$, $f(a \star b) = f(a) * f(b)$.

A function that satisfies the above is called an **isomorphism**. If G and H are isomorphic, then we denote this by the symbol $G \cong H$.

Thus $U(1) \cong SO(2)$ with the isomorphism

$$f : e^{ix} \mapsto \begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix}.$$

We also know that $\mathbb{R}$ under the addition operation is isomorphic with the interval $(0, \infty)$ under the multiplication operation, for there is an isomorphism exp. Indeed, $\exp(x + y) = \exp(x) \cdot \exp(y)$.

Challenge 48 For a group $(G, \star)$ define the **opposite group** $(G^{\text{op}}, *)$ to have the same underlying set G with operation $*$ defined by $g * g' =: g' \star g$ for each $g, g' \in G$. Show that $(G^{\text{op}}, *)$ is a group and show that $G \cong G^{\text{op}}$.

It is true that each complex number on the unit circle e^{ix} is fully specified by a real number x, but $U(1)$ is *not* isomorphic with $\mathbb{R}$ under the addition operation. The reason is that the specification is not unique: $e^{0i} = e^{2\pi i} = e^{-2\pi i}$ and so on. This means that no isomorphism can exist between the set of complex numbers in $U(1)$ and $\mathbb{R}$. It will be useful to relax the condition on bijectivity.

Definition 57. Let $(G, \star)$ and $(H, *)$ be groups. A function f is a **homomorphism** if for each $a, b \in G$ we have $f(a \star b) = f(a) * f(b)$. If $f : G \to H$ is a homomorphism, then the **kernel** of f, written $\ker f$ is the set of elements in G that map to the identity of H.

Thus the map $f : x \mapsto e^{ix}$ is a homomorphism for the group $\mathbb{R}$ under the addition operation and the group $U(1)$. The set $\ker f$ are the integer products of 2π. We write $\ker f = \{2\pi k : k \in \mathbb{Z}\}$, where the symbol k was chosen arbitrarily.

Proposition 58. Let $(G, \star)$ be a group with identity e_G and let $(H, *)$ be a group with identity e_H. Let $f : G \to H$ be a homomorphism and let $a \in G$. Then $f(e_G) = e_H$ and $f(a^{-1}) = f(a)^{-1}$.

Proof. Let $a \in G$ be arbitrary. Thent $f(e_G) = f(e_G \star a) = f(e_G) * f(a)$ and $f(e_G) = f(a \star e_G) = f(a) * f(e_G)$. Therefore $f(e_G) = e_H$. Since $f(a^{-1}) * f(a) = f(a^{-1} \star a) = f(e_G) = e_H$ and $f(a) * f(a^{-1}) = f(a \star a^{-1}) = f(e_G) = e_H$, we see that $f(a^{-1}) = f(a)^{-1}$. □

Challenge 49

(a) Show that if a, b are elements of group $(G, \star)$ with $a \star b = 1$ then $b = a^{-1}$.
(b) Show that the set of nonzero real numbers with the multiplication operation form a group. This group is denoted by the symbol $\mathbb{R}^\times$.
(c) Show that the set of invertible $n \times n$ real matrices with matrix multiplication form a group. This group is called $GL_n(\mathbb{R})$, the **general linear group** over $\mathbb{R}$. [*Hint:* for each $A, B \in GL_n(\mathbb{R})$ show that $(AB)^{-1} = B^{-1}A^{-1}$.]
(d) We know that if A, B are 2×2 matrices, then $\det(AB) = \det A \det B$. Thus the determinant function provides a homomorphism between the groups $GL_2(\mathbb{R})$ and $\mathbb{R}^\times$. Show that the determinant function is *not* an isomorphism. The kernel of $\det : GL_2(\mathbb{R}) \to \mathbb{R}^\times$ with the matrix multiplication operation is called $SL_2(\mathbb{R})$, a **special linear group**.

Group action

The concept of a group is incredibly fundamental. Our first example of a group was the group $U(1)$. This is the group of rotations on the unit circle. The rotations are important because they rotate complex numbers. Groups are so fundamental because they are the natural type of objects that *act* on things. For example, the elements of group $U(1)$ rotate complex numbers. We say that the group $U(1)$ *acts* on the set $\mathbb{C}$ (of complex numbers).

Definition 59. Let $(G, \star)$ be a group with identity e and let S be a set. A **(left) group action** α is a map that takes an element of G and an element of S and returns an element of S (as a shorthand, we write $\alpha : G \times S \to S$) with the following properties.

(a) For each $s \in S$, we have $\alpha(e, s) = s$. For example, rotating a complex number z by applying the zero rotation e^{0i} does nothing.
(b) For each $a, b \in G$ and for each $s \in S$, we have $\alpha(a, \alpha(b, s)) = \alpha(a \star b, s)$. For example, rotating a complex number z by e^{ix} and then e^{iy} has the same effect as rotating z by $e^{i(x+y)}$.

If the above hold, we say that "(group) G **acts** on (set) S".

We are going to upgrade from rotating circles to playing around with triangles. Consider an equilateral triangle centered at the origin with vertex 1 at $e^{i\pi/2}$, vertex 2 at $e^{i7\pi/6}$, and vertex 3 at $e^{i11\pi/6}$ as shown in Figure 5.36. We will consider the transformation on this triangle such that the

triangles shape remains the same, with the only change being the placement of the vertices. For example, rotating the triangle by angles of integer multiples $2\pi/3$ leaves the shape intact, but any other angles are not allowed.

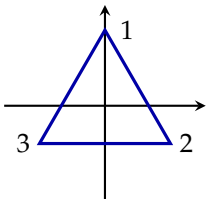


Figure 5.36: An equilateral triangle centered at the origin with vertices 1, 2, and 3.

Rotating the triangle by the angles 0, $2\pi/3$, and $4\pi/3$ gives the results as shown in Figure 5.37. Observe that rotation by the angle $-2\pi/3$ is the same as rotation by angle $4\pi/3$. Similarly, rotation by angle $6\pi/3$ is the same as rotation by angle 0. Therefore, the below three are the only possible unique rotations.

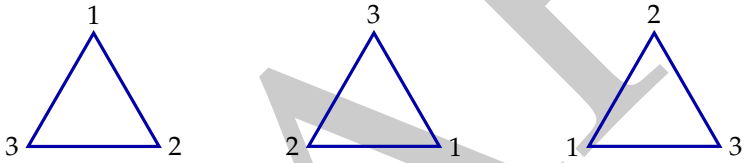


Figure 5.37: The equilateral triangle from Figure 5.36 rotated by angles $0, 2\pi/3$, and $4\pi/3$.

Another way we can play around with our triangle is to turn the triangle upside down by flipping the triangle around one of the edges. For example, if we flip the original triangle from Figure 5.36 about the bottom edge (edge $\overline{23}$), then we get the leftmost triangle in Figure 5.38.

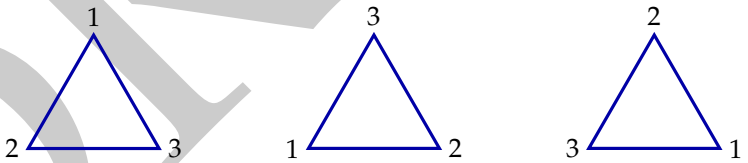


Figure 5.38: The equilateral triangle from Figure 5.36 flipped on edge $\overline{23}$, edge $\overline{13}$, and edge $\overline{12}$.

The group of transformations on the vertices of an equilateral triangle is called the **dihedral group** D_6. Equivalently, we can forget about the triangle and consider a group, called the **symmetric group** S_3, that switches around (or permutes) the set $\{1, 2, 3\}$. Thus the elements of S_3 are the bijective functions $f : \{1, 2, 3\} \to \{1, 2, 3\}$. A bijective function that maps a set to itself is called a **permutation**.

Let us list out the six permutations of the set $S = \{1, 2, 3\}$. Each permutation σ is completely specified by how σ acts on the three elements, that is: $\sigma : 1 \mapsto \sigma(1)$, $\sigma : 2 \mapsto \sigma(2)$, and $\sigma : 3 \mapsto \sigma(3)$.

We will write down these three pieces of information in a table, as shown below, where the inputs to σ are on the top row and the corresponding outputs of σ are at the bottom row.

$$\begin{pmatrix} 1 & 2 & 3 \\ \sigma(1) & \sigma(2) & \sigma(3) \end{pmatrix}$$

Permutations ρ_1, ρ_2, ρ_3 that correspond to rotating our triangle by angles 0, $\frac{2\pi}{3}$, and $\frac{4\pi}{3}$ respectively are:

$$\rho_1 := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} \qquad \rho_2 := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \qquad \rho_3 := \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

Permutations τ_1, τ_2, τ_3 that correspond to the flipping of our triangle shown in Figure 5.38 in that order are:

$$\tau_1 := \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \qquad \tau_2 := \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \qquad \tau_3 := \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

These correspondences define an isomorphism from D_6 to S_3.

Challenge 50 Check that $\rho_1 = \tau_k \circ \tau_k$ for $k \in \{1, 2, 3\}$. Find $i, j \in \{1, 2, 3\}$ such that $\rho_2 = \tau_i \circ \tau_j$. Find $m, n \in \{1, 2, 3\}$ such that $\rho_3 = \tau_m \circ \tau_n$. Find $f, g \in S_3$ such that $f \circ g \neq g \circ f$ and conclude that S_3 is **non-abelian**.

The permutations τ_1, τ_2, and τ_3 are called **transpositions** because they simply swap a pair of elements in $\{1, 2, 3\}$. Transpositions are the most basic type of permutations in because each permutation is a function composition of transpositions (Challenge 50).

Challenge 51 Suppose we encoded our equilateral triangle with vertices 1, 2, and 3 as the vector $v := \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$. Under this encoding a rotation by $2\pi/3$ results in a triangle of the vector $\begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$. For each permutation in S_3, find a 3×3 matrix that implements the transformation when acting on the vector v.[37] These matrices are called **permutation matrices**. Check that each permutation matrix is invertible.

Definition 60. An n-dimensional **matrix representation** ρ of a group G is a homomorphism $\rho : G \to GL_n(\mathbb{R})$.

Thus Challenge 51 gives a 3 dimensional matrix representation of S_3. This representation is not unique, and we give another representation of S_3. We placed our equilateral triangle in an Argand diagram in Figure 5.36 so that each vertices are complex numbers. Interpreting each vertex as a vector of dimension 2 gives the following 2-dimensional matrix representation of S_3.

$$\rho_1 \mapsto \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \rho_2 \mapsto \begin{pmatrix} \cos\frac{2\pi}{3} & -\sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{pmatrix} \qquad \rho_3 \mapsto \begin{pmatrix} \cos\frac{4\pi}{3} & -\sin\frac{4\pi}{3} \\ \sin\frac{4\pi}{3} & \cos\frac{4\pi}{3} \end{pmatrix}$$

$$\tau_1 \mapsto \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \tau_2 \mapsto \begin{pmatrix} -\cos\frac{2\pi}{3} & \sin\frac{2\pi}{3} \\ \sin\frac{2\pi}{3} & \cos\frac{2\pi}{3} \end{pmatrix} \qquad \tau_3 \mapsto \begin{pmatrix} -\cos\frac{4\pi}{3} & \sin\frac{4\pi}{3} \\ \sin\frac{4\pi}{3} & \cos\frac{4\pi}{3} \end{pmatrix}$$

[37] *Hint:* $Mv = M(1e_1 + 2e_2 + 3e_3) = 1(Me_1) + 2(Me_2) + 3(Me_3)$.

A natural question to ask is whether our matrix representations form a group. They had better! The six matrices above form a subset of $GL_2(\mathbb{R})$. Do the six matrices above with the matrix multiplication operation form a group?

Definition 61. If $(G, \star)$ is a group, then a nonempty subset H of G is a **subgroup** if H under the operation $\star$ is a group. If H is a subgroup of G we write $H \leq G$.

To check if a nonempty subset H is a (sub)group under the $\star$ operation, we need to check the four conditions in Definition 51 of a group. However, associativity (condition C4) is guaranteed because each element of H is an element of G. In fact, if closure of operation $\star$ (condition C1) is satisfied, then the condition on the existence of inverses (condition C3) guarantees that the identity e of G is an identity element of H (condition C2). This establishes the following simple criteria for checking whether a subset is a subgroup.

Proposition 62. If G under the operation $\star$ is a group, then a nonempty subset H of G is a subgroup if (1) closure of $\star$ is satisfied: for each $a, b \in H$ we have $a \star b \in H$; and (2) each element of H has an inverse: for each $a \in H$ we have $a^{-1} \in H$.

Challenge 52 Let G be a group under operation $\star$ and let H be a subset of G.

(a) Show that if $H \leq G$, then (1) H is nonempty and (2) for each $a, b \in H$ we have $a \star b^{-1} \in H$.
(b) Suppose that (1) H is nonempty and (2) for each $a, b \in H$ we have $a \star b^{-1} \in H$. Show that $H \leq G$. [*Hint:* first show that H contains the identity e of G; put $a := e$ on (2) to show that H contains inverses. Use the fact that $(b^{-1})^{-1} = b$ to show closure.]

Challenge 53 Let $(G, \star)$ and $(H, *)$ be groups and let $f : G \to H$ be a homomorphism.

(a) Show that $\ker f \leq H$, where the former is to be interpreted as a group with the $*$ operation.
(b) Show that $f(G) \leq H$, where the former is a group with the $\star$ operation. The group $f(G)$ is called the **image** of homomorphism f. We also denote the image of f using the notation $\mathrm{im} f$.

This result confirms that a matrix representation forms a group under the matrix multiplication operation.

Our action on S_3 is *atomic* in the sense that we cannot rotate halfway and stop; if we are rotating, we must complete the rotation operation. This is best seen in our 2-dimensional matrix representations of S_3 where the rotation matrices are of discrete angles: 0, $2\pi/3$, and $4\pi/3$.

However, we know that actions can also be continuous. Indeed, the first action that we saw was a continuous rotation of numbers on the complex plane. Let g be the rotation matrix $g(t) := \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$. We will consider a *passive* rotation (rotation of the axis) and thus $g(-t)$. As this rotation is continuous, we can take the derivative. Since $(e^{-it})' = -ie^{-it}$, we know that

$$g'(-t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} g(-t).$$

By the definition of the derivative at $t = 0$,

$$g(\alpha) = g(0) + g'(0)\alpha + o(\alpha).$$

Write $\alpha := \begin{pmatrix} x & y \end{pmatrix}^t$ and we see that

$$g'(0)\alpha = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix}.$$

Thus the map $g'(0)$ gives $x \mapsto y$ and $y \mapsto -x$. But there is another way to obtain the same map. Indeed, $y\partial_1 : x \mapsto y$ and $-x\partial_2 : y \mapsto -x$. Therefore

$$\left[y\partial_1 - x\partial_2\right]\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix}.$$

If we put in a factor of $i\hbar$ to make this "quantum", we may define $L := i\hbar(y\partial_1 - x\partial_2)$. Recall that the momentum operator P_x is defined by $P_x := -i\hbar\partial_1$.[38] Similarly, we may define the momentum operator $P_y := -i\hbar\partial_2$. Therefore L is a momentum on *rotation*, an **angular momentum** operator.

We can make a jump to three dimensions by adding a z-axis. In the above, the invisible z-axis was fixed (ignored) and no motion occurred along the z-axis. So we may interpret $L = i\hbar(y\partial_1 - x\partial_2)$ as a rotation along the z-axis and define $L_z := i\hbar(y\partial_1 - x\partial_2)$.

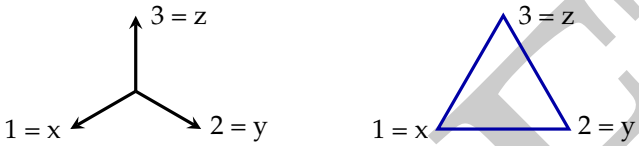


Figure 5.39: The three axis in 3 dimension made to form a triangle.

How about rotation along the x-axis (axis 1) and the y-axis (axis 2)? We may represent each axis as vertices in a triangle as shown in Figure 5.39. Rotating the triangle by $2\pi/3$ gives $1 = x \mapsto 2 = y \mapsto 3 = z \mapsto 1 = x$. Therefore, rotation along the x-axis should be described by the operator

$$L_x := i\hbar(z\partial_2 - y\partial_3).$$

Similarly, rotating the triangle by $4\pi/3$ gives $1 = x \mapsto 3 = z \mapsto 2 = y \mapsto 1 = x$. Therefore, rotation along the y-axis should be described by the operator

$$L_y := i\hbar(x\partial_3 - z\partial_1).$$

Before we close our discussion of groups, we note that a group is associated with one operation. However, we know of $\mathbb{R}$ (and also $\mathbb{C}$) which have two associated operations, addition and multiplication.[39] Now $(\mathbb{R}, +)$ is a group, but $(\mathbb{R}, \times)$ is *not* a group. In particular, the number 0 does not have an inverse (this is what we mean by: we cannot divide by zero). However, if we define $\mathbb{R}\backslash\{0\}$ to be the set of nonzero elements of $\mathbb{R}$, then $(\mathbb{R}\backslash\{0\}, \times)$ *is* an abelian group. To package the two groups $(\mathbb{R}, +)$ and $(\mathbb{R}\backslash\{0\}, \times)$ as one object, we will need to link the addition and multiplication operation together. In order to achieve this, we will require that the product of addition works as we expect: for each real numbers a, b, c, we have $a \times (b + c) = (a \times b) + (a \times c)$. These conditions are precisely what we mean by a field.

First, here are some preliminary definitions. A **binary function** is a function that takes in two inputs. A **binary operation** defined on set S is a binary function that takes in two elements of S and outputs an element of S. If S is a set with element a, then $S\backslash\{a\}$ is the collection of elements of S that are not a.

[38] We have upgraded from motion in one dimension to motion in two dimensions and thus have a momentum along the x-axis P_x and a momentum along the y-axis P_y.

[39] Subtraction by a is addition by the inverse of a, and division by b is multiplication by the inverse of b.

Definition 63. A **field** is a set F with binary operation + (addition) and binary operation $\times$ (multiplication) such that:

(a) $(F, +)$ is an abelian group with the identity element (called **additive identity**) denoted 0,
(b) $(F\backslash\{0\}, \times)$ is an abelian group with the identity element (called **multiplicative identity**) 1,[40]
(c) for each $a, b, c \in F$ the **distributive law** $a \times (b + c) = (a \times b) + (a \times c)$ holds.

In addition to $\mathbb{R}$, we have also seen the field of rational numbers $\mathbb{Q}$ and the complex field $\mathbb{C}$.

How about the real vector space $\mathbb{R}^n$ with vector addition and scalar multiplication? Notice that $\mathbb{R}^n$ cannot be a field as there is no division of vectors. To accommodate, let us try considering the set of vectors and the scalars separately. First, we have an abelian group $(\mathbb{R}^n, +)$ where + is the vector addition operation. Next, we have a field $\mathbb{R}$, with which we wish to "scale" the vectors in $\mathbb{R}^n$. Such a "scaling" of vectors is accomplished by the group action of scalar multiplication $a \cdot v := (av_1, av_2, \ldots, av_n)^\intercal$ by the group $(\mathbb{R}\backslash\{0\}, \times)$ on the set $\mathbb{R}^n$. Finally, we need to ensure that vector addition and scalar multiplication are both compatible with each other. To do this we need to ensure that (a) scalar multiplication of vector addition works as we would expect and (b) scalar multiplication by field addition works as we would expect. We now spell out these conditions.

Definition 64. A **vector space** over a field F (with additive identity 0 and multiplicative identity 1) is a set V with a binary operation + (**vector addition**) defined on V and a binary function $\cdot$ (**scalar multiplication**) defined on the pair F and V whose output is an element of V, such that:

(a) $(V, +)$ is an abelian group with identity denoted 0_V (to distinguish from 0),
(b) $1 \cdot v = v$ for each $v \in V$,
(c) for each $a, b \in F$ and $v \in V$ we have $a \cdot (b \cdot v) = (a \times b) \cdot v$,
(d) for each $a, b \in F$ and $v, w \in V$ we have $a \cdot (u + v) = a \cdot u + a \cdot v$ and $(a + b) \cdot v = a \cdot v + b \cdot v$.[41]

An element of V is called a **vector** and an element of F is called a **scalar**. For $a \in F$ and $v \in V$, we write av instead of $a \cdot v$.

Of course, $\mathbb{R}^n$ and $\mathbb{C}^n$ (the set of real vectors or complex vectors of dimension n under the vector addition operation and the scalar multiplication operation) are both vector spaces.

[40] Observe that 0 and 1 are necessarily distinct. Thus a field must have at least two elements

[41] The last condition gives $0 \cdot v = 0_V$ for each $v \in V$. Indeed, $0 \cdot v = (0 + 0) \cdot v = 0 \cdot v + 0 \cdot v$, and so $0 \cdot v = 0_V$.

6

Multivariables

6.1 Gaussian Integrals

We are back to working with real numbers. The goal of this Appendix is to show that you could readily generalize what we have discovered so far and apply to *multivariable* functions. We begin by calculating several integrals that are crucial in quantum theory. They center around the most important integral: $\int_{-\infty}^{\infty} e^{-ax^2}\,dx$, where a is a positive real number. First, let us try and see what the answer should look like. Let us assign the dimension Length to input x. Because an input to the exponential function must be dimensionless,[1] the constant a will have to take dimension Length^{-2}. Recall that the derivative of f has the dimension of f divided by the dimension of the input x. The inverse operation of integration will thus take the dimension of f and multiply by the dimension of the input x. Therefore, the dimension of $\int_{-\infty}^{\infty} e^{-ax^2}\,dx$ is Length. The only dimensionful quantity we have is a (of dimension Length^{-2}) and to form a dimension of Length, the simplest solution is $\frac{c}{\sqrt{a}}$, for some constant c. It turns out that the dimensionless constant c is $\sqrt{\pi}$. Therefore,

$$\int_{-\infty}^{\infty} e^{-ax^2}\,dx = \sqrt{\frac{\pi}{a}}. \tag{6.1}$$

In this Appendix, we will show that c is $\sqrt{\pi}$.

Before we do this, let us extend Equation 6.1 to the integral $\int_{-\infty}^{\infty} e^{-ax^2+bx}\,dx$, where a is positive and b is some real number. Just as we calculated ellipses by reducing it to a circle, which we reduced to a unit circle, we will reduce this complicated integral into a simpler one.

The trick we will need is a very useful one called *completing the square*.

Proposition 65 (Quadratic Formula)**.** A **quadratic equation** $ax^2 + bx + c = 0$ with nonzero a is solved by the formula

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

[1] This was done in Challenge 12, but let's do it again. Suppose e^x has dimension Y and x has dimension X. Then $(e^x)'$ has dimension Y/X. But $(e^x)' = e^x$, so $Y/X = Y$, and x must be dimensionless.

Proof. We use the the technique of **completing the square**. Dividing by a and subtracting c/a on both sides of the equation gives

$$x^2 + \frac{b}{a}x = -\frac{c}{a}.$$

The idea is that we want the left side to be of the form $(x+\alpha)^2$, for some α. To do this, we add $\left(\frac{b}{2a}\right)^2$ to both sides:

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

The left side of the equation is now a square, as you should verify. Combining the two terms on the right gives

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

Taking the square root on both sides gives us the formula:

$$x + \frac{b}{2a} = \pm\frac{\sqrt{b^2-4ac}}{2a}.$$

□

The symbol ± means there are two solutions $\frac{b}{2a} + \frac{\sqrt{b^2-4ac}}{2a}$ and $\frac{b}{2a} - \frac{\sqrt{b^2-4ac}}{2a}$. To see the necessity of two roots, observe that if $a = \pm 2$ then $a^2 = 4$. But if we take the square root $\sqrt{a^2}$, then we are only left with the positive solution $a = 2$. To fix this, we add the ± symbol and write $\pm\sqrt{a^2}$.

Theorem 66 (Gaussian Integral)**.**

$$\int_{-\infty}^{\infty} e^{-ax^2+bx}\,dx = e^{b^2/(4a)}\sqrt{\frac{\pi}{a}}$$

Proof. First, you should check that the answer makes sense dimensionally. We are going to reduce this integral into the integral in Equation 6.1. If we turn $-ax^2 + bx$ into $-au^2 + c$ for some constant c, then Equation 6.1 gives

$$\int_{-\infty}^{\infty} e^{-au^2+c}\,du = \int_{-\infty}^{\infty} e^c e^{-au^2}\,du = e^c \int_{-\infty}^{\infty} e^{-au^2}\,du = e^c\sqrt{\frac{\pi}{a}}.$$

In order to do this, we complete the square by adding a constant:

$$-ax^2 + bx = -ax^2 + bx - \frac{b^2}{4a} + \frac{b^2}{4a} = -a\left(x^2 - \frac{b}{2a}\right)^2 + \frac{b^2}{4a}.$$

So we should take $c := b^2/4a$ and $u : x \mapsto x - b/(2a)$. Since $u' = 1$, the substitution rule gives

$$\int_{-\infty}^{\infty} e^{-ax^2+bx}\,dx = e^{b^2/(4a)}\int_{-\infty}^{\infty} e^{-ax^2}\,dx = e^{b^2/(4a)}\sqrt{\frac{\pi}{a}}.$$

□

Let us return to the integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx$. First, notice that because of the square, the function e^{-x^2} is an even function. This means that $\int_0^{\infty} e^{-ax^2}\,dx = \int_{-\infty}^{0} e^{-ax^2}\,dx$ and so $\int_{-\infty}^{\infty} e^{-x^2}\,dx = 2\int_0^{\infty} e^{-x^2}\,dx$.

One of the endpoints of our integral $\int_0^{\infty} e^{-x^2}\,dx$ is not finite. An **improper integral** $\int_o^{\infty} f(x)\,dx$ for some real number o is defined by the following limit, if it exists.

$$\int_o^{\infty} f(x)\,dx := \lim_{t\to\infty}\int_o^{t} f(x)\,dx$$

If the above limit defining an improper integral exists, we will say that the improper integral **converges**.

As an aside, we have also seen another type of an improper integral in Challenge 17. The integral $\int_{-1}^{1} 1/x^2\,dx$ is an improper integral because $x \mapsto 1/x^2$ is undefined at $x = 0$. The integral $\int_0^1 1/x\,dx$ is also an improper integral because $1/x \to \infty$ as $x \to 0$. These two integrals are called **improper integrals of the second type**. When we speak of an improper integral in this book, we mean integrals of the form $\int_a^{\infty} f(x)\,dx$ where a is a real number.

Taking a limit always requires greater care. Nevertheless, many of our previous results port over to improper integrals. Here is an example.

Theorem 67 (Integration by Parts)**.** Let f and g be differentiable functions defined on $[a,\infty)$ such that f' and g' are continuous. If $\lim_{b\to\infty} f(b)g(b)$ exists and the integral $\int_a^{\infty} f(x)g'(x)\,dx$ converges then $\int_a^{\infty} f'(x)g(x)\,dx$ converges with

$$\int_a^{\infty} f'(x)g(x)\,dx = \lim_{b\to\infty}\big[f(b)g(b) - f(a)g(a)\big] - \int_a^{\infty} f(x)g'(x)\,dx.$$

Proof. For each $b > a$ integration by parts gives

$$\int_a^{b} f'(x)g(x)\,dx = \big[f(b)g(b) - f(a)g(a)\big] - \int_a^{b} f(x)g'(x)\,dx. \tag{6.2}$$

By assumption the following limit exists.

$$\lim_{b\to\infty}\big[f(b)g(b) - f(a)g(a)\big] - \int_a^{\infty} f(x)g'(x)\,dx = \lim_{b\to\infty}\left\{\big[f(b)g(b) - f(a)g(a)\big] - \int_a^{b} f(x)g'(x)\,dx\right\}$$

Equation 6.2 tells us that $\lim_{b\to\infty}\int_a^b f'(x)g(x)\,dx$ also exists. Since $\lim_{b\to\infty}\int_a^b f'(x)g(x)\,dx = \int_a^{\infty} f'(x)g(x)\,dx$, the integral $\int_a^{\infty} f'(x)g(x)\,dx$ converges with

$$\int_a^{\infty} f'(x)g(x)\,dx = \lim_{b\to\infty}\big[f(b)g(b) - f(a)g(a)\big] - \int_a^{\infty} f(x)g'(x)\,dx.$$

□

Now that we have some idea of what we are dealing with, let us go on ahead and calculate. Not so fast! It turns out that the integral $\int_0^{\infty} e^{-x^2}\,dx$ is special and *very* difficult to calculate.

This is hard to imagine. Suppose we changed the function a little bit by removing the square and putting back the positive constant a:

$$\int_0^\infty e^{-ax}\,dx.$$

This is a relatively straightforward integral because $(-e^{-ax}/a)' = e^{-ax}$ and $\lim_{x\to\infty} e^{-x} = 0$:

$$\lim_{t\to\infty}\int_0^t e^{-ax}\,dx = \lim_{t\to\infty}\left(-\frac{e^{-ax}}{a}\bigg|_0^t\right) = \lim_{t\to\infty}\left(-\frac{1}{ae^{at}} + \frac{e^0}{a}\right) = 0 + \frac{1}{a} = \frac{1}{a}.$$

Challenge 54 We can think of the integral $\int_0^\infty e^{-ax}\,dx$ as a function of a and let $f : a \mapsto \int_0^\infty e^{-ax}\,dx$ be a function defined on the interval $(0,\infty)$. From our previous calculations we know that $f(a) = 1/a$ and so $f'(a) = -1/a^2$. But $f(a)$ is an integral over the variable x, which is independent from variable a. So we may take the derivative inside to obtain $f'(a) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}a}e^{-ax}\,dx$. Hence

$$-\frac{1}{a^2} = f'(a) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}a}e^{-ax}\,dx = -\int_0^\infty xe^{-ax}\,dx$$

and so $\int_0^\infty xe^{-ax}\,dx = 1/a^2$. This technique is called **differentiation under the integral sign**.

(a) Use the well-ordering principle to show that

$$\int_0^\infty x^n e^{-ax}\,dx = \frac{n!}{a^{n+1}}$$

and conclude that

$$n! = \int_0^\infty x^n e^{-x}\,dx.$$

Define the **gamma function** Γ on the interval $(0,\infty)$ by

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx.$$

Recall from the end of Chapter 4 that $\lim_{x\to\infty} x^a/e^x = 0$ for each real a. Use integration by parts (Theorem 67) to show that $\Gamma(t+1) = t\Gamma(t)$. Since $\Gamma(n+1) = n!$ for each positive integer n, the gamma function extends the factorial function $n!$ to positive real numbers.

(b) Apply a differentiation under the integral sign on Equation 6.1 to show that

$$\int_{-\infty}^\infty x^2 e^{-ax^2}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}.$$

Continuing to differentiate under the integral sign gives the formula for $\int_{-\infty}^\infty x^{2n}e^{-ax^2}\,dx$.

