Calculus

By

Dayeol Choi

2024

https://dayeolchoi.com/calculus

Compiled on November 9, 2024

Contents

Welcome!

Calculus is situated right at the center of numerous exciting worlds, many undergoing intense investigation, and countless more still uncharted, waiting for an intrepid explorer to set foot on it. Pioneers who discovered the worlds currently known and the numerous adventurers that followed, began their remarkable journey into the unknown as newcomers to calculus. Indeed, although many tracts of calculus are still mysterious, the basic area every explorer is expected to be familiar with is incredibly well understood. Rough foot paths into deep jungles are now paved roads, rivers previously filled with piranhas are now bridged, and one of the thousands of well experienced tour guides will escort you gently from start to finish, judiciously avoiding the strenuous bits, hitting all the essential landmarks so that the busy traveller can move on to the next destination with far greater ease than once before.

These improvements have been wonderfully effective, and calculus is no longer the terrifying domain it once was. However, during all this progress, something had to make way—the *spirit* of calculus. The unprecedented and immeasurable progress in all fronts over the past three centuries were spearheaded by pioneers who embodied the spirit of calculus—taking the basic principles and instincts developed from calculus, and using them to chart out new worlds that everyone had overlooked, inviting other explorers to join them in their investigations into each new paradise.

It is in the spirit of calculus that we shall pursue it. We will uncover the underlying ideas, not as visitors on a tour, but as discoverers on an expedition, uncovering the mysteries of an unfamiliar land. We will venture into the core to see what lies underneath, into caverns only fully understood two centuries after the initial discovery of calculus. We will take an excursion into a region off limits for first timers, ascending up onto a mountaintop to take a glance at a magnificent view of the quantum world. We will see truly spectacular sights, beyond the wildest dreams of pioneers from not so long ago. Yet all of this will pale in comparison to the real prize of this expedition—you. Our mission is no less than to discover your inner discoverer and unleash it! Your adventure begins, now.

1

Arithmetic

Let's start at the very beginning. If you are very confident in your abilities in arithmetic and basic algebra, 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 $\times$

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

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

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

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),

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

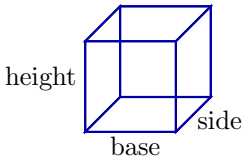


Figure 1.1: A cube with a base, side, and height.

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

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

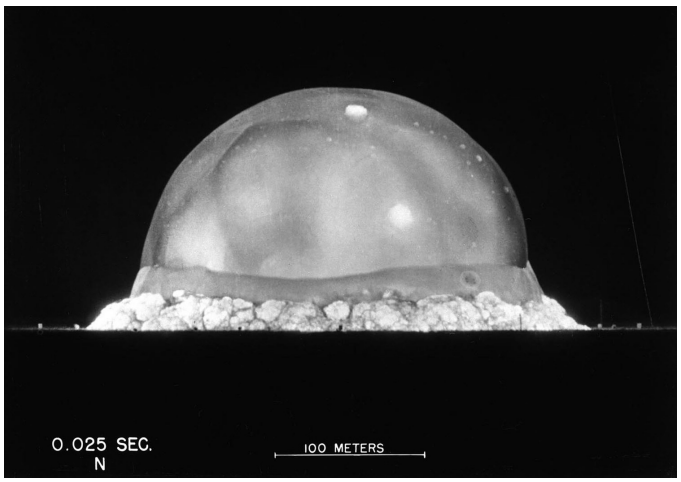
makes no sense. Similarly, we cannot add a **L**ength to **T**ime. Identify which of the following arithmetic expressions are valid:

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

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.

Hot and cold

Below is an image of the Trinity nuclear test, the first detonation of a nuclear weapon in history. The campaigns at Iwo Jima, Okinawa, and many others, were proving to be far too brutal and deadly to its participants (including civillians) during World War II. There was a need for a weapon so dangerous that the other side would have no choice but to surrender, thus ending the war. The atomic bombing of Hiroshima and Nagasaki killed over a hundred thousand people. The bombing also immediately led to the end of the Second World War; the surrender of Japan meant a full scale invasion of the island (Operation Downfall) was cancelled. The counterpart to Operation Downfall was Japan's Operation Ketsugō, which called for "the glorious death of one hundred million." Whether this was realistic or not, an invasion of the island to end the war would have killed several millions of people from both sides.



Trinity ushered in the atomic age and ended the Second World War. However, another one was to follow immediately: the First Cold War. A fear of an imminent end of humanity always loomed in the air, nuclear warfare could happen at any moment. However, many of us were born after the dissolution of the Soviet Union, and we have little idea how bad things were back then.

Nevertheless, you and I reap the benefits of those times. I am writing these words on a computer, first developed during World War II to facilitate calculations for ballistics and explosives. I use a program to compile the words I write into an image file that you can read; the first reprogrammable computers were developed to do numerical calculations for the hydrogen bomb. The internet originates from ARPANET, a project designed for sustained communications during a nuclear war.[2] Our phones (a miniaturized computer) have a long battery life because it uses lithium ion batteries, developed by the CIA during the heat of the Cold War for use in spy gadgets. Cavity magnetrons were top secret technology that had to be approved by Prime Minister Winston Churchill before being shared to the Americans during WWII. The technology completely shocked American scientists as it was a thousands times better than anything they had, and it was used to develop radars. If you have a microwave oven, a cavity magnetron is what vibrates the water molecules in your food to heat it. If I'm feeling hungry were to order food online at home, I know my order will arrive correctly because of GPS technology, developed to pinpoint ballistic missile submarines and mobile launch platforms. A GPS user finds their location by receiving geolocation information from satellites. Satellites were developed by the Soviet Union because they lacked airborne bombing capabilities and needed to develop intercontinental ballistic missiles (ICBM) that could reach continental US.[3]

We return to Trinity, the test that started the atomic age. 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.[4] 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 '$\times$' symbol.

variable	meaning	unit type	unit shorthand
R	Radius (of a Blast)	**L**ength	L
E	Energy (of a Bomb)	**M**ass $\times \frac{\mathbf{L}\text{ength}^2}{\mathbf{T}\text{ime}^2}$	$\mathrm{ML}^2/\mathrm{T}^2$
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

[2] The first email was sent over the ARPANET, for example.

[3] A satellite with a nuclear weapon payload, reduced angle of launch, and protective cap for reentry is an ICBM.

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

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

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

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.

$$E = \beta\frac{R^5\rho}{t^2} \tag{1.1}$$

Challenge 2

(a) Since we do not know what β is, let us assume $\beta = 1$ for now. Does equation 1.1 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 1.1 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 $\text{kg}\cdot\text{m}^2/\text{s}^2$.[5] Just eyeball the value for radius R, and use the fact that air density ρ is about 1.2kg/m^3.

(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.[6]

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.

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

[6] 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.

1.2 Exponentiation

The multiplication $13 \cdot 9$ can be done relatively easily in our heads if we remember that[7]

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

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

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

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.

- $\boxed{a^0 = 1}$ 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 $\mathbf{T}\text{ime}^{-2} = \frac{1}{\mathbf{T}\text{ime}^2}$,[8]
- $\boxed{a^b \cdot a^c = a^{b+c}}$ 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 $(\mathbf{L}\text{ength}^2)^4 = \mathbf{L}\text{ength}^8$.

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

[8] 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: $\frac{3}{1/8} = 3 \cdot 8 = 24$.

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

$$R = d \cdot E^a \cdot t^b \cdot \rho^c, \tag{1.2}$$

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 1.1 was a dimensionless constant.

For instance, notice that R is a length, so it has no units of time T.[9] However, on the right side of equation 1.2, 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 1.2 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

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

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

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

$$E = \beta \frac{R^5 \cdot \rho}{t^2}, \text{ where } \beta \text{ is a dimensionless constant.} \tag{1.3}$$

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.

Although our process for finding Equation 1.3 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.[10]

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**[11]. 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.

[10] 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).

[11] 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.

A search problem

If there are 5 pigeons in pigeonholes, but only 4 pigeonholes, then one of the pigeonholes must have at least 2 pigeons. More generally, if there are m pigeons, and n pigeonholes, with $m > n$, then at least one pigeonhole has 2 or more pigeons. This is called the **pigeonhole principle**.

Challenge 4 Here is a very simple algorithm a device could use for finding music titles after hearing the first few segments of some music. A music begins with a starting pitch; check if the next pitch is the same, higher, or lower. If lower, register the number 0 in a box; if the same, register the number 1; if higher, register the number 2. Check the next pitch, and place the evaluation of the pitch difference in a box to the right. Repeat for each subsequent pitch. Once we are done, we might be able to obtain a long sequence of boxes with numbers in them called a **signature** of a song. An example is:

1	0	2	1	2	2	2	0	$\cdots$.

Whenever someone needs to search for a song, all we need to do is to compare the signature of the song to an existing database of signatures. On a first examination, this algorithm seems to throw away far too much information about a song to work. For example, why are we not recording the first pitch? Why aren't we recording more fine grained information about each subsequent pitch? On a closer examination, this is more effective than it seems, and the simplicity affords advantages.

(a) Each song length varies widely, and it is costly to store too many boxes. If we allocated 0 boxes per song (each song has a signature of length 0), then how many unique signatures are possible? What if we allowed 1 box per song? Repeat for 2, 5, and 8 boxes.

(b) Our algorithm can be thought of placing each song signature (pigeon) into a pigeonhole. We want to ensure that we have enough pigeonholes to make it less likely that pigeons (song signatures) occupy the same hole. What is the minimum number of boxes we need to store the signature of one music, if we wish to distinguish between 100 million songs?

(c) We use a base 10 system, because most humans have 10 fingers. However, hours and minutes are divided into 60 segments (60 minutes = 1 hour, 60 seconds = 1 minute). Below is an illustration of the base systems:

$$127 = 1 \cdot 10^2 + 2 \cdot 10^1 + 7 \cdot 10^0 \text{ (base 10)}, \ 120 = 2 \cdot 60^1 + 0 \cdot 60^0 \text{ (base 60)}.$$

Each box is only allowed to store the numbers 0, 1, 2. Thus our boxes operate in base 3. What is the largest power of 3 whose multiple fits in 12? Express the number 12 in base 3.

(d) Suppose we were to classify pitches into 12 different categories. What is the fewest number of boxes required to record the 12 different pitches? We are assuming a box can only store a single natural number between 0 and 2, inclusive.

(e) The assumption from part (d) is still in place. Suppose we modified the algorithm to keep the starting pitch (by categorizing them into 1 out of 12, and storing the category number into boxes). Thereafter, everything is the same (store 0/1/2 in a box based on the difference in subsequent pitch). Under this scheme, how many boxes would you need to store the signature of one music, given we wish to distinguish between 100 million songs?

(f) In terms of number of boxes needed, is it better to keep the starting pitch or is it worse? How many more/fewer boxes in total would you need to store starting pitches of 100 million songs? Would your answer change if there were 200 million songs in total?

(g) For reasons other than storage space, why might it be better to discard the starting pitch?

We are now ready for calculus, which we begin in the next Chapter!

2

Differentiation

2.1 Arithmetic of Velocities

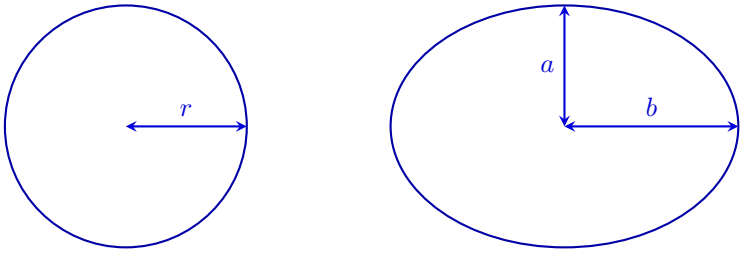


Figure 2.1: A circle of radius r, and an ellipse of height $2a$ and width $2b$.

