Some mathematics from my time at Highgate School

Damon Falck (Class of 2018)

July 2018

Introduction

This is a short collection of some work I have accumulated during my two wonderful years in the Sixth Form at Highgate. It's mainly for my future reference; this is a point in my life at which I'd like to collate and organise what I've done so far, so that I can look over it and have it in one place for the future. It's also so that I can look back and remember my brilliant teachers and how much I learned here.

I have taken care to exclude mundane things from this document. With the exception of my two coursework projects and one particular homework set, nothing here is from the core A-level course.

Of course, so much else has happened (we've managed to run about 60-70 talks at the Vaccaro Society for one), so this is just the stuff that happened to be in a nicely written down format lending itself to compilation.

Contents

Part I

Projects and write-ups

Chapter 1

An introduction to the calculus of variations

I wrote this during the October 2017 half-term break after coming across the brachistochrone problem and being fascinated by the 'calculus of variations' that was apparently behind it. Of course I followed along standard arguments quite closely in deriving the main equations, but I tried to be quite independent in applying them to a few different problems.

An Introduction to the Calculus of Variations

Damon Falck

October 2017

Abstract

This paper is intended as an accessible introduction to the fundamentals of the calculus of variations, a sort of generalisation of calculus. After covering some basic prerequisites in multivariable calculus, I'll explain the derivation of the most important results in the field and then demonstrate their application to a few interesting and historic problems. The reader is assumed to have fluency in single-variable calculus including integration by parts.

Contents

1 Introduction

So far, we've come across the concept of optimising (finding maxima and minima of) functions in one or more dimensions with respect to some variable. However, often in mathematics and physics there comes the need to find some sort of *path* which either maximises or minimises a value — to find the best possible *function* for a given situation. This is the main task with which the calculus of variations is concerned.

Originally, the calculus of variations was Euler's response to the Brachistochrone problem posed by Johann Bernoulli in 1696, which I discuss later on in section 7. Lagrange and others then took this and developed it further into a rich field of interesting and useful mathematics.

Among the problems we can solve with the calculus of variations are:

- Finding the curve between two points that an object will slide down most quickly
- Finding the shape of a given perimeter that encloses the largest area
- Finding the shortest path between two points over some surface
- Finding the best shape of a container to retain heat, subject to certain constraints
- Many other problems

Most problems that involve finding some sort of optimal path or shape can be solved using the calculus of variations.

In this paper I start by touching on a couple of useful concepts from multivariate calculus, before going on to derive the Euler-Lagrange equation, the cornerstone of the calculus of variations. The actual derivation is quite dense and may be skipped by the reader. I'll then show how a few of the historic problems with which variational calculus is concerned can be solved.

Although the calculus of variations is normally a second-year option, its basics are actually easy to understand without much prior knowledge. I've tried to make this as accessible as I can.

2 Some prerequisites

While I'm assuming a comfortable knowledge of single-variable calculus, I just want to quickly mention a couple of key concepts from multivariable calculus that I'll use later on. I won't prove them here: this is just for reference.

2.1 Partial and total derivatives

The first is the partial derivative. Given a function $f(x, y, z)$, the partial derivative $\dfrac{\partial f}{\partial x}$ is defined as the ratio of the change in f to the change in x if we change x very slightly *while* keeping y and z constant. It is calculated just like a regular derivative.

The reason we use a special symbol is there is a distinction between this and the *total* derivative of a multivariate function. The total derivative of f with respect to x is written, familiarly, as $\dfrac{\mathrm{d}f}{\mathrm{d}x}$. If x, y and z are all independent then they're the same:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial x}\,.$$

However, if y and/or z have some *implicit* — that is, not explicitly written — dependency on x (for example, if they're both parameterised by the same variable), then the two are separate: changing x may result in a change in y and/or z as well. The total derivative takes this into account, while the partial derivative does not. This distinction will come into use later on.

2.2 The multivariate chain rule

The second thing I want to mention is how we extend the multivariate chain rule to multiple dimensions. If we have some function $f(x, y)$ but we parameterise x and y using some third variable t, so we write $x(t)$ and $y(t)$, then f is really just a function of t. We can then calculate its derivative as follows:

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}\,.$$

Intuitively this makes sense: the amount that f changes is made up of the change due to the change in x and the change due to the change in y. In general the partial derivative of a multivariate function with respect to a particular variable is just the sum of the partial derivatives of all inputs to that function with respect to that variable.

3 The shortest path between two points

Let's start by taking a look at one of the fundamental problems with which the calculus of variations is concerned. Say we have two points, A and B, with coordinates (x_1, y_1) and (x_2, y_2) respectively (see fig. 1). The task is to find the shortest path between A and B. This can be *any smooth curve* that passes through A and B — it could go to Saturn and back (but we're interested in the one with the shortest length).

At this point your intuition should probably be crying out in despair: *surely it's just a straight line between them!* It seems completely obvious that to get from A to B as quickly as possible, you need to go 'as the crow flies' — that is, in a straight line — but to actually prove this, as we'll soon find out, is decidedly non-trivial. Just think about how many ways you could get between the two points: how is it possible to show that a straight line is better than *absolutely everything else*?

To begin with, we need to formalise this problem algebraically. We're going to limit the possible curves to well-defined smooth functions[1] $y(x)$ between A and B. In order for $y(x)$ to pass

[1] The fact that we can do this is actually not immediately obvious, but we'll assume we can so as to make our lives easier.

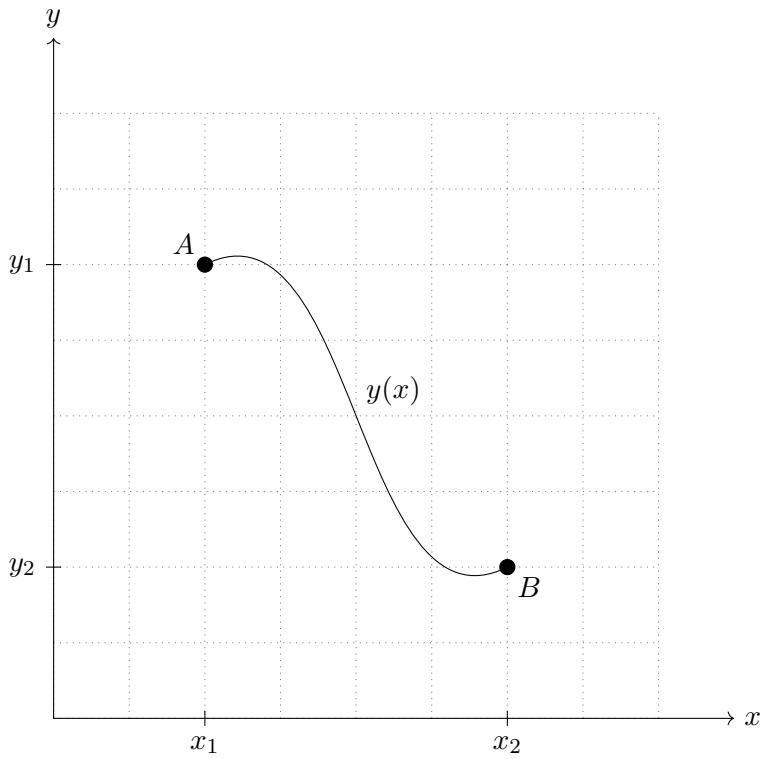


Figure 1: We want to find the shortest path between A and B.

through both points, we must have the constraints

$$y(x_1) = y_1 \quad \text{and} \quad y(x_2) = y_2.$$

Now take a small part of the curve, as shown in fig. 2, and call its length $\mathrm{d}s$. As $\mathrm{d}s$ approaches zero in length, it becomes a straight line, and so if the height of that section of the curve is $\mathrm{d}y$ and its width is $\mathrm{d}x$, then we can apply Pythagoras' theorem to say

$$(\mathrm{d}s)^2 = (\mathrm{d}x)^2 + (\mathrm{d}y)^2.$$

Taking the square root and pulling out a factor of $\mathrm{d}x$ then brings us to

$$\begin{aligned} \mathrm{d}s &= \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2} \\ &= \sqrt{1 + \left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,\mathrm{d}x \end{aligned}$$

but the derivative $\frac{\mathrm{d}y}{\mathrm{d}x}$ can be written in terms of our function $y(x)$:

$$\mathrm{d}s = \sqrt{1 + y'^2(x)}\,\mathrm{d}x.$$

Now, if we want the total length of the whole curve $y(x)$ between $x = x_1$ and $x = x_2$ we just want to add up all these infinitesimal lengths $\mathrm{d}s$, meaning we take the integral:

$$L = \int_{x_1}^{x_2} \sqrt{1 + y'^2(x)}\,\mathrm{d}x.$$

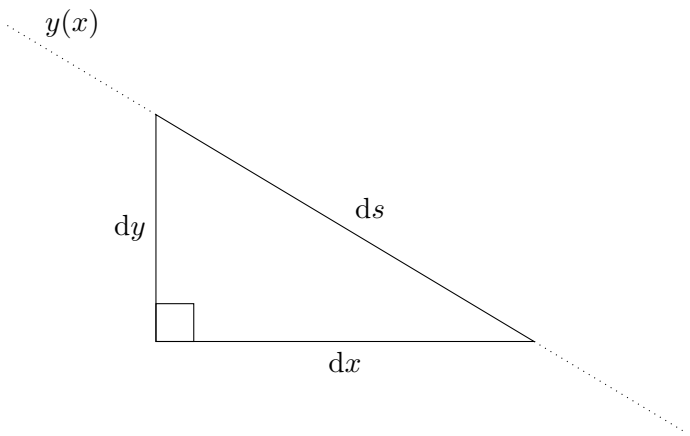


Figure 2: An infinitesimal section of the curve.

So, this total length L is what we want to minimise. Something strikes us as different, however, to minimisation problems we've come across before. Normally we just take the derivative of a function and set it to zero, but here L is dependent upon *every value of* the function $y(x)$ on the interval $[x_1, x_2]$. So L isn't a normal function: it takes a *function* as its input and it outputs a real number, and so we have to somehow differentiate with respect to all possible functions between A and B to find the function $y(x)$ that minimises L. The next section discusses just how we do that.

4 Introducing functionals and their variations

In the previous section we met a new type of function which is dependent upon *every* value of another function in a given interval. In fact, we call such a construct a *functional*.