(c) Apply differentiation under the integral sign on Theorem 66 (on the variable b) to show that

$$\int_{-\infty}^\infty xe^{-ax^2+bx}\,dx = \frac{b}{2a}e^{b^2/(4a)}\sqrt{\frac{\pi}{a}}.$$

Polar coordinates

Calculating the integral of e^{-ax} was simple because $(-e^{-ax}/a)' = e^{-ax}$. But differentiating e^{-x^2} gives an extra term $-2x$, which is *not* a constant. To fix this, we will try to work out some sort of substitution. For example, we can find the integral $\int_0^\infty xe^{-ax^2}\,dx$ with the substitution $g(x) := x^2$:

$$\int_0^\infty xe^{-ax^2}\,dx = \frac{1}{2}\int_0^\infty e^{-ax^2}2x\,dx = \frac{1}{2}\int_0^\infty e^{-ag(x)}g'(x)\,dx = \frac{1}{2}\int_0^\infty e^{-au}\,du = \frac{1}{2a}. \tag{6.3}$$

The substitution we will need is actually a change in coordinate system. Because our integral is too difficult to do on the regular (x, y) coordinate system, we will use our knowledge from trigonometric functions to come up with a new way of identifying points on the plane.

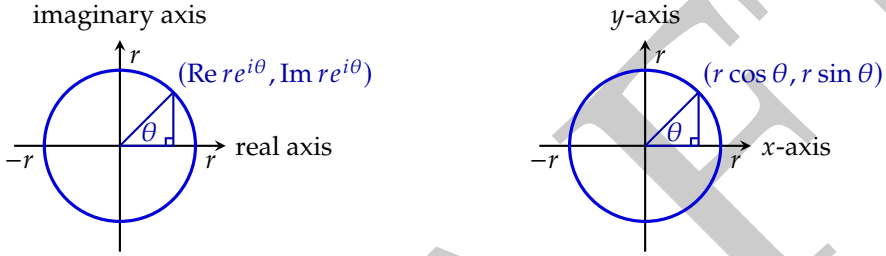
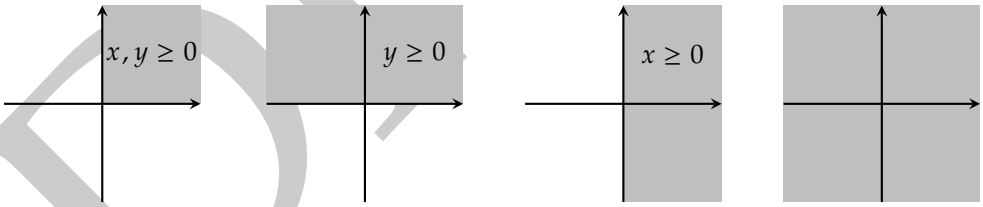


Figure 6.4: Argand diagram of $re^{i\theta}$ (left), which is the same as $(r\cos\theta, r\sin\theta)$ (right).

Although we are working with real numbers, there is no reason we cannot use insights from complex numbers. We will simply replace the real axis with the label "x-axis" and the imaginary axis with the label "y-axis". Then the complex number $z := re^{i\theta}$ on the Argand diagram corresponds to the coordinate $(r\cos\theta, r\sin\theta)$ on our x-y plane, which is the same as $(\operatorname{Re} z, \operatorname{Im} z)$ on the Argand diagram. Since the real and imaginary parts of a complex number are real, all is good!

The representation of points on the x-y plane using $(r\cos\theta, r\sin\theta)$ is called **polar coordinates**. The usual representation (x, y) is called **rectangular coordinates**.



We measure angle θ in radians. The region $x \geq 0$ and $y \geq 0$ (first diagram above) in polar coordinates is the region where $r \geq 0$ and $\theta \in [0, \pi/2]$. The region $y \geq 0$ (second diagram) in polar coordinates is the region $r \geq 0$ and $\theta \in [0, \pi]$. The region $x \geq 0$ (third diagram) in polar coordinates is the region $r \geq 0$ and $\theta \in [-\pi/2, \pi/2]$. The entirety of the x-y plane (final diagram) is represented in polar coordinates by the region $r \geq 0$ and $\theta \in [0, 2\pi]$.

Now that we have a new way of representing coordinates, let us figure out how to make a substitution for our integral!

6.2 Change of Variables

The formula

We recall differentiation with dual numbers. Suppose we have some function f that is differentiable at t. Then by the definition of the derivative, the equation $f(t + a\epsilon) = f(t) + f'(t)(a\epsilon)$ holds. We are free to choose our origin of measurement, so define t to be the origin of the x-axis and $f(t)$ to be the origin of the y-axis so that $t = 0$ and $f(t) = 0$. Then the equation simplifies to

$$f(a\epsilon) = f'(t)(a\epsilon).$$

Everything except the number a in this equation is taken as fixed: function f, the number t, the dual number ϵ. However, the number a is a variable. Take another number $\tilde{a} > a$ and observe that $f(\tilde{a}\epsilon) = f'(t)(\tilde{a}\epsilon)$ also holds. Subtracting one equation from another gives the following.

$$\underbrace{f(\tilde{a}\epsilon) - f(a\epsilon)}_{\text{rise in value}} = f'(t)(\tilde{a}\epsilon) - f'(t)(a\epsilon) = f'(t) \cdot \underbrace{\big([\tilde{a} - a]\,\epsilon\big)}_{\text{change along } x\text{-axis}}$$

We will denote the function's rise by df and the change of inputs along the x-axis by dx and write

$$df = f'(t)\,dx. \tag{6.5}$$

Notice that dx and df are functions that take in a and output a real number. The outputs of dx and df satisfy the relationship given in Equation 6.5.

Let us extend this idea to functions of two variables and three variables. Suppose function f takes two inputs x and y. The relationship between the function's rise and the increase in variable x is described precisely by $\partial_x f(t)$. Similarly, the relationship between the function's rise and the increase in the variable y is given by the number $\partial_y f(t)$. Therefore,

$$df = \partial_x f(t)\,dx + \partial_y f(t)\,dy. \tag{6.6}$$

Repeating this for a function f of three variables, we have $df = \partial_x f(t)\,dx + \partial_y f(t)\,dy + \partial_z f(t)\,dz$.

We have new objects, so let's do some arithmetic with it! As with the dual numbers, we will interpret the symbols $d\square$ to be nonzero quantities that square to zero. The difference is that there was only one ϵ, but now we have lots of $d\square$, so this rule is not enough. The rule that $(d\square)^2 = 0$ is a rule about products of these symbols; we need a rule about addition. But we need our addition rule to be compatible with the squaring rule we already have. The simplest way we can achieve this is to tie the addition rule to the squaring rule: summing the symbols $d\square$ is fine, but if we try to square that sum, then it also becomes zero.

As an example, let $X := dx + dy$. Then $X^2 = 0$, and so

$$0 = X^2 = (dx + dy)(dx + dy) = (dx)^2 + dx\,dy + dy\,dx + (dy)^2 = 0 + dx\,dy + dy\,dx + 0. \tag{6.7}$$

We see that $dx\,dy = -dy\,dx$. How about a product of linear combinations?

$$(\alpha\,dx + \beta\,dy)(\gamma\,dx + \delta\,dy) = 0 + (\alpha\delta)dx\,dy + (\beta\gamma)dy\,dx + 0 = (\alpha\delta - \beta\gamma)dx\,dy. \tag{6.8}$$

We now apply our new algebra to do calculus. The polar coordinates are described by the rule

$$x = r\cos\theta \qquad \text{and} \qquad y = r\sin\theta. \tag{6.9}$$

Observe that we can regard x and y as functions of r and θ. In particular, let us write

$$x = g_1(r, \theta) := r\cos\theta \quad \text{and} \quad y = g_2(r, \theta) := r\sin\theta.$$

We already worked out a function's rise with respect to increases in each of its inputs in Equation 6.6. We see that

$$dx = \partial_r g_1\, dr + \partial_\theta g_1\, d\theta \quad \text{and} \quad dy = \partial_r g_2\, dr + \partial_\theta g_2\, d\theta.$$

Using the derivatives of cosines and sines, the partial derivatives are as follows.

$$\partial_r g_1 = \cos\theta \qquad \partial_\theta g_1 = -r\sin\theta \qquad \partial_r g_2 = \sin\theta \qquad \partial_\theta g_2 = r\cos\theta$$

Their product is then the product of linear combinations from Equation 6.8

$$dx\, dy = \left(\partial_r g_1 \partial_\theta g_2 - \partial_\theta g_1 \partial_r g_2\right) dr\, d\theta = (r\cos^2\theta + r\sin^2\theta)\, dr\, d\theta = r\, dr\, d\theta \tag{6.10}$$

where we have used the Pythagorean theorem: $\cos^2\theta + \sin^2\theta = 1$. How about we use this to calculate the area of a circle once more?

An area of a circle of radius $\bar{r}$ can be calculated with the integral $\iint_A dx\, dy$, where A is the set of points (x, y) on the x-y plane such that $x^2 + y^2 \leq \bar{r}^2$. Instead of taking the integral over A, we will take an integral over a new region $\tilde{A}$, where each point in A corresponds to a point $(r\cos\theta, r\sin\theta) \in \tilde{A}$ such that $r \in [0, \bar{r}]$ and $\theta \in [0, 2\pi]$. The catch is that when integrating over the new region, we need to substitute $dx\, dy$ with our result from Equation 6.10. Applying our procedure gives the expected answer, as shown below.

$$\iint_A dx\, dy = \iint_{\tilde{A}} r\, dr\, d\theta = \int_0^{2\pi}\int_0^{\bar{r}} r\, dr\, d\theta = \int_0^{2\pi} \frac{\bar{r}^2}{2}\, d\theta = \frac{\bar{r}^2}{2}\int_0^{2\pi} d\theta = \pi\bar{r}^2$$

There is one subtlety. Let us recall when we first met the complex field in Section 5.3. We made the choice of $i := \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Suppose someone else decided to define $i := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, a perfectly reasonable choice. As we discussed before, their Argand diagram would have the opposite imaginary axis compared to ours. This means that their polar coordinate is given by the transformation rule $x = r\cos\theta$ and $y = -r\sin\theta$. Their partial derivatives of the transformations will be given by

$$\partial_r g_1 = \cos\theta \qquad \partial_\theta g_1 = -r\sin\theta \qquad \partial_r g_2 = -\sin\theta \qquad \partial_\theta g_2 = -r\cos\theta$$

and so

$$dx\, dy = \left(\partial_r g_1 \partial_\theta g_2 - \partial_\theta g_1 \partial_r g_2\right) dr\, d\theta = (-r\cos^2\theta - r\sin^2\theta)\, dr\, d\theta = -r\, dr\, d\theta.$$

Therefore, using our algebraic rules from Equation 6.8, they will calculate the area of a circle of radius $\bar{r}$ to be

$$\iint_A dx\, dy = \int_0^{2\pi}\int_0^{\bar{r}} (-r)\, dr\, d\theta = -\int_0^{2\pi} \frac{\bar{r}^2}{2}\, d\theta = -\frac{\bar{r}^2}{2}\int_0^{2\pi} d\theta = -\pi\bar{r}^2\text{?!}$$

A circle having negative area is absurd, and they did everything correctly! This means that the algebraic rule from Equation 6.8 must be modified to:

$$(\alpha\, dx + \beta\, dy)(\gamma\, dx + \delta\, dy) = |\alpha\delta - \beta\gamma| dx\, dy.$$

Adding this absolute value gives us the **change of variables formula** in two dimensions:

$$\iint_A f(x, y)\, dx\, dy = \iint_{\tilde{A}} f\left(g_1(u, v), g_2(u, v)\right) \left|\partial_u g_1 \partial_v g_2 - \partial_v g_1 \partial_u g_2\right| du\, dv. \tag{6.11}$$

Calculating the integral

At last, we can calculate the Gaussian integral.[2]

Theorem 68.

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

Proof. Since e^{-x^2} is even, if $I := \int_0^\infty e^{-x^2}\,dx$, then $\int_{-\infty}^{\infty} e^{-x^2}\,dx = 2I$. The trick is to calculate I^2:

$$I^2 = I\int_0^\infty e^{-y^2}\,dy = \int_0^\infty Ie^{-y^2}\,dy = \int_0^\infty \left(\int_0^\infty e^{-x^2}\,dx\right) e^{-y^2}\,dy.$$

Since e^{-y^2} is a constant with respect to the variable x, we push it in:

$$I = \int_0^\infty \int_0^\infty e^{-x^2}e^{-y^2}\,dx\,dy = \int_0^\infty\int_0^\infty e^{-(x^2+y^2)}\,dx\,dy.$$

We are integrating over the region where $x \geq 0$ and $y \geq 0$. Each point in this region corresponds to the polar coordinate $(r\cos\theta, r\sin\theta)$ where $r \geq 0$ and $\theta \in [0, \pi/2]$ (angle θ is between 0 and the right angle). From Equation 6.10 we know that $dx\,dy = r\,dr\,d\theta$.

We make the change of variables $x^2 + y^2 \mapsto r^2$ and use the change of variables formula to get

$$I^2 = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\,dx\,dy = \int_0^{\pi/2}\int_0^\infty e^{-r^2} r\,dr\,d\theta.$$

Since $(-e^{-r^2}/2)' = e^{-r^2}r$ (clean and simple!), we have

$$I^2 = \int_0^{\pi/2}\int_0^\infty e^{-r^2} r\,dr\,d\theta = \int_0^{\pi/2}\left[-\frac{1}{2}e^{-r^2}\Big|_{r=0}^{\infty}\right] d\theta.$$

Since $\lim_{r\to\infty} e^{r^2} = \infty$, we know that $\lim_{r\to\infty} e^{-r^2} = \lim_{r\to\infty} 1/e^{r^2} = 0$. Therefore,

$$I^2 = \int_0^{\pi/2}\left[-\frac{1}{2}e^{-r^2}\Big|_{r=0}^{\infty}\right] d\theta = \int_0^{\pi/2}\left[0 + \frac{e^0}{2}\right] d\theta = \frac{1}{2}\int_0^{\pi/2} d\theta = \frac{\pi}{4}.$$

We conclude that

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = 2I = 2\frac{\sqrt{\pi}}{2} = \sqrt{\pi}$$

as desired. □

Applying the substitution rule with the substitution $x \mapsto \sqrt{a}x$ for positive a gives

$$\int_{-\infty}^{\infty} e^{-ax^2}\,dx = \sqrt{\frac{\pi}{a}}.$$

[2]This important integral can be calculated many different ways, including the differentiation under the integral sign.

At this point, integrals like the following should look quite harmless.

$$\int_0^\infty x^{2n} e^{-x^2/a^2}\,dx = \sqrt{\pi}\,\frac{(2n)!}{n!}\left(\frac{a}{2}\right)^{2n+1} \qquad \int_0^\infty x^{2n+1} e^{-x^2/a^2}\,dx = \frac{n!}{2}a^{2n+2}$$

To obtain the former, apply the substitution rule with the substitution $x \mapsto x/a$ on the integral $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$. Since the function e^{-x^2/a^2} is an even function, we have $\int_0^\infty e^{-x^2/a^2}\,dx = \sqrt{\pi}\,a/2$. Taking a differentiation under the integration sign gives $\int_0^\infty x^2 e^{-x^2/a^2}\,dx = \sqrt{\pi}\,a^3/4$. Let n be the smallest natural number for which the formula does not hold. Since the formula holds for the natural number $n-1$, we know that the following holds.

$$\int_0^\infty x^{2(n-1)} e^{-x^2/a^2}\,dx = \sqrt{\pi}\,\frac{(2[n-1])!}{(n-1)!}\left(\frac{a}{2}\right)^{2(n-1)+1}.$$

Apply differentiation under the integral sign to get

$$\int_0^\infty x^{2(n-1)}(2x^2a^{-3}) e^{-x^2/a^2}\,dx = \sqrt{\pi}\,\frac{(2[n-1])!}{(n-1)!}(2[n-1]+1)\frac{a^{2(n-1)}}{2^{2(n-1)+1}}.$$

Multiply both sides by $\frac{a^3}{2}$ and multiply the right side by $\frac{2n}{2n}$ and tidy up to obtain

$$\int_0^\infty x^{2n} e^{-x^2/a^2}\,dx = \sqrt{\pi}\,\frac{(2[n-1])!}{(n-1)!}(2[n-1]+1)\frac{a^{2(n-1)+3}}{2^{2(n-1)+2}}\frac{2n}{2n} = \sqrt{\pi}\,\frac{(2n)!}{n!}\frac{a^{2n+1}}{2^{2n+1}}$$

as desired.

The latter is similar. Start by applying the substitution rule on the integral $\int_0^\infty xe^{-x^2}\,dx = 1/2$ with the substitution $x \mapsto x/a$.[3] Differentiating this under the integral sign gives us

$$\int_0^\infty (2x^2a^{-3}) xe^{-x^2/a^2}\,dx = \frac{2}{2}a$$

which (after some algebra) is the desired formula for the natural number $n = 1$. If n is the smallest natural number for which our formula does not hold, then the following is true:

$$\int_0^\infty x^{2(n-1)+1} e^{-x^2/a^2}\,dx = \frac{(n-1)!}{2}a^{2(n-1)+2}.$$

Differentiating under the integral sign results in the following equation.

$$\int_0^\infty (2x^2a^{-3}) x^{2(n-1)+1} e^{-x^2/a^2}\,dx = 2n\frac{(n-1)!}{2}a^{2n-1}$$

The above tidies up to the desired form:

$$\int_0^\infty x^{2n+1} e^{-x^2/a^2}\,dx = \frac{n!}{2}a^{2n+2}.$$

Have a look at an introductory quantum mechanics text to see the integrals of this chapter in action.

[3] We know from Equation 6.3 that $\int_0^\infty xe^{-\alpha x^2}\,dx = \frac{1}{2\alpha}$. Take $\alpha = 1$ then substitute to get $\int_0^\infty (x/a)\,e^{-x^2/a^2}(1/a)\,dx = \frac{1}{2}$.

6.3 Determinants

The determinant is an important concept that appears in many places, yet it has a complicated looking formula that makes it difficult to see how one could come up with the concept in the first place. Our goal will be to obtain the determinant for 3×3 matrices and its properties together.

We review the change of variables formula. If we wish to calculate an integral $\iint_A f(x,y)\,dx\,dy$, we can instead do a change of variables $x = g_1(u,v)$ and $y = g_2(u,v)$ to calculate a new integral over the corresponding region $\tilde{A}$ in u, v space (in our case, it was polar coordinates with the variables r and θ). This was the change of variables formula in two dimensions:

$$\iint_A f(x,y)\,dx\,dy = \iint_{\tilde{A}} f\big(g_1(u,v), g_2(u,v)\big)\left|\partial_u g_1 \partial_v g_2 - \partial_v g_1 \partial_u g_2\right| du\,dv. \tag{6.12}$$

The formula above can be tidied up by using a matrix. Each transformation $x = g_1(u,v)$ and $y = g_2(u,v)$ are real valued functions of two variables. Take the gradients of each function and stack their *transpose* together to obtain the **Jacobian matrix** J_g for the transformation $g(u,v) := \begin{pmatrix} g_1(u,v) \\ g_2(u,v) \end{pmatrix}$, defined by

$$J_g := \begin{pmatrix} (\nabla g_1)^\intercal \\ (\nabla g_2)^\intercal \end{pmatrix} = \begin{pmatrix} \partial_u g_1 & \partial_v g_1 \\ \partial_u g_2 & \partial_v g_2 \end{pmatrix}.$$

Observe that the expression inside the absolute values of Equation 6.12 is the determinant of the Jacobian matrix (the determinant of a Jacobian matrix is called the **Jacobian**).

The regions A and $\tilde{A}$ have the relationship $A = g(\tilde{A})$. Renaming $\tilde{A}$ with A allows us to write the change of variables formula in two dimensions as:

$$\iint_{g(A)} f(x,y)\,dx\,dy = \iint_A (f \circ g)(u,v)\left|\det J_g\right| du\,dv. \tag{6.13}$$

If an integral $\int dx$ measures a length, then a double integral $\iint dx\,dy$ measures an area. What does a determinant have to with areas? Recall that the determinant maps a matrix to a real number. An $n \times m$ matrix A transforms a vector of dimension m, where the ith column is Ae_i. So let us visualize a transformation in action.

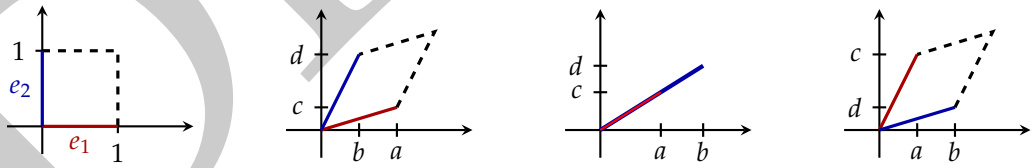


Figure 6.14: The standard basis vectors e_1, e_2 and its transformations form a parallelogram.

The columns of a matrix $M := \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ tells us that $Me_1 = \begin{pmatrix} a \\ c \end{pmatrix}$ and $Me_2 = \begin{pmatrix} a \\ c \end{pmatrix}$. We can visualize a vector as a line emanating from the origin to its coordinates. For example the basis vector e_1 is shown as a red line beginning from the origin to coordinate $(1,0)$ in the first diagram of Figure 6.14.

Similarly the basis vector e_2 is shown as a blue line beginning from the origin to coordinate $(0, 1)$ in the first diagram of Figure 6.14. Using this visualization, the vectors Me_1 (red) and Me_2 (blue) specify a parallelogram, as shown in the second, third, and fourth diagrams of Figure 6.14.

We will assume that both Me_1 and Me_2 are nonzero vectors for now. There are three possibilities: vector Me_1 remains below vector Me_2 (second figure), vectors Me_1 and Me_2 overlap (third figure), or vector Me_1 is above vector Me_2 (fourth figure).

Because rotating an object does not change the object's area, we can easily calculate our parallelogram's area by rotating our parallelogram so that one of the lines lies on the x-axis, then taking the product of the base and height of the rotated parallelogram. Since each nonzero 2×2 matrix specifies a parallelogram, it is sufficient to apply the rotation on matrix M.

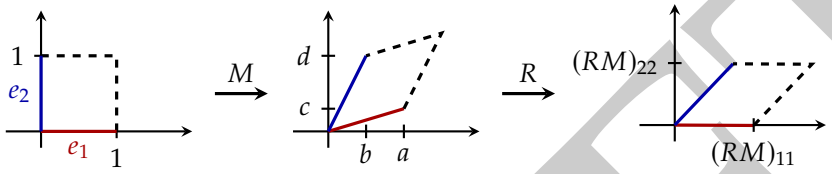


Figure 6.15: Transformation of e_1 and e_2 by M, followed by a rotation R.

First, suppose that vector Me_1 is above vector Me_2. Let $-x$ be the angle between the x axis and the vector Me_2 (red line). We multiply the matrix M from the left by the rotation matrix R corresponding to e^{ix} so that $(RM)e_2$ lies on the x-axis (see Figure 6.15):

$$\begin{pmatrix} \cos x & -\sin x \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a\cos x - c\sin x & b\cos x - d\sin x \\ a\sin x + c\cos x & b\sin x + d\cos x \end{pmatrix}.$$

Then the length of the base of the parallelogram is given by $(RM)_{11} = a\cos x - c\sin x$ while the height of the parallelogram is given by $(RM)_{22} = b\sin x + d\cos x$. The area of the parallelogram is thus

$$\begin{aligned} \text{base} \times \text{height} &= (a\cos x - c\sin x)(b\sin x + d\cos x) \\ &= -bc\sin^2 x + ad\cos^2 x + ab\cos x\sin x - cd\sin x\cos x. \end{aligned}$$

As $(RM)e_1$ lies on the x-axis without a y-component, $(RM)_{21} = a\sin x + c\cos x = 0$. We thus obtain

$$\begin{aligned} \text{base} \times \text{height} &= -bc\sin^2 x + ad\cos^2 x + ab\cos x\sin x - cd\sin x\cos x \\ &= -bc\sin^2 x + ad\cos^2 x - bc\cos^2 x + ad\sin^2 x \\ &= -bc(\sin^2 x + \cos^2 x) + ad(\cos^2 x + \sin^2 x) = ad - bc \end{aligned}$$

which is simply the determinant of M.

If vectors Me_1 and Me_2 overlap (as in the third diagram of Figure 6.14), then the parallelogram's area is 0.[4] This is also correctly given by the determinant of M. Indeed, if at least one of Me_1 and Me_2 is zero, then at least one of the columns of M is zero, which means that $\det M = 0$. Otherwise, if Me_1 and Me_2 overlap and are both nonzero, then Me_2 is a multiple of Me_1 for some scalar α. This means that $M = \begin{pmatrix} a & \alpha a \\ c & \alpha c \end{pmatrix}$ and so $\det M = a \cdot \alpha c - \alpha a \cdot c = 0$, as desired.

[4]This case covers the possibility that at least one of Me_1 and Me_2 is the zero vector.

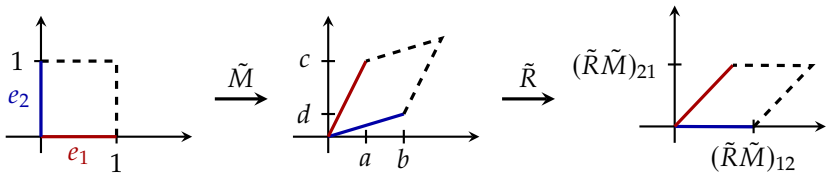


Figure 6.16: Transformation of e_1 and e_2 by $\tilde{M}$, followed by rotation $\tilde{R}$.

Challenge 55

(a) Let $\tilde{M}$ be a matrix such that vector $\tilde{M}e_1$ lies above vector $\tilde{M}e_2$ (second diagram of Figure 6.16). Let $-y$ be the angle between the x axis and the vector $\tilde{M}e_2$ (blue line). Multiply the matrix $\tilde{M}$ from the left by the rotation matrix $\tilde{R}$ corresponding to e^{iy} so that $(\tilde{R}\tilde{M})e_2$ lies on the x-axis. Show that the area of the parallelogram is given by $-\det \tilde{M}$.

(b) Check that multiplying the matrix $\tilde{M}$ from the right by the rotation matrix $\tilde{R}$ does not change the conclusion of part (a). Hence the order that we apply the transformation and rotation does not matter.

We see that the determinant of M by itself does correspond to the area of the parallelogram specified by matrix M. The area of the parallelogram is given by the absolute value $|\det M|$ and the determinant of M is said to give the *signed* area of the parallelogram specified by the columns of matrix M.

Challenge 56 We generalize to functions of three variables.

(a) Use our rules for the symbols $d\square$ to obtain the following relations.

$$du\,dv = -dv\,du \qquad du\,dw = -dw\,du \qquad dv\,dw = -dw\,dv$$

(b) Use part (a) to conclude that

$$du\,dv\,dw = -dv\,du\,dw = dv\,dw\,du = -dw\,dv\,du = dw\,du\,dv = -du\,dw\,dv.$$

(c) Calculate $dx\,dy\,dz$ where $dx := \partial_u g_1\,du + \partial_v g_1\,dv + \partial_w g_1\,dw$, $dy := \partial_u g_2\,du + \partial_v g_2\,dv + \partial_w g_2\,dw$, and $dz := \partial_u g_3\,du + \partial_v g_3\,dv + \partial_w g_3\,dw$. There are six terms because $(du)^2 = (dv)^2 = (dw)^2 = 0$. Don't forget to put the coefficients inside an absolute value to prevent negative volumes arising because of the conversion factor.

(d) Define

$$g(u,v,w) := \begin{pmatrix} g_1(u,v,w) \\ g_2(u,v,w) \\ g_3(u,v,w) \end{pmatrix} \quad \text{and} \quad g' := \begin{pmatrix} (\nabla g_1)^\intercal \\ (\nabla g_2)^\intercal \\ (\nabla g_3)^\intercal \end{pmatrix}.$$

The Jacobian matrix J_g is a matrix of dimension 3×3. In order to make the **change of variables formula** in three dimensions below to work

$$\iiint_{g(A)} f(x,y,z)\,dx\,dy\,dz = \iiint_A (f \circ g)(u,v,w)|\det J_g|\,du\,dv\,dw$$

show that the **determinant** of a matrix of dimension 3×3 should be defined to be

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} := aei - afh + bfg - bdi + cdh - ceg.$$

Due to the absolute value in the change of variables formula, the negative of the above also works. Which do we choose? Show that the above gives $\det 1 = 1$ and is the correct choice.

In part (b) of Challenge 56, we see that every time we do a transposition, we multiply by a negative sign. If we think of a permutation as corresponding to a sign, we have the following.

$$\begin{pmatrix} u & v & w \\ u & v & w \end{pmatrix} \mapsto 1 \qquad \begin{pmatrix} u & v & w \\ v & u & w \end{pmatrix} \mapsto -1 \qquad \begin{pmatrix} u & v & w \\ v & w & u \end{pmatrix} \mapsto 1$$

$$\begin{pmatrix} u & v & w \\ w & v & u \end{pmatrix} \mapsto -1 \qquad \begin{pmatrix} u & v & w \\ w & u & v \end{pmatrix} \mapsto -1 \qquad \begin{pmatrix} u & v & w \\ u & w & v \end{pmatrix} \mapsto -1$$

Notice that the permutations that correspond to a negative signs are the transpositions in S_3, while all non transpositions corresponds to a positive sign.

Challenge 57 Let ρ be the 3-dimensional representation of S_3 as permutation matrices. Show that the permutation matrices corresponding to transpositions have determinant -1 and the others have determinant $+1$. Show that if $f : G \to G'$ and $g : G' \to \tilde{G}$ are homomorphisms then $h := f \circ g$ is a homomorphism. The homomorphism $\rho \circ \det : S_3 \to \{\pm 1\}$ tells us the **sign** of a permutation.

We can formalize this relation in a similar manner as the Kronecker delta δ_{ij}. Let l, m, and n each take a value between 1, 2, and 3. For example, we may take $l = 1, m = 3, n = 2$ or $l = 2, m = 2, n = 1$. Define the **Levi-Civita symbol** ϵ_{lmn} as follows. For each $\sigma := \begin{pmatrix} 1 & 2 & 3 \\ l & m & n \end{pmatrix}$,

$$\epsilon_{lmn} := \begin{cases} 0 & \text{if } \sigma \notin S_3, \\ -1 & \text{if } \sigma \in S_3 \text{ is a transposition,} \\ +1 & \text{if } \sigma \in S_3 \text{ is not a transposition.} \end{cases}$$

Thus $\epsilon_{123} = \epsilon_{231} = \epsilon_{312} = 1$ and $\epsilon_{132} = \epsilon_{213} = \epsilon_{321} = -1$, while $\epsilon_{221} = 0$.

Challenge 58 Let A be a 3×3 matrix whose entry in the i-row and j-th column is denoted by A_{ij}.