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** 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).

[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.

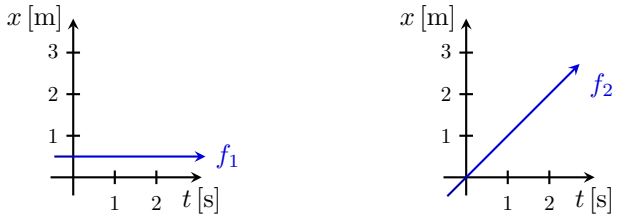


Figure 2.2: (Left) A position function f_1 of a stationary object at position 0.5 m. (Right) A position function f_2 of an object moving at a constant velocity of 1 m/s.

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

$$\boxed{\textbf{Constant Rule: } f' : t \mapsto 0.}$$[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.

$$\boxed{\textbf{Sum Rule: } (f+g)' = f' + g'.}$$

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 $(f-g)' = (f+h)' = f' + h' = f' + (-g') = f' - g'$.

$$\boxed{\textbf{Subtraction Rule: } (f-g)' = f' - g'.}$$

[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 $f' = 0$.

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 $\diamondsuit$ and outputs quantities of dimension $\heartsuit$. Then the rate of change (the derivative) of the function as we vary inputs (of unit $\diamondsuit$) will have the dimension $\heartsuit/\diamondsuit$. 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:

$$c_1(f^2)' + c_2(g^2)', \quad c_3 ff' + c_4 gg', \quad c_5 f'g + c_6 fg'.$$

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:

$$\boxed{\textbf{Product Rule: } (fg)' = f'g + fg'.}$$

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

$$(x^k)' = kx^{k-1}. \tag{2.1}$$

[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 2.1 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 2.1 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 2.1 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

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

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 5 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 $\sum_{\diamondsuit=p}^{q} h(\diamondsuit)$ means $h(p)+h(p+1)+\cdots+h(q)$. In particular, $\sum_{\diamondsuit=p}^{p} h(\diamondsuit) := h(p)$. Using this notation, a polynomial of degree m on the variable $\square$ may be written compactly as $\sum_{\diamondsuit=0}^{m} c_{\diamondsuit}\square^{\diamondsuit}$, or equivalently as $\sum_{\diamondsuit=0}^{m} c_{m-\diamondsuit}\square^{m-\diamondsuit}$. 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 $(t^i)' = it^{i-1}$, and so

$$f' \left(\sum_{i=0}^{m} c_i t^i\right)' = \sum_{i=0}^{m} (c_i i) t^{i-1}. \tag{2.2}$$

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 2.2 is saying.

Challenge 6

(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) Verify that $(1 + 2 + 3 + \cdots + n)^2 = 1^3 + 2^3 + \cdots n^3$ holds when $n = 1$ and $n = 2$.
(d) Use the well-ordering principle to show that the equation $(1+2+3+\cdots+n)^2 = 1^3+2^3+\cdots n^3$ holds for each natural number n. [*Hint:* use the two identities from part (b).]

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

$$\underbrace{f(t+\alpha)}_{\text{future pos.}} = \underbrace{f(t)}_{\text{current pos.}} + \underbrace{v \cdot \alpha + X}_{\text{projected travel dist.}} \tag{2.3}$$

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 2.3 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.

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

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

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 2.3 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

$$f(t+\alpha) = f(t) + f'(t)\alpha + |\alpha| o_\alpha(1). \tag{2.4}$$

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 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

[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.

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

switches the sign of negative inputs. Because the value $|a|$ changes only when a is negative, writing the definition of the derivative as Equation 2.4 reminds us that α can be negative.

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

$$f(t+\alpha) = f(t) + f'(t)\alpha + |\alpha| o(1). \tag{2.5}$$

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).

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 $2 \cdot o(1) = o(1)$.

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

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:

$$f(t + \alpha)g(t + \alpha) = f(t)g(t) + \big[f'(t)g(t) + f(t)g'(t)\big]\,\alpha + f'(t)g'(t)\alpha^2 + |\alpha|Ao(1) \tag{2.6}$$

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 2.6 is thus

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

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

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

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 *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:

$$f(t + a\epsilon) = f(t) + af'(t)\epsilon. \tag{2.7}$$

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

[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.

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 2.7 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 2.7, 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 2.7 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 $\epsilon^2 = 0$.[18] Using dual numbers, the derivative of a function is defined by Equation 2.7 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.

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

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

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

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

[18] For now we will have to take the existence of the object ϵ on faith.

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 $f\big(g(t) + ag'(t)\epsilon\big)$. 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, $f\big(\bar{t} + \bar{a}\epsilon\big) = f\big(\bar{t}\big) + \bar{a}f'\big(\bar{t}\big)\epsilon = f\big(g(t)\big) + ag'(t)f'\big(g(t)\big)\epsilon$. We reorganize what we have found in the following line.

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

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. Since we do not yet have the means to understand exactly what such an object is, we will stick to our previous definition of the derivative using $o(1)$ for the rest of the book.

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 $|a + b| \leq |a| + |b|$ 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 $a + b \geq 0$ 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 result is clear. 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 7

(a) By exhausting the cases (as in the proof of the triangle inequality), show that $|ab| = |a||b|$.
(b) If b is nonzero, show that $|a/b| = |a|/|b|$.
(c) 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 ♣ that satisfies the equation below?

$$(f \circ g)(t + \alpha) = (f \circ g)(t) + \clubsuit \cdot \alpha + |\alpha| o_\alpha(1) \tag{2.8}$$

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

$$g(t + \alpha) = g(t) + g'(t)\alpha + |\alpha| o_\alpha(1), \tag{2.9}$$
$$f(s + \beta) = f(s) + f'(s)\beta + |\beta| o_\beta(1) \tag{2.10}$$

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 Equations 2.9 and 2.10.

By Equation 2.9, $\beta := g(t + \alpha) - g(t) = g'(t)\alpha + |\alpha| o_\alpha(1)$, 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 2.8, 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 $|a + b| \le |a| + |b|$ and homogeneity $|ca| = |c||a|$, we have

$$|\beta| = |g'(t)\alpha + |\alpha| o_\alpha(1)| \le |g'(t)||\alpha| + |\alpha||o_\alpha(1)| = |g'(t)||\alpha| + |\alpha| o_\alpha(1).$$

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

$$\frac{|\beta| o_\beta(1)}{|\alpha|} \le \Big(|g'(t)||\alpha| + |\alpha| o_\alpha(1)\Big) \frac{o_\beta(1)}{|\alpha|} = o_\beta(1) + o_\alpha(1) o_\beta(1) = o_\alpha(1) + o_\alpha(1) o_\alpha(1) = o_\alpha(1).$$

Recall that if $|f| \le o(1)$, then $f = o(1)$. Since $|\beta o_\beta(1)/\alpha| \le o_\alpha(1)$, we have $|\beta| o_\beta(1)/|\alpha| = o_\alpha(1)$, as desired. Equation 2.8 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 $[f \circ (g \circ h)](x) = f((g \circ h)(x)) = f(g(h(x)))$. But this is the same as $[(f \circ g) \circ h](x) = (f \circ g)(h(x)) = f(g(h(x)))$.

Therefore, to take the derivative of a composition of three functions $f \circ g \circ h$, we may use the chain rule to get $([f \circ g] \circ h)' = ([f \circ g]' \circ h) \cdot h'$ or equivalently $(f \circ [g \circ h])' = (f' \circ [g \circ h]) \cdot [g \circ h]'$. 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

No one is stopping us from taking 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 8 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, $\binom{n}{k} = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \times \cdots \times 1}$. 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) Suppose f is a polynomial of degree n. Use the well-ordering principle to show the following.[19]

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

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

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

(b) Let $f : x \mapsto (x+b)^n$ so that $f(0) = b^n$. Use the well-ordering principle to show that

$$f^{(k)}(0) = k!\binom{n}{k}b^{n-k}.$$

[*Hint:* as with all applications of the well-ordering principle, start by finding $f'(0)$ and $f''(0)$.]

(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

Now that we have discussed some examples of derivatives, let us examine some nonexamples.

Nonexample 1: the absolute value function

Consider the absolute value function $f : x \mapsto |x|$ shown in Figure 2.3.

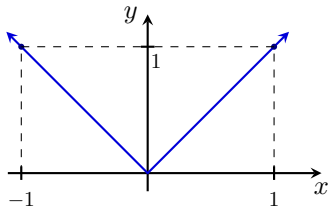


Figure 2.3: A graph of the absolute value function.

Consider 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 2.4).

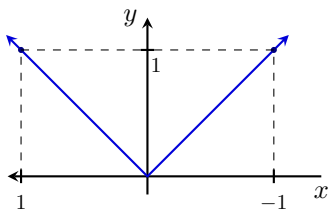


Figure 2.4: The same figure as before, but with the x-axis swapped. Everything is the same, except the labels of the inputs.

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 9 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 on $(-\infty, 0)$, the derivative of relu is zero, while for $x \in (0, +\infty)$, $\text{relu}'(x) = 1$. 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 2.5? 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.

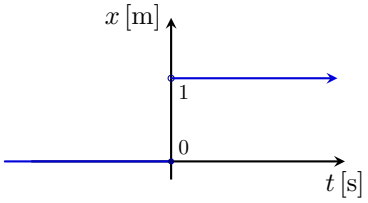


Figure 2.5: Graph of a function defined by $t \mapsto 1$ for $x > 0$ and $t \mapsto 0$ for $t \leq 0$.

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 2.5 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 10 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 $t = 1$.

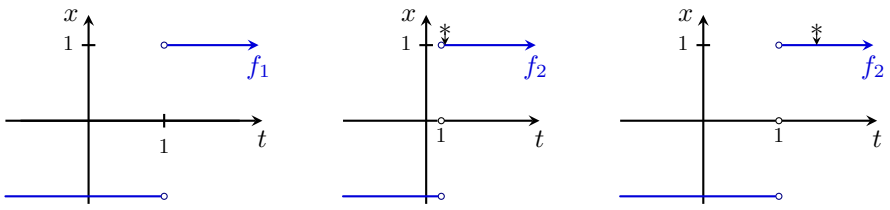


Figure 2.6: A step function f_1 and a step function f_2 defined on a defective time axis.

Let us examine the step function once more. The leftmost graph in Figure 2.6 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 2.6. 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 2.6. 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.[20] Now there is no problem in seeing that function f_2 is also continuous at point $*$.

Well, 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

[20] 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,[21] 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. 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 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.

[21] 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 $*$.

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 11 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).

Intervals

We will use the notion of an *interval* as we discuss displacement from motion starting at one time and ending at another. Here is some handy notation. For real numbers a and b with $a < b$:

(a) the symbol (a, b) denotes the set of real numbers x such that $a < x < b$,

[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.

(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 $a < x \leq b$.

The first is called an **open interval** while the second is called a **closed interval**. The final two are called **half open intervals**.

Sometimes, the symbols ∞ and $-\infty$ are used in a similar context. For a 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 $a \leq x$.

The first and second are considered open intervals while the third and fourth are considered half open intervals. Each of the eight symbols 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 one final type of interval $(-\infty, \infty)$ and it is another way to denote the set of real numbers $\mathbb{R}$. This is also considered an open interval.