Functionals can be thought of as 'functions of functions'. We use square brackets (as opposed to parentheses) to enclose their arguments, which are not numbers but functions themselves. For the total arc length we just derived, for example, we would write

$$L[y(x)] = \int_{x_1}^{x_2} \sqrt{1 + y'^2(x)}\,\mathrm{d}x.$$

In fact, most functionals can be expressed as a definite integral of some combination of a function, its arguments and its derivatives: let's consider a general functional $F[y(x)]$ defined as[2]

$$F[y] = \int_{x_1}^{x_2} f(x, y(x), y'(x))\,\mathrm{d}x \tag{1}$$

for some real-valued smooth function f. We keep our constraints that $y(x_1) = y_1$ and $y(x_2) = y_2$.

Our objective is to find the extrema of this functional F — that is, we want to find which function $y(x)$ either maximises or minimises $F[y]$. Just like in regular calculus, this is going to

[2]Note that writing $F[y]$ is entirely equivalent to writing $F[y(x)]$, as most often the actual number x isn't relevant; in the integral it's just the 'dummy variable'.

involve some sort of derivative, but in our case we want to know how much the real number $F[y]$ will vary as a result of a small change in shape of the function y. When this variation is zero, just like in normal calculus, we have found an extremum of F (a stationary point).

Consider our function $y(x)$ and suppose we make a tiny change to its shape by adding to it a small amount ε of some arbitrary smooth function $\eta(x)$. Note that we require the resulting function still to pass through the points (x_1, y_1) and (x_2, y_2) (for instance in the shortest path problem these are the two points our curve *must* connect) and so at $x = x_1$ and $x = x_2$ this new function $\eta(x)$ must not make any difference to the value of y — that is, we require

$$\eta(x_1) = \eta(x_2) = 0.$$

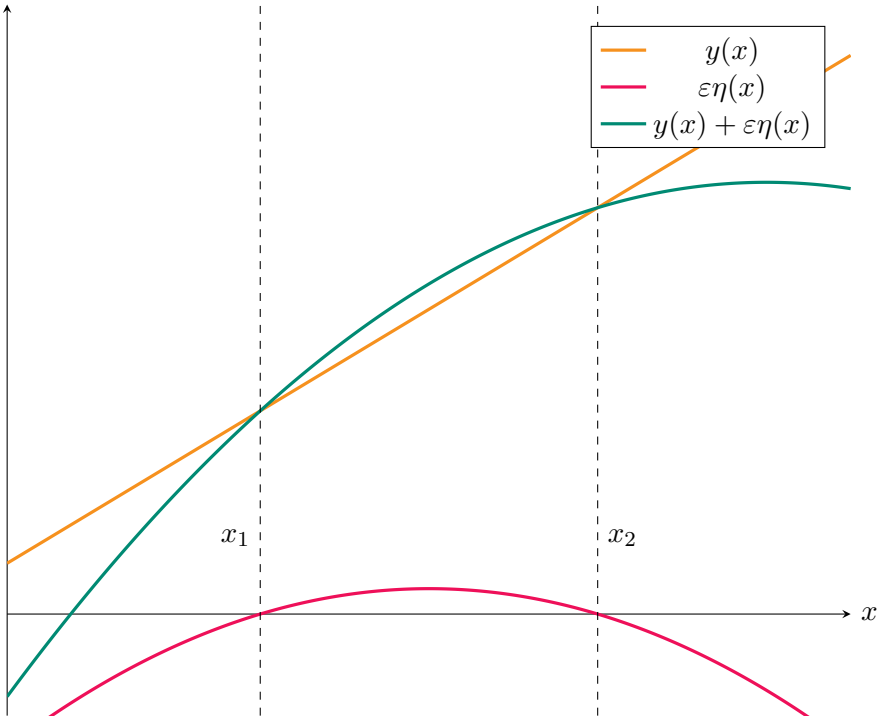


Figure 3: An example of what the functions $y(x)$, $\eta(x)$ and $y(x) + \varepsilon\eta(x)$ might look like.

Then, the difference between the value of the functional when applied to this new, slightly different function $y(x) + \varepsilon\eta(x)$, and its value when applied to $y(x)$, all divided by that amount ε, is defined as the functional's *variation* or *first variation*:

> **Definition.** The variation of a functional $F[y]$ *in the direction of* an arbitrary function η is the ratio of the change in F to the change in ε when the shape of its input function y is changed by an infinitesimal amount ε of η, and is defined as
>
> $$\delta F(y;\eta) \coloneqq \lim_{\varepsilon\to 0} \frac{F[y(x) + \varepsilon\eta(x)] - F[y(x)]}{\varepsilon} = \lim_{\varepsilon\to 0} \frac{F[y + \varepsilon\eta] - F[y]}{\varepsilon}. \tag{2}$$

Compare this with the definition of a regular derivative and there is striking similarity; what we're doing is conceptually very closely related. We're changing the input of something by an

infinitesimal amount and seeing how much the output changes as a result.[3]

We're next going to try to come up with a more useful form for the variation of a functional. Consider the functional $F[y+\varepsilon\eta]$ (as opposed to just $F[y]$ as before). By comparison[4] with eq. (1),

$$F[y+\varepsilon\eta] = \int_{x_1}^{x_2} f(x, y(x)+\varepsilon\eta(x), y'(x)+\varepsilon\eta'(x))\,\mathrm{d}x.$$

If you imagine temporarily that the shapes of $y(x)$ and $\eta(x)$ are fixed but we can freely change the value of ε, then F is just a function of ε:

$$F(\varepsilon) = \int_{x_1}^{x_2} f(x, y(x)+\varepsilon\eta(x), y'(x)+\varepsilon\eta'(x))\,\mathrm{d}x.$$

Differentiating with respect to ε (and using our definition of the derivative) gives

$$\frac{\mathrm{d}F}{\mathrm{d}\varepsilon} = \lim_{h\to 0}\frac{F(\varepsilon+h)-F(\varepsilon)}{h}$$

which is just the same as writing

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}F[y+\varepsilon\eta] = \lim_{h\to 0}\frac{F[y+(\varepsilon+h)\eta]-F[y+\varepsilon\eta]}{h}.$$

So, evaluating this derivative at $\varepsilon = 0$ gives

$$\left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}F[y+\varepsilon\eta]\right|_{\varepsilon=0} = \lim_{h\to 0}\frac{F[y+h\eta]-F[y]}{h}.$$

But, replace[5] every h with an ε and you see from eq. (2) that this is *exactly the same* as our definition of the variation of $F[y]$ in the direction of η! So, our more useful definition is:

> **Definition.** The variation of a functional $F[y]$ in the direction of an arbitrary function η is given by
>
> $$\delta F(y;\eta) = \left.\frac{\mathrm{d}}{\mathrm{d}\varepsilon}F[y+\varepsilon\eta]\right|_{\varepsilon=0}. \tag{3}$$

It is this form that will allow us to derive the all-important Euler-Lagrange equation in the next section.

5 Deriving the Euler-Lagrange equation

Now that we have a nice expression for the variation of a functional $F[y]$, we return to our original problem which is to find the function y that maximises or minimises F. Just like we

[3] For those of you who have studied multivariate calculus, this is actually really just a directional derivative, except here our direction is not a vector but a function. If you like, it's an infinite-dimensional vector.

[4] To find the third argument to f we just had to differentiate $y(x)+\varepsilon\eta(x)$ with respect to x, resulting in $y'(x)+\varepsilon\eta'(x)$.

[5] We can do this as h is just a dummy variable inside the limit.

would set the derivative of a function to zero to find its extrema, we set the variation of a functional to zero to do the same:

$$\begin{aligned} &\delta F(y;\eta) = 0 \\ \implies &\frac{\mathrm{d}}{\mathrm{d}\varepsilon} F[y+\varepsilon\eta]\bigg|_{\varepsilon=0} = 0 \qquad \text{(from eq. (3))}. \end{aligned} \tag{4}$$

In fact, in this situation our expression for the variation makes even more intuitive sense. Suppose y *is indeed* the function which minimises F; then, changing its shape in any way — that is, adding any amount ε of any other function $\eta(x)$ — will surely increase F. That is, the minimum of $F[y+\varepsilon\eta]$ is at $\varepsilon = 0$. Therefore regular calculus tells us that the derivative with respect to ε at $\varepsilon = 0$ *must* be zero — it's a stationary point. This argument is unchanged if y maximises F instead, and either way leads to eq. (4).

So now our task is to somehow use eq. (4) to solve for the function y that minimises or maximises F.

Let's use our definition of F from eq. (1). We clearly require

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int_{x_1}^{x_2} f(x, y(x) + \varepsilon\eta(x), y'(x) + \varepsilon\eta'(x))\,\mathrm{d}x \bigg|_{\varepsilon=0} = 0$$

and to make our lives easier, writing $Y(x) = y(x) + \varepsilon\eta(x)$, this becomes

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \int_{x_1}^{x_2} f(x, Y(x), Y'(x))\,\mathrm{d}x \bigg|_{\varepsilon=0} = 0$$

Using the fact that differentiation and integration of smooth functions can be done in any order[6], we can rewrite this condition as[7]

$$\int_{x_1}^{x_2} \frac{\partial}{\partial\varepsilon} f(x, Y(x), Y'(x)) \bigg|_{\varepsilon=0} \mathrm{d}x = 0. \tag{5}$$

Next, to attack this further, we'll apply the multivariate chain rule we discussed in section 2, which gives

$$\frac{\partial f}{\partial\varepsilon} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial\varepsilon} + \frac{\partial f}{\partial Y}\frac{\partial Y}{\partial\varepsilon} + \frac{\partial f}{\partial Y'}\frac{\partial Y'}{\partial\varepsilon}.$$

Of course x doesn't depend on ε at all, so $\frac{\partial x}{\partial\varepsilon} = 0$ meaning we can get rid of the first term:

$$\frac{\partial f}{\partial\varepsilon} = \frac{\partial f}{\partial Y}\frac{\partial Y}{\partial\varepsilon} + \frac{\partial f}{\partial Y'}\frac{\partial Y'}{\partial\varepsilon}.$$

Now it can be seen that

$$\frac{\partial Y}{\partial\varepsilon} = \frac{\partial}{\partial\varepsilon}\big(y(x) + \varepsilon\eta(x)\big) = \eta(x)$$

[6] This is non-trivial to prove so we'll just take it as an assumption

[7] The reason that we now use a partial derivative is that the only variable on which the value of the whole integral depends is ε and so we can take an ordinary derivative with respect to ε; but the actual function f *inside* the integral depends on x as well and so if we bring our derivative inside the integral we *must* make it a partial derivative to make clear that we're holding x constant and differentiating with respect to ε only.