(a) Show that

$$\det A = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \epsilon_{ijk} A_{1i} A_{2j} A_{3k} \tag{6.17}$$

where the determinant of a 3×3 matrix was defined in Challenge 56. The term on the right is a sum with $3^3 = 27$ terms, only six of which are nonzero, as the group S_3 has six elements.

(b) Let $l, m, n \in \{1, 2, 3\}$ (not necessarily distinct!). Use Equation 6.17 to show that

$$\epsilon_{lmn} \det A = \sum_{i=1}^{3} \sum_{j=1}^{3} \sum_{k=1}^{3} \epsilon_{ijk} A_{li} A_{mj} A_{nk}. \tag{6.18}$$

[*Hint:* this is far less work than it may seem. If $\sigma \notin S_3$, we may assume without loss of generality that $l = m$. Similarly, for the case that $\sigma \in S_3$ is a transposition, we may assume without loss of generality that σ is the transposition $1 \mapsto 2, 2 \mapsto 1$.]

(c) Multiplying both sides of the Equation 6.18 by ϵ_{lmn} and then summing over l, m, n gives

$$\det A \sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3}\epsilon_{lmn}\epsilon_{lmn} = \sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{lmn}\epsilon_{ijk}A_{li}A_{mj}A_{nk}.$$

Show that

$$3!\det A = \sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{lmn}\epsilon_{ijk}A_{li}A_{mj}A_{nk}. \tag{6.19}$$

Proposition 69. If A is a 3×3 matrix, then $\det A^{\mathsf{T}} = \det A$.

Proof. From Equation 6.19 we have

$$3!\det A^{\mathsf{T}} = \sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{lmn}\epsilon_{ijk}A_{il}A_{jm}A_{kn}.$$

Interchange the label i with l, the label j with m, and the label k with n to get

$$3!\det A^{\mathsf{T}} = \sum_{l=1}^{3}\sum_{m=1}^{3}\sum_{n=1}^{3}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{lmn}\epsilon_{ijk}A_{li}A_{mj}A_{nk} = 3!\det A. \qquad \square$$

Theorem 70. Let A be a 3×3 matrix whose entry in the i-row and j-th column is denoted by A_{ij}.

(a) For the 3×3 identity matrix 1, $\det 1 = 1$.
(b) If any of the two rows or columns are a scalar multiple of another, then $\det A = 0$.
(c) Interchanging any two rows or columns of A changes the sign of the determinant.
(d) Multiplying a row or column of A by c multiplies the determinant by c.
(e) Adding a scalar multiple of a row (column) into a different row (column) leaves the determinant unchanged.

Proof. As $\det A^{\mathsf{T}} = \det A$, the results for columns follows from the corresponding result for rows. Let c be a scalar.

(a) Immediate from Equation 6.17.
(b) Immediate from Equation 6.18: $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{ijk}A_{li}(cA_{lj})A_{nk} = 0$.
(c) Also immediate from Equation 6.18.
(d) Immediate from Equation 6.17: $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\epsilon_{ijk}(cA_{1i})A_{2j}A_{3k} = c\det A$.
(e) Add a scalar multiple of the nth row to the mth row to modify Equation 6.18 as follows (the symbol $\sum_{ijk}$ means $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}$).

$$\begin{aligned}\sum_{ijk}\epsilon_{ijk}A_{li}(A_{mj}+cA_{nj})A_{nk} &= \sum_{ijk}\epsilon_{ijk}\left(A_{li}A_{mj}A_{nk} + A_{li}[cA_{nj}]A_{nk}\right)\\ &= \sum_{ijk}\epsilon_{ijk}A_{li}A_{mj}A_{nk} + \sum_{ijk}\epsilon_{ijk}A_{li}(cA_{nj})A_{nk}\\ &= \sum_{ijk}\epsilon_{ijk}A_{li}A_{mj}A_{nk} + c\epsilon_{lnn}\det A = \sum_{ijk}\epsilon_{ijk}A_{li}A_{mj}A_{nk}\end{aligned} \qquad \square$$

Proposition 71. If A and B are 3×3 matrices, then $\det(AB) = \det A \det B$.

Proof. Since $(AB)_{ij} = \sum_{k=1}^{3} A_{ik}B_{kj}$, we have

$$\det(AB) = \sum_{ijk} \epsilon_{ijk} \left(\sum_{l=1}^{3} A_{1l}B_{li} \right) \left(\sum_{m=1}^{3} A_{2m}B_{mj} \right) \left(\sum_{n=1}^{3} A_{3n}B_{nk} \right).$$

We will pull out the terms that only depend on l, m, n to get

$$\det(AB) = \sum_{l,m,n} A_{1l}A_{2m}A_{3n} \sum_{ijk} \epsilon_{ijk} B_{li}B_{mj}B_{nk}.$$

By Equation 6.18, the sum over i, j, k is $\epsilon_{lmn} \det B$:

$$\det(AB) = \sum_{l,m,n} A_{1l}A_{2m}A_{3n} \left(\epsilon_{lmn} \det B \right) = \left(\sum_{l,m,n} \epsilon_{lmn} A_{1l}A_{2m}A_{3n} \right) \det B.$$

But the term in the brackets is simply $\det A$. Therefore, $\det(AB) = \det A \det B$. □

As with 2×2 matrices, the determinant gives the criterion for invertibility.

Corollary 72. A 3×3 matrix A is not invertible if $\det A = 0$.

Proof. If not, there is some A^{-1} such that $AA^{-1} = 1$, giving the following contradiction.

$$1 = \det 1 = \det \left(AA^{-1} \right) = \det A \det A^{-1} = 0 \det A^{-1} = 0$$ □

Challenge 59 If C is a matrix, then $C_{\not i \not j}$ is the matrix formed by removing row i and column j. For example,

$$C := \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \implies C_{\not 2 \not 1} = \begin{pmatrix} b & c \\ h & i \end{pmatrix}.$$

Let A be a 3×3 matrix with nonzero determinant.

(a) Show that $\det A = \sum_{j=1}^{3} (-1)^{1+j} A_{1j} \det A_{\not 1 \not j}$. This is called the **Laplace expansion** along row 1.

(b) Show that for each row number $i \in \{1, 2, 3\}$, the following equation holds.

$$\det A = \sum_{j=1}^{3} (-1)^{i+j} A_{ij} \det A_{\not i \not j}$$

This is the Laplace expansion along row i.

(c) Show that for each column number $j \in \{1, 2, 3\}$, the following equation holds.

$$\det A = \sum_{i=1}^{3} (-1)^{i+j} A_{ij} \det A_{\not i \not j}$$

This is the Laplace expansion along column j. The term $(-1)^{i+j} \det A_{\not i \not j}$ is called the (i, j)-**cofactor** of A.

(d) Let A and B be $n \times n$ matrices. Show that $(AB)^\intercal = B^\intercal A^\intercal$. [*Hint:* $[AB]_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$.]

(e) The **adjugate** matrix of A, denoted adj A, is the transpose of the cofactor matrix of A. Thus $[\text{adj } A]_{ij} := (-1)^{j+i} \det A_{\not j\not i}$. Check that $(\text{adj } A)^\intercal = \text{adj } A^\intercal$.

(f) Observe that

$$[A(\text{adj } A)]_{ii} = \sum_{k=1}^{3} A_{ik}(\text{adj } A)_{ki} = \sum_{k=1}^{3} (-1)^{k+i} A_{ik} \det A_{\not i\not k} = \det A.$$

Show that $[A(\text{adj } A)]_{ij} = (\det A)\delta_{ij}$ (Kronecker delta) and so $A(\text{adj } A) = (\det A)1$. [*Hint:* what happens to the (i, j)-cofactor of A when we replace column j of matrix A with column i?]

(g) Use parts (d), (e), and (f) to show that $(\text{adj } A)A = (\det A)1$ and conclude that A is invertible with $A^{-1} = \frac{1}{\det A}\text{adj } A$. Combining with Corollary 72, we see that a 3×3 matrix A is invertible if and only if $\det A \neq 0$.

One of the first vectors we considered were vectors representing polynomials. We consider the following matrix V called a **Vandermonde matrix**.

$$V := \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{pmatrix}$$

Let us try and calculate $\det V$. Since adding to a row a scalar multiple of another column does not change the determinant, we can add to the third row the middle row times $-x_1$ to get

$$\det V = \det \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 - x_1^2 & x_2^2 - x_2x_1 & x_3^2 - x_3x_1 \end{pmatrix} = \det \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ 0 & x_2(x_2 - x_1) & x_3(x_3 - x_1) \end{pmatrix}$$

Next, we add to the middle row the first row times $-x_1$ to get the following.

$$\det V = \det \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ 0 & x_2(x_2 - x_1) & x_3(x_3 - x_1) \end{pmatrix} = \det \begin{pmatrix} 1 & 1 & 1 \\ 0 & x_2 - x_1 & x_3 - x_1 \\ 0 & x_2(x_2 - x_1) & x_3(x_3 - x_1) \end{pmatrix} \tag{6.20}$$

Let W be the rightmost matrix in Equation 6.20. Using the Laplace expansion along column 1 gives

$$\begin{aligned} \det V = \det W &= \sum_{i=1}^{3} (-1)^{i+1} W_{11} \det W_{\not 1\not 1} = (-1)^2 W_{11} \det W_{\not i\not 1} + 0 + 0 \\ &= (x_2 - x_1)x_3(x_3 - x_1) - (x_3 - x_1)x_2(x_2 - x_1) = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2). \end{aligned}$$

The polynomial $\Delta_3 := (x_2 - x_1)(x_3 - x_1)(x_3 - x_2)$ is called the degree 3 **Vandermonde polynomial**.

Challenge 60 Matrix M is **upper triangular** if $M_{ij} = 0$ whenever $i > j$. Show that if A is an upper-triangular 3×3 matrix then $\det A = A_{11}A_{22}A_{33}$.

Challenge 61

(a) A real matrix O is **orthogonal** if $OO^\intercal = 1$. Show that if O is a 3×3 orthogonal matrix, then $\det O = \pm 1$.

(b) Show that if A is a complex 3×3 matrix, then $\det A^\dagger = (\det A)^*$ (see Challenge 36).

(c) Let complex matrix H be a 3×3 Hermitian matrix. Show that $\det H$ is a real number.

(d) Let complex matrix U be a 3×3 unitary matrix. Show that $|\det U| = 1$.

6.4 Operator Transposition

Continuous functions defined on rectangles

The analogue of an interval in two dimensions is a **rectangle** (Figure 6.21). An **open rectangle** $(a,b)\times(c,d)$ consists of the points $(x,y)\in\mathbb{R}^2$ (the plane) such that $x\in(a,b)$ and $y\in(c,d)$.[5] A **closed rectangle** $[a,b]\times[c,d]$ consists of the points $(x,y)\in\mathbb{R}^2$ such that $x\in[a,b]$ and $y\in[c,d]$.[6] Replacing the intervals with rectangles, in particular, open intervals with open rectangles gives the definition of a continuous function defined on a rectangle.

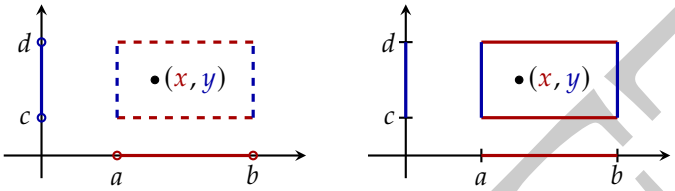


Figure 6.21: An open rectangle $(a,b)\times(c,d)$ and a closed rectangle $[a,b]\times[c,d]$.

Definition 73. A real-valued function f defined on a rectangle R is **continuous at** $p:=(x,y)\in R$ if for each $\epsilon>0$ there is some $\delta(\epsilon)>0$ such that $f(q)\in\big(f(p)-\epsilon,f(p)+\epsilon\big)$ whenever $q\in R$ is an element of the open rectangle $(x-\epsilon,x+\epsilon)\times(y-\epsilon,y+\epsilon)$.[7] Function f is **continuous** if it is continuous on each $p\in R$.

From the definition, if a real-valued function f defined on a closed rectangle $R:=[a,b]\times[c,d]$ is continuous at $p:=(x,y)\in R$, then $g_v:[a,b]\to\mathbb{R}$ defined by $g_v:x\mapsto f(x,v)$ is continuous at x. Similarly, $h_u:[c,d]\to\mathbb{R}$ defined by $h_u:y\mapsto f(u,y)$ is continuous at y. This allows us to port over results we have obtained from real-valued functions defined on intervals. For example, since $[a,b]$ and $[c,d]$ are closed intervals, the Extreme Value Theorem tells us that the functions g_v and h_u attains its maximum and minimum values.

The First Fundamental Theorem of Calculus holds for a continuous real-valued function f defined on a closed rectangle $[a,b]\times[c,d]$.

$$\partial_1\left[\int_a^x f(u,y)\,du\right]=f(x,y)\qquad\qquad\partial_2\left[\int_c^y f(x,v)\,dv\right]=f(x,y)\tag{6.22}$$

To obtain the first equation, apply the First Fundamental Theorem of Calculus to the continuous function f forgetting the y component then adding it in later.

$$\partial_1\left[\int_a^x f(u,y)\,du\right]=\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x g_y(u)\,du=g_y(x)=f(x,y)$$

[5] The endpoints a and c may take the symbol $-\infty$ and the endpoints b and d may take the symbol ∞.

[6] The endpoints a, b, c, and d must all be real numbers, with $a<b$ and $c<d$.

[7] Recall the definition of a function defined on an interval: $f:I\to\mathbb{R}$ is continuous at $p\in I$ if for each $\epsilon>0$ there is some $\delta(\epsilon)>0$ such that $f(q)\in\big(f(p)-\epsilon,f(p)+\epsilon\big)$ whenever $q\in I$ is an element of the open interval $\big(p-\epsilon,p+\epsilon\big)$.

The second equation obtained in a completely analogous fashion.

$$\partial_2 \left[\int_c^y f(x,v)\,dv \right] = \frac{\mathrm{d}}{\mathrm{d}y} \int_a^y h_x(v)\,dv = h_x(y) = f(x,y)$$

The Second Fundamental Theorem of Calculus also holds for a continuous real-valued function f defined on a closed rectangle $[a,b]\times[c,d]$.

$$\int_a^b \partial_1 f(u,y)\,du = f(b,y) - f(a,y) \qquad \int_c^d \partial_2 f(x,v)\,dv = f(x,d) - f(x,c) \tag{6.23}$$

These are also obtained in a similar fashion as shown below.

$$\int_a^b \partial_1 f(u,y)\,du = \int_a^b g_y'(u)\,du = g_y(b) - g_y(a) = f(b,y) - f(a,y)$$
$$\int_c^d \partial_2 f(x,v)\,dv = \int_c^d h_x'(v)\,dv = h_x(d) - h_x(c) = f(x,d) - f(x,c)$$

We return to the definition of a continuous function (Definition 73).[8] As is in the case of functions of a single variable, the value of $\delta(\epsilon)$ will depend on the point p. However, it is an amazing fact that if a real-valued function f is defined on a *closed* rectangle, then the same $\delta(\epsilon)$ will work everywhere f is defined. Such a function is said to be *uniformly continuous*.

Definition 74. A real-valued function f defined on a rectangle R is **uniformly continuous** if for each $\epsilon > 0$ there is some $\delta(\epsilon) > 0$ such that for each $p := (x,y) \in R$ we have $f(q) \in \big(f(p) - \epsilon, f(p) + \epsilon\big)$ whenever $q \in R$ is an element of the open rectangle $(x-\epsilon, x+\epsilon)\times(y-\epsilon, y+\epsilon)$.

Proposition 75. If f is continuous on $[a,b]\times[c,d]$ then the following functions are continuous.

$$g : y \mapsto \int_a^b f(u,y)\,du \qquad h : (x,y) \mapsto \int_a^x f(u,y)\,du$$

Proof. We will take advantage of the fact that f is defined on a closed rectangle and is thus unifirmly continuous. Let $\epsilon > 0$; since f is uniformly continuous there is some $\delta(\epsilon) > 0$ that satisfies uniform continuity of f. Let nonzero α satisfy $|\alpha| < \delta(\epsilon)$; we will show that $\big|g(y+\alpha) - g(y)\big| < \beta\cdot\epsilon$ for some constant β. Linearity of the integral gives

$$\big|g(y+\alpha) - g(y)\big| = \left| \int_a^b f(u,y+\alpha)\,du - \int_a^b f(u,y)\,du \right| = \left| \int_a^b \big[f(u,y+\alpha) - f(u,y)\big]\,du \right|.$$

By continuity $f(u,y+\alpha) - f(u,y) \le \big|f(u,y+\alpha) - f(u,y)\big| < \epsilon$. Setting $\beta := b-a$ gives

$$\big|g(y+\alpha) - g(y)\big| \le \left| \int_a^b \big|f(u,y+\alpha) - f(u,y)\big|\,du \right| < \left| \int_a^b \epsilon\,du \right| = (b-a)\epsilon = \beta\cdot\epsilon.$$

[8] Feel free to skip ahead to the statement of the *Leibniz integral rule* and then continuing to *Fubini's Theorem*.

Next, we show that h is continuous at each point $(x, y) \in [a, b]\times[c, d]$. Let $\epsilon > 0$ and let $\delta_1(\epsilon) > 0$ satisfy uniform continuity of f. Take $\delta \mapsto \min(\delta_1, \epsilon)$. Let $\alpha \in (x - \delta, x + \delta)$ and $\beta \in (y - \delta, y + \delta)$. We use Property (P1) of an integral to break up the first integral then use linearity to get

$$\begin{aligned} h(x, y) - h(\alpha, \beta) &= \int_a^x f(u, y)\, du - \int_a^\alpha f(u, \beta)\, du \\ &= \left[\int_a^\alpha f(u, y)\, du + \int_\alpha^x f(u, y)\, du\right] - \int_a^\alpha f(u, \beta)\, du \\ &= \int_a^\alpha \big[f(u, y) - f(u, \beta)\big]\, du + \int_\alpha^x f(u, y)\, du. \end{aligned}$$

By uniform continuity of f, whenever $|y - \beta| < \delta$ we know that $|f(u, y) - f(u, \beta)| < \epsilon$. Hence

$$\left|\int_a^\alpha |f(u, y) - f(u, \beta)|\, du\right| < \left|\int_a^\alpha \epsilon\, du\right| = (\alpha - a)\epsilon < (b - a)\epsilon.$$

By the Extreme Value Theorem, function f with y fixed is bounded from above by some constant $M > 0$. By construction $|x - \alpha| < \delta \leq \epsilon$ and so

$$\left|\int_\alpha^x f(u, y)\, du\right| \leq \left|\int_\alpha^x M\, du\right| = M|x - \alpha| < M\epsilon.$$

The triangle inequality gives the following.

$$\big|h(x, y) - h(\alpha, \beta)\big| \leq \left|\int_a^\alpha \big|f(u, y) - f(u, \beta)\big|\, du\right| + \left|\int_\alpha^x f(u, y)\, du\right| < (b - a + M)\epsilon$$

Since $(x, y) \in [a, b] \times [c, d]$ was arbitrary, we conclude that h is continuous. □

Switching operators

We now consider the question of switching partial derivatives and integrals. We have already seen the switching of a partial derivative with an integral called *differentiation under the integral sign*. This is also called the **Leibniz integral rule**.

Theorem 76 (Leibniz Integral Rule)**.** Let f and $\partial_2 f$ be continuous on $[a, b] \times [c, d]$. The function $g : y \mapsto \int_a^b f(u, y)\, du$ is differentiable with

$$g'(y) = \int_a^b \partial_2 f(u, y)\, du.$$

Proof. Let $\epsilon > 0$; since $\partial_2 f$ is continuous on $[a, b]\times[c, d]$, we know that $\partial_2 f$ is uniformly continuous and there is some $\delta(\epsilon) > 0$ that satisfies uniform continuity of f. We will show that for each nonzero α with $|\alpha| < \delta(\epsilon)$ we have

$$\left|\frac{g(y + \alpha) - g(y)}{\alpha} - \int_a^b \partial_2 f(u, y)\, du\right| < \beta \cdot \epsilon$$

for some constant β. By linearity of the integral,

$$\frac{g(y+\alpha)-g(y)}{\alpha} = \frac{1}{\alpha}\left(\int_a^b f(u,y+\alpha)\,du - \int_a^b f(u,y)\,du\right) = \int_a^b \frac{f(u,y+\alpha)-f(u,y)}{\alpha}\,du$$

and so

$$\frac{g(y+\alpha)-g(y)}{\alpha} - \int_a^b \partial_2 f(u,y)\,du = \int_a^b \left[\frac{f(u,y+\alpha)-f(u,y)}{\alpha} - \partial_2 f(u,y)\right]\,du.$$

We want to turn the quotient $\frac{f(u,y+\alpha)-f(u,y)}{\alpha}$ into a partial derivative. Let $I := (x, x+\alpha)$ if $\alpha > 0$ and $I := (x+\alpha, x)$ otherwise. We know from the Mean Value Theorem (Theorem 40) that if f is differentiable on I, then there will be a point $p \in I$ where the velocity will attain the average velocity $\frac{f(u,y+\alpha)-f(u,y)}{\alpha}$. Thus

$$\int_a^b \left[\frac{f(u,y+\alpha)-f(u,y)}{\alpha} - \partial_2 f(u,y)\right]\,du = \int_a^b \left[\partial_2 f(u,p) - \partial_2 f(u,y)\right]\,du.$$

By uniform continuity of $\partial_2 f$ we know that $\left|\partial_2 f(p,y) - \partial_2 f(u,y)\right| < \epsilon$ whenever $(p,y) \in (u-\delta, u+\delta)\times(y-\delta, y+\delta)$, which is satisfied. Set $\beta := (b-a)$ and we obtain the desired result.

$$\left|\frac{g(y+\alpha)-g(y)}{\alpha} - \int_a^b \partial_2 f(u,y)\,du\right| < \left|\int_a^b \epsilon\,dy\right| = (b-a)\epsilon = \beta\cdot\epsilon \qquad \square$$

Next, we consider the switching of integrals. We will call this result **Fubini's Theorem** (the name really belongs to a far stronger result).

Theorem 77 (Fubini's Theorem)**.** If f is continuous on $[a,b]\times[c,d]$ then

$$\int_c^d \left(\int_a^b f(u,v)\,du\right)\,dv = \int_a^b \left(\int_c^d f(u,v)\,dv\right)\,du. \tag{6.24}$$

Proof. We will take an approach similar to that of the Second Fundamental Theorem of Calculus. The function $y \mapsto \int_a^b f(u,y)\,du$ is continuous and so the Fundamental Theorem of Calculus gives

$$\partial_2\left[\int_c^y \left(\int_a^b f(u,v)\,du\right)\,dv\right] = \int_a^b f(u,y)\,du. \tag{6.25}$$

The function $(x,y) \mapsto \int_a^x f(u,y)\,du$ is continuous and so we may apply the Fundamental Theorem of Calculus to get $\partial_2 \int_c^y f(u,v)\,dv = f(u,y)$. Integrate both sides over the first variable u on the interval $[a,b]$ and then apply the Leibniz integral rule to pull out the partial derivative to get

$$\partial_2\left[\int_a^b \left(\int_c^y f(u,v)\,dv\right)\,du\right] = \int_a^d f(u,y)\,du. \tag{6.26}$$

The partial derivatives in Equation 6.25 and Equation 6.26 are equal and so they can only differ by a constant γ by Corollary 41 and so

$$\int_c^y \left(\int_a^b f(u,v)\,du \right) dv = \int_a^b \left(\int_c^y f(u,v)\,dv \right) du + \gamma.$$

Taking $y := c$ and using Property (P3) of an integral gives us

$$\underbrace{\int_c^y \left(\int_a^b f(u,v)\,du \right) dv}_{=0} = \int_a^b \underbrace{\left(\int_c^y f(u,v)\,du \right)}_{=0} dv + \gamma.$$

Therefore $\gamma = 0$. Making the substitution of the symbol $y \mapsto d$ gives Equation 6.24. □

That covers switching integrals and switching a partial derivative with an integral. All that is left is to consider the case of switching *mixed* partial derivatives $\partial_1\partial_2$ and $\partial_2\partial_1$. We will use the symbol $\partial_{i,j}$ to mean $\partial_j\partial_i$ (first take the partial derivative with respect to the i-th variable, then take the partial derivative with respect to the j-th variable). This result is called **Clairaut's Theorem**.

Theorem 78 (Clairaut's Theorem)**.** Let $\partial_2 f$ and $\partial_{2,1} f$ be continuous on $[a,b]\times[c,d]$ and let $\partial_1 f(x,c)$ exist for each $x \in (a,b)$. Then both $\partial_1 f$ and $\partial_{1,2} f$ exist on the open rectangle $(a,b)\times(c,d)$ with $\partial_{1,2} f = \partial_{2,1} f$.

Proof. Since $\partial_2 f$ is continuous, the Fundamental Theorem of Calculus gives

$$f(x,y) = \int_c^y \partial_2 f(x,v)\,dv + f(x,c).$$

The partial derivative $\partial_1 f(x,c)$ exists by assumption and the function $y \mapsto \int_c^y \partial_2 f(x,v)$ is differentiable by the Leibniz integral rule. Therefore, $\partial_1 f(x,y)$ exists for each point $(x,y) \in (a,b)\times(c,d)$ and we have

$$\partial_1 f(x,y) = \partial_1 \int_c^y \partial_2 f(x,v)\,dv + \partial_1 f(x,c).$$

Since $\partial_2 f$ and $\partial_{2,1} f$ are continuous, we may apply the Leibniz integral rule to get

$$\partial_1 f(x,y) = \int_c^y \partial_{2,1} f(x,v)\,dv + \partial_1 f(x,c). \tag{6.27}$$

The first term on the right side is differentiable with respect to the second variable v by the Fundamental Theorem of Calculus with $\partial_2 \int_c^y \partial_{2,1} f(x,v)\,dv = \partial_{2,1} f(x,y)$. The second term on the right side is a constant function with respect to the second variable and is thus differentiable with respect to the second variable with $\partial_2 \left[\partial_1 f(x,c)\right] = 0$. Therefore, the left side is differentiable with respect to the second variable, demonstrating that $\partial_{1,2} f$ exists on the open rectangle $(a,b)\times(c,d)$. Taking a partial derivative with respect to the second variable on both sides of Equation 6.27 gives the desired equality

$$\partial_{1,2} f(x,y) = \partial_{2,1} f(x,y).$$ □

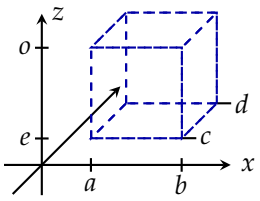


Figure 6.28: An open rectangle in $\mathbb{R}^3$.

The analogue of an open interval in three dimensions is an **open rectangle** $(a,b)\times(c,d)\times(e,o)$, which are the collection of points $(x,y,z)\in\mathbb{R}^3$ such that $x\in(a,b)$, $y\in(c,d)$, and $z\in(e,o)$. The definition of a continuous function defined on $\mathbb{R}^3$ is analogous to the two dimensional case.

If f is a real-valued function defined on an open rectangle R in $\mathbb{R}^3$, we say that $f\in C^1$ on R if $\partial_1 f, \partial_2 f, \partial_3 f$ all exist and are continuous. If in addition $\partial_{i,j} f$ exists for each $i,j\in\{1,2,3\}$ and are all continuous, then we say that $f\in C^2$ on R. In a similar manner, we say that $f\in C^r$ on R if for each $k\in\{1,2,3\}$, $\partial_k f\in C^{r-1}$ on R.

If $f\in C^3$, then we may use Clairaut's Theorem repeatedly to conclude that the mixed partials $\partial_{i,j,k}$ for distinct $i,j,k\in\{1,2,3\}$ all exist and must equal.[9] This is familiar from the algebra of $d\square$, as the sequence i,j,k describes a permutation and each permutation is a product of transpositions. Thus each $\partial_i\partial_j\partial_k f$ can be attained by applying Clairaut's Theorem to the sequence of transpositions that turn σ into the identity permutation in S_3. For example, $\partial_{3,2,1} f = \partial_{1,2,3} f$ because

$$\partial_1\partial_2\partial_3 f = \partial_1\partial_3\partial_2 f = \partial_3\partial_1\partial_2 f = \partial_3\partial_2\partial_1 f. \tag{6.29}$$

This generalizes to $f\in C^r$ where $r>3$, and of course permutation of integrals also works in a similar fashion.

6.5 Grad, Curl, Div, and Laplacian

Forms in $\mathbb{R}^3$

We have been dealing with scalar-valued functions so far.[10] We turn to **vector fields**, which are vector-valued functions $f: R\to\mathbb{R}^n$, where R is some open rectangle in $\mathbb{R}^n$. For what follows we always take $n=3$. The simplest nontrivial vector field we could consider are the following.

$$p:\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\mapsto\begin{pmatrix}x_1\\0\\0\end{pmatrix} \qquad q:\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\mapsto\begin{pmatrix}0\\x_2\\0\end{pmatrix} \qquad r:\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\mapsto\begin{pmatrix}0\\0\\x_3\end{pmatrix}$$

These vector fields simply zero out two components of each input vector $x\in\mathbb{R}^3$. The derivative of p is given by $p'=\begin{pmatrix}1&0&0\end{pmatrix}$. Notice the derivative is (a function and) *not* a vector. Similarly, $q'=\begin{pmatrix}0&1&0\end{pmatrix}$ and $r'=\begin{pmatrix}0&0&1\end{pmatrix}$. The derivatives p', q', and r' are constant functions that are essentially the standard basis vectors e_1, e_2, and e_3, respectively. So we can write each vector $x\in\mathbb{R}^3$

[9] In analogy to the two dimensional case, the notation $\partial_{i,j,k}$ means $\partial_k\partial_j\partial_i$.

[10] This section a little algebra-heavy. Feel free to skip ahead to Section 6.7.

as the linear combination $x = x_1 p' + x_2 q' + x_3 r'$. We denote p', q', and r' by the symbols dx_1, dx_2, and dx_3 respectively, so that each vector $x \in \mathbb{R}^3$ can be written as

$$x = \sum_{i=1}^{3} x_i \, dx_i. \tag{6.30}$$

Summing two vectors $u, v \in \mathbb{R}^3$ and applying a scalar multiplication by c are straightforward:

$$u + v = \sum_{i=1}^{3} u_i \, dx_i + \sum_{i=1}^{3} v_i \, dx_i = \sum_{i=1}^{3} (u_i + v_i) \, dx_i \qquad c \cdot u = c \sum_{i=1}^{3} u_i \, dx_i = \sum_{i=1}^{3} c \cdot u_i \, dx_i.$$

How about multiplying two vectors? We do not know of a way to multiply two vectors in $\mathbb{R}^3$. However, we have worked with vector multiplication in $\mathbb{R}^2$ through the complex numbers and dual numbers (Section 5.3). In fact, we have already extended the dual numbers by using it to obtaining the Change of Variables formula in two and three dimensions.