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 a car uses to determine 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]

[2]This is only an analogy, and if we stretch it a little bit it breaks down. We are studying *displacements* of objects that motion back and forth along a line. When calculating displacement, we can cancel out our accumulated displacement by having negative velocity, in other words, reversing our car. Unfortunately we cannot reduce an odometer reading by driving our car in reverse.

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 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 $v(t) \le w(t)$ for each $t \in (t_i, t_f)$, then

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

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:

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

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

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

An equivalent way to put this is:

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

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

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

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

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

[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.

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

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

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.

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

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

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

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 $F'(t) = f(t)$.

[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 $F(t+\epsilon) - F(t) - f(t)\epsilon = |\epsilon| o(1)$

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 $I : x \mapsto \int_a^x f(t)\,dt$.

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

[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.

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).

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

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

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

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

3.3 Area Under a Curve

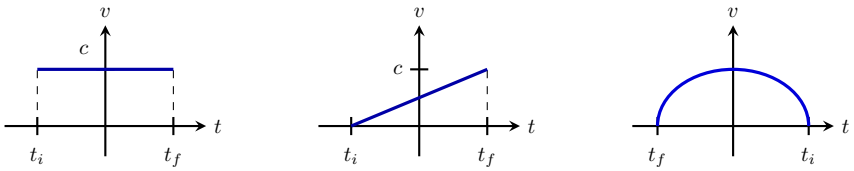


Figure 3.1: Velocities of objects from time t_i to t_f. The displacement is the area under the curve.

Consider the three diagrams in Figure 3.1. 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 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 3.1. 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

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

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!

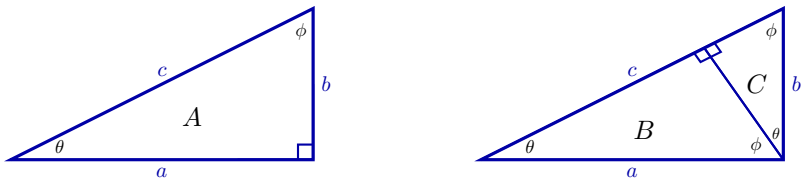


Figure 3.2: A single line is sufficient to prove the Pythagorean theorem.

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 3.2. 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 3.2). 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: the angle Θ 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 (see the diagram on the right in Figure 3.2). 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 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 3.1? 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).

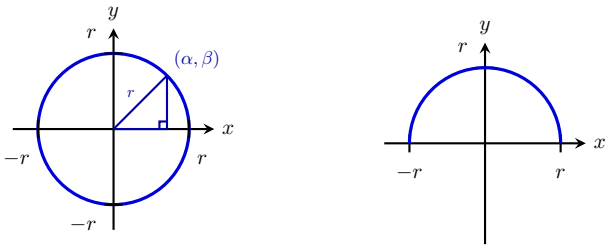
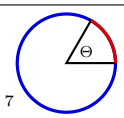


Figure 3.3: A circle of radius r at the origin is given by the equation $x^2 + y^2 = r^2$ (left) and the top semicircle is given by the equation $y = \sqrt{r^2 - x^2}$ (right).

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 3.3). 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



7

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 3.4, 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 3.4 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).

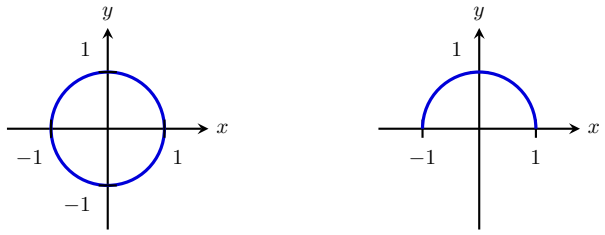


Figure 3.4: The graphs of equation $x^2 + y^2 = 1$ (left) and equation $y = \sqrt{1 - x^2}$ (right).

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 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

$$A(r) = 2\int_{-r}^{r} \sqrt{r^2 - x^2}\, dx = 2r^2 \int_{-r}^{r} \sqrt{1-(x/r)^2} \cdot \frac{1}{r}\, dx \tag{3.5}$$

$$= \left(2\int_{g(-r)}^{g(r)} \sqrt{1-g(x)^2} \cdot g'(x)\, dx\right) r^2 = \left(2\int_{-1}^{1} \sqrt{1-u^2}\, du\right) r^2 = \pi r^2. \tag{3.6}$$

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.

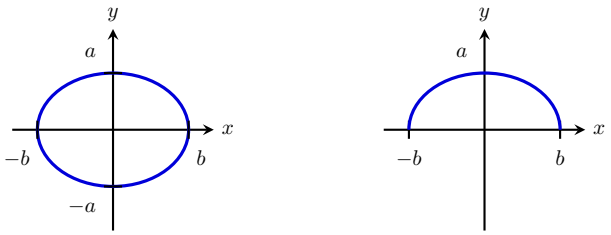


Figure 3.5: The graphs of equation $\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1$ (left) and equation $y = \sqrt{a^2 - (ax/b)^2}$ (right).

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 3.5). 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?

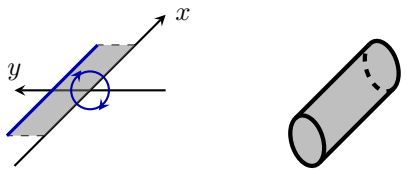


Figure 3.6: Rotating the area underneath the constant function defined on a finite interval sweeps out a cylinder.

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. In Figure 3.6 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

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

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]

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 12

(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 $\frac{4}{3}\pi r^3$.

(c) An **ellipsoid** is a stretched out sphere. Consider an ellipsoid with depth $2a$, width $2b$, and height $2c$. Use the fact that a sphere of radius r has volume $\frac{4}{3}\pi r^3$ to guess the volume of an ellipsoid. Use the solid revolution on the top half of an ellipse and a change of units (or the substitution rule) to verify your guess. The change of units is necessary because applying the solid of revolution on an ellipse will calculate the volume of an ellipsoid with depth $2a$, width $2b$, and height of either $2a$ or $2b$, but not $2c$.

[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.

3.4 Exponentiation Revisited

The logarithm function

Question: what is an antiderivative of the function $1/x$? 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 9. For each $x \in (0, \infty)$, the (**natural**) **logarithm** function is defined as

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

The Fundamental Theorem of Calculus gives $\log'(x) = 1/x$ for each $x \in (0, \infty)$.[10]

By construction, the logarithm function is **monotone 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 10. 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.[11]

[10] As an easy exercise, you should verify that $x \mapsto 1/x$ is continuous on the interval $(0, \infty)$ and bounded in the interval from 1 to x. There are three possibilities for the latter: $(1, 1)$ if $x = 1$, $(x, 1)$ if $0 < x < 1$, and $(1, x)$ if $x > 1$.

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

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 11 (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 $nx > y$.

Proof. In order to derive 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 (it contains 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 has to be a natural number m with $mx > u - x$. 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 the set S. □

The exponential function

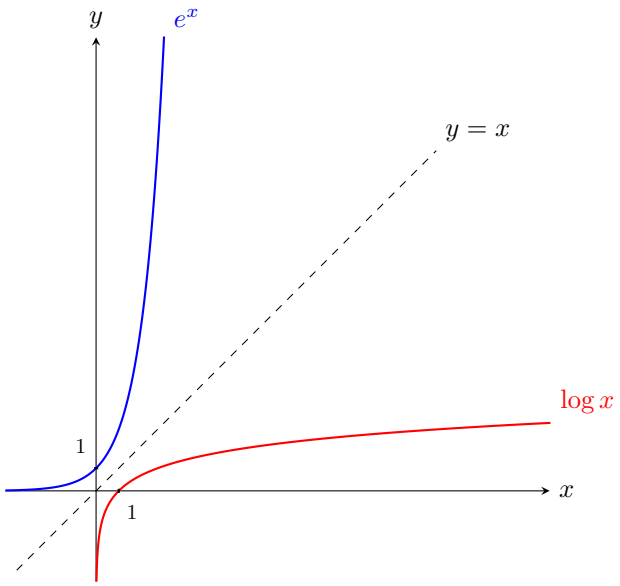
Because the logarithm function is monotone increasing with $\log 1 = 0$, for any $\alpha > 0$, we have $\log \alpha > 0$. By Proposition 10 part (a), $\log(\alpha^n) = n \log \alpha$ for each natural number n. By the Archimedean property, the logarithm function is unbounded, and so each positive real number x 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 12. The **exponential** function exp is defined on $\mathbb{R}$ such that $\exp \circ \log$ and $\log \circ \exp$ are the identity maps $x \mapsto x$.[12] 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, 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.

[12] 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 13. 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 10 part (a).

Theorem 14. If x and y are real numbers, then $e^x \cdot e^y = e^{x+y}$.[13]

Challenge 13 Prove Theorem 14 and deduce that $e^0 = 1$.

Hyperbolic functions

Definition 15. 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 14

(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. We know that such a *decomposition* is always possible because of part (a). We now 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. Conclude that $f_1 = f_e$ and $f_2 = f_o$ as defined in part (a).

[13] *Hint:* start with $\log(e^x e^y)$.

Definition 16 (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.[14]

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

Challenge 15 Use the definitions to show that

(a) $\cosh^2 x - \sinh^2 x = 1$,
(b) $\sinh(x + y) = \sinh x \cosh y + \cosh x \sinh y$,
(c) $\sinh x = \frac{\tanh x}{\sqrt{1-\tanh^2 x}}$,[15]
(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 17. 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 18. 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

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

□

We now obtain Proposition 10 part (c): for real p, the equality $\log x^p = p \log x$ holds.

Proof. We use the exponentiation rule $(a^\alpha)^\beta = a^{\alpha\beta}$.[16] 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) $\log(x^p) = yp$. But (i) = (ii), and we are done. □

[14] Notice that $\cosh x$ is the even function of $\exp(x)$, while $\sinh x$ is the odd function of $\exp(x)$.
[15] The expression $\tanh^2 x$ means $(\tanh x)^2$.
[16] This is Proposition 18 part (e).

Challenge 16

(a) Prove Proposition 18.

(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 17 (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 real $a \neq -1$, $\int x^a \, dx = \frac{x^{a+1}}{a+1} + c$. Notice that $\int x^{-1} \, dx = \log|x| + c$, for if $x < 0$, then using the chain rule, we have $(\log|x|)' = (\log(-x))' = \frac{1}{-x} \cdot (-1) = x^{-1}$. The case of $x = 0$ is undefined because $1/0$ is undefined, while the positive case follows from the definition of the logarithm function.[17]

Challenge 18

(a) Show that the function $f : x \mapsto 1/x$ defined on the nonzero reals is continuous by showing that f is continuous at each nonzero a (this justifies our earlier use of the Fundamental Theorem of Calculus to find $\log'$).

(b) Differentiable functions are an especially nice class of continuous functions. This does *not* mean differentiable functions can necessarily be integrated! Let $f : x \mapsto 1/x$ be defined on the interval $(0, 1)$. Show that f is differentiable and that f can be integrated on the interval $(\alpha, 1)$, for each α such that $0 < \alpha < 1$. Argue that f cannot be integrated on the interval $(0, 1)$.

(c) 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.

[17] 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.

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 are also the result of some limiting process, but a first course in calculus is not the place for that.

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 19. 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 $|f(x)-l| < \epsilon$.

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 $|f(x)-l| < \epsilon$.[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 20. 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. But first, we need to check that a limit of a function at a point is unique, otherwise we will be in trouble!

Proposition 21 (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,

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

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

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

[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, $\min(-10, 2) = -10$.