and similarly

$$\frac{\partial Y'}{\partial \varepsilon} = \frac{\partial}{\partial \varepsilon}\left(y'(x) + \varepsilon\eta'(x)\right) = \eta'(x).$$

Thus, our partial derivative of f with respect to ε is simply

$$\frac{\partial f}{\partial \varepsilon} = \frac{\partial f}{\partial Y}\,\eta(x) + \frac{\partial f}{\partial Y'}\,\eta'(x),$$

making our optimality condition from eq. (5) become

$$\int_{x_1}^{x_2} \left(\frac{\partial f}{\partial Y}\,\eta(x) + \frac{\partial f}{\partial Y'}\,\eta'(x)\right)\Bigg|_{\varepsilon=0} \mathrm{d}x = 0.$$

Remembering that we defined $Y(x)$ as equal to $y(x)+\varepsilon\eta(x)$, at $\varepsilon = 0$ we must have $Y(x) = y(x)$ (and similarly $Y'(x) = y'(x)$), so we can make this replacement and split it into two integrals:

$$\int_{x_1}^{x_1} \frac{\partial f}{\partial y}\,\eta(x)\,\mathrm{d}x + \int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\,\eta'(x)\,\mathrm{d}x = 0. \tag{6}$$

Remember here that we're eventually trying to get rid of any mention of the function η as it is by definition just some arbitrary function. The right-hand integral above can be evaluated using integration by parts. With reference to the parts formula,

$$\int u'v\,\mathrm{d}x = uv - \int uv'\,\mathrm{d}x,$$

we set $u' = \eta'(x)$ so that $u = \eta(x)$, and so our other function is $v = \dfrac{\partial f}{\partial y'}$ meaning $v' = \dfrac{\mathrm{d}}{\mathrm{d}x}\left(\dfrac{\partial f}{\partial y'}\right)$. Then,

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\,\eta'(x)\,\mathrm{d}x = \left[\frac{\partial f}{\partial y'}\,\eta(x)\right]_{x_1}^{x_2} - \int_{x_1}^{x_2} \eta(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\mathrm{d}x.$$

How has this helped us? Well, you may remember that when we introduced the function η, we made sure to mention that it has to be constrained by $\eta(x_1) = \eta(x_2) = 0$ (see fig. 3). Therefore,

$$\left[\frac{\partial f}{\partial y'}\,\eta(x)\right]_{x_1}^{x_2} = 0 - 0 = 0$$

and so coming back to eq. (6), we now need

$$\int_{x_1}^{x_1} \frac{\partial f}{\partial y}\,\eta(x)\,\mathrm{d}x - \int_{x_1}^{x_2} \eta(x)\,\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\mathrm{d}x = 0.$$

Combining the integrals and factoring out $\eta(x)$ leaves us with

$$\int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)\right)\eta(x)\,\mathrm{d}x = 0.$$

We're extremely close now. Since η is arbitrary — that is, it could be *any* function as long as $\eta(x_1) = \eta(x_2) = 0$ — we can conclude that the only way this integral is *always* zero, for *any* function η, is if the other function in the integrand is always zero, that is,

$$\frac{\partial f}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{7}$$

for every value of x on the interval $[x_1, x_2]$. This reasoning is slightly opaque and in fact the argument is known appropriately as 'the fundamental lemma of the calculus of variations'. If you're interested, here's a short formal proof:

Lemma (The Fundamental Lemma of the Calculus of Variations)**.** *Let $\phi(x)$ be a continuous function on the interval $[x_1, x_2]$. Suppose that for all continuous functions $\eta(x)$ with $\eta(x_1) = \eta(x_2) = 0$ we have*

$$\int_{x_1}^{x_2} \phi(x)\eta(x)\,\mathrm{d}x = 0. \tag{8}$$

Then $\phi(x) = 0$ for all $x \in [x_1, x_2]$.

Proof. Suppose that there is some value $\alpha \in [x_1, x_2]$ for which $\phi(\alpha) \neq 0$. Without loss of generality we may say that here $\phi(\alpha) > 0$. Then as ϕ is continuous, there is some interval $[\alpha_1, \alpha_2]$, where $\alpha_1 \leqslant \alpha \leqslant \alpha_2$, on which ϕ is always positive. Suppose η is positive on this interval and zero elsewhere; then

$$\int_{x_1}^{x_2} \phi(x)\eta(x)\,\mathrm{d}x = \int_{\alpha_1}^{\alpha_2} \phi(x)\eta(x)\,\mathrm{d}x > 0$$

which contradicts eq. (8). Hence there cannot exist any $\alpha \in [x_1, x_2]$ for which $\phi(\alpha) \neq 0$. □

Therefore, in eq. (7) we end up at a final, beautiful differential equation which is called the Euler-Lagrange equation. It guarantees the following:

Theorem (The Euler-Lagrange Equation)**.** *Any functional*

$$F[y] = \int_{x_1}^{x_2} f(x, y(x), y'(x))\,\mathrm{d}x$$

will take an extremal value if and only if the function y satisfies the differential equation

$$\frac{\partial f}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) \tag{9}$$

for every $x \in [x_1, x_2]$.

This is exactly what we were hoping for: now given any functional we want to maximise or minimise, we need only solve this differential equation for the function y at which such a stationary point will occur.

6 Returning to the shortest path

In this section we return to the shortest path problem introduced in section 3; we now have all the tools to solve it. As we found out, the total length of a curve $y(x)$ between two points

(x_1, y_1) and (x_2, y_2) is given by

$$L[y] = \int_{x_1}^{x_2} \sqrt{1 + y'^2(x)}\,\mathrm{d}x,$$

and we want to use the Euler-Lagrange equation derived in section 5 to find the curve y which minimises this length. In this case, the function f referred to by the Euler-Lagrange equation is

$$f(y'(x)) = \sqrt{1 + y'^2(x)}.$$

The function f depends only on y', not y itself, and so $\frac{\partial f}{\partial y} = 0$. Differentiating with respect to y',

$$\begin{aligned}\frac{\partial f}{\partial y'} &= \frac{1}{2}(1 + y'^2)^{-\frac{1}{2}} \cdot 2y' \\ &= \frac{y'}{\sqrt{1 + y'^2}},\end{aligned}$$

and so the Euler-Lagrange equation, eq. (9), gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y'}{\sqrt{1 + y'^2}}\right).$$

The derivative with respect to x is always zero — which is the same as saying there is some constant k such that for all x,

$$\frac{y'}{\sqrt{1 + y'^2}} = k.$$

Rearranging this leads us to

$$\begin{aligned} y' &= k\sqrt{1 + y'^2} \\ \implies y'^2 &= k^2(1 + y'^2) \\ \implies y'^2 &= \frac{k^2}{1 - k^2} \\ \implies y' &= \frac{k}{\sqrt{1 - k^2}}.\end{aligned}$$

Therefore it can be seen that $y'(x)$, the derivative of the curve with respect to x, is a constant. In other words, that curve is a straight line.

Of course, it seems quite superfluous to have gone through so much abstract mathematics to get to this point, but there is real value in proving this rigorously. While our intuition is correct in this case, a lot of the time it may not be and we *must* resort to mathematics for the answer. For instance, we know that if these two points are on a flat plane, the shortest path between them is a straight line — but what if they're on a sphere, or a pseudosphere (a surface with outward curvature)? What route should an aeroplane take from London to New York to get there in the shortest time? These are questions which we need the calculus of variations to answer.

7 The brachistochrone problem

Brachistochrone literally means 'least time' in Greek. The problem is an historic one, and perhaps the most perfect example of a useful application of the calculus of variations. Indeed, it was Johann Bernoulli's raising of the problem in 1696 that eventually led to Euler's 1756 publication *Elementa Calculi Variationum* which kick-started the field.

We start with two points A and B in the same vertical plane. Imagine releasing a ball or other massive particle at point A. We want to find the smooth curve from A to B down which a particle will slide, without friction, in the shortest time possible. That is, we want to minimise the total time taken between the particle being released from rest at A and arriving at B.

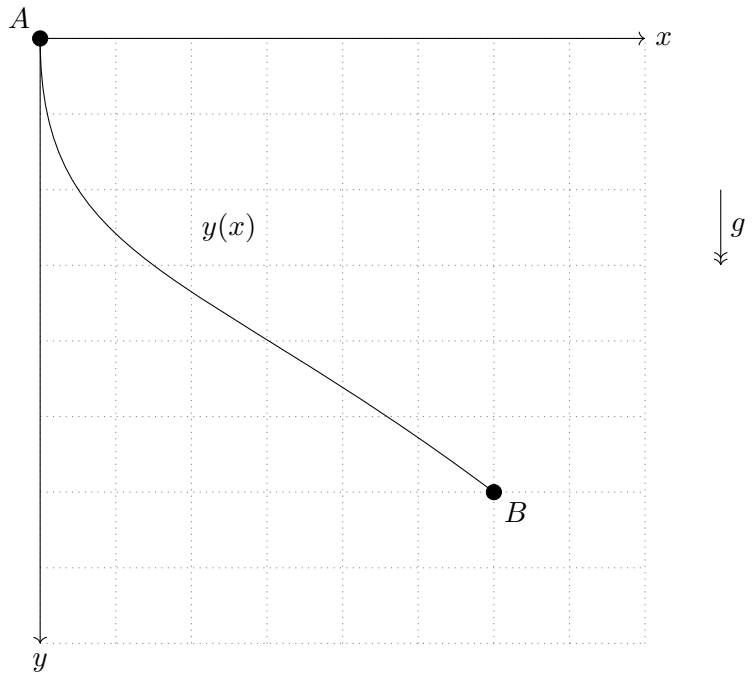


Figure 4: For this problem we want to find the 'brachistochrone' curve from A to B down which an object acting under gravity and without friction will slide in the least time.

Unlike the shortest path problem discussed in sections 3 and 6, the solution is not immediately obvious. Sliding down a straight line, it turns out, is surprisingly slow, and there are all sorts of other types of curves we could draw from A to B. We resort then to our newly acquired skills in the calculus of variations.[8]

[8]There are actually several other ways to solve this problem, perhaps the nicest using an analogy with optics. Fermat's principle states that light will always take the path of least time between two points, and so by shining a ray of light at an angle through a medium of constantly changing optical density (such that the light experiences a downward acceleration of g), we can use Snell's law of refraction to find what path the light will take. This solution was originally proposed by Johann Bernoulli in 1697, long before Euler's paper on the calculus of variations was published.

7.1 Setting up the integral

Let A be at the origin and let B have coordinates (x_0, y_0). We set our axes such that *downwards* is the positive y-direction; this will save us some hassle later on. Our task is to minimise the total time taken for the particle to slide from A to B, so we first want an expression for the total time in terms of things we know.