So let us try and apply the algebra of $d\square$ to vector multiplication.[11] We will denote the vector multiplication by the symbol $\wedge$ called a **wedge product**. As we have done with the algebra of $d\square$, we will build up from simple vectors in $\mathbb{R}^3$. Recall that $d\square d\square = 0$. Therefore, if two vectors u and v only have a single i-th component, then

$$u \wedge v := (u_i \, dx_i) \wedge (v_i \, dx_i) = (u_i v_i) \, dx_i \wedge dx_i = 0.$$

As we want the group action of scalar multiplication to work nicely with the $\wedge$ product, we require that $(u_i \, dx_i) \wedge (v_j \, dx_j) = (u_i v_j) \, dx_i \wedge dx_j$. For example, $u_i \wedge v_i = u_i v_i$.

On the other hand, if $i \neq j$ and $v = v_j dx_j$, we use the fact that $d\spadesuit \, d\clubsuit = -\, d\clubsuit \, d\spadesuit$ to get

$$u \wedge v := (u_i \, dx_i) \wedge (v_j \, dx_j) = (u_i v_j) \, dx_i \wedge dx_j = -(u_i v_j) \, dx_j \wedge dx_i.$$

For $i, j, k \in \{1, 2, 3\}$ and $u = u_i \, dx_i$, $v = v_j \, dx_j$, $w = w_k \, dx_k$ we have

$$(u \wedge v) \wedge w = \big[(u_i v_j) \, dx_i \wedge dx_j\big] \wedge w_k \, dx_k = (u_i v_j w_k) \, dx_i \wedge dx_j \wedge dx_k = u \wedge (v \wedge w)$$

Of course if at least one pair of i, j, k are equal in the above then $(u \wedge v) \wedge w = w \wedge (u \wedge v) = 0$.

Manipulations with dual numbers gave us $(d\clubsuit + d\spadesuit) \, d\heartsuit = d\clubsuit \, d\heartsuit + d\spadesuit \, d\blacktriangledown$ and $d\heartsuit(d\clubsuit + d\spadesuit) = d\heartsuit \, d\clubsuit + d\heartsuit \, d\spadesuit$. Hence if ω_1, ω_2 are forms then the following distributive laws hold.

$$(dx_i + dx_j) \wedge dx_k = dx_i \wedge dx_k + dx_j \wedge dx_k \qquad dx_i \wedge (dx_j + dx_k) = dx_i \wedge dx_j + dx_i \wedge dx_k$$

Adding a coefficient in front of the symbols gives us the following distributive laws.

$$(f \, dx_i + g \, dx_j) \wedge h \, dx_k = (fg) \, dx_i \wedge dx_k + (fh) \, dx_j \wedge dx_k$$
$$f \, dx_i \wedge (g \, dx_j + h \, dx_k) = (fg) \, dx_i \wedge dx_j + (fh) \, dx_i \wedge dx_k$$

To distinguish dx_1, dx_2, and dx_3 from our regular vectors, we will call our objects using the following terminology. Let f, g, and h be real-valued functions defined on a rectangle in $\mathbb{R}^3$.

[11] This is why we denoted the functions p', q', and r' by the symbols dx_1, dx_2, and dx_3.

(a) A **1-form** denotes objects of the form $f\,dx_1 + g\,dx_2 + h\,dx_3$.
(b) A **2-form** denotes objects of the form $f\,dx_1 \wedge dx_2 + g\,dx_1 \wedge dx_3 + h\,dx_2 \wedge dx_3$.
(c) A **3-form** denotes objects of the form $f\,dx_1 \wedge dx_2 \wedge dx_3$.

Hence a k**-form** (for $k \in \{1,2,3\}$) is the linear combination of terms that are formed by taking k number of $\wedge$ products. A **0-form** will thus simply denote a real-valued function f.

Repeating the previous manipulations tells us that if ω and η are k-forms then

$$(\omega + \eta) \wedge h\,dx_i = \omega \wedge (h\,dx_i) + \eta \wedge (h\,dx_i)$$
$$(\omega + \eta) \wedge h\,dx_i \wedge dx_j = \omega \wedge (h\,dx_i \wedge dx_j) + \eta \wedge (h\,dx_i \wedge dx_j).$$

A 1-form has a simple interpretation, as it is equivalent to a vector field by Equation 6.30. Indeed, each vector field $f := (f_1, f_2, f_3)$ can be turned into a 1-form $\omega := f_1\,dx_1 + f_2\,dx_2 + f_3\,dx$. Conversely, each 1-form $\omega := g_1\,dx_1 + g_2\,dx_2 + g_3\,dx$ can turned into the vector field $g := (g_1, g_2, g_3)$.

Exterior derivative

Recall the key equation from our discussion of the Change of Variables formula,

$$df = \partial_x f\,dx + \partial_y f\,dy + \partial_z f\,dz.$$

This is an application of a "derivative" d on a 0-form to get a 1-form. We will thus define the **exterior derivative** of a 0-form f to be the 1-form df given by the following.[12]

$$df := \sum_{i=1}^{3} \partial_i f\,dx_i$$

Thus if $f : (x_1, x_2, x_3) \mapsto x_1^2 x_3 + x_2 x_3$ then the exterior derivative of the 0-form f is the 1-form $df = 2x_1x_3\,dx_1 + x_3\,dx_2 + (x_1^2 + x_2)\,dx_3$. Similarly, for the 0-form $x_i : (x_1, x_2, x_3) \mapsto x_i$ the exterior derivative $d(x_i) = dx_i$.

The exterior derivative of a 0-form is given by taking derivates, tacking on the symbols $d\square$ and summing. Analogously, we will define the **exterior derivative** of a 1-form ω_f corresponding to the vector field $f := (f_1, f_2, f_3)$ as follows.[13]

$$d\omega_f = \sum_{1 \le i \le 3} df_i \wedge dx_i$$

We can evaluate the above explicitly using the fact that $df_i := \sum_{j=1}^{3} \partial_j f_i\,dx_j$ and $dx_i \wedge dx_j = -dx_j \wedge dx_i$.

$$\begin{aligned} d\omega_f &= \sum_{1 \le i \le 3} df_i \wedge dx_i = \sum_{1 \le i \le 3} \sum_{j=1}^{3} \partial_j f_i\,dx_j \wedge dx_i \\ &= \cancel{\partial_1 f_1\,dx_1 \wedge dx_1} + \partial_2 f_1\,dx_2 \wedge dx_1 + \partial_3 f_1\,dx_3 \wedge dx_1 \\ &\quad + \partial_1 f_2\,dx_1 \wedge dx_2 + \cancel{\partial_2 f_2\,dx_2 \wedge dx_2} + \partial_3 f_2\,dx_3 \wedge dx_2 \\ &\quad + \partial_1 f_3\,dx_1 \wedge dx_3 + \partial_2 f_3\,dx_2 \wedge dx_3 + \cancel{\partial_3 f_3\,dx_3 \wedge dx_3} \\ &= \left(\partial_2 f_3 - \partial_3 f_2\right)\,dx_2 \wedge dx_3 + \left(\partial_3 f_1 - \partial_1 f_3\right)\,dx_3 \wedge dx_1 + \left(\partial_1 f_2 - \partial_2 f_1\right)\,dx_1 \wedge dx_2 \end{aligned}$$

[12] We will always assume that $f \in C^r$ for sufficiently high r. That is, enough continuous partial derivatives exist.
[13] (1) Take an exterior derivative of each f_i (2) do a wedge product with dx_i and (3) sum.

Challenge 62

(a) The ordering of the terms above may seem out of order, but it is written to be easier to recognize. Let $\star dx_1 := dx_2 \wedge dx_3$, $\star dx_2 := dx_3 \wedge dx_1$, and $\star dx_3 := dx_1 \wedge dx_2$. Check that $dx_i \wedge \star dx_i = dx_1 \wedge dx_2 \wedge dx_3$. Use the Laplace expansion over row 1 to verify that

$$d\omega_f = \det\begin{pmatrix} \star dx_1 & \star dx_2 & \star dx_3 \\ \partial_1 & \partial_2 & \partial_3 \\ f_1 & f_2 & f_3 \end{pmatrix}.$$

(b) Use the fact that $\partial_i \partial_j h = \partial_j \partial_i h$ to show that if f is a 0-form then $d(df) = 0$. [*Hint:* it suffices to check that $d\omega_g = 0$ where $\omega_g := \sum_1^3 g_i \, dx_i$ with $g_i := \partial_i f$.]

Let f_{12}, f_{13}, and f_{23} be real-valued functions defined on a rectangle in $\mathbb{R}^3$ with the necessary continuous partial derivatives and let $f := (f_{12}, f_{13}, f_{23})$. We define the **exterior derivative** of a 2-form $\eta_f := f_{12}\, dx_1 \wedge dx_2 + f_{13}\, dx_1 \wedge dx_3 + f_{23}\, dx_2 \wedge dx_3$, denoted $d\eta_f$ by the following.

$$d\eta_f := \sum_{1 \le i < j \le 3} df_{ij} \wedge dx_i \wedge dx_j$$

The above must reduce down to a single term because all $dx_i \wedge dx_j \wedge dx_k$ are equivalent (up to a sign) or 0 in $\mathbb{R}^3$. We do the simplification as we did with the exterior derivative of a 1-form.

$$\begin{aligned} d\eta_f &:= \sum_{1 \le i < j \le 3} df_{ij} \wedge dx_i \wedge dx_j = \sum_{1 \le i < j \le 3} \sum_{k=1}^{3} \partial_k f_{ij}\, dx_k \wedge dx_i \wedge dx_j \\ &= \partial_3 f_{12}\, dx_3 \wedge dx_1 \wedge dx_2 + \partial_2 f_{13}\, dx_2 \wedge dx_1 \wedge dx_3 + \partial_1 f_{23}\, dx_1 \wedge dx_2 \wedge dx_3 \\ &= \big(\partial_1 f_{23} - \partial_2 f_{13} + \partial_3 f_{12}\big)\, dx_1 \wedge dx_2 \wedge dx_3 \end{aligned}$$

The reason we have two indices on f_{ij} is to help us remember that the function pairs with $dx_i \wedge dx_j$. We further simplify into a single index. A 2-form in $\mathbb{R}^3$ can be written as $\eta_g := g_1 \star dx_1 + g_2 \star dx_2 + g_3 \star dx_3$. This is equivalent to our previous representation $\eta_f := \sum_{1 \le i < j \le 3} f_{ij}\, dx_i \wedge dx_j$ with $g_1 := f_{23}$, $g_2 = -f_{13}$, and $g_3 := f_{12}$. Using these substitutions, the exterior derivative of η_g is then

$$d\eta_g = \big(\partial_1 g_1 + \partial_2 g_2 + \partial_3 g_3\big)\, dx_1 \wedge dx_2 \wedge dx_3.$$

The **exterior derivative** of a 3-form $\alpha := f_{123}\, dx_1 \wedge dx_2 \wedge dx_3$ is defined analogously as

$$d\alpha := \sum_{1 \le i < j < k \le 3} df_{ijk} \wedge dx_i \wedge dx_j \wedge dx_k.$$

Proposition 79. Each k-form ω in $\mathbb{R}^3$ satisfies $d(d\omega) = 0$.

Proof. The case for $k = 0$ was left for you in Challenge 62. Let ω_f be a 1-form corresponding to the vector field $f := (f_1, f_2, f_3)$. Define

$$g_1 := \big(\partial_2 f_3 - \partial_3 f_2\big) \qquad g_2 := \big(\partial_3 f_1 - \partial_1 f_3\big) \qquad g_3 := \big(\partial_1 f_2 - \partial_2 f_1\big)$$

so that

$$\partial_1 g_1 = \partial_1\partial_2 f_3 - \partial_1\partial_3 f_2 \qquad \partial_2 g_2 = \partial_2\partial_3 f_1 - \partial_2\partial_1 f_3 \qquad \partial_3 g_3 := \partial_3\partial_1 f_2 - \partial_3\partial_2 f_1.$$

Then

$$\begin{aligned} d(d\omega_f) &= d\left[\left(\partial_2 f_3 - \partial_3 f_2\right)\, dx_2 \wedge dx_3 + \left(\partial_3 f_1 - \partial_1 f_3\right)\, dx_3 \wedge dx_1 + \left(\partial_1 f_2 - \partial_2 f_1\right)\, dx_1 \wedge dx_2\right] \\ &= d\left[g_1\, dx_2 \wedge dx_3 + g_2\, dx_3 \wedge dx_1 + g_3\, dx_1 \wedge dx_2\right] \\ &= \left[\partial_1 g_1 + \partial_2 g_2 + \partial_3 g_3\right]\, dx_1 \wedge dx_2 \wedge dx_3 \\ &= \left[\partial_1\partial_2 f_3 - \partial_1\partial_3 f_2 + \partial_2\partial_3 f_1 - \partial_2\partial_1 f_3 + \partial_3\partial_1 f_2 - \partial_3\partial_2 f_1\right]\, dx_1 \wedge dx_2 \wedge dx_3. \end{aligned}$$

By Clairaut's Theorem (equality of mixed partials) the terms in the brackets vanish and so $d(d\omega_f) = 0$. The exterior derivative of a 3-form $\alpha := f_{123}\, dx_1 \wedge dx_2 \wedge dx_3$ in $\mathbb{R}^3$ is always zero:

$$d\alpha := \sum_{1\le i<j<k\le 3} df_{ijk} \wedge dx_i \wedge dx_j \wedge dx_k = \sum_{1\le i<j<k\le 3} \sum_{l=1}^{3} \partial_l f_{ijk} dx_l \wedge dx_i \wedge dx_j \wedge dx_k = 0$$

and since the exterior derivative of a 2-form η was shown to be a 3-form, $d(d\eta) = 0$. □

We cannot have 4-forms in $\mathbb{R}^3$, but in $\mathbb{R}^n$, we can have r-forms for $r \in \{0, 1, 2, \ldots, n\}$. If $n \ge 4$ we can define the **exterior derivative** of a 4-form $\alpha := \sum_{1\le i<j<k<l\le n} f_{ijkl}\, dx_i \wedge dx_j \wedge dx_k \wedge dx_l$ by

$$d\alpha := \sum_{1\le i<j<k<l\le n} df_{ijkl} \wedge dx_i \wedge dx_j \wedge dx_k \wedge dx_l.$$

where the exterior derivative of a 0-form is $df := \sum_{i=1}^{n} \partial_i f\, dx_i$. The **exterior derivative** of a k-form $\alpha := \sum_{1\le i_1<\cdots<i_k\le n} f_{i_1\cdots i_k}\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}$ for $k \le n$ should then be defined as the following.

$$d\alpha := \sum_{1\le i_1<\cdots<i_k\le n} df_{i_1\cdots i_k} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \tag{6.31}$$

To recap, in $\mathbb{R}^3$ the exterior derivative of a 0-form f is

$$df = \partial_1 f\, dx_1 + \partial_2 f\, dx_2 + \partial_3 f\, dx_3. \tag{6.32}$$

The exterior derivative of a 1-form $\omega_f := \sum_{i=1}^{3} f_i\, dx_i$ is

$$d\omega_f = \left(\partial_2 f_3 - \partial_3 f_2\right)\, dx_2 \wedge dx_3 + \left(\partial_3 f_1 - \partial_1 f_3\right)\, dx_3 \wedge dx_1 + \left(\partial_1 f_2 - \partial_2 f_1\right)\, dx_1 \wedge dx_2. \tag{6.33}$$

Finally, the exterior derivative of a 2-form $\eta_f := \sum_{i=1}^{3} f_i \star dx_i$ is

$$d\eta_f = \left(\partial_1 f_1 + \partial_2 f_2 + \partial_3 f_3\right)\, dx_1 \wedge dx_2 \wedge dx_3. \tag{6.34}$$

The coefficients of 1-form df vectorized is the gradient of a real-valued function f, written ∇f. For a vector field $\boldsymbol{f} = (f_1, f_2, f_3)$, the coefficients of 2-form $d\omega_f$ vectorized is called the **curl** of $\boldsymbol{f}$ and is written $\nabla \times \boldsymbol{f}$. For a vector field $\boldsymbol{f} = (f_1, f_2, f_3)$, the coefficient of the 3-form $d\eta_f$ is called the **divergence** of $\boldsymbol{f}$ and is written $\nabla \cdot \boldsymbol{f}$.

$$\nabla f := \begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \partial_3 f \end{pmatrix} \qquad \nabla \times \boldsymbol{f} := \begin{pmatrix} \partial_2 f_3 - \partial_3 f_2 \\ \partial_3 f_1 - \partial_1 f_3 \\ \partial_1 f_2 - \partial_2 f_1 \end{pmatrix} \qquad \nabla \cdot \boldsymbol{f} := \partial_1 f_1 + \partial_2 f_2 + \partial_3 f_3$$

Challenge 63 Use the fact that $d(d\alpha) = 0$ to conclude that $\nabla \times (\nabla f) = 0$ and $\nabla \cdot (\nabla \times \boldsymbol{f}) = 0$.

The notation $\star dx_i$ was defined by the equation $dx_i \wedge \star dx_i = dx_1 \wedge dx_2 \wedge dx_3$. We now apply the $\star$ symbol (**Hodge star**) to k-forms in $\mathbb{R}^3$. If $\alpha := dx_i \wedge dx_j$ for distinct i, j, then $\star\alpha$ is defined by the equation $\alpha \wedge \star\alpha = dx_1 \wedge dx_2 \wedge dx_3$. Thus

$$\star(dx_1 \wedge dx_2) = dx_3 \qquad \star(dx_2 \wedge dx_3) = dx_1 \qquad \star(dx_3 \wedge dx_1) = dx_2.$$

A Hodge star only acts on the objects dx_i. For example,

$$\star\left(f_1\, dx_1 \wedge dx_2 + f_2\, dx_3 \wedge dx_1\right) = f_1 \star(dx_1 \wedge dx_2) + f_2 \star(dx_3 \wedge dx_1).$$

As a 0-form has no dx_i and a 3-form has all the symbols dx_j, for each 0-form f and 3-form α_g,

$$\star f = f\, dx_1 \wedge dx_2 \wedge dx_3 \qquad \star\alpha_g = g.$$

Challenge 64 Show that if $\omega_{\boldsymbol{f}}$ is a 1-form where $\boldsymbol{f} := (f_1, f_2, f_3)$ then the following hold.

$$\star d\omega_{\boldsymbol{f}} = \sum_i \left[\nabla \times \boldsymbol{f}\right]_i\, dx_i := \omega_{\nabla \times \boldsymbol{f}} \qquad \star d\star\omega_{\boldsymbol{f}} = \sum_i \left[\nabla \cdot \boldsymbol{f}\right]\, dx_i := \omega_{\nabla \cdot \boldsymbol{f}}$$

Challenge 65 How about using the wedge product to define a vector multiplication? By Equation 6.30 a 1-form corresponds to a vector. But taking the wedge product of 1-forms gives us a 2-form, so we must take a Hodge star to get it back to a 1-form. Let α be a 1-form corresponding to the vector $(a_1, a_2, a_3)^\intercal$ and let β be a 1-form corresponding to the vector $(b_1, b_2, b_3)^\intercal$. Check that

$$\star(\alpha \wedge \beta) = \det\begin{pmatrix} dx_1 & dx_2 & dx_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix}.$$

We write $\alpha \times \beta := \star(\alpha \wedge \beta)$. Let v and w be the vectors corresponding to 1-forms α and β, respectively. Then the vector corresponding to $\omega_{\alpha\times\beta}$ is called the **cross product** of v and w. Check that $v \times w = -w \times v$ and $v \times w = 0$ if $w = c \cdot w$ for some constant c. Let $X := (x, y, z)$ and $P := (-i\hbar\partial_1, -i\hbar\partial_2, -i\hbar\partial_3)$. Show that the **angular momentum** operator in three dimensions $L := L_x + L_y + L_z$ is equal to $L = X \times P$ (relevant definitions given in Page 127).

Laplacian

We saw that there are three fundamental operators in $\mathbb{R}^3$: the *gradient* of a real-valued function, the *curl* of a vector field, and the *divergence* of a vector field. Let us see what happens when we combine these together.

There are nine possible pairs, but not all of them are permissible. For example, the curl of a divergence of a vector field makes no sense because the divergence of a vector field is a real-valued function, and we cannot take the curl of a real-valued function. Similarly we cannot take the divergence of a divergence of a vector field because the divergence is not defined for a real-valued function in $\mathbb{R}^3$. As a gradient is only defined for a real-valued function, it also makes no sense to consider the gradient of a gradient and the gradient of a curl.

Therefore, there are a grand total of five possibilities: gradient of a divergence, curl of a gradient, curl of a curl, divergence of a gradient, and divergence of a curl. In symbols, for a real-valued function f and a vector field $\boldsymbol{f}$, the possibilities are the following.

$$\nabla(\nabla \cdot \boldsymbol{f}) \qquad \nabla\times(\nabla f) \qquad \nabla\times(\nabla \times \boldsymbol{f}) \qquad \nabla\cdot(\nabla f) \qquad \nabla\cdot(\nabla \times \boldsymbol{f})$$

We know from Challenge 63 that the curl of a gradient and the divergence of a curl always vanish. This is an immediate consequence of the fact that the exterior derivative of an exterior derivative vanishes. Alternatively, we can invoke Clairaut's Theorem directly as shown below.

$$\nabla \times (\nabla f) = \nabla \times \begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \partial_3 f \end{pmatrix} = \begin{pmatrix} \partial_2\partial_3 f - \partial_3\partial_2 f \\ \partial_3\partial_1 f - \partial_1\partial_3 f \\ \partial_1\partial_2 f - \partial_2\partial_1 f \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\nabla \cdot (\nabla \times f) = \nabla \cdot \begin{pmatrix} \partial_2 f_3 - \partial_3 f_2 \\ \partial_3 f_1 - \partial_1 f_3 \\ \partial_1 f_2 - \partial_2 f_1 \end{pmatrix} = \partial_1(\partial_2 f_3 - \partial_3 f_2) + \partial_2(\partial_3 f_1 - \partial_1 f_3) + \partial_3(\partial_1 f_2 - \partial_2 f_1) = 0$$

Challenge 66 For a real-valued function f and a vector field $f := (f_1, f_2, f_3)$, show that

$$\nabla(\nabla \cdot f) = \begin{pmatrix} \partial_{1,1} f_1 + \partial_{2,1} f_2 + \partial_{3,1} f_3 \\ \partial_{1,2} f_1 + \partial_{2,2} f_2 + \partial_{3,2} f_3 \\ \partial_{1,3} f_1 + \partial_{2,3} f_2 + \partial_{3,3} f_3 \end{pmatrix} \qquad \nabla \times (\nabla \times f) = \begin{pmatrix} \partial_{1,2} f_2 - \partial_{2,2} f_1 - \partial_{3,3} f_1 + \partial_{1,3} f_3 \\ \partial_{2,3} f_3 - \partial_{3,3} f_2 - \partial_{1,1} f_2 + \partial_{2,1} f_1 \\ \partial_{3,1} f_1 - \partial_{1,1} f_3 - \partial_{2,2} f_3 + \partial_{3,2} f_2 \end{pmatrix} \tag{6.35}$$

The gradient of a divergence and a curl of a curl do not appear to have a nice form. However, calculating the divergence of gradient shows that it is none other than the Laplacian

$$\nabla \cdot (\nabla f) = \nabla \cdot \begin{pmatrix} \partial_1 f \\ \partial_2 f \\ \partial_3 f \end{pmatrix} = \partial_{1,1} f + \partial_{2,2} f + \partial_{3,3} f = \nabla^2 f.$$

We have seen the Laplacian in the three dimensional Schrödinger Equation $i\hbar\partial_t \Psi = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi$.

If we subtract the curl of a curl from a gradient of a divergence using Equation 6.35, we have

$$\nabla(\nabla \cdot f) - \nabla \times (\nabla \times f) = \begin{pmatrix} \partial_{1,1} f_1 + \partial_{2,2} f_1 + \partial_{3,3} f_1 \\ \partial_{1,1} f_2 + \partial_{2,2} f_2 + \partial_{3,3} f_2 \\ \partial_{1,1} f_3 + \partial_{2,2} f_3 + \partial_{3,3} f_3 \end{pmatrix} \tag{6.36}$$

which also looks like a Laplacian, albeit for a vector field. We will thus define the **Laplacian of a vector field** $\nabla^2 f$ by Equation 6.36.

6.6 Integral Theorems

Oriented integrals

As we have talked about derivatives, it is natural to talk about integration.[14] Consider a 3-form $\omega_f := f(x)\, dx_1 \wedge dx_2 \wedge dx_3$ defined on an open rectangle U in $\mathbb{R}^3$. For each closed rectangle R contained in U, we define the **(oriented) integral** of ω_f over R as

$$\int_R \omega_f := \iiint_R f(x)\, dx_1\, dx_2\, dx_3 = \int_R f.$$

[14] This section is not essential for the rest of the book. Fell free to skip ahead to Section 6.7.

In general, a 3-form in $\mathbb{R}^3$ can take the form $f(x)\,dx_i \wedge dx_j \wedge dx_k$, where distinct $i, j, k \in \{1, 2, 3\}$ are in any order. However, this is not a problem, as we can permute the indices into the order $dx_1 \wedge dx_2 \wedge dx_3$ by adding in the necessary sign. For example,

$$f(x)\,dx_2 \wedge dx_3 \wedge dx_1 = -f(x)\,dx_2 \wedge dx_1 \wedge dx_3 = f(x)\,dx_1 \wedge dx_2 \wedge dx_3.$$

Therefore,

$$\iiint_R f(x)\,dx_i\,dx_j\,dx_k = \operatorname{sgn}\begin{pmatrix} i & j & k \\ 1 & 2 & 3 \end{pmatrix} \int_R f$$

where "sgn σ" of a permutation σ is the sign of permutation σ.

As an example, suppose we have a 3-form $\omega_f := f(x)\,dx_2 \wedge dx_3 \wedge dx_1$ defined on an open rectangle U in $\mathbb{R}^3$. If we want to integrate ω_f over a closed rectangle $R := [a, b] \times [c, d] \times [e, o]$ contained in the rectangle U, we calculate the integral

$$\int_R \omega_f = \operatorname{sgn}\begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \end{pmatrix} \int_e^o \left(\int_c^d \left(\int_a^b f(x)\,dx_1 \right) dx_2 \right) dx_3.$$

We may use Fubini's Theorem to switch the order of integration if it makes life easier. For example,

$$\int_e^o \left(\int_c^d \left(\int_a^b f(x)\,dx_1 \right) dx_2 \right) dx_3 = \int_e^o \left(\int_a^b \left(\int_c^d f(x)\,dx_2 \right) dx_1 \right) dx_3.$$

Let us take a step back and think about what an oriented integral is in one dimension. The analogue of a closed rectangle is of course a closed interval $I := [a, b]$, where $a < b$. So if we wish to integrate a 1-form $f(x)\,dx$ in $\mathbb{R}$, then

$$\int_I f = \int_a^b f(x)\,dx.$$

Recall that we can also integrate from b to a. This was the analogue of calculating displacement of a marathon runner by running the video *backwards*. Since our runner is now running backwards, we need a minus sign. Thus if $J := [b, a]$ with $b < a$, then

$$\int_J f = -\int_I f = -\int_a^b f(x)\,dx.$$

If function f has an antiderivative, we can use the Fundamental Theorem of Calculus to evaluate the integral.

In fact, we can be more direct. Adapting the definition of an exterior derivative of a 0-form in $\mathbb{R}^3$ into the one-dimensional case tells us that if f is a differentiable 0-form in $\mathbb{R}$, then $df = f'(x)\,dx$. We can integrate this over the interval $I := [a, b]$ to get

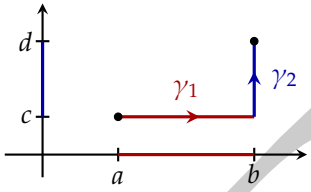
$$\int_I df = \int_I f'(x)\,dx = \int_a^b f'(x)\,dx.$$

(Of course if $b > a$, we can integrate over the interval $[b, a]$ and add in a minus sign.) The Fundamental Theorem of Calculus then tells us that

$$\int_a^b f'(x)\,dx = f(b) - f(a).$$

With the notable exception of the Gaussian integrals, all functions we integrated had an antiderivative. Thus it will be quite sufficient to only consider integrals of exterior derivatives.

Now let us try the same, but in two dimensions. Suppose we have two closed intervals $I := [a, b]$ and $J := [c, d]$. We can place these intervals on the plane as shown in the figure below. The interval I becomes a set $\gamma_1 := I \times \{c\}$ on the plane and the interval J becomes a set $\gamma_2 := \{b\} \times J$ on the plane. If we consider these sets as paths, then we can connect the two paths and call their sum γ. Of course, we not actually taking a sum $\gamma_1 + \gamma_2 = \gamma$, we are simply noting that joining the two paths makes one path.



If we have a 0-form $f \in C^1$ defined on the plane, then its exterior derivative is given by

$$df = \partial_1 f(x, y)\, dx + \partial_2 f(x, y)\, dy.$$

We can take the integral of this 1-form over the path γ using the linearity of the integral and considering the two pieces of the path separately as follows.

$$\int_\gamma df = \int_{\gamma_1+\gamma_2} \partial_1 f(x, y)\, dx + \partial_2 f(x, y)\, dy = \int_{\gamma_1} \partial_1 f(x, y)\, dx + \int_{\gamma_2} \partial_2 f(x, y)\, dy$$

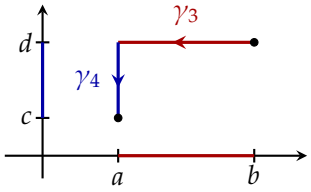
On the path γ_1, the y-value is fixed at c. Furthermore, $f \in C^1$ and so $\partial_1 f(x, y)$ is continuous. We can apply the Fundamental Theorem of Calculus to get

$$\int_{\gamma_1} \partial_1 f(x, y)\, dx = \int_a^b \partial_1 f(x, c)\, dx = f(b, c) - f(a, c).$$

Repeating for the path γ_1 gives $\int_{\gamma_2} \partial_2 f(x, y)\, dx = f(b, d) - f(b, c)$. Therefore,

$$\int_\gamma df = f(b, c) - f(a, c) + f(b, d) - f(b, c) = f(b, d) - f(a, c)$$

which is just like the Fundamental Theorem of Calculus, except at the plane!



Challenge 67

(a) Consider the paths $\gamma_3 := I \times \{d\}$ and $\gamma_4 := \{a\} \times J$ with the orientations shown above. Notice the orientations are *opposite* from the usual orientation. Let $\bar{\gamma} := \gamma_3 + \gamma_4$ be the path formed by joining the paths γ_3 and γ_4. Show that $\int_{\bar{\gamma}} df = f(a,c) - f(b,d)$, just as we would expect.

(b) Let $\tilde{\gamma} := \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$ (see Figure 6.37). Show that $\int_{\tilde{\gamma}} df = 0$.