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 22 (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 23 (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.[7] By applying the triangle inequality and homogeneity on a sneaky addition and subtraction of the term $f(x)l_2$, the following holds.

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

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

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

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

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: $|a| - |b| \leq |a - b|$.[8] 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.

Proof. Suppose $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} g(x) = l_2$. 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 $|(fg)(x) - l_1 \cdot l_2| < c{\cdot}\epsilon$ 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 $|f(x) - l_1| < \epsilon$. 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 $|g(x) - l_2| < \epsilon$. The triangle inequality and homogeneity gives the following inequality, as we discussed before.

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

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

```
def δ(ε):
    if ε ≤ 1:
        d₁ ← δ₁(ε)
    else:
        d₁ ← δ₁(1)
    return min(d₁, δ₂(ε))
```

[8] This was left for you in Challenge 7, but I will prove it again. 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!

[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.

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

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

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

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

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. The first was the sneaky manipulation

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

The second is the realization that defining $\delta : \epsilon \mapsto \min\big(\delta_1(\epsilon), \delta_2(\epsilon)\big)$ 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 19 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 $\delta' : \epsilon \to \delta(\epsilon)/2$. For each $x_{\neq t}$ within $\delta'(\epsilon)$ of t, can we guarantee that $|f(x) - l| < \epsilon$ holds? Repeat for $\delta'' : \epsilon \to \delta(\epsilon)/c$, where $c > 1$ is a constant.

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

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

(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

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

Challenge 20

(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 $|f(x) - l| < \epsilon$. 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 $\lim_{x\to 1} f(x) \neq c$.[12]

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

Quotient rule

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

Proposition 24 (Quotient Rule)**.** If $\lim_{x\to t} f(x) = l_1$ and g is 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 one 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 $|(1/g)(x) - 1/l_2| < c \cdot \epsilon$ 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 $\epsilon = X$.

[10] Here is one possible answer. Yours will be better. 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 $1 \cdot 10^{-34}$. 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: $(\exists \epsilon > 0)(\forall \delta(\epsilon) > 0)(\exists x \in \mathbb{R})\,(0 < |x - t| < \delta(\epsilon) \Rightarrow |f(x) - l| \geq \epsilon)$.

[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.

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 $\lim_{x\to t}(1/g)(x) = 1/l_2$. 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

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

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 $f \cdot (1/g)$. □

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. First we will prove the reciprocal rule $(1/f)' = -f'/f^2$.

Proposition 25 (Reciprocal Rule)**.** Suppose f is differentiable at t and $f(t)$ is nonzero. Then

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

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

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

□

We once again prove 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 26 (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.

Challenge 21 (Power rule for rational numbers) We check that the power rule applies to rational powers. All functions discussed below are not defined at zero.[14]

(a) Use the reciprocal rule to show that x^{-1} is differentiable and that $(x^{-1})' = -x^{-2}$.

(b) (Power rule for integers) If m is a negative integer, show that $(x^m)' = mx^{m-1}$.

(c) Show that $1/\sqrt{x}$, defined on $(0, \infty)$ is differentiable and find the derivative. [*Hint:* Use the reciprocal rule and the chain rule.]

(d) Let $f : x \mapsto x^{1/n}$, where n is a positive integer. By definition, $f(x)^n = x$. Differentiate both sides (one side needs the chain rule) and solve for $f'(x)$ to show that $f'(x) = \frac{1}{n}x^{1/n-1}$.

(e) Let $f : x \mapsto x^{m/n}$, where m is an integer and n is a positive integer. Apply the chain rule to show that $f'(x) = \frac{m}{n}x^{m/n-1}$. Conclude that if p is rational, then for the function $f : x \mapsto x^p$ defined on $(0, \infty)$, we have $f'(x) = px^{p-1}$.

The following is yet another application of the quotient rule for limits.

Theorem 27 (Bernoulli's rule)**.** Suppose f and g are differentiable at t with $g'(t) \neq 0$. Furthermore, assume that $f(t) = g(t) = 0$. Then

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

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

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

□

4.3 Further Notions

Little oh

For a function f with the property $\lim_{x\to 0} f(x) = 0$, we have used the shorthand $f = o(1)$. The object $o(1)$ is the collection of functions g with the property $\lim_{x\to 0} g(x) = 0$. 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

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = 0.$$

[14] We cannot divide by 0, and so $1/(0^p)$ for rational p makes no sense.

If $g : x \mapsto 1$, then we recover $f = o(1)$. Actually, from the definition, for each positive $c > 0$, $o(c) = o(1)$. Once again, the object $o(g)$ is a 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 22

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

(b) 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 $o(\alpha)o(\alpha) = o(\alpha)$.

I should note that if you see $f = o(g)$ in the wild, it will mean the following limit is satisfied.

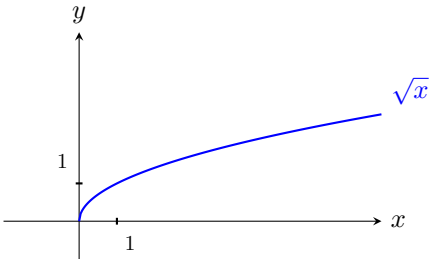
$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$$

The meaning remains 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 $|f(x) - l| < \epsilon$.

One sided limits

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.

A 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 $x \in (t, t + \delta(\epsilon))$ is guaranteed to satisfy $f(x) \in \big(l - \epsilon, l + \epsilon\big)$. We write this using the notation $\lim_{x\to t+} f(x) = l$.

Notice that the statement $f(x) \in \big(l-\epsilon, l+\epsilon\big)$ is equivalent to the statement $|f(x)-l| < \epsilon$. Both statements mean the same thing: $f(x)$ is within ϵ of l.

Challenge 23

(a) Verify that $\lim_{x\to 0+} \sqrt{x} = 0$.[15]

(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) Conversely, 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.[16] 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.

Limits and inequalities

Whenever we use properties like if $|f| \le o(1)$, then $f = o(1)$, we are using limits with inequalities. We now check that limits work as expected with inequalities.

Proposition 28. If $f \le g$ with $\lim_{x\to t} f(x) = l_1$ and $\lim_{x\to t} g(x) = l_2$, then $l_1 \le l_2$.

Proof. We wish to show that $l := l_2 - l_1 \ge 0$. In order to derive a contradiction, suppose $l < 0$.

Let $h : x \mapsto g(x) - f(x)$ and observe that h 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 29 (Squeeze Theorem)**.** Suppose we have functions f, g and h with $f(x) \le h(x) \le g(x)$. If $\lim_{x\to t} f(x) = \lim_{x\to t} g(x) = l$, then $\lim_{x\to t} h(x) = l$.

Challenge 24 Prove the squeeze theorem. [*Hint:* By the triangle inequality, $|h(x) - l| \le \underbrace{|h(x) - f(x)|}_{\le g(x)-f(x)} + \underbrace{|f(x) - l|}_{<\epsilon}$. Apply the triangle inequality to $|g(x) - f(x)| = |g(x) - l + l - f(x)|$.]

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.

What closeness?

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

[15] *Hint:* we want $\delta(\epsilon)$ such that if $0 < |x - 0| < \delta(\epsilon)$, then $|\sqrt{x} - 0| < \epsilon$. Absolute values can be discarded. Why?

[16] *Hint:* there is no need to get particularly creative with the strictly positive function δ.

satisfied as long as the outputs are close by $c \cdot \epsilon$ for some positive constant c. As an example, in Challenge 24, 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 25 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 23 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 $\delta_2(\epsilon_2) > 0$ 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 $x \in I_i$. We will also write $f(I_i) \subset I_o$ to mean that the set $f(I_i)$ is **contained** within I_o. We also say that $f(I_i) \subset I_o$ means that the set $f(I_i)$ is a **subset** of I_o.

Theorem 30. 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. □

This disposal of "closeness" and "distances" in favor of working with open sets leads to the development of the field of *topology*. Quite a few difficult theorems in the rigorous study of calculus (introductory analysis) are easy corollaries of some of the elementary but nevertheless abstract results from topology. You are ready for a dive into either subject.

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. Now that we have done some calculus, let us switch gears. Instead of taking the motion of objects, their velocities and displacements as a given, let us 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**,[1] 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 $\text{Mass} \times \text{Length} \times \text{Time}^{-2}$.

Because acceleration is the second derivative of position, Newton's second law is an example of a **differential equation**, which is 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

After studying velocities (derivatives) and displacements (integrals), two natural questions arise. (1) How much total effort must we exert in order to get an object to attain a certain velocity? (2) How much total effort must we exert in order to displace an object by a certain distance? For

[1] We have stated the law for objects constrained to motion along a line, as this is sufficient for our use.

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 we exerted, we simply 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 $\times$ Length, that is, Mass $\times$ Length^2 $\times$ 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, $\frac{\mathrm{d}V}{\mathrm{d}x} = -F(x)+F(o)$. 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

$$F(x) = -\frac{\mathrm{d}V}{\mathrm{d}x}.$$

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 31 (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 $\frac{\mathrm{d}}{\mathrm{d}t}\square$. 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, $F(x) = -\frac{\mathrm{d}V}{\mathrm{d}x}$ and so by the chain rule,

$$0 = m\ddot{x}\dot{x} + \frac{\mathrm{d}V(x)}{\mathrm{d}x}\dot{x} = \frac{\mathrm{d}}{\mathrm{d}t}\left(m\frac{\dot{x}^2}{2} + V(x)\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(K + V\right).$$

Notice that $\frac{\mathrm{d}}{\mathrm{d}t}V(x) = \frac{\mathrm{d}V}{\mathrm{d}x}\dot{x}$ and *not* $\frac{\mathrm{d}V}{\mathrm{d}t}$, 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

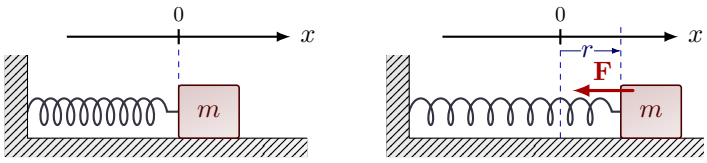


Figure 5.1: A mass and spring system at rest and displaced by r. (By Izaak Neutelings at `tikz.net`)

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. Then at a displacement of r, the potential energy V of our system is

$$V = \int_o^r -F(x)\,dx = \int_0^r kx\,dx = \frac{1}{2}kr^2.$$

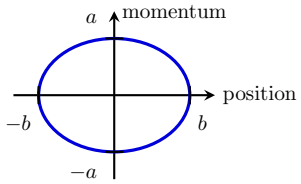
Therefore, the total energy E of our mass and spring system is given by

$$E := K + V = \frac{p^2}{2m} + \frac{1}{2}kx^2.$$

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 $\frac{x^2}{2E/k} + \frac{p^2}{2mE} = 1$.[2] Setting $a^2 := 2E/k$ and $b^2 := 2mE$ gives us an equation of an ellipse!

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



The figure above illustrates 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**.

Phase space

Let us try to get some intuition about phase space. Consider Figure 5.2, where an object has been displaced from position x_i to x_f. We say that our object has undergone a **(spatial) translation**.

[2] We have used the fact that $\frac{b}{1/a} = ab$.

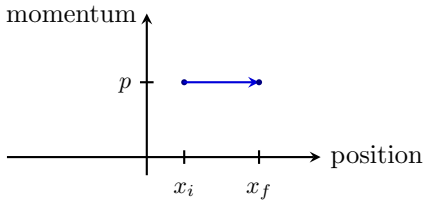


Figure 5.2: Momentum p is applied to an object, displacing it from x_i to x_f.