The particle is released from rest at A so has zero kinetic energy at this point. Let's also set the gravitational potential energy to be zero here, so that the particle has a total energy of zero. Then at some arbitrary height y below A, if the particle has speed v then it has kinetic energy $\frac{1}{2}mv^2$ and potential energy $-mgy$ (as its elevation has decreased) — but by energy conservation the total energy must still be zero, so

$$\begin{aligned} \frac{1}{2}mv^2 - mgy &= 0 \\ \implies \frac{1}{2}v^2 &= gy \\ \implies v &= \sqrt{2gy}. \end{aligned}$$

This is the *speed* of the particle in the direction of the curve. Since we don't know the length of the curve, we want to find either the vertical or horizontal velocity instead. The horizontal velocity of the particle at this point is given by

$$v_x = v \sin\theta = \sqrt{2gy}\sin\theta$$

where θ is the angle the curve makes to the vertical.

Now take a small length $\mathrm{d}s$ of the curve, as in fig. 5.

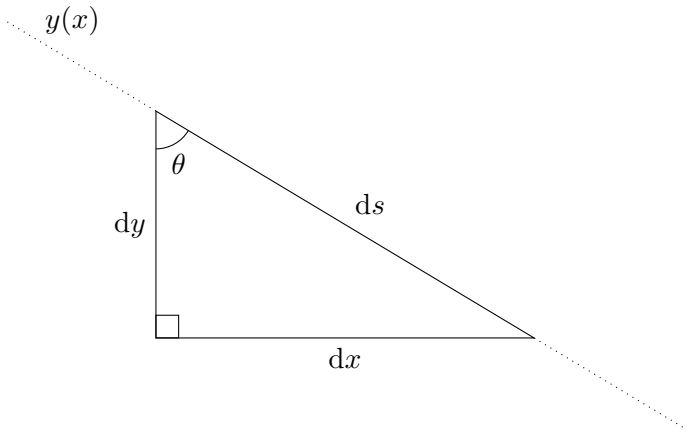


Figure 5: An infinitesimal section of the curve.

It's clear from this diagram that $\sin\theta = \frac{\mathrm{d}x}{\mathrm{d}s}$ and so we can simplify our expression for the horizontal velocity to

$$v_x = \sqrt{2gy}\,\frac{\mathrm{d}x}{\mathrm{d}s}\,.$$

but looking again at fig. 5, Pythagoras' theorem tells us that $\mathrm{d}s = \sqrt{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}$ and so our expression becomes

$$\begin{aligned} v_x &= \sqrt{\frac{2gy(\mathrm{d}x)^2}{(\mathrm{d}x)^2 + (\mathrm{d}y)^2}} \\ &= \sqrt{\frac{2gy(x)}{1 + \left(\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)^2}} \\ &= \sqrt{\frac{2gy}{1 + y'^2(x)}}. \end{aligned}$$

We now have a nice expression for the horizontal velocity at every point, and we know that the total horizontal distance the particle needs to travel is x_0, so by analogy with the linear formula

$$\text{time} = \frac{\text{displacement}}{\text{velocity}},$$

to find the total time taken T we take the definite integral:[9]

$$\begin{aligned} T &= \int_0^{x_0} \frac{\mathrm{d}x}{v_x} \\ \implies T[y] &= \int_0^{x_0} \sqrt{\frac{1 + y'^2}{2gy}}\, \mathrm{d}x. \end{aligned}$$

This is the functional we want to minimise over all functions $y(x)$. We know everything we need to know to do this now!

7.2 Solving using the Euler-Lagrange equation

The Euler-Lagrange differential equation we derived in section 5 tells us that for the function $y(x)$ to minimise T we must have

$$\frac{\partial f}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) \tag{10}$$

where in this case the function f referred to is the integrand,

$$f(y, y') = \sqrt{\frac{1 + y'^2}{2gy}}.$$

We need therefore to work out $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial y'}$. The former is

$$\begin{aligned} \frac{\partial f}{\partial y} &= \frac{1}{2}\sqrt{\frac{2gy}{1 + y'^2}}\left(-\frac{1 + y'^2}{2gy^2}\right) \\ &= -\frac{\sqrt{2}}{4}\sqrt{\frac{1 + y'^2}{gy^3}}, \end{aligned} \tag{11}$$

[9] Another way to see that we have to take this integral is by looking at how much time the particle takes to traverse the small distance $\mathrm{d}x$ and adding up all these times.

and the latter is

$$\begin{aligned}\frac{\partial f}{\partial y'} &= \frac{1}{2}\sqrt{\frac{2gy}{1+y'^2}}\left(\frac{y'}{gy}\right)\\ &= \frac{\sqrt{2}}{2}\left(\frac{y'}{\sqrt{gy(1+y'^2)}}\right).\end{aligned} \tag{12}$$

Looking again at eq. (10), we need to differentiate this with respect to x. Of course, we have simply

$$\frac{\mathrm{d}y}{\mathrm{d}x} = y' \qquad \text{and} \qquad \frac{\mathrm{d}y'}{\mathrm{d}x} = y'',$$

so using the quotient rule with reference to eq. (12) we come to

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) = \frac{\sqrt{2}}{2}\left[\frac{y''\sqrt{gy(1+y'^2)} - y'\cdot\frac{1}{2\sqrt{gy(1+y'^2)}}\cdot\left(gy'(1+y'^2)+gy(2y'y'')\right)}{gy(1+y'^2)}\right].$$

This is particularly nasty, but we can simplify it down a bit:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) = \frac{\sqrt{2}}{2}\left[\frac{y''}{\sqrt{gy(1+y'^2)}} - \frac{y'^2}{2y^{\frac{3}{2}}\sqrt{g(1+y'^2)}} - \frac{y'^2y''}{\sqrt{gy}(1+y'^2)^{\frac{3}{2}}}\right].$$

So, using this and eq. (11), the Euler-Lagrange equation gives

$$\begin{aligned}&-\frac{\sqrt{2}}{4}\sqrt{\frac{1+y'^2}{gy^3}} = \frac{\sqrt{2}}{2}\left[\frac{y''}{\sqrt{gy(1+y'^2)}} - \frac{y'^2}{2y^{\frac{3}{2}}\sqrt{g(1+y'^2)}} - \frac{y'^2y''}{\sqrt{gy}(1+y'^2)^{\frac{3}{2}}}\right]\\ \implies &-\frac{1}{2}\sqrt{\frac{1+y'^2}{gy^3}} = \frac{y''}{\sqrt{gy(1+y'^2)}} - \frac{y'^2}{2y\sqrt{gy(1+y'^2)}} - \frac{y'^2y''}{(1+y'^2)\sqrt{gy(1+y'^2)}}.\end{aligned}$$

Multiplying through by $\sqrt{gy(1+y'^2)}$, we get

$$-\frac{1+y'^2}{2y} = y'' - \frac{y'^2}{2y} - \frac{y'^2y''}{1+y'^2}$$

which simplifies down to

$$\begin{aligned}-\frac{1}{2y} &= y'' - \frac{y'^2y''}{1+y'^2}\\ &= \frac{y''(1+y'^2) - y''y'^2}{1+y'^2}\\ &= \frac{y''}{1+y'^2}\end{aligned}$$

or, rearranging slightly,

$$y'^2 + 2yy'' + 1 = 0.$$

This is a second-order differential equation that we can solve to find the curve $y(x)$ which minimises the total time! Multiplying by y' as our integrating factor, we have

$$y'^3 + 2yy'y'' + y' = 0$$

and after staring at the left hand side for a short while while thinking about the product rule, we see that this is the same as saying

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(yy'^2 + y\right) = 0.$$

So,

$$yy'^2 + y = a$$

for some constant a. Rearranging slightly and re-writing y' as $\frac{\mathrm{d}y}{\mathrm{d}x}$, this becomes

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 = \frac{a-y}{y}. \tag{13}$$

We've reduced the problem to a first-order differential equation! Of course, we could slog through and find the general solution by separating the variables, and then use the constraint that the curve must pass through points A and B to find the specific solution. But, when Bernoulli arrived at this equation from his own method, he instantly recognised it as the equation of a *cycloid*.

7.3 Showing that the optimal curve is a cycloid

A cycloid is what we call the path that a point on the circumference of a circle traces out as the circle rolls along a flat surface. A standard such curve is shown in fig. 6.

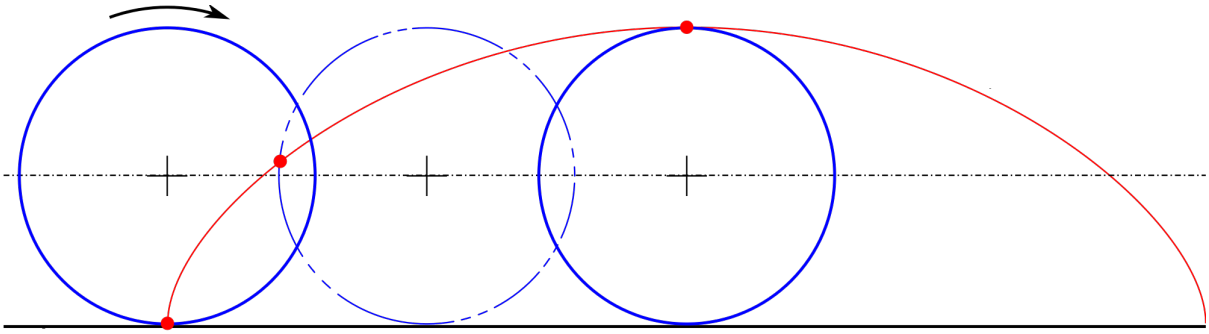


Figure 6: A cycloid curve traced out by a circle rolling along a horizontal surface. (Image source: Wikimedia Foundation.)

In the case of the brachistochrone, of course, the imaginary 'circle' that we roll to get our cycloid path will be rolling along the *ceiling* — the x-axis. It'll look something like fig. 7.

Let's see why such a cycloid gives the same differential equation as eq. (13). We want to find the square of the derivative of such a cycloid, so that we can directly compare the two. In fact, we'll just use Euclidean geometry for this part of the proof.

Point A is, as before, at the origin, as shown in fig. 8. Let our circle of radius r have centre C and be tangential to the x-axis at D. The point on the circumference of the circle that is drawing our cycloid curve is E and has coordinates (x, y). (Remember that y gets larger

Figure 7: The approximate shape of the cycloid path between A and B.