Green's Theorem

We upgrade from integrating over paths to integrating over rectangles. Merging the paths γ_1, γ_2, γ_3, γ_4 into one path γ encloses a surface (rectangle) $S := [a,b] \times [c,d]$ (Figure 6.37). The orientation of a surface is determined by the orientation of the enclosing path. We will take the orientation of γ to be counterclockwise so that beginning from the bottom left corner, we first move to the right along the orientation of the x-axis, then move up along the orientation of the y-axis.

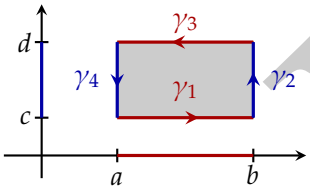


Figure 6.37: A surface S enclosed in a closed path $\gamma := \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$.

Suppose we have a 1-form $\omega := f\,dx + g\,dy$ with $f, g \in C^1$ defined in an open rectangle U in $\mathbb{R}^2$ that contains S. The exterior derivative of ω is as follows.

$$d\omega = df \wedge dx + dg \wedge dy = \cancel{\partial_x f\, dx \wedge dx} + \partial_y f\, dy \wedge dx + \partial_x g\, dx \wedge dy + \cancel{\partial_y g\, dy \wedge dy} = \left(\partial_x g - \partial_y f\right) dx \wedge dy$$

Adapting the definition of an oriented integral of a 3-form in $\mathbb{R}^3$ to an oriented integral of a 2-form in $\mathbb{R}^2$ gives the following.

$$\int_S d\omega = \int_c^d \left(\int_a^b \left[\partial_x g - \partial_y f\right] dx \right) dy$$

Let us use the linearity of the integral to calculate each term in the integral separately. By the Fundamental Theorem of Calculus and the linearity of the integral,

$$\int_c^d \left(\int_a^b \partial_x g\, dx \right) dy = \int_c^d g(b,y) - g(a,y)\, dy = \int_c^d g(b,y)\, dy + \int_d^c g(a,y)\, dy.$$

Notice each integral on the right is an oriented integral on the path γ_2 and γ_4, respectively. Hence

$$\int_c^d \left(\int_a^b \partial_x g\, dx \right) dy = \int_{\gamma_2} g(x,y)\, dy + \int_{\gamma_4} g(x,y)\, dy. \tag{6.38}$$

In fact, we say a bit more. By property (P3) of an integral, $\int_{\gamma_1} g(x,y)\, dy = \int_c^c g(x,c)\, dy = 0$ and $\int_{\gamma_3} g(x,y)\, dy = \int_d^d g(x,d)\, dy = 0$. Adding terms that are zero to the right side of Equation 6.38

changes nothing. Therefore,

$$\int_c^d \left(\int_a^b \partial_x g \, dx \right) dy = \int_{\gamma_1+\gamma_2+\gamma_3+\gamma_4} g(x,y)\, dy = \int_\gamma g(x,y)\, dy.$$

Challenge 68 Use Fubini's Theorem (Theorem 77), the Fundamental Theorem of Calculus, and the linearity of the integral to show that

$$\int_c^d \left(\int_a^b -\partial_y f(x,y)\, dx \right) dy = \int_{\gamma_3} f(x,y)\, dx + \int_{\gamma_1} f(x,y)\, dx.$$

Since $\int_{\gamma_2} f(x,y)\, dx + \int_{\gamma_4} f(x,y)\, dx = 0$, we see that $\int_c^d \left(\int_a^b -\partial_y f(x,y)\, dx \right) dy = \int_\gamma f(x,y)\, dx$. Conclude that **Green's Theorem**, shown below, holds.

$$\boxed{\int_S d\omega = \int_\gamma \omega} \tag{6.39}$$

Kelvin-Stokes Theorem

Just as we combined intervals in a plane to make a path, and then combined paths to obtain a surface, we will now try merging surfaces on the plane into a surface in $\mathbb{R}^3$.

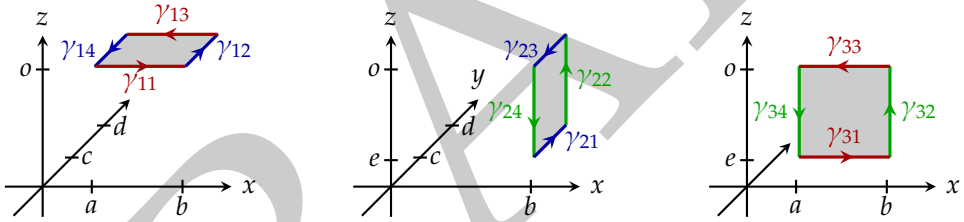


Figure 6.40: Surfaces S_1, S_2, and S_3 with z values, x values, and y values held constant, respectively.

Each surface in Figure 6.40 are essentially rectangles on a plane. The first surface $S_1 := [a,b] \times [c,d] \times \{o\}$ is a rectangle with z-values fixed at $z = o$. The second surface $S_2 := \{b\} \times [c,d] \times [e,o]$ is a rectangle with x-values fixed at $x = b$. The third surface $S_3 := [a,b] \times \{c\} \times [e,o]$ is a rectangle with y-values fixed at $y = c$. Let us assume that each rectangle S_i is enclosed by path $\gamma_i := \gamma_{i1} + \gamma_{i2} + \gamma_{i3} + \gamma_{i4}$ with the orientations shown in Figure 6.40.

We calculate the integral of the exterior derivative of 1-form $\omega := f\, dx + g\, dy + h\, dz$, with $f, g, h \in C^1$, over each surface. For the rectangle S_1, we have

$$\int_{S_1} d\omega = \iint_{S_1} \cancel{(\partial_y h - \partial_z g)\, dy\, dz} + \cancel{(\partial_z f - \partial_x h)\, dz\, dx} + (\partial_x g - \partial_y f)\, dx\, dy$$

where the integrands integrated over z vanish because z is a fixed constant. The integral

$$\iint_{S_1} (\partial_x g - \partial_y f)\, dx\, dy = \int_c^d \int_a^b (\partial_x g(x,y,o) - \partial_y f(x,y,o))\, dx\, dy$$

is the surface integral of the 2-form $\left(\partial_x g - \partial_y f\right)\, dx \wedge dy$ on the surface S_1 with the orientation of surface S_1 as shown in Figure 6.40. By Green's Theorem,

$$\int_{S_1} d\omega = \iint_{S_1} \left(\partial_x g - \partial_y f\right)\, dx\, dy = \int_{\gamma_1} f(x,y,z)\, dx + g(x,y,z)\, dy.$$

Since we know that $\int_{\gamma_1} h(x,y,z)\, dz = \int_o^o h(x,y,o)\, dz = 0$, we have

$$\int_{S_1} d\omega = \int_{\gamma_1} f(x,y,z)\, dx + g(x,y,z)\, dy + h(x,y,z)\, dz = \int_{\gamma_1} \omega.$$

The calculations over the surfaces S_2 and S_3 are completely analogous. We simply note that $\iint_{S_2} \phi(x,y,z)\, dy\, dz = \int_e^o \int_c^d \phi(b,y,z)\, dy\, dz$ is the surface integral of the 2-form $\phi\, dy \wedge dz$ over the surface $S_2 = \{b\} \times [c,d] \times [e,o]$ with the orientation of surface S_2 as shown in Figure 6.40. Similarly, $\iint_{S_3} \psi(x,y,z)\, dx\, dz = \int_e^o \int_a^b \psi(x,d,z)\, dx\, dz$ is the surface integral of the 2-form $\psi\, dz \wedge dx$ over the surface $S_2 = [a,b] \times \{d\} \times [e,o]$ with the orientation of surface S_3 as shown in Figure 6.40. As each rectangle S_i are rectangles on a plane, we can apply Green's Theorem to get

$$\int_{S_2} d\omega = \int_{\gamma_2} \omega \qquad\qquad \int_{S_3} d\omega = \int_{\gamma_3} \omega.$$

Now consider the surface S obtained by merging the rectangles S_1, S_2, and S_3 as shown in Figure 6.41 below. By linearity, the integral $\int_S d\omega$ is the sum $\int_{S_1} d\omega + \int_{S_2} d\omega + \int_{S_3} d\omega$.

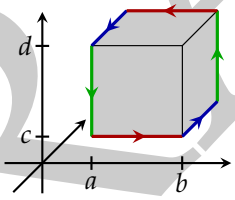


Figure 6.41: Surface S formed by merging surfaces S_1, S_2, and S_3.

Using the equalities obtained before, we have

$$\int_S d\omega = \int_{S_1} d\omega + \int_{S_2} d\omega + \int_{S_3} d\omega = \int_{\gamma_1} \omega + \int_{\gamma_2} \omega + \int_{\gamma_3} \omega.$$

But we can go one step further. Observe that the paths γ_{11} and γ_{33} overlap completely, with *opposite* orientation. This means that $\int_{\gamma_{11}} \omega + \int_{\gamma_{33}} \omega = 0$. The same goes for the paths γ_{12}, γ_{23} and the paths γ_{24}, γ_{32}. Incorporating these cancellations, we have the equality

$$\int_{\gamma_1} \omega + \int_{\gamma_2} \omega + \int_{\gamma_3} \omega = \int_{\gamma_{31}} \omega + \int_{\gamma_{21}} \omega + \int_{\gamma_{22}} \omega + \int_{\gamma_{13}} \omega + \int_{\gamma_{14}} \omega + \int_{\gamma_{34}} \omega.$$

The paths listed in the integrals on the right are precisely those that enclose surface S in Figure 6.41. Taking γ to be the sum of the paths that enclose surface S, we obtain the **Kelvin-Stokes Theorem**.

$$\boxed{\int_S d\omega = \int_\gamma \omega}$$

Divergence Theorem

Finally, we consider merging surfaces to enclose a *volume*. Once we start merging surfaces to enclose volumes, there is a fixed convention for the orientation of the surfaces. Each orientation of the surfaces will be shown in the figures to follow.

We consider the exterior derivative of a 2-form $\eta := f\,dy \wedge dz + g\,dz \wedge dx + h\,dx \wedge dy$, which we will integrate over a volume. The exterior derivative of η is

$$d\eta = \partial_1 f\,dx \wedge dy \wedge dz + \partial_2 g\,dx \wedge dy \wedge dz + \partial_3 h\,dx \wedge dy \wedge dz.$$

The volume integral of $d\eta$ over the cube $V := [a,b] \times [c,d] \times [e,o]$ (see Figure 6.42) is given by

$$\int_V d\eta = \int_V \left(\partial_1 f + \partial_2 g + \partial_3 h\right)\,dx \wedge dy \wedge dz.$$

As before, we break up the integral and consider each of the three terms separately. Using the definition of an oriented integral of a 3-form in $\mathbb{R}^3$ and applying the Fundamental Theorem of Calculus gives

$$\begin{aligned}\int_V \partial_1 f\,dx \wedge dy \wedge dz &:= \int_e^o \int_c^d \left(\int_a^b \partial_1 f\,dx\right) dy\,dz = \int_e^o \int_c^d f(b,y,z) - f(a,y,z)\,dy\,dz \\ &= \int_e^o \int_c^d f(b,y,z)\,dy\,dz - \int_e^o \int_c^d f(a,y,z)\,dy\,dz.\end{aligned}$$

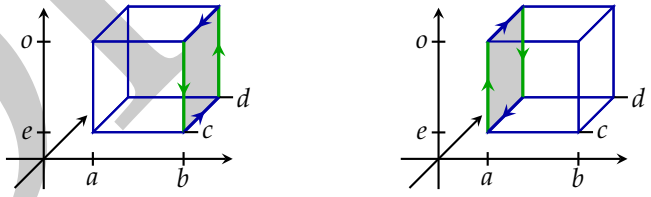


Figure 6.42: Faces S_1 and S_2 of a cube in $\mathbb{R}^3$ where x values are held constant.

The integral $\int_e^o \int_c^d f(b,y,z)\,dy\,dz$ is the surface integral of the 2-form $f\,dy \wedge dz$ over the face $S_1 := \{b\} \times [c,d] \times [e,o]$. Similarly, the integral $\int_e^o \int_c^d f(a,y,z)\,dy\,dz$ is the surface integral of the 2-form $f\,dy \wedge dz$ over the face $S_2 := \{a\} \times [c,d] \times [e,o]$. However, surface S_2 is (by convention) oriented in the opposite direction of surface S_1. Thus the integral of $f\,dy \wedge dz$ over surface S_2 needs

an extra minus sign. Putting these together, we have

$$\begin{aligned}\int_V \partial_1 f\, dx \wedge dy \wedge dz &= \int_e^o \int_c^d f(b,y,z)\, dy\, dz - \int_e^o \int_c^d f(a,y,z)\, dy\, dz \\ &= \int_{S_1} f\, dy \wedge dz + \int_{S_2} f\, dy \wedge dz = \int_{S_1 \cup S_2} f\, dy \wedge dz.\end{aligned}$$

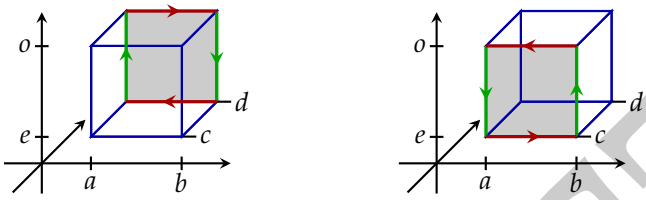


Figure 6.43: Faces S_3 and S_4 of a cube in $\mathbb{R}^3$ where y values are held constant.

Next is the volume integral of $\partial_2 g\, dx \wedge dy \wedge dz$. We use Fubini's Theorem to switch the order of integration so that we can integrate over y first.

$$\int_V \partial_2 g\, dx \wedge dy \wedge dz = \int_e^o \int_c^d \left(\int_a^b \partial_2 g\, dx \right) dy\, dz = \int_e^o \int_a^b \left(\int_c^d \partial_2 g\, dy \right) dx\, dz$$

We can now use the Fundamental Theorem of Calculus to get

$$\begin{aligned}\int_V \partial_2 g\, dx \wedge dy \wedge dz &= \int_e^o \int_a^b g(x,d,z) - g(x,c,z)\, dx\, dz \\ &= \int_e^o \int_a^b g(x,d,z)\, dx\, dz - \int_e^o \int_a^b g(x,c,z)\, dx\, dz.\end{aligned}$$

The integral $\int_e^o \int_a^b g(x,d,z)\, dx\, dz$ is the surface integral of the 2-form $g\, dz \wedge dx$ over the surface $S_3 := [a,b] \times \{d\} \times [e,o]$. Similarly, the integral $\int_e^o \int_a^b g(x,c,z)\, dx\, dz$ is the surface integral of the 2-form $g\, dz \wedge dx$ over the surface $S_4 := [a,b] \times \{c\} \times [e,o]$. Once again, because the surface S_4 is oriented in the opposite direction of S_3, it needs an extra minus sign. We thus have

$$\int_V \partial_2 g\, dx \wedge dy \wedge dz = \int_e^o \int_a^b g(x,d,z)\, dx\, dz - \int_e^o \int_a^b g(x,c,z)\, dx\, dz = \int_{S_3 \cup S_4} g\, dz \wedge dx.$$

Finally, we calculate the volume integral of the 3-form $\partial_3 h\, dx \wedge dy \wedge dz$. The calculation is analogous to the previous ones. Applying Fubini's Theorem, the Fundamental Theorem of Calculus, and the linearity of the integral gives

$$\begin{aligned}\int_V \partial_3 h\, dx \wedge dy \wedge dz &= \int_e^o \int_c^d \left(\int_a^b \partial_3 h\, dx \right) dy\, dz = \int_c^d \int_a^b \left(\int_e^o \partial_3 h\, dz \right) dx\, dy \\ &= \int_c^d \int_a^b h(o,y,z)\, dx\, dy - \int_c^d \int_a^b h(e,y,z)\, dx\, dy.\end{aligned}$$

The former is a surface integral of the 2-form $h\,dx \wedge dy$ over the surface $S_5 := [a,b] \times [c,d] \times \{o\}$. The latter is a surface integral of the 2-form $h\,dx \wedge dy$ over the surface $S_6 := [a,b] \times [c,d] \times \{e\}$ with the orientation of S_5. We conclude that

$$\int_V \partial_3 h\,dx \wedge dy \wedge dz = \int_{S_5 \cup S_6} h\,dx \wedge dy.$$

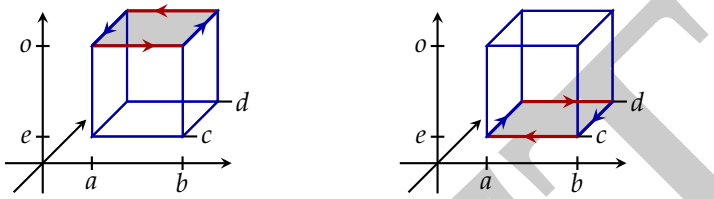


Figure 6.44: Faces S_5 and S_6 of a cube in $\mathbb{R}^3$ where z values are held constant.

Let $S := S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6$ be the surface that encloses the cube V. The integrals $\int_{S_3} f\,dy \wedge dz$ and $\int_{S_4} f\,dy \wedge dz$ vanish because y is assumed to be a constant in S_3 and S_4, but we are integrating over y. The integrals $\int_{S_5} f\,dy \wedge dz$ and $\int_{S_6} f\,dy \wedge dz$ similarly vanish and so

$$\int_{S_1} f\,dy \wedge dz + \int_{S_2} f\,dy \wedge dz = \int_S f\,dy \wedge dz.$$

The same reasoning gives the following equalities.

$$\int_{S_3 \cup S_4} g\,dz \wedge dx = \int_S g\,dz \wedge dx \qquad \int_{S_5 \cup S_6} h\,dx \wedge dy = \int_S h\,dx \wedge dy$$

By the linearity of the integral,

$$\begin{aligned}\int_V d\eta &= \int_{S_1 \cup S_2} f\,dy \wedge dz + \int_{S_3 \cup S_4} g\,dz \wedge dx + \int_{S_5 \cup S_6} h\,dx \wedge dy \\ &= \int_S f\,dy \wedge dz + \int_S g\,dz \wedge dx + \int_S h\,dx \wedge dy = \int_S \eta\end{aligned}$$

and we obtain the **Divergence Theorem**.

$$\boxed{\int_V d\eta = \int_S \eta}$$

Continuity equation

The Divergence Theorem relates an integral over a surface to an integral over a volume. As an application, consider a **charge density** ρ which is a real-valued function that describes the electric charge per unit volume at each position. The total charge in the volume V enclosed within a closed rectangular surface S is given by the integral $\iiint_V \rho\,dx\,dy\,dz$.

The charge density may change over time, and the **current density** $\boldsymbol{J} := (J_1, J_2, J_3)$ describes the charge per unit time that passes each position per unit area. The total amount of charge that passes through the surface S is then given by the integral $\int_S J_1\, dy \wedge dz + J_2\, dy \wedge dz + J_3\, dx \wedge dy$.

Charge is conserved. Moreover, they cannot move instantaneously. This means that charge passing by our closed surface S must equal the negative rate of change of the total charge contained in volume V.[15] That is

$$-\frac{\mathrm{d}}{\mathrm{d}t} \iiint_V \rho\, dx\, dy\, dz = \int_S J_1\, dy \wedge dz + J_2\, dy \wedge dz + J_3\, dx \wedge dy. \tag{6.45}$$

We use the Divergence Theorem to turn the integral over the surface on the right side of Equation 6.45 into an integral over a volume

$$\int_S J_1\, dy \wedge dz + J_2\, dy \wedge dz + J_3\, dx \wedge dy = \int_V (\partial_1 J_1 + \partial_2 J_2 + \partial_3 J_3)\, dx \wedge dy \wedge dz.$$

On the other hand, the Leibniz integral rule allows us to push in the time derivative in the left side of Equation 6.45. We obtain the following.

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t} \iiint_V \rho\, dx\, dy\, dz &= -\iiint_V \partial_t \rho\, dx\, dy\, dz \\ &= \iiint_V (\boldsymbol{\nabla} \cdot \boldsymbol{J})\, dx\, dy\, dz \\ &= \int_V (\partial_1 J_1 + \partial_2 J_2 + \partial_3 J_3)\, dx \wedge dy \wedge dz \\ &= \int_S J_1\, dy \wedge dz + J_2\, dy \wedge dz + J_3\, dx \wedge dy \end{aligned}$$

The above will certainly have to hold for each arbitrary cube V, and so we require

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = -\partial_t \rho$$

which is called the **continuity equation**.

Of course, the continuity equation holds in many other situations. Consider the transfer of heat (thermal energy) from warm to cooler zones. Let T is the temperature of each position at each time and let $\boldsymbol{J}$ be the corresponding current density. The speed in which the heat flows will depend on the properties of the material, and much of the material we have.

To simplify, we may assume that the thermal energy density ρ (corresponding to charge density) is proportional to the mass density ϱ of the material (we will assume this to be a constant), and the temperature T. Thus $\rho = c\varrho T$, where the dimensionful constant c is called the *specific heat*. The fact that thermal energy goes from warm to cooler and is proportional to their difference is expressed in calculus by $\boldsymbol{J} = -k\boldsymbol{\nabla} T$ (*Fourier's law of thermal conduction*),[16] where the dimensionful constant k is the *thermal conductivity* of the material. The continuity equation then gives $\partial_t \rho = c\varrho \partial_t T = k\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} T) = k\nabla^2 T$. Define the *thermal diffusivity* constant $\alpha := k/(c\varrho)$ and we obtain the **heat equation**

$$\partial_t T = \alpha \nabla^2 T.$$

[15] Negative because the more charges exit, the less charge remains.

[16] Remember that $\boldsymbol{\nabla} T = (\partial_x T, \partial_y T, \partial_z T)^\mathsf{T}$ where each (partial) derivative is a difference taken to the limit.

6.7 Maxwell's Equations

The equations

A *vector field* in $\mathbb{R}^3$ assigns to each point in $\mathbb{R}^3$ a vector of dimension 3. An example is a vector field that describes the wind blowing at an instant. Each vector v at location (x, y, z) tells us the direction the wind is blowing towards at that instant. Observe that 3-coordinates are needed to specify where the wind is blowing towards

We discovered in Section 6.5 that a vector field $\boldsymbol{f} := (f_1, f_2, f_3)$ in $\mathbb{R}^3$ has the associated objects $\boldsymbol{\nabla} \cdot \boldsymbol{f}$, $\boldsymbol{\nabla} \times \boldsymbol{f}$, and $\nabla^2 \boldsymbol{f}$ defined as follows.

$$\boldsymbol{\nabla} \cdot \boldsymbol{f} := \partial_1 f_1 + \partial_2 f_2 + \partial_3 f_3 \qquad \boldsymbol{\nabla} \times \boldsymbol{f} := \begin{pmatrix} \partial_2 f_3 - \partial_3 f_2 \\ \partial_3 f_1 - \partial_1 f_3 \\ \partial_1 f_2 - \partial_2 f_1 \end{pmatrix} \qquad \nabla^2 \boldsymbol{f} := \begin{pmatrix} \partial_{1,1} f_1 + \partial_{2,2} f_1 + \partial_{3,3} f_1 \\ \partial_{1,1} f_2 + \partial_{2,2} f_2 + \partial_{3,3} f_2 \\ \partial_{1,1} f_3 + \partial_{2,2} f_3 + \partial_{3,3} f_3 \end{pmatrix}$$

The scalar $\boldsymbol{\nabla} \cdot \boldsymbol{f}$ is called the *divergence* of vector field $\boldsymbol{f}$, the vector $\boldsymbol{\nabla} \times \boldsymbol{f}$ is called the *curl* of vector field $\boldsymbol{f}$, and the vector $\nabla^2 \boldsymbol{f}$ is called the *Laplacian* of vector field $\boldsymbol{f}$.

In this section, we will be concerned with the **electric** field $\boldsymbol{E} := (E_1, E_2, E_3)$ and the **magnetic** field $\boldsymbol{B} := (B_1, B_2, B_3)$. Like the vector field of winds, the electric and magnetic fields can change over time. So we will manually add a distinguished time axis t as the 4-th axis, which is independent from the other three positional axis.

Here is the situation before 1861. The divergence of an electric field obeys **Gauss's law**

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho / \epsilon_0 \tag{6.46}$$

where ϵ_0 is the constant from Coulomb's law and ρ describes the electric charge per unit volume (*charge density*). The divergence of a magnetic field obeys the "no magnetic monopoles law"

$$\boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \tag{6.47}$$

and so there is no analogue of a magnetic charge. The curl of an electric field obeys **Faraday's law**

$$\boldsymbol{\nabla} \times \boldsymbol{E} = -\partial_t \boldsymbol{B}. \tag{6.48}$$

The curl of a magnetic field obeys **Ampère's law**

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} \tag{6.49}$$

where the constant μ_0 satisfies the equation $\mu_0 \epsilon_0 = 1/c^2$. The constant c is the **speed of light** (in vacuum) which takes the value of approximately 3×10^8 m/s. The *current density* $\boldsymbol{J}$ describes the amount of charge flowing through a unit area per unit time. As the total current flowing out $\boldsymbol{\nabla} \cdot \boldsymbol{J}$ is precisely the reduction of charge over time $-\partial_t \rho$, the current density $\boldsymbol{J}$ and charge density ρ satisfies the *continuity equation*

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = -\partial_t \rho.$$

In the simplest case where the charge density ρ and current density $\boldsymbol{J}$ are both zero, we can use Equation 6.35 to calculate the Laplacian of the electric field to be zero:

$$\nabla^2 \boldsymbol{E} = \boldsymbol{\nabla} (\boldsymbol{\nabla} \cdot \boldsymbol{E}) - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = \boldsymbol{\nabla} 0 + \boldsymbol{\nabla} \times (\partial_t \boldsymbol{B}) = 0 + \partial_t (\boldsymbol{\nabla} \times \boldsymbol{B}) = \partial_t 0 = 0$$

where we have used the fact that all permutations of $\partial_{i,j,k} f$ are equal (see Equation 6.29) to pull ∂_t out. If $\rho = 0$ and $\boldsymbol{J} = 0$, the Laplacian of the magnetic field is also zero:

$$\nabla^2 \boldsymbol{B} = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{B}) - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{B}) = \boldsymbol{\nabla} 0 + \boldsymbol{\nabla} \times 0 = 0 + 0 = 0.$$

This is what was known before 1861. Let us play around a little. The simplest thing we can try is taking the divergence of a curl (recall that it is not possible to take the curl of a divergence). By Challenge 63, we have $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{f}) = 0$. Taking the divergence of both sides of Faraday's law gives

$$0 = \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{E}) = -\boldsymbol{\nabla} \cdot (\partial_t \boldsymbol{B}) = -\partial_t (\boldsymbol{\nabla} \cdot \boldsymbol{B}) = -\partial_t 0 = 0$$

as expected. On the other hand, taking the divergence of both sides of Ampère's law gives

$$0 = \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{B}) = \mu_0 \boldsymbol{\nabla} \cdot \boldsymbol{J}.$$

By the continuity equation, the above asserts that $\partial_t \rho = 0$. The charge density is forbidden from changing over time!

How we can fix this? We will need to modify Ampère's law into $\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \boldsymbol{X}$ such that $\boldsymbol{\nabla} \cdot (\mu_0 \boldsymbol{J} + \boldsymbol{X}) = 0$. By linearity of partial derivatives $\boldsymbol{\nabla} \cdot (\mu_0 \boldsymbol{J} + \boldsymbol{X}) = \mu_0 \boldsymbol{\nabla} \cdot \boldsymbol{J} + \boldsymbol{\nabla} \cdot \boldsymbol{X}$. To find $\boldsymbol{X}$ we use the continuity equation $\mu_0 \boldsymbol{\nabla} \cdot \boldsymbol{J} + \mu_0 \partial_t \rho = 0$. Gauss's law $\rho = \epsilon_0 \boldsymbol{\nabla} \cdot \boldsymbol{E}$ gives

$$\mu_0 \partial_t \rho = \mu_0 \partial_t (\epsilon_0 \boldsymbol{\nabla} \cdot \boldsymbol{E}) = \boldsymbol{\nabla} \cdot (\mu_0 \epsilon_0 \partial_t \boldsymbol{E}).$$

Therefore $\boldsymbol{X} = \mu_0 \epsilon_0 \partial_t \boldsymbol{E}$ and we have obtained the **Ampère-Maxwell law**

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \mu_0 \epsilon_0 \partial_t \boldsymbol{E}. \tag{6.50}$$

As $\mu_0 \epsilon_0 = 1/c^2$ is about 9×10^{-16} s^2/m^2 this addition (called *Maxwell's correction*) is incredibly tiny!

Equations 6.46, 6.47, 6.48, and 6.50 are collectively known as **Maxwell's equations (in vacuum)**. In the special case where charge density $\rho = 0$ and current density $\boldsymbol{J} = 0$, the four equations are called the **source-free** Maxwell's equations.

We now recalculate the Laplacians of $\boldsymbol{E}$ and $\boldsymbol{B}$ using the source-free Maxwell's equations.

$$\nabla^2 \boldsymbol{E} = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{E}) - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{E}) = \boldsymbol{\nabla} 0 + \boldsymbol{\nabla} \times (\partial_t \boldsymbol{B}) = 0 + \partial_t (\boldsymbol{\nabla} \times \boldsymbol{B}) = \mu_0 \epsilon_0 \partial_{t,t} \boldsymbol{E}$$

$$\nabla^2 \boldsymbol{B} = \boldsymbol{\nabla}(\boldsymbol{\nabla} \cdot \boldsymbol{B}) - \boldsymbol{\nabla} \times (\boldsymbol{\nabla} \times \boldsymbol{B}) = \boldsymbol{\nabla} 0 - \boldsymbol{\nabla} \times (\mu_0 \epsilon_0 \partial_t \boldsymbol{E}) = -\mu_0 \epsilon_0 \partial_t (\boldsymbol{\nabla} \times \boldsymbol{E}) = \mu_0 \epsilon_0 \partial_{t,t} \boldsymbol{B}.$$

Where have we seen such equations before? Recall that the one-dimensional wave equation for a wave with speed v is given by

$$\partial_{1,1} f = \frac{1}{v^2} \partial_{t,t} f.$$

The analogue of $\partial_{1,1}$ for a vector field is the Laplacian ∇^2. Thererefore, the three-dimensional **wave equation** for wave traveling at speed v is given by

$$\nabla^2 \boldsymbol{f} = \frac{1}{v^2} \partial_{t,t} \boldsymbol{f}.$$

Since $\mu_0 \epsilon_0 = 1/c^2$ we see that in a vacuum, electromagnetic waves propagate at the speed of light. This led Maxwell to postulate that light is an electromagnetic wave. The discovery of electromagnetic waves changed *everything*, and it comes out of the theory because of one tiny imperceptible correction! It took over two decades from Maxwell's predictions for the existence of electromagnetic waves to be demonstrated conclusively. This led to the universal acceptance of Maxwell's equations and marked the beginning of telecommunication as we know it.