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 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 5.3. What is going on there?

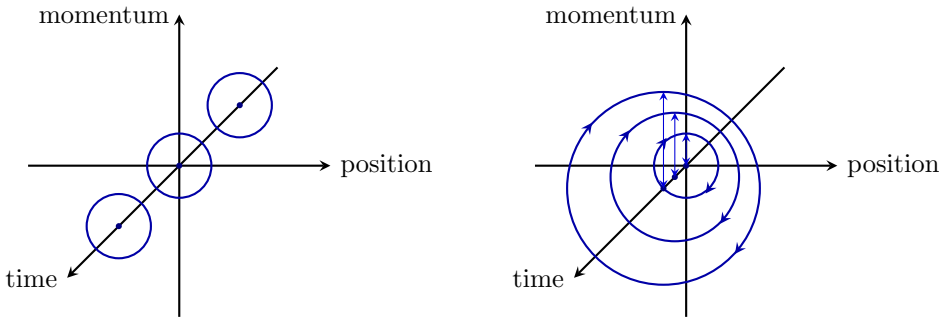


Figure 5.3: The phase space of an oscillator and another with energy being added.

Let us review the simpler case of momentum being added to the object (Figure 5.2). 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 5.3). 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 5.3). 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.[3] 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:

$$(A + B)^o = A^o + B^o \text{ and } (cA)^o = cA^o. \tag{5.1}$$

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

$$(f + g)' = f' + g', \text{ and } (cf)' = cf'.$$

[3] Nevertheless, we will revisit this matter in a restricted setting in Section 5.3.

Ditto for integration of bounded continuous functions f and g defined on an interval $[t_i, t_f]$:

$$\int_{t_i}^{t_f} [f(x) + g(x)]\, dx = \int_{t_i}^{t_f} f(x)\, dx + \int_{t_i}^{t_f} g(x)\, dx, \text{ and } \int_{t_i}^{t_f} cf(x)\, dx = c\int_{t_i}^{t_f} f(x)\, dx.$$

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 5.1 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,

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix}' = \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix}' + \begin{pmatrix} 0 \\ b \\ 0 \end{pmatrix}' + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}' = a\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}' + b\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' + c\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' \tag{5.2}$$

which has expressed our need to use the sum rule $(f+g)' = f' + g'$ and the product rule $(cf) = cf'$. By the power rule,

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}' = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}' = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.3}$$

Plugging these values back into Equation 5.2, 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 5.3, 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 **Euclidean 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 Euclidean space are the n vectors $e_1, e_2, \ldots, e_n$ defined by the following, and listed in that order.

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \quad \cdots, \quad e_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

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,

$$e_1' = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix}, e_2' = \begin{pmatrix} 0 \\ 0 \\ n-2 \\ \vdots \end{pmatrix}, \cdots, e_n' = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

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.

$$D := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ n-1 & 0 & \cdots & 0 \\ 0 & n-2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \tag{5.4}$$

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

$$v = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = a_1 e_1 + a_2 e_2 + \cdots a_n e_n.$$

Then the derivative can be taken by looking at each column of the matrix D in 5.4:

$$v' = a_1 e_1' + a_2 e_2' + \cdots a_n e_n' = a_1 \text{ column 1 of } D + a_2 \text{ column 2 of } D + \cdots + a_n \text{ column } n \text{ of } D. \tag{5.5}$$

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

$$e_1^o = \begin{pmatrix} 1/1000 \\ 0 \\ 0 \end{pmatrix}, e_2^o = \begin{pmatrix} 0 \\ 1/1000 \\ 0 \end{pmatrix}, e_3^o = \begin{pmatrix} 0 \\ 0 \\ 1/1000 \end{pmatrix}.$$

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

$$r^o = r_1 \text{ column 1 of } C + r_2 \text{ column 2 of } C + r_3 \text{ column 3 of } C. \tag{5.6}$$

Since it gets rather tedious to right out the expressions in Equations 5.5 and Equations 5.6, we will use the shorthand Dv and Cr, respectively.

Thus if we have a matrix A and a vector v defined by

$$A := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \text{ and } v := \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

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

$$Av := \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1n}v_n \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2n}v_n \\ \vdots \\ a_{m1}v_1 + a_{m2}v_2 + \cdots a_{mn}v_n \end{pmatrix}. \tag{5.7}$$

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:[4]

$$\partial_x f(x, y) = \frac{1}{xy}\partial_x(xy) - \frac{1}{x} = \frac{y}{xy} - \frac{1}{x} = 0.$$

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.[5]

A popular notation that we will use is that if we have a function f which has inputs denoted by the variables $\clubsuit, \spadesuit$, then we will write $\frac{\partial f}{\partial \clubsuit}$ and $\frac{\partial f}{\partial \spadesuit}$ to denote the partial derivatives with respect to $\clubsuit$ and $\spadesuit$, respectively. If the function f is twice partial differentiable with respect to the $\clubsuit$ 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 $\boldsymbol{\nabla} f(t)$, is defined to be

$$\boldsymbol{\nabla} f(t) := \begin{pmatrix} \partial_1 f(t) \\ \vdots \\ \partial_n f(t) \end{pmatrix}.$$

There is also a **Laplacian** operator, denoted by the symbol ∇^2 or Δ, defined by

$$\nabla^2 f := \sum_{i=1}^{n} \partial_i f.$$

[4] The notation ∂_x is used in place of ∂_1 because we know that x is the first variable.

[5] 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

Recall that units do not support chaining, but 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 5.7, 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 5.4 by concatenating (squashing) vectors together.

$$e_1' = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix}, e_2' = \begin{pmatrix} 0 \\ 0 \\ n-2 \\ \vdots \end{pmatrix}, \cdots, e_n' = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \Longrightarrow D := \begin{pmatrix} 0 & 0 & \cdots & 0 \\ n-1 & 0 & \cdots & 0 \\ 0 & n-2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

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! Taking our derivative matrix D as an example,

$$D_1 = \begin{pmatrix} 0 \\ n-1 \\ 0 \\ \vdots \end{pmatrix} := e_1', \quad \cdots . \quad D_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} := e_n'$$

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 26

(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, but generalized beyond derivative matrices. 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 32. 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.[6] 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,

$$I := \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

[6] 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 27 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 5.7 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".[7] 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.[8] 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 $\alpha \cdot v = \alpha$.[9] 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:

$$\alpha \cdot \alpha = \alpha,^{10} \quad \alpha \cdot \beta = \beta \cdot \alpha = \beta,^{11} \quad \beta \cdot \beta = a\alpha + b\beta$$

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.[12] 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

$$\bar{\beta} \cdot \bar{\beta} = (a + c^2)\alpha + (b - 2c)\beta.$$

[7] Obviously no one *created* quantum phenomena, but someone had to work out the *theory* of quantum mechanics.

[8] We already know how to multiply two vectors, each with one real entry. Two entries is the next simplest.

[9] 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.

[10] This is the analogue of $1 \cdot 1 = 1$.

[11] This is the analogue of $1 \cdot \beta = \beta \cdot 1 = \beta$.

[12] 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 27 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.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ a_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{pmatrix}$$

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 27: 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

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} ac-bd & ad+bc \\ -bd-ad & -bd+ac \end{pmatrix} = \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix}.$$

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 28 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 $AB = BA = 1$.[13] The matrix B is called the **inverse matrix** of A, and is denoted by the symbol A^{-1}.[14]

[13] 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).

[14] For this Challenge, it may be helpful to recall that for real numbers a and b, with nonzero a, we have $\frac{b}{1/a} = ab$.

(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 $B = C$.[15]

(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:

$$aw + by = 1, \quad ax + bz = 0, \quad cw + dy = 0, \quad cx + dz = 1.$$

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.[16]

(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) 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?

(f) Verify that your answer from part (e) matches our intuition from real numbers by setting $y := 0$ and checking that it is the same as that of real numbers.

(g) 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 $d := 0$.

From your work in Challenge 28, we know that we can divide complex numbers by other nonzero complex numbers,[17] 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 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.

[15] *Hint:* $B = 1B = (CA)B$.

[16] *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!

[17] A nonzero complex number is a complex number with at least one nonzero real part or imaginary part.

Further concepts

The sum of the squares of the real part and imaginary part of a complex number appeared numerous times in Challenge 28, 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.

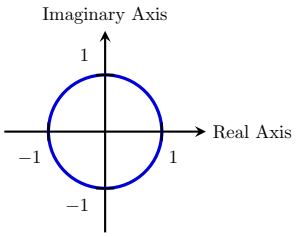


Figure 5.4: The set of complex numbers z with $|z| = 1$ form a unit circle (the blue circle).

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**.[18] Figure 5.4 is an example of an Argand diagram.

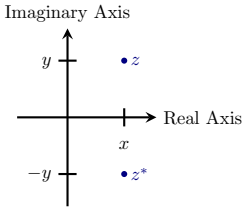


Figure 5.5: A complex number $z := x + yi$ and its complex conjugate $z^* = x - yi$.

[18]The identification of complex numbers as geometric objects (points on a plane) was apparently done first in 1799 by the mathematician Caspar Wessel.

From Figure 5.5, 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 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. At this point, we are comfortable with making such choices from calculus, so we see that there was no loss in generality by making one choice over the other.

Challenge 29 Let z and z' be complex numbers.

(a) Show that

$$\operatorname{Re} z = \frac{1}{2}(z + z^*), \qquad \operatorname{Im} z = \frac{1}{2i}(z - z^*).$$

(b) Show that the division of complex numbers can be done by a division by a real number:

$$z'/z = z' \left(\frac{z^*}{zz^*} \right).$$

Challenge 30

(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).[19] What is $|uw|$?

(d) Let $z_1 := a + bi$ and $z_2 := c + di$ be complex numbers, with $|z_1| = 1$. Show that $z_3 := z_1 z_2$ 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 $|u_z| = 1$ 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

[19] 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 $\alpha - \epsilon < \sqrt{2} \leq \alpha$. This naive but simple procedure will give better estimates of $\sqrt{2}$ for smaller ϵ, but will take longer.

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 (assuming they have compatible dimensions). Using matrix inverses, as discussed in Challenge 28, 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. There is however, one operation that we can do with complex numbers that we do not have a matrix analogue. This is the complex conjugation operation.

Let us examine the conjugation operation one more time. 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 a matrix with dimension $m \times n$ as defined below.

$$A := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \tag{5.8}$$

The **transpose** of matrix A, denoted A^t, is the matrix of dimension $n \times m$ defined by

$$A^t := \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2m} & \cdots & a_{mn} \end{pmatrix}.$$

For example, for the real matrix $z := \begin{pmatrix} x & y \\ -y & x \end{pmatrix}$, its transpose matrix is

$$z^t = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}.$$

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 defined in 5.8 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 $H^\dagger = H$.[20] 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 30). 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 $UU^\dagger = U^\dagger U = 1$.[21] 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 began this section by trying to work out the product of two real vectors. The result of our labour was the complex numbers. We then made a connection with real matrices to work out the arithmetic of this new number system. Now that we know about complex vectors and complex matrices, can we discover a number system even better?

Unfortunately, this procedure will not produce new fields, and we will have to give up some essential properties that we expect numbers to have. We have been driven from the start to extend our sense of what a number is, and this is where this journey ends.[22] However, we have built up so much machinery that it would be a shame not to put it into action before we part. Let us 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. Be warned, we will be rather cavalier about applying previously obtained results in more general settings, and so our methods will be even less rigorous than before.