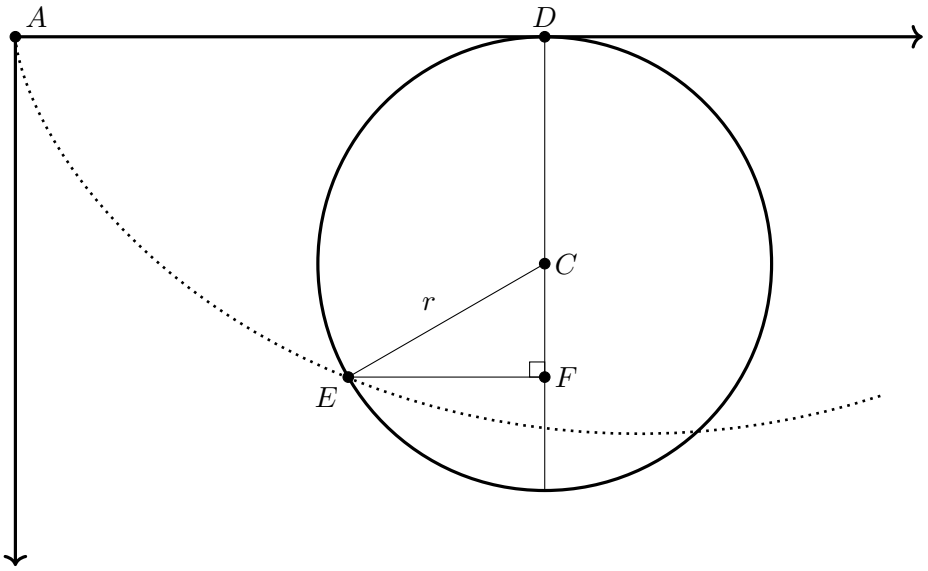


Figure 8: A snapshot of the circle in motion drawing our brachistochrone cycloid.

downwards.) Now drop a perpendicular from E to the vertical diameter of the circle and call this new point F.

The circle started with point E at the origin and rolled along the x-axis, so the total distance rolled is clearly

$$\overline{AD} = \overline{DE},$$

where $\overline{DE}$ means the arc length between D and E.

Therefore, the x-coordinate of point E is just

$$x = \overline{AD} - \overline{EF} = \overline{DE} - \overline{EF}, \tag{14}$$

and the y-coordinate is

$$y = \overline{DF}. \tag{15}$$

So, let's see what happens when the circle rotates very slightly and E is advanced to a new point G.

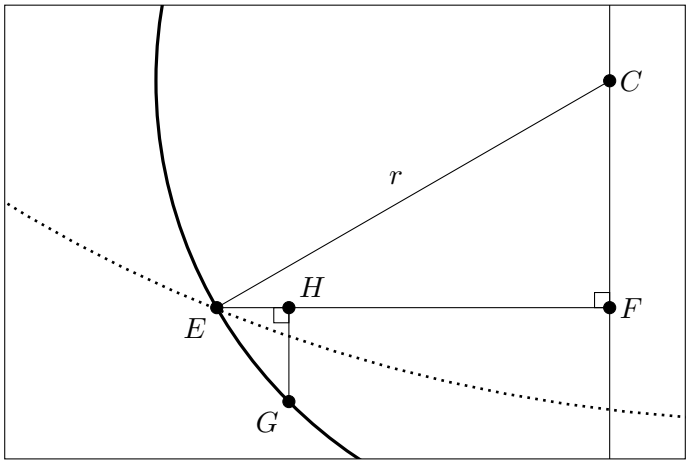


Figure 9: Point E is advanced slightly to G.

Using point H marked in fig. 9, the change in E's y-coordinate is

$$\mathrm{d}y = \overline{GH}$$

and the change to its x-coordinate is

$$\mathrm{d}x = \overline{EG} - (-\overline{EH})$$

by analogy with eqs. (14) and (15). So, we can say that the derivative of the cycloid path E makes is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \lim_{G\to E} \frac{\overline{GH}}{\overline{EG} + \overline{EH}}.$$

However, as G approaches E, the arc EG becomes a straight line and the triangle $\triangle EGH$ approaches similarity with $\triangle ECF$. So, our derivative becomes

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\overline{EF}}{\overline{CE} + \overline{CF}}.$$

We wanted to find the *square* of the derivative, so squaring this gives

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 = \frac{\overline{EF}^2}{(\overline{CE} + \overline{CF})^2}.$$

By Pythagoras, $\overline{EF}^2 = \overline{CE}^2 - \overline{CF}^2$ and so

$$\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 = \frac{\overline{CE}^2 - \overline{CF}^2}{(\overline{CE} + \overline{CF})^2}.$$

Looking at this, though, the distance $\overline{CE}$ is just the radius r of the circle, and similarly

$$\overline{CF} = y - r.$$

So, finally,

$$\begin{aligned}\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2 &= \frac{r^2-(y-r)^2}{(r+y-r)^2} \\ &= \frac{r^2-y^2+2ry-r^2}{y^2} \\ &= \frac{2ry-y^2}{y^2} \\ &= \frac{2r-y}{y}.\end{aligned}$$

Compare this to the differential equation for the brachistochrone we derived in eq. (13); if our arbitrary constant is $a = 2r$, they are exactly the same! Thus, the least-time curve between two points is a cycloid.

How do we find the *exact* cycloid that will get us from A to B in the shortest time? Just draw a cycloid starting from point A (the origin) as in fig. 7 and then scale it until it intersects point B. There's only one that will!

This is a wonderful proof and its consequences are rather unintuitive. As can be seen, depending on the positions of A and B the curve may actually slope upwards at the end towards B. Thinking about why this may occur helps: sometimes building greater speed by accelerating earlier will outweigh the slowing effect of the upwards climb at the end.

8 Proving the fundamental theorem of calculus

Another corollary hidden away in the powers of the Euler-Lagrange equation is a quick proof of the fundamental theorem of calculus for continuous functions! Suppose we want find the continuous function y through points (x_1, y_1) and (x_2, y_2) that minimises the functional

$$I = \int_{x_1}^{x_2} y'(x)\,\mathrm{d}x.$$

Applying the Euler-Lagrange equation, where now the function f it talks about is y', we get

$$\begin{aligned}&\frac{\partial f}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) \\ \iff &\frac{\mathrm{d}y'}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial y'}{\partial y'}\right) \\ \iff &0 = \frac{\mathrm{d}}{\mathrm{d}x}(1) \\ \iff &0 = 0.\end{aligned}$$

This tells us that y minimises $\int_{x_1}^{x_2} y'(x)\,\mathrm{d}x$ if and only if $0 = 0$. Since the latter is always true, *all* functions y minimise that functional. That is, the value of the definite integral $\int_{x_1}^{x_2} y'(x)\,\mathrm{d}x$ is the same for all functions y that pass through (x_1, y_1) and (x_2, y_2).

Suppose $y(x)$ is a straight line of equation

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1).$$

So, $y'(x)$ is

$$y' = \frac{y_2 - y_1}{x_2 - x_1},$$

a constant. Thus in this case, the definite integral of y' between x_1 and x_2 is just the area of a rectangle of width $x_2 - x_1$ and height y'. That is,

$$I = (x_2 - x_1)y' = (x_2 - x_1)\left(\frac{y_2 - y_1}{x_2 - x_1}\right) = y_2 - y_1.$$

But we just showed that the value of this integral is the same for *all* functions y, and so we have the general identity

$$\int_{x_1}^{x_2} y'(x)\,\mathrm{d}x = y_2 - y_1.$$

By definition $y_1 = y(x_1)$ and $y_2 = y(x_2)$, so we have

$$\int_{x_1}^{x_2} y'(x)\,\mathrm{d}x = y(x_2) - y(x_1).$$

Let $g(x) = y'(x)$ and $G(x) = y(x)$ be an antiderivative of $g(x)$. Then we have the first fundamental theorem of calculus in its usual form:

$$\int_{x_1}^{x_2} g(x)\,\mathrm{d}x = G(x_2) - G(x_1).$$

The second fundamental theorem of calculus, or the Newton-Leibniz axiom, is really just a generalisation of this to non-continuous functions.

9 The isoperimetric problem

Here's another classic problem the calculus of variations can be used to solve. We want to find the plane figure of a fixed perimeter ('isoperimetric') that encloses the largest possible area. The Greeks correctly convinced themselves that this was a circle, but what follows is a rigorous proof.

Most textbooks begin by taking a portion of the curve between $(0, 0)$ and $(1, 1)$ and finding the Cartesian equation of the curve between the two points that maximises the area, given a fixed arc length.

Instead, I'm going to use polar coordinates as I think this way the result will be much nicer. Let our shape be given by the polar function $r = f(\theta)$. (Here, r is the distance from the origin and θ is the angle made anticlockwise from the positive x-axis.)

Let's start by finding an expression for the area of the shape. Take a tiny wedge from the shape as in fig. 11.

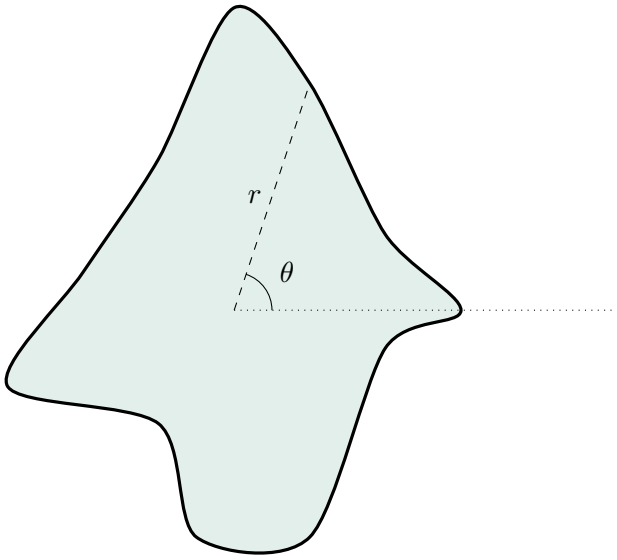


Figure 10: A generalised, blobby, polar curve.

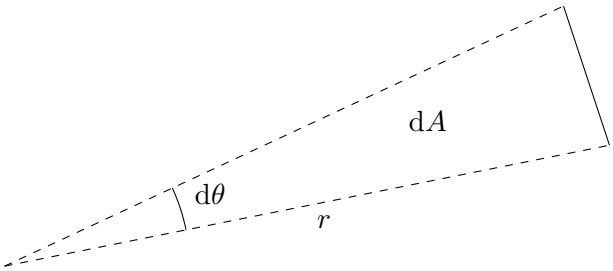


Figure 11: A small portion of the area enclosed by a polar curve.