Potentials (optional)

Recall that in one dimensions a force F is conservative if we can unambiguously define a potential energy function V. We saw in Section 5.1 that $F = -\frac{\mathrm{d}V}{\mathrm{d}x}$. The generalization to three dimensions is that $\boldsymbol{F}$ is conservative if there is some potential energy function ϕ satisfying $\boldsymbol{F} = -\boldsymbol{\nabla}\phi$. That is,

$$\boldsymbol{F} = (F_1, F_2, F_3) = \left(-\partial_1\phi, -\partial_2\phi, -\partial_3\phi\right).$$

By Challenge 63, the curl of a gradient vanishes and so $\boldsymbol{\nabla} \times \boldsymbol{F} = 0$ whenever $\boldsymbol{F}$ is conservative. Is the converse true? That is, if $\boldsymbol{\nabla} \times \boldsymbol{F} = 0$ can we conclude that $\boldsymbol{F}$ is conservative?

Proposition 80 (Curl Test)**.** Let vector field $\boldsymbol{f} := (f_1, f_2, f_3)$ be defined on an open rectangle R in $\mathbb{R}^3$ and suppose $f_1, f_2, f_3 \in C^1$ (all partial derivatives $\partial_i f_j$ exist and are continuous). If $\boldsymbol{\nabla} \times \boldsymbol{f} = 0$ then there is some real-valued function ϕ such that $\boldsymbol{\nabla}\phi = \boldsymbol{f}$ on the rectangle R.

Proof. Let (a, c, e) be a point in the rectangle R and let

$$\phi : (x, y, z) \mapsto \int_a^x f_1(u, y, z)\, du + \int_c^y f_2(a, v, z)\, dv + \int_e^z f_3(a, c, w)\, dw$$

where the second and third terms are constant with respect to the first variable so that they vanish when we take the partial derivative ∂_1. By the Fundamental Theorem of Calculus (Equation 6.22 adapted to three variables)

$$\partial_1\phi(x, y, z) = f(x, y, z) + 0 + 0.$$

The third term in ϕ is constant with respect to the second variable and vanishes when we take the partial derivative ∂_2. Differentiating under the integral sign (Leibniz integral rule) in the first term, while applying the Fundamental Theorem of Calculus to the second term gives

$$\begin{aligned}\partial_2\phi(x, y, z) &= \partial_2 \int_a^x f_1(u, y, z)\, du + \partial_2 \int_c^y f_2(a, v, z)\, dv + \partial_2 \int_e^z f_3(a, c, w)\, dw \\ &= \int_a^x \partial_2 f_1(u, y, z)\, du + f_2(a, y, z) + 0.\end{aligned}$$

Since $\boldsymbol{\nabla} \times \boldsymbol{f} = 0$, in particular, $\partial_1 f_2 - \partial_2 f_2$. Therefore,

$$\int_a^x \partial_2 f_1(u, y, z)\, du + f_2(a, y, z) = \int_a^x \partial_1 f_2(u, y, z)\, du + f_2(a, y, z)$$

and applying the Fundamental Theorem of Calculus (Equation 6.23) gives

$$\partial_2\phi(x, y, z) = \int_a^x \partial_1 f_2(u, y, z)\, du + f_2(a, y, z) = f_2(x, y, z) - f_2(a, y, z) + f_2(a, y, z) = f_2(x, y, z).$$

It suffices to check that $\partial_3\phi(x, y, z) = f(x, y, z)$; the steps are similar to before. Applying a differentiation under the integral sign to the first two terms and a Fundamental Theorem of Calculus to the third term gives

$$\partial_3\phi(x, y, z) = \int_a^x \partial_3 f_1(u, y, z)\, du + \int_c^y \partial_3 f_2(a, v, z)\, dv + f_3(a, c, z).$$

Because $\nabla \times \boldsymbol{f} = 0$ we know that $\partial_3 f_1 = \partial_1 f_3$ and $\partial_3 f_2 = \partial_2 f_3$. Making this switch and applying the Fundamental Theorem of Calculus gives

$$\partial_3 \phi(x,y,z) = f_3(x,y,z) - f_3(a,y,z) + f_3(a,y,z) - f_3(a,c,z) + f_3(a,c,z) = f_3(x,y,z).$$

Therefore, $\nabla \phi = \boldsymbol{f}$ on rectangle R. □

By Hooke's law, a simple harmonic oscillator moving along the x-axis only obeys the force law $\boldsymbol{F} := (-kx, 0, 0)$. Since $\nabla \times \boldsymbol{F} = 0$, the curl test verifies that the force has a potential energy function ϕ satisfying $\boldsymbol{F} = -\nabla \phi$. Therefore, Hooke's law describes a conservative force.

Challenge 69 The Coulomb force felt on a particle with charge q_2 at location (x,y,z) due to a charge q_1 at the origin is given by Coulomb's law. By varying (x,y,z) such that at least one of x,y,z is nonzero, we can think of the electrostatic force as the following vector field

$$\boldsymbol{F} : (x,y,z) \mapsto \left(\frac{q_1 q_2}{4\pi\epsilon_0} \frac{x}{r^3}, \frac{q_1 q_2}{4\pi\epsilon_0} \frac{y}{r^3}, \frac{q_1 q_2}{4\pi\epsilon_0} \frac{z}{r^3} \right)$$

where $r := \sqrt{x^2 + y^2 + z^2}$. Check that the Coulomb force is conservative.

In Challenge 63 we also saw that the divergence of a curl always vanishes: $\nabla \cdot (\nabla \times \boldsymbol{f}) = 0$. Is there an analogous result to Proposition 80 that if $\nabla \cdot \boldsymbol{F} = 0$ there is a vector field $\boldsymbol{A}$ such that $\nabla \times \boldsymbol{A} = \boldsymbol{F}$?

Proposition 81. If vector field $\boldsymbol{f} := (f_1, f_2, f_3)$ defined on an open rectangle R in $\mathbb{R}^3$ satisfies $\nabla \cdot \boldsymbol{f} = 0$, then there is a vector field $\boldsymbol{A} := (A_1, A_2, A_3)$ such that $\nabla \times \boldsymbol{A} = \boldsymbol{f}$.

Proof. As before, we construct $\boldsymbol{A}$ such that $\nabla \times \boldsymbol{A} = \boldsymbol{f}$. Let $(a,c,e) \in R$ and define A_i as follows.

$$\begin{aligned}
A_1 &: (x,y,z) \mapsto \int_c^y -f_3(x,v,e)\, dv + \int_e^z f_2(x,y,w)\, dw \\
A_2 &: (x,y,z) \mapsto \int_e^z -f_1(x,y,w)\, dw \\
A_3 &: (x,y,z) \mapsto (0,0,0)
\end{aligned}$$

By the Fundamental Theorem of Calculus, $\partial_2 A_3 - \partial_3 A_2 = 0 + f_1(x,y,z) = f_1(x,y,z)$. The first integral in A_1 is a constant with respect to the third variable. Applying the Fundamental Theorem of Calculus to the second term of A_1 gives

$$\partial_3 A_1 - \partial_1 A_3 = 0 + f_2(x,y,z) - 0 = f_2(x,y,z).$$

Since $\nabla \cdot \boldsymbol{f} = 0$ we know that $\partial_3 f_3 = -\partial_1 f_1 + \partial_2 f_2$. Therefore,

$$\begin{aligned}
\partial_1 A_2 - \partial_2 A_1 &= \int_e^z -\partial_1 f_1(x,y,w)\, dw - \left(-f_3(x,y,e) + \int_e^z \partial_2 f_2(x,y,w)\, dw \right) \\
&= \int_e^z \left[-\partial_1 f_1(x,y,w) - \partial_2 f_2(x,y,w) \right] dw + f_3(x,y,e) - f_3(x,c,e) \\
&= \int_e^z \partial_3 f_3(x,y,w)\, dw + f_3(x,y,e) = f_3(x,y,z) - f_3(x,y,e) + f_3(x,y,e) \\
&= f_3(x,y,z)
\end{aligned}$$

and we have verified that $\nabla \cdot \boldsymbol{A} = \boldsymbol{f}$ on the rectangle R. □

From the "no magnetic monopole law" we know that $\nabla \cdot \boldsymbol{B} = 0$. By Proposition 81 there is a vector field $\boldsymbol{A}$ called a **(magnetic) vector potential** defined by $\nabla \times \boldsymbol{A} = \boldsymbol{B}$. Then Faraday's law gives

$$0 = \nabla \times \boldsymbol{E} + \partial_t \boldsymbol{B} = \nabla \times \boldsymbol{E} + \partial_t (\nabla \times \boldsymbol{B}) = \nabla \times (\boldsymbol{E} + \partial_t \boldsymbol{E})$$

where we have pushed in the time derivative using (the generalized) Clairaut's Theorem. Proposition 80 tells us that there is a scalar-valued function ϕ such that $-\nabla\phi = \boldsymbol{E} + \partial_t \boldsymbol{A}$.[17] The function ϕ is called a **scalar potential**.

The constructions used to obtain Proposition 80 and Proposition 81 are easily seen to be not unique, and so we know that the same electric and magnetic field can arise from distinct scalar and vector potentials. Indeed, as the curl of a gradient is zero (Challenge 63), for a vector potential $\boldsymbol{A}$, we have

$$\nabla \times (\boldsymbol{A} + \nabla\psi) = \nabla \times \boldsymbol{A} + \cancel{\nabla \times (\nabla\psi)} = \nabla \times \boldsymbol{A} = \boldsymbol{B}.$$

Defining $\boldsymbol{A}' := \boldsymbol{A} + \nabla\psi$ and applying it to the definition of the scalar potential gives

$$\boldsymbol{E} = -\nabla\phi - \partial_t \boldsymbol{A} = -\nabla\phi - \partial_t (\boldsymbol{A}' - \nabla\psi) = -\nabla\phi - \partial_t \boldsymbol{A}' + \partial_t \nabla\psi = -\nabla (\phi - \partial_t \psi) - \partial_t \boldsymbol{A}'.$$

Hence making the simultaneous substitutions

$$\boldsymbol{A} \mapsto \boldsymbol{A} + \nabla\psi \qquad\qquad \phi \mapsto \phi - \partial_t \psi$$

leaves the electric field $\boldsymbol{E}$ and the electric field $\boldsymbol{B}$ invariant. The pair of transformations above are called a **gauge transformation**. That a gauge transformation leaves Maxwell's equations invariant is called **gauge invariance**.

Potentials and forms (optional)

Challenge 70 We can rephrase Proposition 80 and Proposition 81 in the language of Section 6.5. A k-form η is **closed** if $d\eta = 0$. For $k \geq 1$, a k-form η is **exact** if there is a $k-1$ form ω such that $d\omega = \eta$. Since a k-form α in $\mathbb{R}^3$ satisfies $d(d\alpha) = 0$, each exact form is closed. Prove **Poincaré's lemma** in $\mathbb{R}^3$: for $k \geq 1$, each closed k-form on an open rectangle R in $\mathbb{R}^3$ is exact on R.

Let us pretend that we knew about forms in $\mathbb{R}^3$, but we did not know about Maxwell's equations. Being bored with working on $\mathbb{R}^3$, we consider working on $\mathbb{R}^4$. We will think of the extra axis as representing time t because it is the only thing that makes sense to add as an extra dimension. We will add time as the zeroth dimension so that $dx_0 := dt$, while dx_1, dx_2, dx_3 are as before.

The most basic object on $\mathbb{R}^4$, if a vector field defined on $\mathbb{R}^4$, which is equivalent to a 1-form in $\mathbb{R}^4$. So consider a 1-form $A := \phi\, dt + A_1\, dx_1 + A_2\, dx_2 + A_3\, dx_3$, where the zeroth element has a different symbol because time is different from space. The whole reason we are considering $\mathbb{R}^4$ is so that we can do calculus on it. So we take an exterior derivative of A using Equation 6.31. We put $A_0 := \phi$ and apply the definition of the exterior derivative to get the following.

$$\begin{aligned} dA &= \sum_{0 \leq i \leq 4} dA_i \wedge dx_i = \sum_{0 \leq i \leq 4} \sum_{j=0}^{4} \partial_j A_i\, dx_j \wedge dx_i \\ &= (\partial_t A_1 + \partial_1 \phi)\, dt \wedge dx_1 + (\partial_t A_2 + \partial_2 \phi)\, dt \wedge dx_2 + (\partial_t A_3 + \partial_3 \phi)\, dt \wedge dx_3 \\ &\quad + (\partial_2 A_3 - \partial_3 A_2)\, dx_2 \wedge dx_3 + (\partial_3 A_1 - \partial_1 A_3)\, dx_3 \wedge dx_1 + (\partial_1 A_2 - \partial_2 A_1)\, dx_1 \wedge dx_2 \end{aligned}$$

[17] The minus sign in front of $\nabla\phi$ is simply due to convention.

Notice that we have separated out the time-space forms $dt \wedge dx_i$ and space-space forms $dx_j \wedge dx_k$.

For convenience, we will denote the coefficients of the time-space forms $dt \wedge dx_i$ by $-E_i$. We will also denote the coefficients of the space-space forms by B_i so that dA can be written as follows.

$$dA = -E_1\, dt \wedge dx_1 - E_2\, dt \wedge dx_2 - E_3\, dt \wedge dx_3 + B_1\, dx_2 \wedge dx_3 + B_2\, dx_3 \wedge dx_1 + B_3\, dx_1 \wedge dx_2$$

This allows us to define a 3-component vector field $\boldsymbol{E} := (E_1, E_2, E_3)$ and a 3-component vector field $\boldsymbol{B} := (B_1, B_2, B_3)$.

Because we know about forms in $\mathbb{R}^3$, we know about gradients, curls, and divergences. So we know that $\boldsymbol{E} = -\boldsymbol{\nabla}\phi - \partial_t \boldsymbol{A}$ if we define the 3-component vector field $\boldsymbol{A} := (A_1, A_2, A_3)$. Similarly, we also know that $\boldsymbol{\nabla} \times \boldsymbol{A} = \boldsymbol{B}$. We know from Challenge 63 that the curl of a gradient is zero and the divergence of a curl is zero. Therefore,

$$\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{\nabla} \times \big(-\boldsymbol{\nabla}\phi - \partial_t \boldsymbol{A}\big) = \boldsymbol{\nabla} \times \big(-\boldsymbol{\nabla}\phi\big) - \partial_t \boldsymbol{\nabla} \times \boldsymbol{A} = -\partial_t \boldsymbol{B} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{B} = \boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{A}) = 0$$

So far we have been able to obtain an equation for the curl of an object called $\boldsymbol{E}$ and the divergence of an object called $\boldsymbol{B}$. It makes sense to try and cook up an equation for the divergence of $\boldsymbol{E}$ and an equation for the curl of $\boldsymbol{B}$. The simplest option would be to assert that $\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0$, but let us try and come up with a simple nontrivial equation. We will say that there is a density of "stuff" called ρ with a dimensionful constant ϵ_0 for flexibility with defining what this "stuff" is. We have $\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0$. Notice that by introducing a "stuff" density $\boldsymbol{E}$ is now dimensionful, and so are $\boldsymbol{A}$ and $\boldsymbol{B}$ due to the equations $\boldsymbol{E} = -\boldsymbol{\nabla}\phi - \partial_t \boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$.

If the "stuff" is constant at all times, our equations would be too boring. So let us assume that the density of the "stuff" is allowed to change. We will describe the change over time of our "stuff" by the 3-component vector field $\boldsymbol{J}$.

All that is left is to come up with an equation for the curl of $\boldsymbol{B}$. We could define it as $\boldsymbol{\nabla} \times \boldsymbol{B} = -\partial_t \boldsymbol{E}$ to keep things symmetric with the equation $\boldsymbol{\nabla} \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$. However, we will not do so, because we want ρ and $\boldsymbol{J}$ to obey the continuity equation. To ensure that the continuity equation $\partial_t \rho + \boldsymbol{J} = 0$ holds, as we will verify soon, it is sufficient to define $\boldsymbol{J}$ as the difference of the curl of $\boldsymbol{B}$ and $\partial_t \boldsymbol{E}$, along with some constants to match the units. We will thus define $\mu_0 \boldsymbol{J} := \boldsymbol{\nabla} \times \boldsymbol{B} - \mu_0 \epsilon_0 \partial_t \boldsymbol{E}$.

Here are our four equations.

$$\boldsymbol{\nabla} \cdot \boldsymbol{E} = \rho/\epsilon_0 \qquad \boldsymbol{\nabla} \times \boldsymbol{E} = -\partial_t \boldsymbol{B} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \qquad \boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \mu_0 \epsilon_0 \partial_t \boldsymbol{E}$$

These have the exact same appearances as Maxwell's equations! Repeating what we did earlier word for word shows that the equations above also results in wave equations for waves traveling at the speed $c := 1/\sqrt{\mu_0 \epsilon_0}$ in vacuum. In particular, all the results in the subsequent section can be obtained even if we knew nothing about Maxwell's equations and light.

6.8 Paradigm Shattering

Born rule

Maxwell's correction to Ampère's law gives an unexpected benefit that the continuity equation is built right into Maxwell's equations. As only two of the four equations contain information on sources ρ and $\boldsymbol{J}$, we can ignore half of Maxwell's equations. Since the divergence of a curl is

zero, taking the divergence of both sides of the Ampère-Maxwell law and using the linearity of the derivative and equality of mixed partials gives $0 = \mu_0 \nabla \cdot \boldsymbol{j} + \mu_0 \epsilon_0 \partial_t (\nabla \cdot \boldsymbol{E})$. Gauss's law provides the substitution $\nabla \cdot \boldsymbol{E} = \rho/\epsilon_0$. Dividing through by μ_0 gives the continuity equation $\nabla \cdot \boldsymbol{J} + \partial_t \rho = 0$.

We can guess that the temperature (energy) distribution of an object and the heat flow (energy current) of the heat equation should also obey the continuity equation. Since Schrödinger's equation is a heat equation (Section 5.5 and Section 6.6), it should also hold a continuity equation.

Left multiplying Ψ^* on Schrödinger's equation and left multiplying Ψ on the complex conjugate of Schrödinger's equation gives the following (notice V is real because H must be Hermitian).

$$\Psi^* \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \Psi = \Psi^* i\hbar \partial_t \Psi \qquad\qquad \Psi \left(-\frac{\hbar^2}{2m} \nabla^2 + V \right) \Psi^* = -i\hbar \partial_t \Psi^*$$

Subtracting the latter from the former, we have

$$-\frac{\hbar^2}{2m} \left[\Psi^* \nabla^2 \Psi - \Psi \nabla^2 \Psi^* \right] = i\hbar \left[\Psi^* \partial_t \Psi + \Psi \partial_t \Psi^* \right].$$

The right side is an application of the product rule on $i\hbar\partial_t (\Psi^* \Psi)$. On the other hand, we can pull out a divergence on the left side to get $-\frac{\hbar^2}{2m} \left[\Psi^* \nabla^2 \Psi - \Psi \nabla^2 \Psi^* \right] = -\frac{\hbar^2}{2m} \nabla \cdot [\Psi^* \nabla \Psi - \Psi \nabla \Psi^*]$ as you can verify. Therefore, the Schrödinger equation gives us the continuity equation

$$\partial_t \rho = -\nabla \cdot \boldsymbol{J}$$

with $\rho := |\Psi|^2$ and $\boldsymbol{J} := \frac{\hbar}{2im} (\Psi^* \nabla \Psi - \Psi \nabla \Psi^*)$. Thus $|\Psi|^2$ is an analogue of a "charge density". How can a point particle have a density? Since $|\Psi|^2$ must give the propensity of the particle to be each location, it follows that once we scale the wavefunction Ψ such that $\iiint_{\mathbb{R}^3} |\Psi|^2 \, dx \, dy \, dx = 1$, then $\int_a^b \int_c^d \int_e^f \left| \Psi(x, y, z, t) \right|^2 \, dx \, dy \, dz$ is the probability that the particle is to be found in the rectangle $[a, b] \times [c, d] \times [e, f]$ at time t. This postulate is called the **Born Rule**. We call ρ a **probability density** and $\boldsymbol{J}$ a **probability current**.

A crazy proposition, but this is what calculus is telling us! As the calculus of probability requires a different mindset, we will take ourselves back to 1904 and forget about Schrödinger's equation. Surely if we stick to good old Newtonian mechanics and Maxwell's equations, all will be alright and sensible. Unlike quantum theory which needs complex numbers, Newtonian mechanics and Maxwellian theory (**electromagnetism**) only require real numbers. We now discuss the space in which Newtonian mechanics and electromagnetism plays out.

Euclidean space

Our discussion of forms began with the wedge product which gave us a way to combine forms.[18] What forms can we get with wedge products? Taking the wedge product of a form with a 0-form is not so interesting, and taking the wedge product of a 3-form with a k-form with $k \geq 1$ gives zero, so it suffices to combine 1-forms and 2-forms. We already saw that taking the wedge product of two 1-forms and then taking the Hodge star gives the cross product $\times$ (Challenge 65). Taking the wedge product of two 2-forms gives zero, so all that is left is to consider the wedge product of a 1-form with a 2-form, and vice versa.

[18] The references to forms can be ignored if you skipped Section 6.5.

We know that such a wedge product will return a 3-form, so we might as well take a Hodge star to get a 0-form. Let ω_u be the 1-form corresponding to the vector (u_1, u_2, u_3) and let $\eta_v := v_1\,dx_2 \wedge dx_3 + v_2\,dx_3 \wedge dx_1 + v_3\,dx_1 \wedge dx_2$. As you should verify,

$$\star\left(\omega_u \wedge \eta_v\right) = \sum_{i=1}^{3} u_i v_i.$$

If we take $v_i := u_i$ then we have the number $\sum_{i=1}^{3} u_i^2$. This number measures how far away the point $(u_1, u_2, u_3) \in \mathbb{R}^3$ is located from the origin. Indeed, if $u_3 = 0$, we recover the distance from the origin given by the Pythagorean theorem. The binary function that takes two vectors $u, v \in \mathbb{R}^3$ and returns the scalar $\sum_{i=1}^{3} u_i v_i$ is called the **dot product** of vectors u and v. More generally, if u, v are vectors in $\mathbb{R}^n$, then $u \cdot v := u^\intercal v$. We denote the dot product of two vectors u and v by $u \cdot v$ and we call $\mathbb{R}^n$ equipped with the dot product operation the **Euclidean space** $\mathbb{E}^n$. Observe that the dot product is a commutative operation.

Challenge 71 All vectors and forms are assumed to be in $\mathbb{E}^3$.

(a) Let $a := (a_1, a_2, a_3)^\intercal$, $b := (b_1, b_2, b_3)^\intercal$, and $c := (c_1, c_2, c_3)^\intercal$. Verify that

$$a \cdot (b \times c) = \det\begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}.$$

Conclude that $a \cdot (b \times c) = b \cdot (c \times a) = c \cdot (a \times b)$ and deduce that $a \cdot (b \times c) = (a \times b) \cdot c$. Vector u is **perpendicular** to vector v if $u \cdot v = 0$. For example, distinct pairs of standard basis vectors e_1, e_2, e_3 are perpendicular to each other. Check that $a \times b$ is perpendicular to both a and b.

(b) Pretend $\nabla := (\partial_1, \partial_2, \partial_3)^\intercal$ is a vector in $\mathbb{E}^3$ and let f be a vector field in $\mathbb{E}^3$. Check that the cross product of ∇ and f is the curl of f. Check that the dot product of ∇ and f is the divergence of f. This explains the notation of the curl and divergence operators.

(c) Let α, β be 1-forms and let η be a 2-form. Show that $\alpha \wedge \eta = \eta \wedge \alpha$ and $\alpha \wedge \beta = -\beta \wedge \alpha$. Conclude that if ω_1 is a k-form and ω_2 is an r-form, then $\omega_1 \wedge \omega_2 = (-1)^{kr}\omega_2 \wedge \omega_1$.

Obvious assumptions

Here are the most basic assumptions we use to apply calculus to study of the world around us.

Space and time To do calculus, we need to have a space for functions and numbers to live in. The arena for our laws of nature, which in 1904 are Newton's second law and Maxwell's equations, is the three dimensional Euclidean space and a separate one dimensional time axis. In particular, an absolute time evidently exists, regardless of how good we are at keeping times, progressing at an equal rate for everyone. Indeed, if everyone had accurate clocks that were uniformly synchronized, all will agree with the time at which each event occurs.

Galileo's principle of relativity To assign positions of objects in the Euclidean space $\mathbb{E}^3$ and the corresponding times in the time axis, we need to make a choice of origin. For the Euclidean space $\mathbb{E}^3$ we need to make a further choice of axis and orientation of the axis. Such a choice is called a **coordinate system** or a **reference frame**. We assume that there are **inertial** reference frames such that: (i) the laws of nature are the same for all time in all inertial reference frames, and (ii) each

reference frame in *uniform* one-dimensional motion along a line with respect to an inertial reference frame is also inertial.[19]

Newton's principle of determinacy This is the assumption that knowing the initial state of a physical system and applying calculus allows us to uniquely determine the dynamics of the system. We shall say no more as it was already invalidated by the probabilistic nature of Schrödinger's equation from the Born rule.

Transverse waves

As a concrete example of the application of calculus to electromagnetism, we find a solution to the source free Maxwells equations in vacuum shown below.

$$\nabla \cdot \boldsymbol{E} = 0 \qquad \nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B} \qquad \nabla \cdot \boldsymbol{B} = 0 \qquad \nabla \times \boldsymbol{B} = 1/c^2 \partial_t \boldsymbol{E}$$

To simplify things, we will assume $\boldsymbol{E}$ has a y-component only, that is, $\boldsymbol{E} := (0, E_y, 0)$. Gauss's law $\cancel{\partial_x E_x} + \partial_y E_y + \cancel{\partial_z E_z} = 0$ is satisfied if E_y is not a function of y. To simplify further, we will assume that E_y is a function of x and t only. In particular, as we know that the source free Maxwell's equations gives a wave equation for $\boldsymbol{E}$ and for $\boldsymbol{B}$ for a wave traveling at the speed of light (in a vacuum) c, we will take a page from the one-dimensional wave equation and put $E_y := f(x - ct)$ for some twice differentiable real-valued function f.

Next, we turn to Faraday's law $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$. We know from Challenge 71 that the field $\boldsymbol{E}$ is perpendicular to $\boldsymbol{B}$. The choice $\boldsymbol{B} := (0, 0, B_z)$ satisfies this because the standard basis vectors e_2 and e_3 are perpendicular. Furthermore, we require that $\nabla \cdot \boldsymbol{B} = 0$. Just as we did with the electric field $\boldsymbol{E}$, we assume that B_y is a function of x and t only with the form $B_z := \xi f(x - ct)$ for some nonzero constant ξ. With these choices, Faraday's law $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$ becomes

$$\begin{pmatrix} 0 \\ 0 \\ \partial_x f(x - ct) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -\xi \partial_t f(x - ct) \end{pmatrix}.$$

The above is familiar from our derivation of the one dimensional wave equation in Section 5.4. Indeed, the chain rule gives

$$\partial_x f(x - ct) = f'(x - ct) \qquad -\partial_t f(x - ct) = c f'(x - ct)$$

and so for Faraday's law to hold, we require that $\xi := 1/c$.

The Ampère-Maxwell law is all that is left to check. Indeed,

$$\nabla \times \boldsymbol{B} = \begin{pmatrix} 0 \\ -\frac{1}{c}\partial_x f(x - ct) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{c} f'(x - ct) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{c^2}\partial_t f(x - ct) \\ 0 \end{pmatrix} = \frac{1}{c^2}\partial_t \boldsymbol{E}$$

and the Ampère-Maxwell law holds. As all four equations have been satisfied, we see that

$$\boldsymbol{E} = (0, f(x - ct), 0) \qquad \boldsymbol{B} = (0, 0, f(x - ct)/c)$$

[19] Thus if we move along a line with respect to an inertial reference frame, but with non-zero acceleration (in a car/-train/plane/etc), our laws of nature will appear different from that of an inertial reference frame. Indeed, an apple on the floor of a plane during take off will slide towards the back of the plane instead of staying put.

is a solution to the source free Maxwell's equation. In fact, repeating the above steps but starting with the guess that $\boldsymbol{E} = (0, g(x + ct), 0)$ for some twice differentiable function g gives another solution to the source free Maxwell's equation

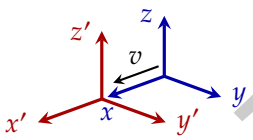
$$\boldsymbol{E} = (0, g(x + ct), 0) \qquad \boldsymbol{B} = (0, 0, -g(z + ct)/c).$$

As the source free Maxwell's equations are linear partial differential equations, the sum of solutions

$$\boldsymbol{E} = (0, f(x - ct) + g(x + ct), 0) \qquad \boldsymbol{B} = (0, 0, f(x - ct)/c - g(x + ct)/c).$$

is also a solution.

Principle of relativity



Let us apply Galileo's principle of relativity to our solution to the source free Maxwell's equations. We begin by considering an inertial observer O whose choice of coordinate system $\mathcal{S}$ is (x, y, z, t), as shown in the diagram above (colored blue). Suppose we have another observer O' whose choice of coordinate system $\mathcal{S}'$ is (x', y', z', t). To keep things simple yet nontrivial, we will assume that this coordinate system's origin is moving at a constant speed v along the shared x/x'-axis as shown in the diagram above (colored red), with both origins aligning exactly (that is, $x = x'$, $y = y'$, $z = z'$) at time $t = t' = 0$. By Galileo's principle of relativity, the coordinate system $\mathcal{S}'$ of O' is also an inertial reference frame.

Each *event* witnessed by observers O and O' takes place at different coordinates because these observers do not share the same coordinate systems. In particular, the coordinates of each event in $\mathcal{S}$ and $\mathcal{S}'$ are related by the equations below, called the **Galilean transformations**.

$$x' = x - vt \qquad y' = y \qquad z' = z \qquad t' = t$$

By construction, the second and third coordinates of each event will always be agreed upon by both observers. Time will always be agreed upon because we are assuming the observer's clocks are functional and synchronized. The only disagreement will be in the event's first coordinate because observer O' is moving ahead of observer O along that axis.

The source free Maxwell's equation had a solution $E_y(x, t) = f(x - ct) + g(x + ct)$. This is true for the observer O. For observer O', applying the Galilean transformations turns our solution into

$$E_y(x, t) = f(x' - [c - v]t') + g(x' + [c + v]t').$$

Observe that a wave traveling at speed c has now turned into a wave that is *not* traveling at speed c, unless v is zero!

This is a huge problem because the same laws of nature must be equally valid for all inertial observers, and the laws of nature (Maxwell's equations) dictate that electromagnetic waves in a vacuum propagates at the speed of light.

By our calculations, there is only one reference frame in which electromagnetic waves in vacuum propagate at the speed c. It follows that there is exactly one distinguished absolute reference frame in which Maxwell's equations are valid. But this conclusions is highly unsatisfactory! We have been given an amazing toy called Maxwell's equations, only to have it immediately yanked away from our hands because we are (very likely) not in this absolute reference frame, and are forbidden from appreciating it for what it is, unless we nail down what this absolute reference frame is.