5.4 Quantum Dynamics

The Schrödinger equation

The only mechanical system we know of is the 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 $|f\rangle$, which exists in the physical world without us writing down its contents as a vector in some choice

[20] We can deduce that H must be a square matrix.

[21] Observe that $U^\dagger$ is the matrix inverse of U. Hence a unitary matrix is always a square matrix with an inverse.

[22] There is still *much* left to discover about numbers, fields or otherwise, and so I do hope that with the end of this journey, you will begin another!

of units to quantify its ingredients. If someone pressed us for the ingredients, then we could present a representation of $|f\rangle$ 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 $U(\alpha)$.

$$|\Psi(t+\alpha)\rangle = U(\alpha)\,|\Psi(t)\rangle \tag{5.9}$$

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.

$$U(\alpha)=U(0)+o(1)=U(0)-\alpha H+o(\alpha).$$

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 $-H$.[23]

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 (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 5.9 and use linearity to get

$$\begin{aligned}|\Psi(t+\alpha)\rangle &= U(\alpha)\,|\Psi(t)\rangle = (1-i\alpha H+o(\alpha))\,|\Psi(t)\rangle\\ &= |\Psi(t)\rangle - i\alpha H\,|\Psi(t)\rangle + o(\alpha).\end{aligned}$$

[23]Notice that $-H$ is the matrix $(-1)H$, where -1 is a real number.

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 $\times$ Time. Then we will put $U(\alpha) = 1 - \frac{i}{\hbar}\alpha H + o(\alpha)$ from which we deduce $|\Psi(t+\alpha)\rangle = |\Psi(t)\rangle - \frac{iH}{\hbar}\alpha|\Psi(t)\rangle + o(\alpha)$, giving us the equation $\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = -\frac{i}{\hbar}H|\Psi\rangle$, or equivalently:

$$\boxed{i\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = H|\Psi\rangle}$$

the **Schrödinger equation**.

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 $\left(1 - \alpha\frac{\mathrm{d}}{\mathrm{d}x}\right)$.[24] 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

$$T(\alpha) = 1 - \frac{i}{\hbar}\alpha\left(-i\hbar\frac{\mathrm{d}}{\mathrm{d}x}\right).$$

[24] Modulo some terms negligible compared to α.

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!

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.

Challenge 31 The spirit of our approach is starting from our intuition gained from previous chapters and generalizing. In particular, we will assume that all instances of e^X for some X (real, complex, complex matrices, etc) obey the same rules as the real exponential function e^x.

(a) In Section 3.4, we defined the real exponential function e^x whose derivative was itself. That is, $e^{x+\alpha} = e^x + e^x\alpha + |\alpha|o(1)$. Show that an input to the exponential function must be dimensionless.

(b) Taking $x := 0$ gives $e^\alpha = e^0 + e^0\alpha + |\alpha|o(1)$. Let A be a dimensionless constant. Using the fact that $e^0 = 1$ and $|c|o(1) = o(1)$ for a constant c, show that $e^{\alpha A} = 1 + \alpha A + o(1)$. For a scalar c and matrix A, we define the **matrix exponential** as $e^{cA} := 1 + cA + o(1)$.

(c) Use the fact that $e^{a+b} = e^a e^b$ to show that $\frac{\mathrm{d}}{\mathrm{d}x}\left(e^{Ax}\right) = Ae^{Ax}$.

(d) We began with the relation $|\Psi(t)\rangle = U(t)\,|\Psi(0)\rangle$. Plugging this into the Schrodinger equation gives $i\hbar\frac{\mathrm{d}}{\mathrm{d}t}U(t)\,|\Psi(0)\rangle = HU(t)\,|\Psi(0)\rangle$. This is true for any $|\Psi(0)\rangle$, and we are free to pick whichever initial value of $|\Psi(0)\rangle$ we want, so we will ignore those terms and write $i\hbar\frac{\mathrm{d}}{\mathrm{d}t}U(t) = HU(t)$. Assuming that H is a time-independent Hamiltonian (that is, H is not a function of time), and thus treating H like a constant, this is a differential equation that looks familiar from part (c). Find the factor A so that $U(t) = e^{At}$ is a solution to the differential equation $i\hbar\frac{\mathrm{d}}{\mathrm{d}t}U(t) = HU(t)$. In particular, notice that the matrix exponential e^{iX} rotates complex vectors, because it is a unitary matrix.

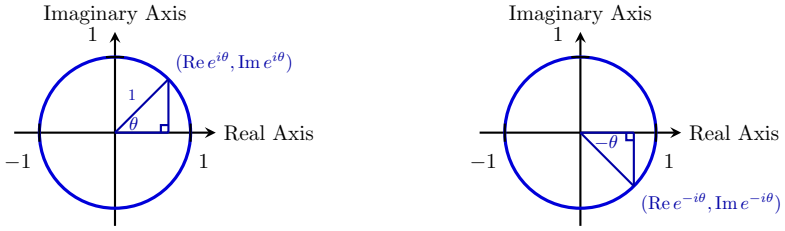


Figure 5.6: Argand diagram of $e^{i\theta}$ (left) and $e^{-i\theta}$ (right).

The interpretation that e^{iX} rotates complex vectors is incredibly useful. Let us stick to the simplest case of the rotation by $e^{i\theta}$ where θ is a real number. Taking the complex number 1 on

the unit circle and multiplying it by $e^{i\theta}$ for positive θ rotates 1 into another complex number from the x-axis upwards in the unit circle (left diagram in Figure 5.6). How much does the function $e^{i\theta}$ rotate a complex number by? Recall that an angle is dimensionless; the simplest formula for the amount of rotation would be $c \cdot \theta$. The dimensionless constant c is very easy to remember, it is 1! So the function $e^{i\theta}$ rotates a complex number by angle θ. If θ is negative, then taking the complex number 1 on the unit circle and multiplying it by $e^{i\theta}$ rotates 1 by angle θ from the x-axis downwards (right diagram in Figure 5.6). Observe that $(e^{i\theta})^* = (e^{-i\theta})$. From Section 3.2, we saw that the Pythagorean theorem allows us to identify each point on the unit circle with a right triangle.

Definition 33. The **cosine** function is defined to be $\cos : \theta \mapsto \operatorname{Re} e^{i\theta}$. The **sine** function is defined to be $\sin : \theta \mapsto \operatorname{Im} e^{i\theta}$. The **tangent** function is defined to be $\tan : \theta \mapsto \sin\theta / \cos\theta$.

By the Pythagorean theorem, $\cos^2\theta + \sin^2\theta = 1$.[25] Since $e^{i\theta} = \operatorname{Re} e^{i\theta} + i \operatorname{Im} e^{i\theta}$, we know that

$$e^{i\theta} = \cos\theta + i\sin\theta. \tag{5.10}$$

Equation 5.10 is called **Euler's formula**.

A rotation of a nonzero complex number z by angle $(\theta_1 + \theta_2)$ is achieved by taking the product $z \cdot e^{i(\theta_1+\theta_2)}$. The same rotation can be achieved by rotating it first by angle θ_1 with $z \cdot e^{i\theta_1}$ and then rotating the result by angle θ_2 by taking a second product $(z \cdot e^{i\theta_1}) \cdot e^{i\theta_2}$. Therefore,

$$e^{i(\theta_1+\theta_2)} = e^{i(\theta_1+\theta_2)}. \tag{5.11}$$

Challenge 32

(a) Use the definition of the cosine function and Equation 5.11 to show that

$$\cos(\theta_1 + \theta_2) = \cos\theta_1 \cos\theta_2 - \sin\theta_1 \sin\theta_2.$$

[*Hint:* after following the steps given, use Euler's formula (there is not a whole lot to try!).]

(b) Use the definition of the sine function and Equation 5.11 to show that

$$\sin(\theta_1 + \theta_2) = \cos\theta_1 \sin\theta_2 + \sin\theta_1 \cos\theta_2.$$

The equations of parts (a) and (b) are called the trigonometric **addition formulas**.

(c) Use Equation 5.11 to obtain the **double angle formulas**:

$$\cos(2\theta) = \cos^2\theta - \sin^2\theta, \qquad \sin(2\theta) = 2\sin\theta\cos\theta.$$

(d) Use the well-ordering principle on Equation 5.11 to show that for each natural number n:

$$(e^{i\theta})^n = e^{in\theta}$$

(e) Use part (d) to obtain **de Moivre's formula**:

$$(\cos\theta + i\sin\theta)^n = \cos(n\theta) + i\sin(n\theta).$$

(f) Use Challenge 29 and the fact that $(e^{i\theta})^* = (e^{-i\theta})$ to show that:

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \qquad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

[25] The expressions $\cos^2\theta$ and $\sin^2\theta$ mean $(\cos\theta)^2$ and $(\sin\theta)^2$ respectively.

Challenge 33 Consider a particle of mass m moving on a line located at position x_0 at initial time $t = 0$. Put

$$U(x,t) := \sqrt{\frac{m}{2\pi i\hbar t}} e^{im(x-x_0)^2/(2\hbar t)}.$$

We will think of $e^{im(x-x_0)^2/(2\hbar t)}$ as the real exponential function e^{cx}. That is to say, $(e^{cx})' = ce^{cx}$.

(a) As a warm up, use Challenge 32 part (f) to show that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\cos\theta = -\sin\theta, \qquad \frac{\mathrm{d}}{\mathrm{d}\theta}\sin\theta = \cos\theta, \qquad \frac{\mathrm{d}}{\mathrm{d}\theta}\tan\theta = \frac{1}{\cos^2\theta}.$$

(b) Show that

$$\frac{\partial}{\partial x}U = \frac{im(x-x_0)}{\hbar t}U.$$

(c) Show that

$$\frac{\partial^2}{\partial x^2}U = \frac{im}{\hbar t}U - \frac{m^2(x-x_0)^2}{\hbar^2 t^2}U.$$

(d) Show that

$$\frac{\partial}{\partial t}U = -\frac{1}{2t}U - \frac{im(x-x_0)^2}{2\hbar t^2}U.$$

(e) 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 $V = 0$.

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.[26] 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 was one of the crowning jewels of mathematical physics of the 19th Century: the **Hamilton–Jacobi equation**

$$-\frac{\partial S}{\partial t} = H\left(x, p := \frac{\partial S}{\partial x}, t\right). \tag{5.12}$$

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.

[26] Newton's experimental work on optics and light involved a very famous experiment that nearly blinded him.

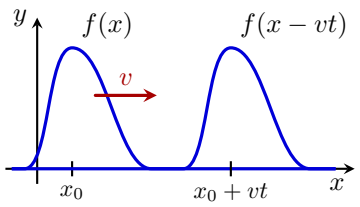


Figure 5.7: A wave traveling at speed v to the right.

Imagine a wave which we represent by a function f that is traveling to the right at some speed v (see Figure 5.7). 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,

$$f(x, t) = g(x - vt).$$

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 stipulates 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.

$$\boxed{\frac{\partial^2 f}{\partial t^2} = v^2 \frac{\partial^2 f}{\partial x^2}} \tag{5.13}$$

This is a **linear differential equation** because linear combinations of solutions to the wave equation are also solutions.[27]

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

$$-\frac{i}{\hbar}\frac{\partial \Psi}{\partial t} = \frac{1}{2m}\frac{\partial^2 \Psi}{\partial x^2}.$$[28]

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

$$\frac{\mathrm{d}}{\mathrm{d}t}|\Psi\rangle = -\frac{i}{\hbar}H|\Psi\rangle. \tag{5.14}$$

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 5.14 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 34 From Challenge 30, we saw that each complex number z is as a product of the real number $|z|$ and a complex number in the unit circle u_z. Just as the number u_z rotates a complex number, the unitary matrix U rotates our states to create time evolution on our oscillator. In the case of our simple system where the Hamiltonian H stays the same over time, we know that the unitary matrix U can take the form $U(t) = e^{-iH/\hbar}$ (Challenge 31). So e^{iX} rotates complex vectors.