We can actually approximate the area of this wedge by imagining it's a sector of a circle. Then, it has area $\frac{1}{2}r^2\,\mathrm{d}\theta$. Summing over all of these wedges therefore, the total area enclosed by the shape is

$$A = \int_0^{2\pi} \frac{1}{2} r^2 \,\mathrm{d}\theta = \int_0^{2\pi} \frac{1}{2} f^2(\theta)\,\mathrm{d}\theta.$$

Looking at the same wedge, and still approximating it as a sector of a circle, the circular arc has length $r\,\mathrm{d}\theta$ and so the perimeter of the entire shape is

$$P = \int_0^{2\pi} r\,\mathrm{d}\theta = \int_0^{2\pi} f(\theta)\,\mathrm{d}\theta.$$

Let's fix this perimeter as l. We see that this is a different kind of problem to previously: we want to minimise the functional $A[f]$ *subject to the constraint* that $P[f] = l$.

Here we resort to using a Lagrange multipliers, a concept from multivariate calculus. I'll leave it unexplained for now, but normally if we want to optimise some multivariate function $g(\vec{v})$ with the constraint that some other function $h(\vec{v})$ of the same variables is equal to a constant k,

we first optimise the function $g(\vec{v}) - \lambda(h(\vec{v}))$ where λ is an arbitrary constant, and then check which solutions match our initial constraint $h(\vec{v}) = k$.[10]

It turns we need to do exactly the same thing when we're dealing with functionals instead. (Imagine the number of dimensions of the input vector $\vec{(v)}$ just tending towards infinity.)

So, we want to minimise

$$\int_0^{2\pi} \left(\frac{1}{2} f^2(\theta) - \lambda f(\theta) \right) \mathrm{d}\theta.$$

Applying the Euler-Lagrange equation (and taking care since our function here is also called f),

$$\begin{aligned} \frac{\partial}{\partial f} \left[\frac{1}{2} f^2(\theta) - \lambda f(\theta) \right] &= \frac{\mathrm{d}}{\mathrm{d}\theta} \left(\frac{\partial}{\partial f'} \left[\frac{1}{2} f^2(\theta) - \lambda f(\theta) \right] \right) \\ \implies f(\theta) - \lambda &= \frac{\mathrm{d}}{\mathrm{d}\theta}(0) \\ \implies f(\theta) &= \lambda. \end{aligned}$$

That is, the radius is some constant λ — a circle! Our constraint was that the perimeter is l and so the radius is just

$$r = \lambda = \frac{l}{2\pi}.$$

Admittedly that took a little bit of hand-waving when talking about constrained optimisation, but I hope to explain what's going on in more detail in a future document. For now, I included it because I think it's just another of many wonderful examples of things the Euler-Lagrange equation can do.

10 A specialisation: the Beltrami identity

Just before I round off, I want to mention the fact that although we derived the Euler-Lagrange equation to find extremals of functionals of the form

$$F[y] = \int_{x_1}^{x_2} f(x, y(x), y'(x)) \, \mathrm{d}x,$$

so far none of our functions f within the integral have explicitly depended upon the actual parameter x at all. It turns out that in many physical problems this is true: we actually only care about the *value* of the function $y(x)$ on the interval (x_1, x_2), not the input to the function.

For such situations, where $\dfrac{\partial f}{\partial x} = 0$, we can find a simplified and less general form of the Euler-Lagrange equation.

[10] This is to do with tangency of the two functions' contour lines — it's very interesting and I hope to do a follow-up explaining this and some of the consequences in Lagrangian and Hamiltonian mechanics.

We start by looking at the total derivative of f with respect to x. The multivariate chain rule gives

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}x} + \frac{\partial f}{\partial y'}\frac{\mathrm{d}y'}{\mathrm{d}x} \\ &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y''\end{aligned}$$

but we're dealing with the situation when $\dfrac{\partial f}{\partial x} = 0$, so

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial y'}y''.$$

Rearranging this leads to

$$\frac{\partial f}{\partial y}y' = \frac{\mathrm{d}f}{\mathrm{d}x} - \frac{\partial f}{\partial y'}y''. \tag{16}$$

Now, the Euler-Lagrange equation says that if y minimises the value of the integral, then

$$\frac{\partial f}{\partial y} = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right)$$

and multiplying this by y' gives

$$\frac{\partial f}{\partial y}y' = y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right). \tag{17}$$

So, we can set the right hand sides of eqs. (16) and (17) equal! Doing so results in

$$\begin{aligned}\frac{\mathrm{d}f}{\mathrm{d}x} - \frac{\partial f}{\partial y'}y'' &= y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right) \\ \implies \frac{\mathrm{d}f}{\mathrm{d}x} &= y''\frac{\partial f}{\partial y'} + y'\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\partial f}{\partial y'}\right).\end{aligned}$$

Looking at the right hand side, though, we can apply the product rule in reverse to say

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}}{\mathrm{d}x}\left(y'\frac{\partial f}{\partial y'}\right).$$

Finally, integrating both sides with respect to x, we come to

$$\begin{aligned}f &= y'\frac{\partial f}{\partial y'} + c \\ \implies f - y'\frac{\partial f}{\partial y'} &= c\end{aligned}$$

for some constant c.

This is Beltrami's identity, and it makes finding extrema of functionals with no explicit x-dependence *much* easier than with the raw Euler-Lagrange equation: computation of only one derivative is involved and the result will generally immediately be a just first-order differential equation. In fact, we could have used this instead in all of the examples given in this paper.

11 Conclusion and questions for thought

We've come a long way and are now armed with the tools to solve all sorts of problems that would have seemed impenetrable before. The true calculus of variations in this paper was all in deriving the Euler-Lagrange equation; that's where we discussed the concepts of functionals and their variations, the concepts that are really at the heart of this field. This is the reason that the Euler-Lagrange equation is so beautiful: it empowers us to solve endless numbers of problems that should require a deep conceptual understanding of functionals, all using just a single differential equation.

I've gone through the use of the calculus of variations in finding the shortest path between two points, and in finding the path of *least time* between two points, as well as in proving the fundamental theorem of calculus and solving the Greeks' isoperimetric problem. Those are just a few of the surprisingly many areas that I've realised the field touches though, and here I include some questions for the reader to ponder. When I have some time, I might answer some of them too.

11.1 Future questions to consider

- We found the shortest path between two points on a plane is a straight line. What if the two points are on a sphere, or a pseudosphere? Investigate the effect of non-Euclidean geometry on the shortest path problem. This is useful, for example, for choosing the route that an airplane should fly between two cities on Earth.
- Even more abstractly, could we use the same technique to find the shortest path between two points on some general smooth surface? What applications could this have in real life?
- We derived the Euler-Lagrange equation for functionals of single-variable functions. Generalise the result to multivariate functions. Does one equation still apply, or do several need to be satisfied simultaneously?
- Similarly, the Euler-Lagrange equation applies to functionals that are integrals of a function of the form $f(x, y(x), y'(x))$. What if the function has dependence on higher derivatives of y? Derive a similar result for functions of the form $f(x, y, y'', y''', ...)$.
- The brachistochrone problem that we solved is accompanied by a closely related problem: is there a curve on which any frictionless particle that is released anywhere from rest will reach the bottom (the minimum) of the curve in the *same* amount of time? Such a curve is called a tautochrone. Is it also a cycloid?
- In solving the isoperimetric problem, we found the 2-dimensional shape that enclosed the largest *area*. What if we want the solid that encloses the largest volume? Solve the problem for n dimensions.
- A similar problem to the isoperimetric one is to find the volume of revolution between two points that has the minimum surface area. How is it related to the isoperimetric problem? Can this be generalised to higher dimensions as well?

- Look further into the use of Lagrange multipliers in both multivariate calculus and the calculus of variations. What is the intuition behind their effect?
- Finally, investigate the application of the calculus of variations to the principle of least action. Can you use the techniques discussed here to prove Newton's second law using just the aforementioned principle? Look into the Lagrangian and Hamiltonian reformulations of classical mechanics — they are alternatives to Newtonian mechanics based heavily on the Euler-Lagrange equation.

12 References

While I tried to derive as much of this as I could by myself (after the initial explanation of the Euler-Lagrange equation), I of course heavily referred to existing resources for inspiration and help when I got stuck. The main ones I used are listed below.

[1] Jose Figueroa-O'Farrill. *Brief Notes on the Calculus of Variations*. University of Edinburgh. URL: http://www.maths.ed.ac.uk/~jmf/Teaching/Lectures/CoV.pdf.

[2] Paul Kunkel. *The Brachistochrone*. Whistler Alley Mathematics. URL: http://whistleralley.com/brachistochrone/brachistochrone.htm.

[3] Charles Byrne. *Notes on The Calculus of Variations*. University of Massachusetts at Lowell. 2009. URL: https://pdfs.semanticscholar.org/a26c/0342b54e8456930bf4d320280c1a08f732a5.pdf.

[4] Eric W. Weisstein. *Beltrami Identity*. Wolfram MathWorld. URL: http://mathworld.wolfram.com/BeltramiIdentity.html.

[5] Markus Grasmair. *Basics of Calculus of Variations*. Norwegian University of Science and Technology. URL: https://wiki.math.ntnu.no/_media/tma4180/2015v/calcvar.pdf.

[6] Yutaka Nishiyama. *The Brachistochrone Curve: The Problem of Quickest Descent*. Osaka University of Economics. URL: http://www.osaka-ue.ac.jp/zemi/nishiyama/math2010/cycloid.pdf.

Chapter 2

Simulating the evolution of the velocity distribution in an ideal gas

During a lesson in December 2016, Dr Cheung suggested the idea of simulating the collisions of gas particles and seeing what happens to their speeds. I had a go at this over Christmas and this was the result.

SIMULATING THE EVOLUTION OF THE VELOCITY DISTRIBUTION IN AN IDEAL GAS

Damon Falck

January 8, 2017

1 INTRODUCTION

In an ideal gas, particles collide elastically with one another and the walls of the container, exchanging kinetic energy and momentum but not interacting in any other way. Any attractions or forces between the molecules (such as van der Waals forces) are ignored, and indeed can change the behaviour of the particles dramatically.

Because of the constant collisions between particles, the probability distribution of particle speeds varies over time, tending towards and reaching equilibrium at the Maxwell-Boltzmann speed distribution,

$$f(v) = 4\pi \left(\frac{m}{2\pi k_B T} \right)^{\frac{3}{2}} v^2 \mathrm{e}^{-\frac{mv^2}{2k_B T}} \qquad (1)$$

where f is the probability of a given particle having speed v.