There are two options. The first option is to accept that Galileo's principle of relativity only holds for Newtonian mechanics, but not for electromagnetism. The second option is to declare that *the principle of relativity must hold for all laws of nature, including electromagnetism*. But as we saw, this causes a contradiction, unless we accept the existence of a preferred absolute reference frame, which defeats the purpose of insisting on a principle of relativity in the first place! To wiggle out of this problem, we will add in a second clause to the second option that *the speed of light in vacuum must be the same for all inertial observers* (regardless of how the source of the light is moving). As the Galilean transformation are not compatible with this second clause, we will need to find another set of coordinate transformations that is.

The second option was first laid out by Albert Einstein and is known as the **postulates of special relativity**. The first clause of the postulates is called the **principle of relativity**. The second clause of the postulates is called the **invariance of** c.

Lorentz transformations

We will accept the postulates of special relativity because the postulates allows us to keep the principle of relativity while removing the need for us to (1) accept that there is an absolute reference frame, (2) figure out what this absolute reference frame is, and (3) translate all our calculations into this absolute reference frame. However, this means that we no longer accept the Galilean transformations as describing coordinate transformations from one inertial reference frame to another. We must find another set of transformations.

Our starting point will have to be the invariance of c. As before, we consider observers O and O' whose reference frames have origins that coincide at time $t = t' = 0$, but thereafter move relative to each other along the shared x/x'-axis at a constant speed v. Now suppose a burst of light is emitted at time $t = t' = 0$.

To observer O, the light signal satisfies the equation $ct = \sqrt{x^2 + y^2 + z^2}$, where the right side is the distance the light traveled and the left side is the time the light traveled multiplied by the (constant) speed of light in vacuum c.

As the speed of light in vacuum is universal, observer O' will also say that $ct' = \sqrt{x'^2 + y'^2 + z'^2}$. Therefore,

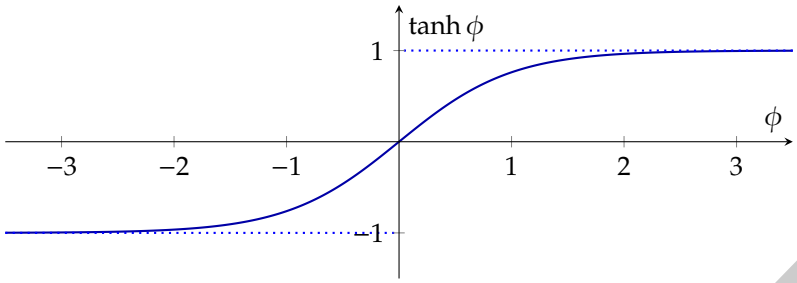
$$(ct)^2 - (x^2 + y^2 + z^2) = 0 = (ct')^2 - (x'^2 + y'^2 + z'^2).$$

Since $y' = y$ and $z' = z$ at all times, we have the simplified equation

$$(ct')^2 - x'^2 = (ct)^2 - x^2. \tag{6.51}$$

For simplicity, let us assume that the coordinate transformation can be done by a real matrix

$$\begin{pmatrix} x' \\ ct' \end{pmatrix} := \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} x \\ ct \end{pmatrix}. \tag{6.52}$$



This means that $x' = \alpha x + \beta(ct)$ and $ct' = \gamma x + \delta(ct)$. Plugging these substitutions into Equation 6.51 gives

$$\left[(\gamma x)^2 + (\delta ct)^2 + 2\gamma\delta cxt\right] - \left[(\alpha x)^2 + (\beta ct)^2 + 2\alpha\beta cxt\right] = (ct)^2 - x^2.$$

After some rearrangement, we have

$$\left[\delta^2 - \beta^2\right](ct)^2 - \left[\alpha^2 - \gamma^2\right]x^2 + \left[2\gamma\delta - 2\alpha\beta\right](cxt) = (ct)^2 - x^2. \tag{6.53}$$

Equation 6.53 is a little too complicated. To simplify, let us study the point $x = y = z = 0$ in the reference frame of observer O and see how that point moves in the reference frame of observer O'. Equation 6.53 then becomes

$$\left[\delta^2 - \beta^2\right](ct)^2 = 1 \cdot (ct)^2. \tag{6.54}$$

We need to find δ and β such that $\delta^2 - \beta^2 = 1$. This is *almost* like the trigonometric identity $\sin^2\phi + \cos^2\phi = 1$. In fact, it is satisfied by the *hyperbolic* identity $\cosh^2\phi - \sinh^2\phi = 1$ (Challenge 14 of Chapter 3). We therefore guess that $\delta := \cosh\phi$ for some ϕ and so there is a transformation rule for time given by $ct' = ct\cosh\phi$.[20] We see that time may not be the same for everyone! This raises serious questions about the concept of absolute time.

Continuing on, we know that $\beta = \pm\sinh\phi$ for some ϕ, with a sign to be determined. Matrix Equation 6.52 with $x = 0$ tells us that

$$x' = \pm ct\sinh\phi.$$

This is a shocking result, as space and time are no longer independent, but intertwined! As our assumption is that the second reference frame moves away from the stationary reference frame at a rate of v to the *left*, the sign of β will be negative. Therefore $x' = -ct\sinh\phi$.

The speed v at which the point (x', y', z') in the second reference frame moves away from the point $(0, 0, 0)$ in the reference frame of observer O is given by dividing $|x'| := ct\sinh\phi$ by $t' := t\cosh\phi$. Hence

$$v = \frac{ct\sinh\phi}{t\cosh\phi} = c\tanh\phi$$

and we define **rapidity** ϕ such that $\tanh\phi := v/c$.

Because the tanh function takes values strictly in the interval $(-1, 1)$ the value $v = c\tanh\phi$ must lie strictly within the interval $(-c, c)$. This means that inertial reference frames cannot move away

[20] We rule out $\delta := -\cosh\phi$ as it would lead to the transformation rule $ct' = -ct\cosh\phi$ where time runs in opposite directions. Notice $\cosh\phi := (e^{\phi} + e^{-\phi})/2$ is a positive function.

from each other with arbitrary speeds. The speed of the relative motion of inertial reference frames must be capped by the speed of light c.

From Challenge 14 we obtained the formulas

$$\sinh\phi = \tanh\phi/\sqrt{1-\tanh^2\phi} \qquad \cosh\phi = 1/\sqrt{1-\tanh^2\phi}$$

which we can combine with the definition of rapidity $\tanh\phi = v/c$ to get the following.

$$x' = -ct\sinh\phi = -\frac{ct\tanh\phi}{\sqrt{1-\tanh^2\phi}} = -\frac{vt}{\sqrt{1-v^2/c^2}}$$
$$t' = t\cosh\phi = \frac{t}{\sqrt{1-\tanh^2\phi}} = \frac{t}{\sqrt{1-v^2/c^2}}$$

Taking $c \to \infty$ so that $v^2/c^2 \to 0$ allows us to recover the Galilean transformations when $x = 0$.

$$x' = -vt \qquad y' = y \qquad z' = z \qquad t' = t$$

This suggest that the Galilean transformations are a special case of this more general transformation rule, where the speeds involved are far smaller than c. Observe that we also recover the correct sign of the parameter β.

Repeating our previous calculations starting from Equation 6.53 with $t = 0$ and $x \neq 0$ tells us that $\alpha = \cosh\phi$ and $\gamma = -\sinh\phi$. The full matrix transformation is then

$$\begin{pmatrix} x' \\ ct' \end{pmatrix} := \begin{pmatrix} \cosh\phi & -\sinh\phi \\ -\sinh\phi & \cosh\phi \end{pmatrix} \begin{pmatrix} x \\ ct \end{pmatrix}. \tag{6.55}$$

Translating Equation 6.55 using the formulas for sinh and cosh as we have done before allows us to recover the full set of transformations. These are the **Lorentz transformations**

$$x' = \gamma(x - vt) \qquad y' = y \qquad z' = z \qquad t' = \gamma\left(t - xv/c^2\right)$$

where the **Lorentz factor** $\gamma := 1/\sqrt{1-v^2/c^2}$. Once again we recover the Galilean transformations by taking $c \to \infty$. For γ to be a real number and thus our transformed coordinates to be real numbers, we require $|v| < c$. Therefore, according to the Lorentz transformations, it is not possible for a physical object to be at or exceed the speed of light.

Challenge 72 Let reference frames $\mathcal{S}$ and $\mathcal{S}'$ be as defined previously.

(a) Show that the Lorentz transformations for transforming coordinates in reference frame $\mathcal{S}'$ to coordinates in reference frame $\mathcal{S}$ are given by the following.

$$x = \gamma(x' + vt') \qquad y = y' \qquad z = z' \qquad t = \gamma\left(t' + x'v/c^2\right)$$

(b) (**Relativity of simultaneity**) Suppose that event A happens at location (x_A, y_A, z_A, t_A) and event B happens at location (x_B, y_B, z_B, t_B) according to reference frame $\mathcal{S}$. If $t_A = t_B$ but $x_A \neq x_B$, use the Lorentz transformation for $t \mapsto t'$ to transform the time coordinate of both events into reference frame $\mathcal{S}'$ and show that $t'_A - t'_B = [x_B - x_A]\,\gamma v/c^2$. Simultaneous events in one reference frame are not simultaneous in a different reference frame!

(c) (**Length contraction**) Suppose we have a stick lying on the x/x'-axis that is moving at speed v to the left along with reference frame $\mathcal{S}'$. Thus the stick is stationary according to reference frame $\mathcal{S}'$ and its length is $l' := x'_r - x'_l$, where x'_r is the location of the right end of the stick in reference frame $\mathcal{S}'$ and x'_l is the location of the left end of the stick in reference frame $\mathcal{S}'$. Let $l := x_r - x_l$ and use the Lorentz transformation for $x \mapsto x'$ to show that $l = l'/\gamma$.[21] Since $\gamma > 1$ when $v \neq 0$, a moving object is shorter along the direction of motion by a factor of γ. Wow!

(d) (**Time dilation**) By part (b), events happening simultaneously in one reference frame at different spatial locations are not simultaneous in other reference frames. To accommodate, each spatial location in each reference frame has a separate clock that keeps track of time in that spatial location, to be transformed as needed. Let us consider the clock at the origin $x' = 0, y' = 0, z' = 0$ in reference frame $\mathcal{S}'$ and suppose it measures the time from $t'_i = 0$ to $t'_f = T'$. Use the Lorentz transformation for the transformation $t' \mapsto t$ from part (a) to show that the clocks in reference frame $\mathcal{S}$ measures a *longer* time interval $T = \gamma T'$.[22] Since the origin of $\mathcal{S}'$ is moving according to $\mathcal{S}$, we see that moving clocks run *slower*!

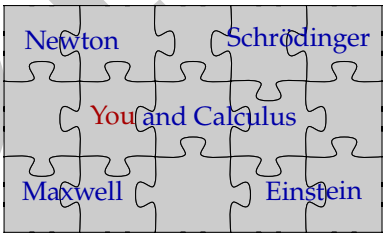
The Lorentz transformations and its consequences (length contraction, the relativity of simultaneity, and time dilation) fly at the face of our normal everyday experiences and sound ridiculous! Yet once we accept the postulates of special relativity, these are the logical conclusions. The math tells us that our intuitive ideas about space and time require such radical reformulations.

You and Calculus

Starting from arithmetic, we have built up the magnificent edifice of calculus with the goal of understanding the world around us. These investigations naturally led us to the revolutionary advances by Newton, Schrödinger, Maxwell, and Einstein. Ironically, our efforts and investigations have led to us to realize how little we truly know. Worse, we have seen that even the most basic things we have assumed known about space, time, and the basic properties of particles were wrong or incomplete.

However, there is yet another complementary point of view, for you could have discovered any of these earth-shattering breakthroughs by investigating the natural questions that arise while sticking to the belief that truth is to be found in simplicity. And the truth is simple.

You could have done it too.



[21] We are considering the measurement of the stick done in reference frame $\mathcal{S}$ and so time t is assumed to be constant. The time t' in reference frame $\mathcal{S}'$ is irrelevant as the stick is stationary at all times in $\mathcal{S}'$.

[22] Notice that x' is fixed at the origin of $\mathcal{S}'$ since we are looking at one particular clock at the origin.

A

Appendix

A.1 The Chain Rule

Consider a real-valued differentiable function f defined in an open interval I and let $s \in I$. Then there is a real number ξ such that the following equation holds for each real number b.

$$f(s + b\epsilon) = f(s) + b\xi\epsilon \tag{A.1}$$

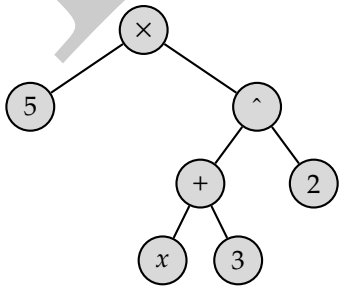
This is the definition of differentiability at s using dual numbers (see Section 2.3), and we denote the number ξ using the symbol $f'(s)$.

One way to interpret Equation A.1 is the following. If f is differentiable at s, then the output of f given input $s + b\epsilon$ takes the form $X + Y\epsilon$, where $X := f(s)$ and $Y := bf'(s)$. As an example, consider the function $f : x \mapsto 5(x + 3)^2$. Using the fact that $\epsilon^2 = 0$ we have

$$f(s + b\epsilon) = 5\left([s + b\epsilon] + 3\right)^2 = 5\left(s^2 + 6s + 9 + [2sb + 6b]\epsilon\right) = 5\left(s^2 + 6s + 9\right) + b\left(10s + 30\right)\epsilon.$$

Therefore $f(s) = 5\left(s^2 + 6s + 9\right)$ and $f'(s) = 10s + 30$, as you should check.

We can be more explicit and introduce a *computational graph* which spells out all the operations contained within a function. The computational graph of the function $f : x \mapsto 5(x + 3)^2$ is shown below.



We can evaluate the value of function f at $s + b\epsilon$ by replacing the input node x by $s + b\epsilon$ and then applying all the computations specified. For example, to find the value of function f at $0 + \epsilon$,

we first add 3 to $0+\epsilon$ to get $3+\epsilon$. Then we square it to get $(3+\epsilon)^2 = 9+6\epsilon$ (recall that $\epsilon^2 := 0$). We then left multiply by the number 5 to get

$$f(0+\epsilon) = 45 + 30\epsilon.$$

By the definition of the derivative, we see that $f(0) = 45$ and $f'(0) = 30$.

Multivariable differentiation

Differentiable functions: $\mathbb{R} \to \mathbb{R}^n$

We now extend our previous discussion to more general functions. First, let us consider a vector-valued function $f := (f_1, f_2, \ldots, f_n)$ defined on an interval I. As was the convention in Chapter 6, each function f_i is a real-valued differentiable function (defined on the interval I). To draw a computational graph for function f, we can draw a computational graph for each function f_i. For example, if $f := (f_1, f_2)$ with $f_1 : x \mapsto 5(x+3)^2$ and $f_2 : x \mapsto \sin(3[x+2])$, then the computational graph is as follows.

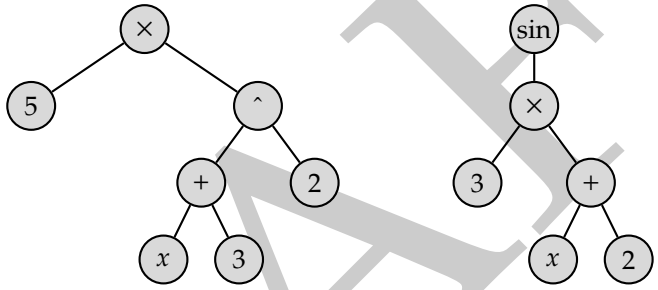


Figure A.2: Computational graph for the function $\big(x \mapsto 5(x+3)^2, x \mapsto \sin(3[x+2])\big)$.

We have already done a computation for the function $f_1 : x \mapsto 5(x+3)^2$ and so we will go through the computational steps for the function $f_2 : x \mapsto \sin(3[x+2])$. We replace the node x with the value $s + b\epsilon$ then add the number 2 to get $(s+2) + b\epsilon$. Then we multiply by 3 to get $(3s+6) + 3b\epsilon$, and then apply the cosine function at the top to obtain $\cos(3s+6+3b\epsilon)$. Since the sine function is differentiable with derivative cos, we see that $\sin(3s+6+3b\epsilon) = \sin(3s+6) + 3b\cos(3s+6)\epsilon$. This means that $f_2(s) = \sin(3s+6)$ and $f_2'(s) = 3\cos(3s+6)$. Therefore, $f(s) = \big(5(s^2+6s+9), \sin(3s+6)\big)^\intercal$ and $f'(s) = (10s+30, 3\cos(3s+6))^\intercal$.

To summarize, a vector-valued function f defined on an interval I is **differentiable** at $s \in I$ if there are real numbers $\xi_1, \xi_2, \ldots, \xi_n$ such that

$$\begin{pmatrix} f_1(s+b\epsilon) \\ f_2(s+b\epsilon) \\ \vdots \\ f_n(s+b\epsilon) \end{pmatrix} = \begin{pmatrix} f_1(s) + b\xi_1\epsilon \\ f_2(s) + b\xi_2\epsilon \\ \vdots \\ f_n(s) + b\xi_n\epsilon \end{pmatrix} \tag{A.3}$$

for each real number b. Each number ξ_i is denoted $f_i'(t)$ and the **derivative** of function f at s is the vector $\big(f_1'(s), f_2'(s), \ldots, f_n'(s)\big)^\intercal$, denoted by $f'(s)$. Equation A.3 can then be written succinctly as

$$f(s+\epsilon b) = f(s) + bf'(s)\epsilon.$$

Differentiable functions: $\mathbb{R}^m \to \mathbb{R}$

We now apply this to multivariable real-valued functions. As a concrete example, consider the function $f : (x_1, x_2, x_3) \mapsto (x_1^2 - x_2)\log(3x_3)$. We can similarly create a computational graph for function f as shown below.

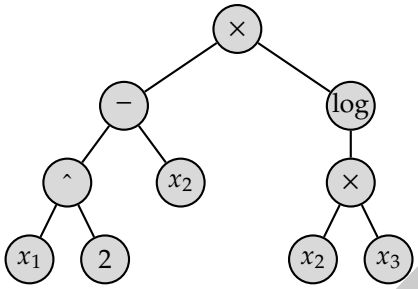


Figure A.4: Computational graph for the function $(x_1, x_2, x_3) \mapsto (x_1^2 - x_2)\log(x_2x_3)$.

As before, we go through the computational graph for an input $s + \epsilon$. The complication is that now there are three different inputs. To handle this, for each x_i we assign the point $s_i + \epsilon e_i$.

Let us calculate the output of function f at $(s_1 + \epsilon(1,0,0), s_2 + \epsilon(0,1,0), s_3 + \epsilon(0,0,1))$. We start from the bottom left and square $x_1 = (s_1 + \epsilon(1,0,0))$ to get

$$x_1^2 = [s_1 + \epsilon(1,0,0)]\,[s_1 + \epsilon(1,0,0)] = s_1^2 + \epsilon(s_1,0,0) + \epsilon(s_1,0,0) + \cancel{\epsilon^2(1,0,0)} = s_1^2 + \epsilon(2s_1,0,0).$$

Next, we subtract $x_2 := s_2 + \epsilon(0,1,0)$ and obtain

$$x_1^2 - x_2 = \left[s_1^2 + \epsilon(2s_1^2,0,0)\right] - [s_2 + \epsilon(0,1,0)] = s_1^2 - s_2 + \epsilon(2s_1^2,-1,0).$$

Shifting to the nodes on the right gives us

$$\log(x_2x_3) = \log\left([s_2 + \epsilon(0,1,0)]\,[s_3 + \epsilon(0,0,1)]\right) = \log\left(s_2s_3 + \epsilon(0,s_3,s_2)\right).$$

Recall that the logarithm function is differentiable and so $\log(x + a\epsilon) = \log(x) + a\frac{1}{x}\epsilon$. Hence

$$\log(x_2x_3) = \log\left(s_2s_3 + \epsilon(0,s_3,s_2)\right) = \log(s_2s_3) + \epsilon(0,1/s_3,1/s_2).$$

Therefore,

$$\begin{aligned} f(x_1,x_2,x_3) &= \left[s_1^2 - s_2 + \epsilon\left(2s_1^2,-1,0\right)\right]\left[\log(s_2s_3) + \epsilon\left(0,1/s_3,1/s_2\right)\right] \\ &= (s_1^2 - s_2)\log(s_2s_3) + \epsilon\left(0,[s_1^2 - s_2]/s_3,[s_1^2 - s_2]/s_2\right) + \epsilon\left(2s_1^2\log(s_2s_3), -\log(s_2s_3), 0\right) \\ &= (s_1^2 - s_2)\log(s_2s_3) + \epsilon\left(2s_1^2\log(s_2s_3),[s_1^2 - s_2]/s_3 - \log(s_2s_3),[s_1^2 - s_2]/s_2\right) \end{aligned}$$

and we see that $f(s_1,s_2,s_3) = (s_1^2 - s_2)\log(s_2s_3)$ with

$$\partial_1 f(s_1,s_2,s_3) = 2s_1^2\log(s_2s_3) \quad \partial_2 f(s) = [s_1^2 - s_2]/s_3 - \log(s_2s_3) \quad \partial_3 f(s_1,s_2,s_3) = [s_1^2 - s_2]/s_2.$$

To summarize, a real-valued function f defined on an open rectangle R in $\mathbb{R}^m$ is **differentiable** at $t := (s_1, s_2, \ldots, s_m) \in R$ if there are real numbers $\xi_1, \xi_2, \ldots, \xi_m$ such that

$$f\begin{pmatrix} s_1 + b_1\epsilon \\ s_2 + b_2\epsilon \\ \vdots \\ s_m + b_m\epsilon \end{pmatrix} = f\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix} + \begin{pmatrix} \xi_1 & \xi_2 & \cdots & \xi_m \end{pmatrix}\begin{pmatrix} b_1\epsilon \\ b_2\epsilon \\ \vdots \\ b_m\epsilon \end{pmatrix}. \tag{A.5}$$

for each collection of real numbers $b_1, b_2, \ldots, b_m$. We denote each ξ_i by $\partial_i f(t)$ and define the **derivative** of f at s by $f'(s) := \big(\partial_1 f(s), \partial_2 f(s), \cdots, \partial_m f(s)\big)$. If we take $b := (b_1, b_2, \cdots, b_m)^\intercal$, then Equation A.5 can be written succinctly as

$$f(s + \epsilon b) = f(s) + f'(s)\epsilon b. \tag{A.6}$$

The second term on the right can be thought of as a matrix vector product between $f'(s)$ and ϵb. In reality, just as the derivative f' of a single variable function is a function, the derivative f' is also a function. However, if we represent $f'(s)$ as a vector, we obtain the *gradient*, which we denote using the symbol $\nabla f(s)$. If all of the real numbers b_i are zero except b_k, then Equation A.5 gives the definition of the **partial derivative** $\partial_k f(s)$.

Differentiable functions: $\mathbb{R}^m \to \mathbb{R}^n$

If we package several multivariable real-valued functions into one, we have a multivariable vector valued function $f := (f_1, f_2, \ldots, f_n)$ defined in a rectangle R in $\mathbb{R}^m$.[1] The only real challenge here is to package things together neatly so that we can simplify the notation.

We begin by reducing to the previous case, considering each scalar-valued function separately. Indeed, we already know that if f_i is differentiable at $s \in R$, then $f_i(s + \epsilon b_i) = f_i(s) + f_i'(s)\epsilon b_i$ for each real vector b_i.[2] We will combine these equations into one, as we did in Equation A.3.

Vector-valued function f is **differentiable** at $s \in R$ if the following equation holds for each real vector $b := (b_1, b_2, \ldots, b_n)$.

$$\begin{pmatrix} f_1(s + \epsilon b) \\ f_2(s + \epsilon b) \\ \vdots \\ f_n(s + \epsilon b) \end{pmatrix} = \begin{pmatrix} f_1(s) + f_1'(s)\epsilon b \\ f_2(s) + f_2'(s)\epsilon b \\ \vdots \\ f_n(s) + f_n'(s)\epsilon b \end{pmatrix} \tag{A.7}$$

The **derivative** of f at $s \in R$, written $f'(s)$, is defined to be the matrix

$$f'(s) := \begin{pmatrix} f_1'(s) \\ f_2'(s) \\ \vdots \\ f_n'(s) \end{pmatrix}.$$

If we write out the matrix $f'(s)$ using the definition of the derivative of a real-valued function $f_i'(s) := \big(\partial_1 f_i(s), \partial_2 f_i(s), \cdots, \partial_m f_i(s)\big)$, we see that the matrix is none other than the Jacobian matrix.

[1] As was the convention in Chapter 6, each function f_i is a real-valued function defined on rectangle R.

[2] This is simply Equation A.6 indexed by $i \in \{1, 2, \ldots, n\}$.

Here is an example of Equation A.7 written out when $m = 2$ and $n = 3$.

$$\begin{pmatrix} f_1(s_1 + b_1\epsilon, s_2 + b_2\epsilon) \\ f_2(s_1 + b_1\epsilon, s_2 + b_2\epsilon) \\ f_3(s_1 + b_1\epsilon, s_2 + b_2\epsilon) \end{pmatrix} = \begin{pmatrix} f_1(s_1, s_2) \\ f_2(s_1, s_2) \\ f_3(s_1, s_2) \end{pmatrix} + \begin{pmatrix} \partial_1 f_1(s_1)b_1\epsilon & \partial_2 f_1(s_2)b_2\epsilon \\ \partial_1 f_2(s_1)b_1\epsilon & \partial_2 f_2(s_2)b_2\epsilon \\ \partial_1 f_3(s_1)b_1\epsilon & \partial_2 f_3(s_2)b_2\epsilon \end{pmatrix}$$

Observe that the second term on the right is the matrix-vector multiplication of matrix $f'(s)$ and vector ϵb where $b := (b_1, b_2)^\intercal$. Indeed,

$$f'(s)\epsilon b = \begin{pmatrix} \partial_1 f_1(s) & \partial_2 f_1(s) \\ \partial_1 f_2(s) & \partial_2 f_2(s) \\ \partial_1 f_3(s) & \partial_2 f_3(s) \end{pmatrix} \begin{pmatrix} b_1\epsilon \\ b_2\epsilon \end{pmatrix} = \begin{pmatrix} \partial_1 f_1(s_1)b_1\epsilon & \partial_2 f_1(s_2)b_2\epsilon \\ \partial_1 f_2(s_1)b_1\epsilon & \partial_2 f_2(s_2)b_2\epsilon \\ \partial_1 f_3(s_1)b_1\epsilon & \partial_2 f_3(s_2)b_2\epsilon \end{pmatrix}.$$

Therefore, Equation A.7 simplifies to

$$f(s + \epsilon b) = f(s) + f'(s)\epsilon b.$$

Challenge 73

(a) The complex function $f : z \mapsto z$ maybe interpreted as a function of two real variables that outputs two real numbers by representing f as $f(x, y) = \begin{pmatrix} x \\ y \end{pmatrix}$ where $x := \operatorname{Re} z$ and $y := \operatorname{Im} z$. Calculate the Jacobian matrix J_f and interpret the real matrix as a complex number.

(b) We can represent the complex function $g : z \mapsto z^2$ by $g(x, y) = \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix}$ where $x := \operatorname{Re} z$ and $y := \operatorname{Im} z$. Calculate the Jacobian matrix J_g and interpret the real matrix as a complex function. Is it as you expected?

(c) If f is a differentiable complex-valued function whose inputs are complex numbers, then J_f must have the form $\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$ because the derivative of f better be complex! Conclude that a complex-valued function $f(x + iy) := u(x, y) + iv(x, y)$ for real x and y is differentiable if the **Cauchy-Riemann equations** $\partial_x u = \partial_y v$ and $\partial_y u = -\partial_x v$ hold.

The chain rule

One of the great things about dual numbers is that once the basic derivatives are known, derivatives of composition of differentiable functions with known derivatives is a straightforward arithmetic. We have seen this in action while going through computational graphs where we calculated derivatives without calculating derivatives!

This means that the chain rule becomes a trivial, self-evident statement. We have already seen this in Section 2.3. As a review, let us re-derive the single variable chain rule. Suppose we have an open interval I and differentiable functions $g : I \to \mathbb{R}$ and $f : g(I) \to \mathbb{R}$. Since g and f are differentiable, the following equations hold for each $s \in I$, $t \in g(I)$, and real numbers a, b.

$$g(s + a\epsilon) = g(s) + ag'(s)\epsilon \qquad\qquad f(t + b\epsilon) = f(t) + f'(t)b\epsilon$$

Put $t := g(s)$ and $b := ag'(s)$ to chain the functions together. Then $g(s + a\epsilon) = t + b\epsilon$ and so

$$f\left(g(s + a\epsilon)\right) = f\left(g(s)\right) + af'\left(g(s)\right)g'(s)\epsilon.$$

Therefore $f \circ g$ is differentiable with $(f \circ g)' : s \mapsto (f' \circ g)(s)g'(s)$. We now generalize.

Theorem 82 (The Chain Rule)**.** Let $g := (g_1, g_2, \ldots, g_n)$ be a function defined on an open rectangle R in $\mathbb{R}^m$ and let $f := \left(f_1, f_2, \ldots, f_l\right)$ be a function defined on $g(R)$. If g is differentiable at s and f is differentiable at $g(s)$, then $f \circ g$ is differentiable at s with $(f \circ g)'(s) = (f' \circ g)(s)g'(s)$.