(a) 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. Use the product rule to show that

$$\frac{\partial \Psi}{\partial t} = \left(\dot{\rho} + i\frac{\dot{\omega}}{\hbar}\rho\right)e^{i\omega/\hbar}, \qquad \frac{\partial \Psi}{\partial x} = \left(\rho' + i\frac{\omega'}{\hbar}\rho\right)e^{i\omega/\hbar}.$$

[27] 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!

[28] We have divided both sides by the nonzero constant $\hbar^2$ and multiplied both sides by -1.

(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

$$i\hbar\dot{\rho} - \rho\dot{\omega} = -\frac{\hbar^2}{2m}\rho'' - i\frac{\hbar}{m}\rho'\omega' - i\frac{\hbar}{2m}\omega''\rho + \frac{1}{2m}(\omega')^2\rho + V\rho. \tag{5.15}$$

(d) Equation 5.15 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 have magically disappeared. Take $\hbar \to 0$ to obtain a much simpler equation and then divide through by R (so R must be nonzero for each x and t, interesting!) to get the following.[29]

$$-\frac{\partial \omega}{\partial t} = \frac{1}{2m}\left(\frac{\partial \omega}{\partial x}\right)^2 + V(x) \tag{5.16}$$

(e) The right side of Equation 5.16 from part (d) is a Hamiltonian (total energy) with momentum $p := \frac{\partial \omega}{\partial x}$. Conclude that

$$-\frac{\partial \omega}{\partial t} = H\left(x, p := \frac{\partial \omega}{\partial x}, t\right). \tag{5.17}$$

Have we seen Equation 5.17 before? It is simply the Hamilton–Jacobi equation (Equation 5.12)! 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 $\hbar$ has the dimension Energy $\times$ Time, where energy is measured in joules (symbol J). 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.

The circle

The dynamics of an oscillator in phase space is that of an ellipse, however its motion is a mass simply moving back and forth. How about we look into a system with something taking the motion of a circle? The development of calculus from Newton's side began with the desire to understand the motion of planets. However, as we saw, these are too big to study using quantum mechanics.

[29] 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.

So let us examine the simplest circular system that is also tiny. An atom, but not just any atom: the **hydrogen atom**. The hydrogen atom is not only by far the most abundant type of atom, it is also the simplest: an electron orbiting a proton (see Figure 5.8).

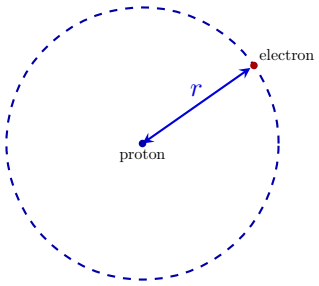


Figure 5.8: A hydrogen atom (diagram not to scale).

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: $q_1 q_2$.

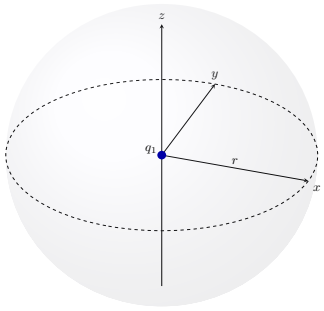


Figure 5.9: All charged particles of equal charge on the boundary of a sphere of radius r centered at a point charge q_1 feels the same electric force.

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 charge.[30] Thus all charged particles with charge q_2 in the boundary of a sphere of radius r (see Figure 5.9) 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

[30] It would be weird if there was some distinguished axis where the force was stronger or weaker.

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.[31]

From Challenge 12, 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

$$F = \frac{q_1 q_2}{4\pi r^2 \epsilon_0} \tag{5.18}$$

where ϵ_0 is a dimensionful constant that allows us match the units in both sides of the equation. Equation 5.18 is called **Coulomb's law**.

Let us calculate the potential energy for the hydrogen atom, where $q_1 q_2 = -e^2$. 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 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 $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ takes the place of $\frac{\partial^2}{\partial x^2}$.[32]

$$i\hbar\frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\Psi + V\Psi$$

For the hydrogen atom, the mass term is now the mass of the electron m_e, and the potential energy V is given by $-\frac{e^2}{4\pi\epsilon_0}\frac{1}{r}$. Therefore, the Schrödinger equation for the hydrogen atom is

$$i\hbar\frac{\partial \Psi}{\partial t} = -\frac{\hbar^2}{2m_e}\nabla^2\Psi - \frac{e^2}{4\pi\epsilon_0}\frac{1}{r}\Psi.$$

There is a lot going on in this equation, so I have written down the dimensionful quantities of the equation in the table below with the 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.

[31] This 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 particles of charge q_2 on the boundary of a general ellipsoid felt the same force, because the rate of change is no longer uniform and differing depending on the direction away.

[32] Energy is a scalar quantity, and so we cannot replace $\frac{\partial^2}{\partial x^2}$ with the vector $\left(\frac{\partial^2}{\partial x^2} \quad \frac{\partial^2}{\partial y^2} \quad \frac{\partial^2}{\partial z^2}\right)^t$.

	Variable	Dimension
Radius of hydrogen atom	a_0	L
Reduced Planck constant	$\hbar$	ML^2/T
Electron mass	m_e	M
Coulomb term	$\frac{e^2}{4\pi\epsilon_0}$	$\mathrm{ML}^3/\mathrm{T}^2$

Challenge 35

(a) The simplest formula for expressing the radius a_0 using the other three variables is

$$a_0 = \beta \cdot \hbar^x \cdot m_e^y \cdot \left(\frac{e^2}{4\pi\epsilon_0}\right)^z,$$

where β is some dimensionless constant. Find integers x, y, and z that satisfy the formula.

(b) Suppose we knew nothing about quantum mechanics, and so we did not know the existence of $\hbar$. Show that it is not possible to find a combination of integers x and y such that $a_0 = \beta \cdot m_e^x \cdot \left(\frac{e^2}{4\pi\epsilon_0}\right)^y$.

(c) By part (b) we know that $\hbar$ is crucial in our formula in part (a). However, the value of $\hbar$ is far too small and so there is a danger we ignore it when rounding things while we do our calculations. Furthermore, our formula from part (a) contains too many constants. Let us fix both problems. We introduce a new constant c, the *speed of light in vacuum* and the **fine-structure constant** $\alpha := \frac{e^2}{4\pi\epsilon_0\hbar c}$. Find integers x', y', and z' such that $a_0 = \beta \cdot (\hbar c)^{x'} \cdot (m_e c^2)^{y'} \cdot \alpha^{z'}$.[33]

(d) The value of $\hbar c$ is about 200 eV nm, the value of $m_e c^2$ is about 0.5×10^6 eV, and α is about 1/137. The unit of energy eV is called an *electronvolt* which, if we did everything correctly in part (c), should cancel out (because a_0 has the dimension of Length, and an electron*volt* is not a Length!). The unit nm is **nanometers**, which is defined to be 1×10^{-9} meters. Calculate the value of a_0 using the values given, then include the dimensionless constant β at the end.

(e) We introduced the "speed of light in vacuum" c in order to simplify calculations. But perhaps we should have included it from the beginning in part (a)? We did not include c because it did not show up in the Schrödinger equation. Nevertheless, form a velocity v_0 using dimensional analysis on the variables $\hbar, m_e, \frac{e^2}{4\pi\epsilon_0}, a_0$ and the new constant c.[34] Conclude that v_0 is proportional to αc (about $c/137$). Since the velocity involved is quite small compared to the speed of light, *relativistic* effects can be ignored, and we were justified in not including c at the start.

(f) An **angstrom**, denoted by the symbol Å, is defined to be 0.1 nm. The *diameter* of the hydrogen atom is experimentally known to be about 1.1 Å.[35] To a first approximation, what is the dimensionless constant β?

[33] *Hint:* just a glance at our formula from part (a) gives the values of y' and z'. So the only thing to do is match the units with x'.

[34] *Hint:* we are looking for the *simplest* formula. Most of the constants will be unneeded.

[35] Recall that a **diameter** of a circle is twice that of its radius.

All done?

Calculus is not easy. However, if you go back to the first pages, you will discover what follows to be painfully slow baby steps. Allow me to say a few words on baby steps.

When we were very very young, we were at the care of our guardians, fed and cared for, with no harm from natural predators in the wild. The optimal thing to do was to lie on our backs and idle. But that is not what you and I did. We crawled. We tried to make some steps and failed. We would fall, and get up again, and fall. No matter the hurt from the falling and obstacles, we would take our baby steps to wherever our curiosities lead us.

You were born to be a discoverer. You were trained to be one from the youngest age, by yourself! You did not need me to tell you that, but perhaps those baby steps were so long ago that a gentle reminder was in order. Now go forth and discover!

A

Gaußian Integrals

A.1 The Integrals

We are back to working with real numbers. The goal of this Appendix is to calculate the most important integral of them all: $\int_{-\infty}^{\infty} e^{-ax^2}\,dx$, 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,

$$\int_{-\infty}^{\infty} e^{-ax^2}\,dx = \sqrt{\frac{\pi}{a}}. \tag{A.1}$$

In this Appendix, we will show that c is $\sqrt{\pi}$.

Before we dive into the calculation, let us extend this by considering 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 34 (Quadratic Formula)**.** A **quadratic equation** $ax^2 + bx + c = 0$ with nonzero a is solved by the formula

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

Proof. We use the the technique of **completing the square**. Dividing by a and subtracting c/a on both sides of the equation gives

$$x^2 + \frac{b}{a}x = -\frac{c}{a}.$$

[1] This was left for you in Challenge 31, 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.

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 $\pm$ 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 $\pm$ symbol and write $\pm\sqrt{a^2}$.

Theorem 35 (The 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 A.1. If we turn $-ax^2+bx$ into $-au^2+c$ for some constant c, then Equation A.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

$$\int_o^\infty f(x)\,dx := \lim_{t\to\infty}\int_o^t f(x)\,dx.$$

Now that we know 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 36 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 $(0,\infty)$. Taking the derivative of f with respect to a (which is now a variable) shows that

$$-\frac{1}{a^2} = f'(a) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}a}e^{-ax}\,dx = -\int_0^\infty xe^{-ax}\,dx.$$

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 obtain the **gamma function**:

$$n! = \int_0^\infty x^n e^{-x}\,dx.^2$$

(b) Apply a differentiation under the integral sign on Equation A.1 to show that

$$\int_{-\infty}^\infty x^2 e^{-ax^2}\,dx = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}.$$

Observe that repeating the technique gives the formula for $\int_{-\infty}^\infty x^4 e^{-ax^2}\,dx$ and so on.

(c) Apply differentiation under the integral sign on Theorem 35 (on the variable b) to show that

$$\int_{-\infty}^\infty xe^{-ax^2+bx}\,dx = \frac{b}{2a}e^{b^2/(4a)}\sqrt{\frac{\pi}{a}}.$$

[2] This function allows us to calculate factorials for positive real numbers n, beyond natural numbers.