While the speeds of the eventually stabilised system can be described by this probability density function, it is the object of this simulation to numerically model and examine the evolution over time of this distribution of velocities.

To do so requires an accurate simulation of the particle interactions in an ideal gas.

2 IMPLEMENTATION

The simulation was written in Matlab R2015b. There are three main parts to the code:

1. Setup of the initial conditions, including modelling the container and particle properties.
2. Detecting collisions with the walls of the container and changing the particle velocities accordingly.
3. Detecting collisions between particles and changing the particle velocities accordingly.

After every iteration of parts 2 and 3 the position of each particle is updated, and the code loops. Finally when the specified number of iterations has been completed, the program halts and plots can be made from the data.

2.1 INITIAL CONDITIONS

For simplicity, the simulation runs within a cube of volume specified by the ideal gas equation. Every particle is given the same initial speed and a random direction.

2.1.1 PARAMETERS

The following physical parameters are specified by the user:

- Pressure P
- Thermodynamic temperature T
- Particle mass m
- Number of particles N

In addition, the following parameters controlling the nature of the simulation are also specified:

- Particle radius R
- Observation time t
- Number of iterations n_{iter}

Note that this is only a way of setting the approximate initial conditions desired; once the simulation runs pressure and temperature are not fixed and particle mass becomes irrelevant.

2.1.2 PARTICLE VELOCITIES

At the start of the simulation, we must determine the speed of every particle.

Firstly we generate a random unit vector to describe the direction of motion of each particle. To ensure that the vectors are uniformly distributed on the unit sphere, we use spherical coordinates. Two random angles θ and ϕ from the uniform distribution on $[-\pi, \pi]$ are generated and the unit vector is defined as

$$\begin{pmatrix} \cos\theta \sin\phi \\ \sin\theta \sin\phi \\ \cos\phi \end{pmatrix}, \tag{2}$$

a standard conversion from spherical to Cartesian coordinates. Figure 1 shows 1500 such unit vectors.

Rather than directly specifying the speed of the particles (all particles begin with equal speed), we use the kinetic energy equation

$$\frac{1}{2} m \langle v^2 \rangle = \frac{3}{2} k_B T \tag{3}$$

to find the speed, where k_B is Boltzmann's constant and T and m are specified. Hence, the initial velocity of each particle is

$$\overrightarrow{v_{\text{init}}} \coloneqq \sqrt{\frac{3k_B T}{m}} \begin{pmatrix} \cos\theta \sin\phi \\ \sin\theta \sin\phi \\ \cos\phi \end{pmatrix}. \tag{4}$$

(The v_{init} calculated here is the root-mean-square particle velocity of an ideal gas at thermodynamic equilibrium with the temperature and particle mass specified. Its purpose here is only for an approximately correct starting velocity.)

2.1.3 PARTICLE POSITIONS

The position of the particles also needs to be determined. Given pressure P and number of particles N, the volume of the gas to be examined is given by the ideal gas law

$$PV = Nk_B T \tag{5}$$

and so the side length ℓ of the cube we are considering is

$$\ell = \sqrt[3]{\frac{Nk_B T}{P}}. \tag{6}$$

The positions of each particle in the x, y and z axes are drawn randomly from the uniform distribution on $[0, \ell]$. The randomly generated initial positions of 1500 particles are shown in fig. 2.

2.2 MANAGING PARTICLE COLLISIONS

To correctly model oblique collisions between particles, we cannot assume they are point masses; we must give them a radius R. The distribution of possible collision angles does not depend on the radius chosen, however, so the choice of radius will not affect the final velocity distribution, though it will affect the collision frequency: a larger radius will mean that it takes less time for the system to reach equilibrium.

2.2.1 COLLISION DETECTION

We can say that two particles i and j have collided if the distance between their midpoints is less than the sum of their radii. That is, if

$$\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2} - 2R \leq 0. \tag{7}$$

(Note that if R is too small relative to the speed of the particles, they may pass through each

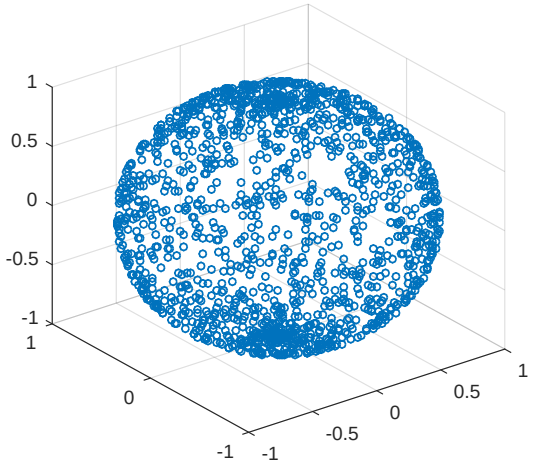


Figure 1: *A 3-dimensional scatter plot showing 1500 unit vectors generated by the method described.*

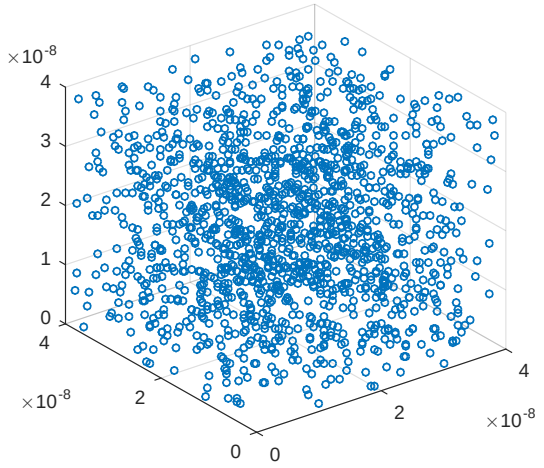


Figure 2: *A 3-dimensional scatter plot showing the initial randomly generated x, y and z positions (in metres) of 1500 particles.*

other without a collision being registered, due to the discrete nature of the simulation. For this reason a good choice of R is crucial.)

To implement this, two nested for loops are used and the presence of a collision is only evaluated when $j > i$:

```
for i = 1:N
    for j = (i+1):N
        % Detect collision
        % Act on collision
    end
end
```

This ensures that each collision is only detected once.

2.2.2 POST-COLLISION VELOCITIES

If a collision is detected between two particles, we must change the velocities of these two particles accordingly. When two particles of equal mass collide elastically in 3-dimensional space, their component velocities in the direction of collision (the vector between their two midpoints) are swapped. The other components of their velocities are unaffected.

This is because in a 1-dimensional elastic collision, momentum is conserved so

$$m_1u_1 + m_2u_2 = m_1v_1 + m_2v_2 \tag{8}$$

and kinetic energy is conserved so

$$\frac{1}{2}m_1u_1^2 + \frac{1}{2}m_2u_2^2 = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2. \tag{9}$$

By collecting terms of mass, taking the difference of two squares and dividing the equations, we come to

$$u_1 + v_1 = u_2 + v_2. \tag{10}$$

Hence if the particles have equal mass then $v_1 = u_2$ and $v_2 = u_1$. This applies along the collision axis in an oblique 3-dimensional collision.

To model this in the program, we first find the unit vector $\hat{r}$ that denotes the direction from the midpoint of particle i to the midpoint of particle j. The vector from i to j is

$$\vec{s}_j - \vec{s}_i \tag{11}$$

where s_i is the 3-dimensional position vector of particle i. Therefore,

$$\hat{r} = \frac{\vec{s}_j - \vec{s}_i}{||\vec{s}_j - \vec{s}_i||}. \tag{12}$$

The vector component $\vec{v}_c$ of each particle's velocity in the direction of the collision can be

found using the dot product:

$$\vec{v}_{ci} = (\hat{r} \cdot \vec{v}_i)\hat{r}, \tag{13}$$
$$\vec{v}_{cj} = (\hat{r} \cdot \vec{v}_j)\hat{r}. \tag{14}$$

Hence, to swap these components of the particles' velocities, we update them as follows:

$$\vec{v}_i \coloneqq \vec{v}_i - \vec{v}_{ci} + \vec{v}_{cj}, \tag{15}$$
$$\vec{v}_j \coloneqq \vec{v}_j - \vec{v}_{cj} + \vec{v}_{ci}. \tag{16}$$

2.3 MANAGING WALL COLLISIONS

Fortunately, dealing with the walls of the container is much simpler. If either

$$x_i < R \tag{17}$$

or

$$\ell - x_i < R, \tag{18}$$

then particle i is in collision with one of the walls at the ends of the x-axis. Then all we need to do is flip the sign of the x-component of i's velocity:

$$v_{xi} \coloneqq -v_{xi}. \tag{19}$$

The same is true for the y and z axes.

2.4 ITERATING THE PROGRAM

The collision checks are run n_{iter} times, with an interval of $\frac{t}{n_{\text{iter}}}$ between each iteration. After every iteration, the positions of every particle are updated according to their determined velocity:

$$\vec{s}_i \coloneqq \vec{s}_i + \vec{v}_i \left(\frac{t}{n_{\text{iter}}}\right). \tag{20}$$

In addition, the velocity and position of each particle is saved after every iteration to 3-dimensional position and velocity matrices.

Once the loop has terminated, we can use the data in these matrices to plot our results.

3 RUNNING THE SIMULATION

The final simulation was run at approximately atmospheric pressure ($P \coloneqq 101\,325\,\text{Pa}$) and room temperature ($T \coloneqq 293\,\text{K}$) with particles of mass $m \coloneqq 4.65 \times 10^{-26}\,\text{kg}$, the mass of an N_2 molecule.

As a compromise between fidelity and computing time, the number of particles was set to be 1500 and the simulation took 200 iterations.

Deciding on the radius of the molecules and the total observation time was rather more difficult. After initially using the van der Waals radius of nitrogen, 155 pm, it became apparent that the majority of collisions were going undetected, because the average distance Δs moved by each particle every iteration was much larger than the particle's radius (so particles would pass through each other and the walls without a collision being detected).

To remedy this, the average total distance travelled by a particle was set at a reasonable 1.5ℓ (which was adjusted to elicit the best rate of evolution of the velocity distribution) and the total time t was then calculated as

$$t \coloneqq \frac{1.5\ell}{v_{\text{init}}}. \tag{21}$$

So that all collisions are detected, the radius must be larger than Δs, and so

$$R \coloneqq \Delta s_{\max} = v_{\max}\left(\frac{t}{n_{\text{iter}}}\right) \tag{22}$$

where $v_{\max} \approx 2 \cdot v_{\text{init}}$.