Proof. Since g is differentiable at s and f is differentiable at $t := g(s)$, the following equations hold for real vectors $a := (a_1, a_2, \ldots, a_m)^\intercal$ and $b := (b_1, b_2, \ldots, b_n)^\intercal$.

$$g(s + \epsilon a) = g(s) + g'(s)\epsilon a \qquad\qquad f(t + \epsilon b) = f(t) + f'(t)\epsilon b$$

Put $b := g'(s)a$ to chain the functions together. Then $g(s + a\epsilon) = t + \epsilon b$ and so

$$f\left(g(s + \epsilon a)\right) = f\left(g(s)\right) + f'\left(g(s)\right) g'(s)\epsilon a.$$

Therefore $f \circ g$ is differentiable at s with $(f \circ g)'(s) = (f' \circ g)(s)g'(s)$. □

Let us unpack what the chain rule says. Let $g := (g_1, g_2)$ be a differentiable function defined on an open rectangle R in $\mathbb{R}^3$ and let $f := (f_1, f_2, f_3)$ be a differentiable function defined on $g(R)$. Since g is differentiable at $s := (s_1, s_2, s_3) \in R$, for each real vector $a := (a_1, a_2, a_3)$, the following equation holds.

$$\begin{pmatrix} g_1(s_1 + a_1\epsilon, s_2 + a_2\epsilon, s_3 + a_3\epsilon) \\ g_2(s_1 + a_1\epsilon, s_2 + a_2\epsilon, s_3 + a_3\epsilon) \end{pmatrix} = \begin{pmatrix} g_1(s) \\ g_2(s) \end{pmatrix} + \begin{pmatrix} \partial_1 g_1(s) & \partial_2 g_1(s) & \partial_3 g_1(s) \\ \partial_1 g_2(s) & \partial_2 g_2(s) & \partial_3 g_2(s) \end{pmatrix} \epsilon \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

Similarly, as f is differentiable at each $t := (t_1, t_2) \in g(R)$, the following equation holds for each collection of real numbers b_1 and b_2.

$$\begin{pmatrix} f_1(t_1 + b_1\epsilon, t_2 + b_2\epsilon) \\ f_2(t_1 + b_1\epsilon, t_2 + b_2\epsilon) \\ f_3(t_1 + b_1\epsilon, t_2 + b_2\epsilon) \end{pmatrix} = \begin{pmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \end{pmatrix} + \begin{pmatrix} \partial_1 f_1(t) & \partial_2 f_1(t) \\ \partial_1 f_2(t) & \partial_2 f_2(t) \\ \partial_1 f_3(t) & \partial_2 f_3(t) \end{pmatrix} \epsilon \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{A.8}$$

To chain the two functions together, define the following

$$\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} := \begin{pmatrix} g_1(s) \\ g_2(s) \end{pmatrix} \qquad\qquad \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} := \begin{pmatrix} \partial_1 g_1(s) & \partial_2 g_1(s) & \partial_3 g_1(s) \\ \partial_1 g_2(s) & \partial_2 g_2(s) & \partial_3 g_2(s) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

so that $t := g(s)$, $b := g'(s)a$, and $g(s + \epsilon a) = t + \epsilon b$. Then Equation A.8 becomes

$$(f \circ g)(s + a\epsilon) = f \circ g(s) + \left[f' \circ g(s)\right]\left[g'(s)\epsilon a\right].$$

Challenge 74

(a) Let $\mathrm{id} : \mathbb{R}^n \to \mathbb{R}^n$ be defined by $\mathrm{id} : v \mapsto v$. Check that $\mathrm{id}'(t) = 1$. Let $A : \mathbb{R}^m \to \mathbb{R}^n$ be *linear* (that is, $A[cv + dw] = cAv + dAw$).[3] Check that A is differentiable with $A'(t) = A$.

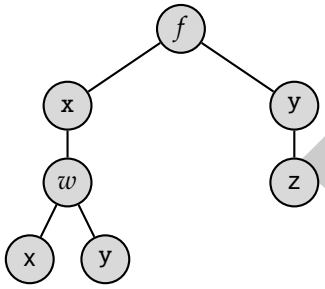
(b) Let $g := (g_1, g_2, \ldots, g_n)$ be a differentiable function defined on an open rectangle R in $\mathbb{R}^m$ and let f be a differentiable real-valued function defined on $g(R)$. If $s \in R$, show that

$$\partial_i(f \circ g)(s) = \sum_{j=1}^{n} \partial_j f\left(g(t)\right) \cdot \partial_i g_j(s). \tag{A.9}$$

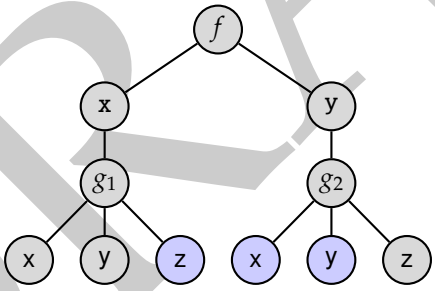
[3] As a linear function $\mathbb{R}^m \to \mathbb{R}^n$ is essentially a matrix, $A(v)$ is the same as Av. For example, $1(t)$ is the same as $1t$.

Although Equation A.9 is a special case of the general chain rule, it really embodies everything one needs to know about the general chain rule. Indeed, if we want to compute the derivative of a multivariable vector-valued function $h := (h_1, h_2, \ldots, h_l)$ that is a composition of functions, it is sufficient to consider each scalar valued function h_i separately and compute the derivative h_i', which in vector form is just the gradient ∇h_i. But to compute the gradient ∇h_i, it is sufficient to compute the partial derivatives $\partial_k h_i$ separately, for which Equation A.9 can be used.

Sometimes we encounter functions with interdependencies that do not admit an easy application of Equation A.9. This is not a problem of course, as we know how to take derivatives of arbitrary compositions of differentiable functions without actually calculating derivatives. Nevertheless, consider the function $h : (\mathsf{x}, \mathsf{y}, \mathsf{z}) \mapsto f\,(w(\mathsf{x}, \mathsf{y}), \mathsf{z})$ whose *dependancy graph* is shown below.



If we wish to apply the chain rule, we encounter a problem. To fix this, define the functions $g_1 : (\mathsf{x}, \mathsf{y}, \mathsf{z}) \mapsto w(\mathsf{x}, \mathsf{y})$ and $g_2 : (\mathsf{x}, \mathsf{y}, \mathsf{z}) \mapsto \mathsf{z}$ so that $h = f \circ g$, where $g := (g_1, g_2)$. The dependancy graph of function h with these new functions is shown below, where the nodes with a blue shade represent the "phantom" inputs which do nothing.



We can then apply Equation A.9 to obtain the following partial derivatives at $s := (s_1, s_2, s_3)$.

$$\partial_{\mathsf{x}}(f \circ g)(s) = \partial_{\mathsf{x}} f\left(g(s)\right) \cdot \partial_{\mathsf{x}} g_1(s) + \cancel{\partial_{\mathsf{y}} f\left(g(s)\right) \cdot \partial_{\mathsf{x}} g_2(s)} = \partial_{\mathsf{x}} f\left(w(s_1, s_2), s_3\right) \cdot \partial_{\mathsf{x}} w(s_1, s_2)$$

$$\partial_{\mathsf{y}}(f \circ g)(s) = \partial_{\mathsf{x}} f\left(g(s)\right) \cdot \partial_{\mathsf{y}} g_1(s) + \cancel{\partial_{\mathsf{y}} f\left(g(s)\right) \cdot \partial_{\mathsf{y}} g_2(s)} = \partial_{\mathsf{x}} f\left(w(s_1, s_2), s_3\right) \cdot \partial_{\mathsf{y}} w(s_1, s_2)$$

$$\partial_{\mathsf{z}}(f \circ g)(s) = \cancel{\partial_{\mathsf{x}} f\left(g(s)\right) \cdot \partial_{\mathsf{z}} g_1(s)} + \partial_{\mathsf{y}} f\left(g(s)\right) \cdot \partial_{\mathsf{z}} g_2(s) = \partial_{\mathsf{y}} f\left(w(s_1, s_2), s_3\right) \cdot 1$$

B

Answers

Below are the answers to the Challenges in the first four chapters.

Challenge 1 (a)

$$5^5 \times \frac{3}{5^3} = (5 \times 5 \times \cancel{5} \times \cancel{5} \times \cancel{5}) \times \frac{3}{\cancel{5} \times \cancel{5} \times \cancel{5}} = 5 \times 5 \times 3 = 5^2 \times 3,$$

$$\frac{5^5 \times \frac{3}{5^3}}{2^2} = \frac{5^2 \times 3}{2^2}.$$

(b) Since density is **M**ass/**L**ength^3, we have

$$\frac{\textbf{L}\text{ength}^5 \times \text{density}}{\textbf{T}\text{ime}^2} = \frac{\textbf{L}\text{ength}^5 \times \textbf{M}\text{ass}/\textbf{L}\text{ength}^3}{\textbf{T}\text{ime}^2} = \frac{\textbf{L}\text{ength}^2 \times \textbf{M}\text{ass}}{\textbf{T}\text{ime}^2}.$$

This is the same as Energy.

(c) Only the second expression is valid.

Challenge 2 (a) We do a few check to verify that the equation $E = R^5\rho/t^2$ makes sense. Indeed, an increase in blast radius R is associated with an increase in energy E. If R and time t were the same, but density ρ was higher, then E must have been higher as well. It time t to reach radius R was smaller, then E must have been smaller. These facts agree with the equation $E = R^5\rho/t^2$.

(b) The radius R of the blast looks to be about 150 meters to me. The time t is given as 0.025 seconds. With $\rho = 1.2\ \text{kg/m}^3$, we have

$$E = \frac{1.5^5 \cdot 100^5\ \text{m}^5 \cdot 1\ \text{kg/m}^3}{0.025^2\ \text{s}} = \frac{1.5^5 \cdot 10^{10}\ \text{m}^5 \cdot 1\ \text{kg/m}^3}{2.5^2 \cdot 10^{-4}\ \text{s}}$$

and a value E of about $1.2 \times 10^{14}\ \text{kg}\cdot\text{m}^2/\text{s}^2$.

(c)

$$\frac{1.2 \times 10^{14}\ \text{joule}}{4.2 \cdot 10^9} = 0.29 \cdot 10^5\ \text{tons of TNT}.$$

Divide by a thousand (10^3) to get that E is about 29 kilotons.

(d) Looking up the yield at `https://en.wikipedia.org/wiki/Trinity_(nuclear_test)`, we can see that the yield was actually about 25 kilotons. The closest integer value of β is 1.

Challenge 3 (a) Using the fact that $(a+b)(c+d) = ac + ad + bc + bd$, we get

$$(10+x)(10+y) = 10\cdot 10 + 10\cdot y + 10\cdot x + x\cdot y.$$

The first three terms are multiples of 10. Therefore,

$$(10+x)(10+y) = 10\cdot 10 + 10\cdot y + 10\cdot x + x\cdot y = 10\cdot(10+x+y) + xy.$$

(b) Using the formula obtained from the previous part, with $x = 6$ and $y = 4$, we need only add a zero to the sum $10+6+4$, then add 24 to get 224.

In order to do the multiplication $116\cdot 114$, we use the same formula from part b, except we switch the '10' with '100':

$$(10+x)(10+y) = 10\cdot(10+x+y) + xy \longrightarrow (100+x)(100+y) = 100\cdot(100+x+y) + xy.$$

Hence we need only add two zeros to the sum $100+16+14$, and add $16\cdot 14$, which we already know to be 224, to get 13224.

Challenge 4 We already know that the equation $(f_1 + f_2 + \cdots + f_n)' = f_1' + f_2' + \cdots + f_n'$ holds if n is 1 or 2. Let S be the collection of positive natural numbers for which the equation does not hold. If S is nonempty, by the well-ordering principle, there is a smallest positive natural number m belong to collection S. Notice that $m > 2$ and that the natural number $m-1$ does not belong in the collection S. Let $X := f_1 + f_2 + \cdots + f_{m-1}$. Since $m-1$ does not belong in the collection S, we know that

$$(f_1 + f_2 + \cdots + f_{m-1})' = X' = f_1' + f_2' + \cdots + f_{m-1}'.$$

By the sum rule, $(X + f_m)' = X' + f_m'$ and so

$$(f_1 + f_2 + \cdots + f_{m-1} + f_m)' = (X + f_m)' = X' + f_m' = f_1' + f_2' + \cdots + f_{m-1}' + f_m'.$$

We see that the equation holds for m, and so natural number m does not belong in collection S. Since S has no smallest element, it must be an empty collection.

Challenge 5 (a) We can write $\sum_{i=1}^{k} i^3$. Of course, the choice of the letter "i" is arbitrary. The expressions $\sum_{a=1}^{k} a^3$ or $\sum_{\tau=1}^{k} \tau^3$, and so on would be perfectly acceptable, as long as the indexing variable used is a single symbol.

(b) Since $(1+X)^3 = (1+X)^2(1+X)$, or equivalently, $(1+X)^3 = (1+X)(1+X)^2$, we have

$$\begin{aligned}(1+X)^3 &= (1+X)^2(1+X) = (1+2X+X^2)(1+X)\\ &= (1+2X+X^2) + X(1+2X+X^2) = 1+3X+3X^2+X^3.\end{aligned}$$

(c) Certainly $1 = 1\cdot 2/2$, $1+2 = 3 = 2\cdot 3/2$, and $1+2+3 = 6 = 3\cdot 4/2$.

(d) Let S be the collection of positive natural numbers n such that $\sum_{k=1}^{n} k \neq n(n+1)/2$. If S is not an empty collection, by the well-ordering principle there is some smallest natural number m, greater than 3, such that m belongs to the collection S. In particular, $m-1$ does not belong in the collection S, and so the following holds.

$$\sum_{k=1}^{m-1} k = \frac{(m-1)m}{2}$$

Then

$$\sum_{k=1}^{m} k = m + \sum_{k=1}^{m-1} k = m + \frac{(m-1)m}{2} = \frac{2m}{2} + \frac{m^2-m}{2} = \frac{m^2+m}{2} = \frac{m(m+1)}{2}$$

which shows that m does not belong to the collection S. Therefore, S is an empty collection, and the equation $\sum_{k=1}^{n} k = n(n+1)/2$ is true for each natural number n.

Challenge 6 (a) If $a \geq 0$ and $b \geq 0$, then $ab \geq 0$ and so

$$|a|b| = ab = |ab|.$$

If $a \leq 0$ and $b \leq 0$, then $ab \geq 0$ and so

$$|a||b| = -a \cdot (-b) = ab = |ab|.$$

The only remaining case is where exactly one is positive and the other is negative. Let us suppose that $a < 0$ and $b > 0$. Then their product ab is negative and we have

$$|ab| = -(ab) = -a \cdot b = |a| \cdot |b|.$$

(b) By part (a) $|1/b||b| = |(1/b) \cdot b|$ and so

$$|1/b||b| = |(1/b) \cdot b| = |1| = 1.$$

Dividing both sides of the equation $|1/b||b| = 1$ by nonzero $|b|$, we have $|1/b| = 1/|b|$.

(c) By part (b), $1/|b| = |1/b|$ and so

$$|a|/|b| = |a| \cdot (1/|b|) = |a| \cdot |1/b|.$$

By part (a), $|a| \cdot |1/b| = |a \cdot (1/b)|$ and so

$$|a|/|b| = |a| \cdot |1/b| = |a \cdot (1/b)| = |a/b|.$$

(d) It is sufficient to show that $|a| \leq |a-b| + |b|$. But this is simply the triangle inequality $|c+d| \leq |c| + |d|$ for $c := a - b$ and $d := b$. Done!

Challenge 7 (a) The terms $f^{(k)}(0)/k!$ are the coefficients of the polynomial f. For example, if f is the polynomial $7x^5 + 2x^3 + 5$, then $f^{(5)}(0)/5! = 7$, $f^{(4)}(0)/4! = 0$, $f^{(3)}(0)/3! = 2$, $f^{(2)}(0)/2! = 0$, $f^{(1)}(0)/1! = 0$, and $f^{(0)}(0) = 5$.

(b) Applying the power rule with the chain rule,

$$f^{(0)} = (x+b)^n, \qquad f^{(1)}(x) = n(x+b)^{n-1}, \qquad f^{(2)}(x) = n(n-1)(x+b)^{n-2}.$$

Since $\binom{n}{0} = 1$, $\binom{n}{1} = \frac{n}{1}$ and $\binom{n}{2} = \frac{n(n-1)}{2\cdot 1}$, we have

$$f^{(0)} = 0!\binom{n}{0}(x+b)^{n-0} \qquad f^{(1)}(x) = 1!\binom{n}{1}(x+b)^{n-1}, \qquad f^{(2)}(x) = 2!\binom{n}{2}(x+b)^{n-2}.$$

Thus if $k \leq n$, we can guess that

$$f^{(k)}(x) = k!\binom{n}{k}(x+b)^{n-k}.$$

(c) The expression $(x+b)^n$ is a polynomial of degree n. Applying the result of part (b) to part (a) gives

$$(x+b)^n = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!}x^k = \sum_{k=0}^{n} \frac{1}{k!}\left(k!\binom{n}{k}b^{n-k}\right)x^k = \sum_{k=0}^{n}\binom{n}{k}x^k b^{n-k}.$$

Substituting x with a, we obtain the binomial formula

$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}.$$

Challenge 8 (a) $(f \circ g)(x) = (g \circ f)(x) = x^2$ and so $(f \circ g)'(0) = (g \circ f)'(0) = 0$.

(b) The relu function is the zero function for negative inputs and $x \mapsto x$ for $x \geq 0$. Thus for negative inputs, the derivative is zero while for positive values, $\text{relu}'(x) = 1$. However, the relu function is *not* differentiable at 0. If $\alpha > 0$, then

$$\text{relu}(0+\alpha) = 0 + \alpha = \text{relu}(0) + 1 \cdot \alpha$$

and so it appears that $\text{relu}'(0) = 1$, but if $\alpha < 0$, then

$$\text{relu}(0+\alpha) = 0 + 0 = \text{relu}(0) + 0 \cdot \alpha$$

and so it appears that $\text{relu}'(0) = 0$, which shows that the relu function is not differentiable at 0.

(c) Let $g := f \circ \text{relu}$, $h := \text{relu} \circ f$ and observe that $g(0) = h(0) = f(0) = 0$. If $\alpha > 0$, then by the power rule:

$$g(0+\alpha) = f(0+\alpha) = f(0) + f'(0)\cdot\alpha + |\alpha|o(1) = g(0) + 0\cdot\alpha + |\alpha|o(1)$$

and if $\alpha < 0$, then

$$g(0+\alpha) = 0 = 0 + 0\cdot\alpha.$$

Therefore, $g'(0) = 0$. Repeating the same argument, replacing the label g with h, gives $h'(0) = 0$. Notice that the argument does *not* work if $f : x \mapsto x$ because $f'(0) = 1$.

(d) If $f'(0)$ exists, then by the sum rule, the relu function is differentiable at 0, which is false by part (b). Thus $f'(0)$ does not exist. Similarly, $g'(0)$ does not exist. However, $f \circ g = g \circ f = 0$, and so $(f \circ g)'(0) = (g \circ f)'(0) = 0$.

Challenge 9 We cannot conclude that h is continuous! If h is a function that it zero everywhere except at the origin, where $h(0) = 1$, then h is not continuous, even though $h(t+\alpha) - h(t-\alpha) = o(1)$ for each input t.

Challenge 10 (a) Dimensional analysis gives $x = c_1 x_i + c_2(v_i + v)t$. The constants are $c_1 = 1$ and $c_2 = 1/2$.

(b) Dimensional analysis gives $x = c_1 x_i + c_2 v_i t + c_3 a t^2$. The constants are $c_1 = 1$, $c_2 = 1$, and $c_3 = 1/2$.

(c) Dimensional analysis gives $v = c_1 v_i + c_2 a t$. The constants are $c_1 = 1$ and $c_2 = 1$.

(d) Dimensional analysis gives $v^2 = c_1 v_i^2 + c_2 a(x - x_i)$. The constants are $c_1 = 1$ and $c_2 = 2$.

(e) Indeed, $v = x' = v_i + at$ and $x'' = a$. Furthermore, following the instructions recovers the formula $v^2 = v_i^2 + 2a(x - x_i)$ of part (d).

Challenge 11 (a)

$$\int_0^h \pi (rx/h)^2 \, dx = \frac{\pi r^2}{h^2} \cdot \frac{h^3}{3} = \frac{\pi r^2 h}{3}.$$

(b)

$$\int_{-r}^{r} \pi \left(\sqrt{r^2 - x^2}\right)^2 \, dx = \left(\pi r^2 x - \pi \frac{x^3}{3}\right)\Bigg|_{-r}^{r} = \frac{4}{3}\pi r^3.$$

Challenge 12 (a) Since $\frac{de^x}{dx} = e^x$, we see that the exponential function is its own velocity. For the units to match up, x must be dimensionless.
Since $\frac{d \log x}{dx} = x^{-1}$, for the units to match up, $\log x$ must be dimensionless.

(b) From the definition $\log(1) = 0$. If the logarithm function accepts inputs with units, then we get the contradiction that $\log(1\text{ m}) = 0$ and $0 = \log(1\text{ km}) = \log(1000\text{ m})$. The number 1 in the definition of log must be dimensionless, and as we are integrating from 1 to x, the latter must also be dimensionless.
Since $\frac{d^2 \cos x}{dx^2} = -\cos x$, the acceleration of $\cos x$ is itself (with a minus sign). To match the units, we need x to be dimensionless. The same reasoning applies to the sin function.

(c) Since $\log xy = \log x + y$,

$$\log e^x e^y = \log e^x + \log e^y = x + y = \log e^{x+y}.$$

Apply exp to both sides of $\log e^x e^y = \log e^{x+y}$. The fact that $e^0 = 1$ follows from:

$$e^x e^0 = e^{x+0} = e^x.$$

Challenge 13 (a) Certainly $f_e(x) = f_e(-x)$ and $f_o(x) = -f_o(-x)$ are both true for each x.

(b) Since f_1 is even and f_2 is odd, we know that

$$f(-x) = f_1(-x) + f_2(-x) = f_1(x) - f_2(x).$$

Solving for f_1 and f_2 gives the following.

$$f_1(x) = f(-x) + f_2(x), \qquad f_2(x) = f_1(x) - f(-x)$$

We use the decomposition $f(x) = f_1(x) + f_2(x)$ to get

$$f_1(x) = f(-x) + f(x) - f_1(x), \qquad f_2(x) = f(x) - f_2(x) - f(-x).$$

Observe that $f_1(x)$ is precisely f_e and $f_2(x)$ is precisely f_o as defined in part (a).

Challenge 14 (a) $\cosh^2 x - \sinh^2 x = (e^x + e^{-x})^2/4 - (e^x - e^{-x})^2/4 = (e^{2x} + 2e^x e^{-x} + e^{-2x} - e^{2x} + 2e^x e^{-x} - e^{-2x})/4 = (4e^x e^{-x})/4 = e^{x-x} = e^0 = 1.$

(b) $\sinh x \cosh y + \cosh x \sinh y = [(e^x - e^{-x})(e^y + e^{-y}) + (e^x + e^{-x})(e^y - e^{-y})]/4 = [e^{x+y} + e^{x-y} - e^{-x+y} - e^{-x-y} + e^{x+y} - e^{x-y} + e^{-x+y} - e^{-x-y}]/4 = (e^{x+y} - e^{-x-y})/2 = \sinh\big(x + y\big).$

(c) $\dfrac{\tanh x}{\sqrt{1-\tanh^2 x}} = \dfrac{\tanh x}{\sqrt{(\cosh^2 x)/(\cosh^2 x) - (\sinh^2 x)/(\cosh^2 x)}} = \dfrac{\tanh x}{\sqrt{1/(\cosh^2 x)}} = \dfrac{\sinh x/\cosh x}{1/\cosh x} = \sinh x.$

(d) $\dfrac{1}{\sqrt{1-\tanh^2 x}} = \dfrac{1}{\sqrt{(\cosh^2 x)/(\cosh^2 x) - (\sinh^2 x)/(\cosh^2 x)}} = \dfrac{1}{\sqrt{1/(\cosh^2 x)}} = \dfrac{1}{1/\cosh x} = \cosh x.$

(e)

$$\sinh' x = \left(\frac{e^x - e^{-x}}{2}\right)' = \cosh x, \qquad \cosh' x = \left(\frac{e^x + e^{-x}}{2}\right)' = \sinh x.$$

By the quotient rule,

$$\tanh' x = \frac{\sinh' x \cosh x - \sinh x \cosh' x}{\cosh^2 x} = \frac{\cosh x \cosh x - \sinh x \sinh x}{\cosh^2 x} = \frac{1}{\cosh^2 x}.$$

Challenge 15 (a) First, $a^0 = e^{0 \log a} = e^0 = 1$. Second,

$$a^x \cdot a^{-x} = e^{x \log a} \cdot e^{-x \log a} = e^{x \log a - x \log a} = e^0 = 1.$$

Third, $a^x a^y = e^{x \log a} e^{y \log a} = e^{(x+y) \log a} = a^{x+y}$. Fourth,

$$(a \cdot b)^x = e^{x \log(ab)} = e^{x \log a + x \log b} = e^{x \log a} e^{x \log b} = a^x b^x.$$

(b) The function $e^{x \log a}$ is differentiable by the chain rule with

$$(a^x)' = (e^{x \log a})' = e^{x \log a} \log a = a^x \log a.$$

The antiderivative of a^x is therefore $\int a^x \, dx = \frac{a^x}{\log a} + c$, where c is an arbitrary constant.

Challenge 16 By definition $x^a = e^{a \log x}$, hence by the chain rule the function $f : x \mapsto e^{a \log x}$ is differentiable. The derivative is $f'(x) = \frac{a}{x} e^{a \log x} = \frac{a x^a}{x} = a x^a x^{-1} = a x^{a-1}$. The antiderivative of x^a is therefore $\int x^a \, dx = \frac{x^{a+1}}{a+1} + c$, where c is an arbitrary constant. Notice that a cannot equal -1 because we cannot divide by zero.

Challenge 17 (a) By our sign convention for integrals, $\int_\alpha^1 \frac{1}{x}\,dx = -\int_1^\alpha \frac{1}{x}\,dx = -\log\alpha$.[1] Because the logarithm function is unbounded from above and below, as we drop $\alpha \to 0$, the value of $-\log\alpha$ continues to increase without an upper bound to constrain it. This shows us that $\int_0^1 \frac{1}{x}\,dx$ cannot exist.

(b) The function $f : x \mapsto 1/x^2$ is undefined for $x = 0$ and so it is *not* continuous. We cannot apply the fundamental theorem of calculus!

Challenge 18 (a) Yes to both.

(b) No to both.

(c) δ_1 and δ_3.

(d) Yes.

(e) Yes.

Challenge 19 (a) $(\forall\delta > 0)(\exists\epsilon > 0)(\forall x \in \mathbb{R})\left(0 < |x - t| < \delta \implies |f(x) - l| < \epsilon\right)$.

(b) $(\exists\epsilon > 0)(\forall\delta(\epsilon) > 0)(\exists x \in \mathbb{R})\left(0 < |x - t| < \delta(\epsilon) \implies |f(x) - l| \not< \epsilon\right)$.

(c) The negation of the proposed definition is

$$(\exists\epsilon > 0)(\forall\delta(\epsilon) > 0)(\exists x \in \mathbb{R})\left(|f(x) - l| < \epsilon \implies |x - t| \geq \delta(\epsilon)\right).$$

Take ϵ to be any positive number and take $x = \delta(\epsilon)$.

Challenge 20 (a) Immediate from L'Hospital's Rule (Theorem 28).

(b) $\left[\int_a^x C(x-a)^k\,dt\right]/(x-a)^k = C\int_a^x dt = C(x-a) \to 0$ as $|x - a| \to 0$.

(c) Replace $o(\alpha)$ with $\alpha \cdot o(1)$. The last equality holds because $\left(\alpha^2 \cdot o(1)\right)/\alpha = o(1)$ by the product rule.

Challenge 21 The one critical point of x^3 at the origin is neither a local maximum or minimum. The one critical point of x^4 at the origin is a local minimum.

Challenge 22 Answers given in the problem statement.

Challenge 23 Adapt the proofs for finite limits.

Challenge 24 (a) It suffices to take $\delta(\epsilon) := \epsilon^2$.

(b) Function f has a **limit** l **from below** at input t, if for each $\epsilon > 0$, there is a $\delta(\epsilon) > 0$ such that each $x \in (t - \delta(\epsilon), t)$ satisfies $f(x) \in \left(l - \epsilon, l + \epsilon\right)$.

(c) By assumption, for each $\epsilon > 0$, there is a corresponding $\delta(\epsilon) > 0$ value which applies to both limits from above and below.

(d) By assumption, there are $\delta_+(\epsilon) > 0$ and $\delta_-(\epsilon) > 0$ from the limit from above and the limit from below, respectively. Take δ to be their minimum.

Challenge 25 (a) See the proof of Rolle's Theorem (Theorem 38).

[1] Notice that since $\alpha < 1$, this is a positive number.

(b) For each $\epsilon > 0$, we know there are $\delta_f(\epsilon) > 0$ and $\delta_g(\epsilon) > 0$ such that each $x_{\neq t}$ with $|x - t| < \delta_f(\epsilon)$ satisfies $|f(x) - l| < \epsilon$ and such that each $x_{\neq t}$ with $|x - t| < \delta_g(\epsilon)$ satisfies $|g(x) - l| < \epsilon$. Take $\delta(\epsilon)$ to be the minimum of $\delta_f(\epsilon)$ and $\delta_g(\epsilon)$. Applying the triangle inequality to the hint gives $|h(x) - l| < |g(x) - l + l - f(x)| + \epsilon \leq 3\epsilon$.

Challenge 26 Given $\epsilon > 0$, continuity of f at $g(t)$ gives a $\delta_f(\epsilon) > 0$. Continuity of g at t in turn gives a $\delta_g\left(\delta_f(\epsilon)\right) > 0$.

Challenge 27 For each open interval $I_o := (p, q)$ containing $f(t)$, the corresponding ϵ is the smaller of $f(t) - p$, $q - f(t)$, from which we obtain an open interval I_i of length $2\delta(\epsilon)$ around t.

Challenge 28 For $t > 0$, if an open interval I_o contains 1 but not 0, the open interval $I_i = (0, +\infty)$ suffices. Otherwise, open interval I_o contains both 0 and 1, in which case the open interval $I_i = (-\infty, +\infty)$ works. The case for $t < 0$ is similar. On the other hand, the function g is not continuous at $t = 0$ because for the open interval $I_o = (-1, 1)$ each open interval I_i containing $t = 0$ will necessarily satisfy $f(I_i) = \{0, 1\}$.

Challenge 29 (a) Apply function g to both sides.

(b) $h(y) = h\left(f(x)\right) = x = g\left(f(x)\right) = g(y)$.

(c) If $g(p) > g(q)$ for $p < q$, since f is increasing $f \circ g(p) > f \circ g(q)$, contradicting $p < q$.

(d) If f defined on an interval I is strictly decreasing with inverse function g, then g is also strictly decreasing. To check, switch p and q in the answer of part (c).

Challenge 30 Replace the exponential function exp with the function g and the logarithm function log with the function f. For part (c), apply the results to function $-f$.

Challenge 31 Parts (a) through (d) are immediate from the Mean Value Theorem. For part (e), apply part (b) to the function $h = g - f$. For part (f), part (e) tells us that $\alpha(x) - \alpha(a) \leq f(x) - f(a)$ and $f(x) - f(a) \leq \beta(x) - \beta a$. The mean value inequality follows.

Challenge 32 (a) Take $i = 2$ and $k = 1$.

(b) $f'(x) = \frac{2}{x^3} e^{-1/x^2}$.

(c) $f''(x) = \left[-\frac{6}{x^4} + \frac{2}{x^3}\right] e^{-1/x^2}$.

(d) Straightforward application of the well-ordering principle.

(e) Routine algebra.

(f)

$$\lim_{x\to 0}\left|\frac{f^{(j-1)}(x) - \cancel{f^{(j-1)}(0)}}{x - 0}\right| = \lim_{x\to 0}\left|\frac{P_{j-1}e^{-1/x^2}}{x}\right| \leq \lim_{x\to 0}\left|\frac{P_{j-1}e^{-1/x}}{x}\right| \leq \lim_{y\to\infty}\left|\frac{yP_{j-1}}{e^y}\right| = 0.$$

Index