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 $g(x) := x^2$:

$$\int_0^\infty xe^{-ax^2}\,dx = \frac{1}{2}\int_0^\infty e^{-ax^2}2x\,dx = \frac{1}{2}\int_0^\infty e^{-ag(x)}g'(x)\,dx = \frac{1}{2}\int_0^\infty e^{-au}\,du = \frac{1}{2a}.$$

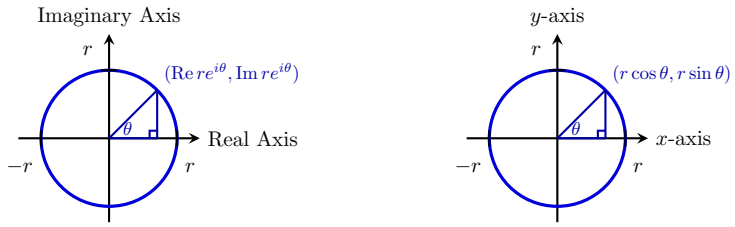
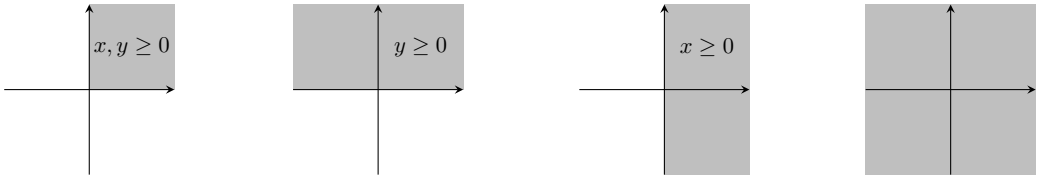


Figure A.1: Argand diagram of $re^{i\theta}$ (left), which is the same as $(r\cos\theta, r\sin\theta)$ (right).

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 coordinates (x, y) on the x-y plane using the representation $(r\cos\theta, r\sin\theta)$ is called **polar coordinates**. One thing to note is that we will measure angle in **radians**. Recall that an angle Θ is the length of the red arc divided by the circumference of the blue circle (see footnote).[3] The angle that covers the full circle is defined in radians to be the circumference of the unit circle 2π. Hence a right angle in radians is $\pi/2$ because we need four right angles to cover the circumference of a circle. The angle corresponding to a semicircle is π because we need two of them to cover the circumference of a circle.



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 $\theta \in [0, 2\pi]$.

3

Now that we have a new way of representing coordinates, let us figure out how to make a substitution for an integral.

A.2 Change of Variables

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 we have

$$\underbrace{f(\tilde{a}\epsilon) - f(a\epsilon)}_{\text{rise in value}} = f'(t)(\tilde{a}\epsilon) - f'(t)(a\epsilon) = f'(t) \cdot \underbrace{\big([\tilde{a} - a]\,\epsilon\big)}_{\text{change along } x\text{-axis}}.$$

We will denote the function's rise by df and the change of inputs along the x-axis by dx and write

$$df = f'(t)\,dx. \tag{A.2}$$

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 A.2.

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,

$$df = \partial_x f(t)\,dx + \partial_y f(t)\,dy. \tag{A.3}$$

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

$$0 = X^2 = (dx + dy)(dx + dy) = (dx)^2 + dx\,dy + dy\,dx + (dy)^2 = 0 + dx\,dy + dy\,dx + 0.$$

We see that $dx\,dy = -dy\,dx$. How about a product of linear combinations?

$$(\alpha\,dx + \beta\,dy)(\gamma\,dx + \delta\,dy) = 0 + (\alpha\delta)dx\,dy + (\beta\gamma)dy\,dx + 0 = (\alpha\delta - \beta\gamma)dx\,dy. \tag{A.4}$$

We now apply our new algebra to do calculus. The polar coordinates are described by the rule

$$x = r\cos\theta \qquad \text{and} \qquad y = r\sin\theta. \tag{A.5}$$

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 respects to increases in each of its inputs in Equation A.3. 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 (Challenge 33) the partial derivatives are:

$$\partial_r g_1 = \cos\theta, \quad \partial_\theta g_1 = -r\sin\theta, \quad \partial_r g_2 = \sin\theta, \quad \partial_\theta g_2 = r\cos\theta.$$

Their product is then the product of linear combinations from Equation A.4

$$dx\, dy = (\partial_r g_1 \partial_\theta g_2 - \partial_\theta g_1 \partial_r g_2)\, dr\, d\theta = (r\cos^2\theta + r\sin^2\theta)\, dr\, d\theta = r\, dr\, d\theta \tag{A.6}$$

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 $\int_A dx\, dy$, where A is the set of points (x, y) on the x-y plane such that $x^2 + y^2 \leq \bar{r}^2$. Each point in A corresponds to a point $(r\cos\theta, r\sin\theta)$, where $r \in [0, \bar{r}]$ and $\theta \in [0, 2\pi]$. We will take the integral $\int_A dx\, dy$ and substitute set A with the region $r \in [0, \bar{r}]$ and $\theta \in [0, 2\pi]$ and substitute $dx\, dy$ with our result from Equation A.6. Applying our procedure gives the expected answer, as shown below.

$$\int_A dx\, dy = \int_0^{2\pi}\int_0^{\bar{r}} r\, dr\, d\theta = \int_0^{2\pi} \frac{\bar{r}^2}{2}\, d\theta = \frac{\bar{r}^2}{2}\int_0^{2\pi} d\theta = \pi\bar{r}^2$$

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

$$\partial_r g_1 = \cos\theta, \quad \partial_\theta g_1 = -r\sin\theta, \quad \partial_r g_2 = -\sin\theta, \quad \partial_\theta g_2 = -r\cos\theta$$

and so

$$dx\, dy = (\partial_r g_1 \partial_\theta g_2 - \partial_\theta g_1 \partial_r g_2)\, dr\, d\theta = (-r\cos^2\theta - r\sin^2\theta)\, dr\, d\theta = -r\, dr\, d\theta.$$

Therefore, using our algebraic rules from Equation A.4, they will calculate the area of a circle of radius $\bar{r}$ to be

$$\int_A dx\, dy = \int_0^{2\pi}\int_0^{\bar{r}} (-r)\, dr\, d\theta = -\int_0^{2\pi} \frac{\bar{r}^2}{2}\, d\theta = -\frac{\bar{r}^2}{2}\int_0^{2\pi} d\theta = -\pi\bar{r}^2 ?!$$

A circle having negative area is absurd, and they did everything correctly! This means that the algebraic rule from Equation A.4 must be modified to:

$$(\alpha\,dx + \beta\,dy)(\gamma\,dx + \delta\,dy) = |\alpha\delta - \beta\gamma| dx\,dy.$$

Let us review what we have found. If we wish to calculate an integral $\int_A f(x,y)\,dx\,dy$, 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 is the **change of variables formula** in two dimensions:

$$\int_A f(x,y)\,dx\,dy = \int_{\tilde{A}} f\big(g_1(u,v), g_2(u,v)\big)\,|\partial_u g_1 \partial_v g_2 - \partial_v g_1 \partial_u g_2|\;du\,dv. \tag{A.7}$$

We can tidy up our formula. Each transformation $x = g_1(u,v)$ and $y = g_2(u,v)$ are real valued functions of two variables. Let us take the gradients of each function and stack their *transpose* together to obtain the **Jacobian matrix** g' for the transformation $g(u,v) := \begin{pmatrix} g_1(u,v) \\ g_2(u,v) \end{pmatrix}$, defined by

$$g' := \begin{pmatrix} (\boldsymbol{\nabla} g_1)^t \\ (\boldsymbol{\nabla} g_2)^t \end{pmatrix} = \begin{pmatrix} \partial_u g_1 & \partial_v g_1 \\ \partial_u g_2 & \partial_v g_2 \end{pmatrix}.$$

Observe that the expression inside the absolute values of Equation A.7 is the determinant of the Jacobian matrix![4] Furthermore, the region A and $\tilde{A}$ have the relationship $A = g(\tilde{A})$. We will rename $\tilde{A}$ with A and write the change of variables formula in two dimensions as:

$$\int_{g(A)} f(x,y)\,dx\,dy = \int_A (f \circ g)(u,v)\,|\det g'|\;du\,dv. \tag{A.8}$$

Challenge 37 Why is the Jacobian matrix of function g denoted as if it was the derivative of g? For a Jacobian matrix of dimension 1×1, the concepts do coincide, and the Jacobian matrix g' is the same as the derivative. For this Challenge only, we will work with the complex numbers.

(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 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 g' and interpret the real matrix as a complex function. Is it as you expected?

(c) We may think of the Jacobian matrix f' as the matrix representation of the derivative of function f. If f is a complex-valued function, then its Jacobian matrix must have the form $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$ because the derivative 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 following equations hold.

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \qquad \text{and} \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \tag{A.9}$$

Equations A.9 are called the **Cauchy-Riemann equations**.

[4] The determinant of a Jacobian matrix is called the **Jacobian**.

Challenge 38 We generalize to functions of three variables.

(a) Use our rules for the symbols $d\square$ to obtain the relations

$$dx\,dy = -dy\,dx, \quad dx\,dz = -dz\,dx, \quad dy\,dz = -dz\,dy.$$

(b) Use part (a) to conclude that

$$dx\,dy\,dz = -dy\,dx\,dz = dy\,dz\,dx = -dz\,dy\,dx = dz\,dx\,dy = -dx\,dz\,dy.$$

(c) Calculate $du\,dv\,dw$ where $du := \partial_x g_1\,dx + \partial_y g_1\,dy + \partial_z g_1\,dz$, $dv := \partial_x g_2\,dx + \partial_y g_2\,dy + \partial_z g_2\,dz$, and $dw := \partial_x g_3\,dx + \partial_y g_3\,dy + \partial_z g_3\,dz$. There are six terms because $(dx)^2 = (dy)^2 = (dz)^2 = 0$. Don't forget to put the coefficients inside an absolute value to prevent negative volumes arising because of the conversion factor.

(d) Define

$$g(u,v,w) := \begin{pmatrix} g_1(u,v,w) \\ g_2(u,v,w) \\ g_3(u,v,w) \end{pmatrix} \quad \text{and } g' := \begin{pmatrix} (\boldsymbol{\nabla} g_1)^t \\ (\boldsymbol{\nabla} g_2)^t \\ (\boldsymbol{\nabla} g_3)^t \end{pmatrix}.$$

The Jacobian matrix g' is a matrix of dimension 3×3. The determinant of a matrix of dimension 3×3 is defined to be

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} := aei - afh + bfg - bdi + cdh - ceg.$$

The determinant here is also defined such that a 3×3 matrix is invertible if its determinant is nonzero. Conclude that the **change of variables formula** in three dimensions is given by

$$\int_{g(A)} f(x,y,z)\,dx\,dy\,dz = \int_A (f \circ g)(u,v,w) |\det g'|\,du\,dv\,dw.$$

Calculating the integral

Finally at last, we will calculate the integral.[5]

Theorem 36.

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

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:

$$I^2 = I \int_0^\infty e^{-y^2}\,dy = \int_0^\infty I e^{-y^2}\,dy = \int_0^\infty \left(\int_0^\infty e^{-x^2}\,dx \right) e^{-y^2}\,dy.$$

Since e^{-y^2} is a constant with respect to the variable x, we push it in:

$$I = \int_0^\infty \int_0^\infty e^{-x^2} e^{-y^2}\,dx\,dy = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\,dx\,dy.$$

[5]This integral can also be calculated using differentiation under the integral sign. I encourage you to look it up!

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 A.6 we know that $dx\,dy = r\,dr\,d\theta$.

We make the change of variables $x^2 + y^2 \mapsto r^2$ 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 $x \mapsto x/\sqrt{a}$ for positive a gives

$$\int_{-\infty}^{\infty} e^{-ax^2}\,dx = \sqrt{\frac{\pi}{a}}.$$

Index