This method worked remarkably well, and the final values used were

$$t \coloneqq 1.2 \times 10^{-10}\,\text{s}, \tag{23}$$
$$R \coloneqq 5.9 \times 10^{-10}\,\text{m}. \tag{24}$$

4 RESULTS

Figure 3 shows a histogram of the velocity distribution after the 200^{th} iteration.

For comparison, figs. 5 and 6 show the velocity distributions after 20 and 70 iterations respectively.

At the start, the speeds were all $v_{\text{init}} = 510.7\,\text{m}\,\text{s}^{-1}$. As the simulation progressed, the distribution approached the Maxwell-Boltzmann distribution, reaching it entirely by the end of the simulation as expected. Matlab R2015b does not have Maxwell distribution fitting function, and so for illustration purposes a Rayleigh distribution was fitted to the velocities after all 200 iterations. This is shown in fig. 4.

The root-mean-square velocity of this final distribution was $510.7\,\text{m}\,\text{s}^{-1}$, exactly the same as the initial velocity v_{init}.

In total, there were 22,712 collisions between particles and 27,492 collisions with the walls.

5 CONCLUSIONS

After a very short amount of time, the particles in an ideal gas reach thermodynamic equilibrium at a Maxwell-Boltzmann distribution of velocities. If all of the particles start at the same initial speed, then this speed is also the final root-mean-square velocity of the particles.

It was interesting to note that the choice of particle radius played a large role in determining the success of the simulation, despite the final distribution not depending on it.

In the way of further analysis of the completed simulation, it could also be possible to

a) compare the final distributions at various different temperatures.

b) build a subroutine to test the actual pressure and temperature and compare to the result predicted by the ideal gas law.

c) run statistical tests throughout to evaluate exactly how the speeds approach the Maxwell-Boltzmann distribution over time.

The simulation could also be run for a much larger number of particles, given more computing power, to decrease random fluctuations in the distribution of speeds.

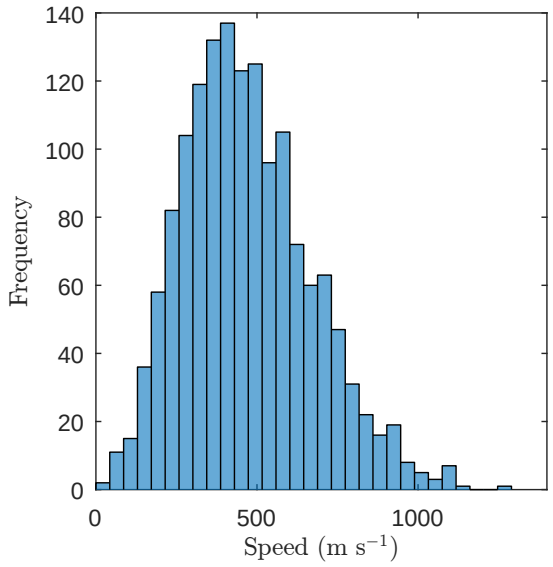


Figure 3: *A 30-bin histogram of the speeds of the particles after the full 200 iterations (*114.9 ps*).*

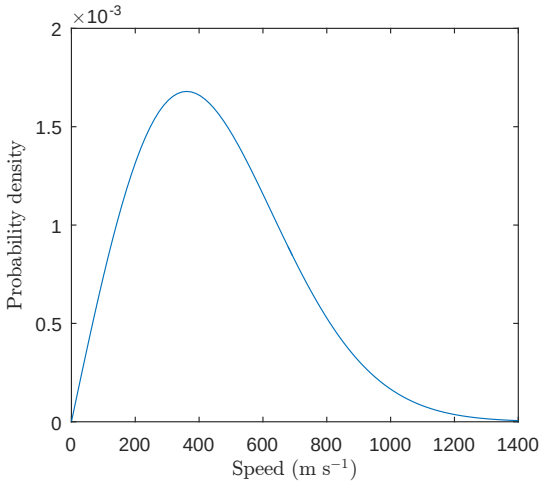


Figure 4: *Probability density function of the final speeds of the particles under a Rayleigh distribution.*

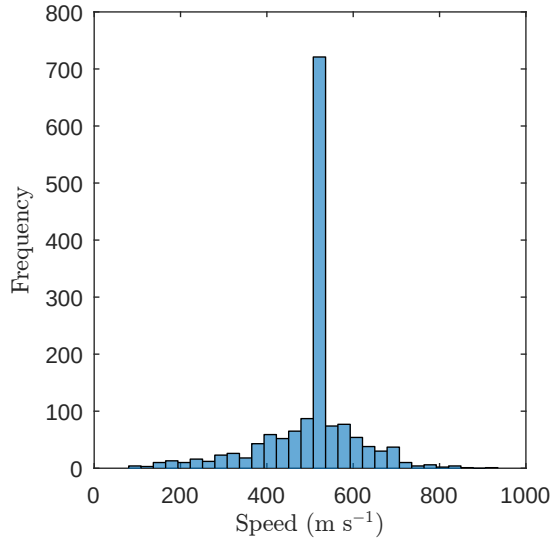


Figure 5: *Speeds of the particles after 20 iterations (*11.49 ps*).*

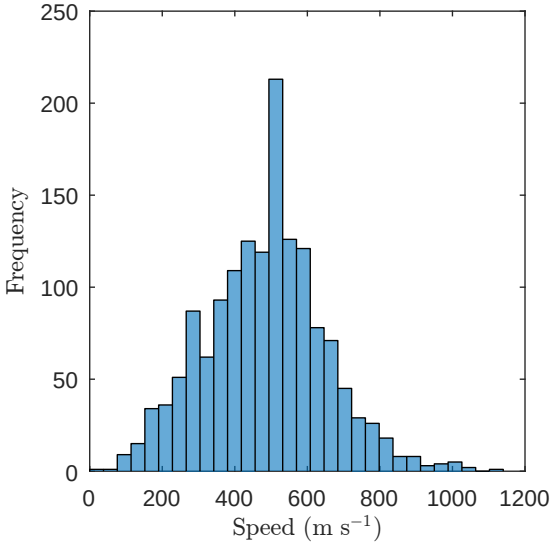


Figure 6: *Speeds of the particles after 70 iterations (*40.22 ps*).*

APPENDIX: SOURCE CODE

(To be run in Matlab or Octave)

```
clear;
close all;

% Simulation parameters

P = 101325; % Pressure of particles / Pa
T = 293; % Thermodynamic temperature of particles / K
m = 4.65e-26; % Mass of particles / kg
N = 1500; % Number of particles to consider

R = 5.87e-10; % Radius of particles / m
t = 1.15e-10; % Total time to observe for / s
n_iter = 200; % Resolution of the simulation

%----------------------------------------

% Setup

k_B = 1.38e-23; % Boltzmann's constant
v_init = sqrt(3*k_B*T/m); % Initial (uniform) speed of every particle (using
↪  3/2 k_B T = 1/2 m v^2) / m s^-1
V = N*k_B*T/P; % Volume of container / m^3
side = nthroot(V,3); % Side length of the cubic box we're considering / m

dt = t/n_iter; % Time per step / s

s = side*rand(3,N); % Random x,y,z positions for each particle / m

angles = -pi + 2*pi*rand(2,N); % Two random angles per particle (between -pi
↪  and pi) / rad
unit_vect = [cos(angles(1,:)).*sin(angles(2,:)); ...
                sin(angles(1,:)).*sin(angles(2,:)); ...
                cos(angles(2,:))]; % A random uniformly distributed
↪  3-dimensional unit vector per particle
v = v_init*unit_vect; % Random velocities with uniform specified magnitude / m
↪  s^-1



% Data storage

wallcollisions = 0; % Collision counters
particlecollisions = 0;
firsttimeparticlecollisions = 0;

```

```
s_history = zeros(3,N,n_iter+1); % 3D matrices to hold the entire history of particle positions and velocities
v_history = zeros(3,N,n_iter+1);

s_history(:,:,1) = s; % Fill in the initial values
v_history(:,:,1) = v;

animation = struct('cdata',[],'colormap',[]); % Movie array

% Main loop

for iter = 1:n_iter

    for i = 1:N % For each particle

        % Check for collision with wall

        for dim = 1:3 % Repeat for x,y,z

            if s(dim,i) <= R || side - s(dim,i) <= R % Check distance from each wall

                v(dim,i) = -v(dim,i); % Reverse velocity
                wallcollisions = wallcollisions + 1; % Increment collision counter

            end

        end

        % Check for collision with another particle

        for j = i+1:N % Check against all particles with a higher index (to avoid duplicate collision checks)

            if (s(1,j)-s(1,i))^2 + ...
                    (s(2,j)-s(2,i))^2 + ...
                    (s(3,j)-s(3,i))^2 <= 4*R^2 % Detect collision

                % Unit vector between the two particles' centres
                difference_vector = (s(:,j)-s(:,i));
                difference_magnitude = sqrt((s(1,j)-s(1,i))^2 + ...
                                            (s(2,j)-s(2,i))^2 + ...
                                            (s(3,j)-s(3,i))^2);
                direction = difference_vector/difference_magnitude;

                % Component velocities in the collision direction

                v_i = dot(v(:,i),direction)*direction;
                v_j = dot(v(:,j),direction)*direction;

```

```
                delta_v = v_j - v_i;

                % Swap these velocities in this direction

                v(:,i) = v(:,i) + delta_v;
                v(:,j) = v(:,j) - delta_v;

                particlecollisions = particlecollisions + 1; % Increment collision counter

                if iter == 1

                    firsttimeparticlecollisions = firsttimeparticlecollisions + 1;

                end

            end
        end
    end

    % Update all positions

    s = s + v*dt;

    % Add to history

    s_history(:,:,iter+1) = s;
    v_history(:,:,iter+1) = v;

end

speeds = sqrt((v_history(1,:,:)).^2+(v_history(2,:,:)).^2+(v_history(3,:,:)).^2); % Matrix of history of speeds

% The matrix 'speeds' can then be used to plot histograms etc. of the speed
% distribution at different times.
```

Chapter 3

Deriving the Maxwell-Boltzmann distribution

A few months after showing the Maxwell-Boltzmann distribution arises numerically, I became interested in the analytic derivation and wrote this argument up.

Deriving the Maxwell-Boltzmann distribution

Damon Falck

August 24, 2017

Maxwell's symmetry argument

We wish to derive the fraction of particles in an ideal gas with speed between v and $v + \mathrm{d}v$.

Let v_x, v_y and v_z be the components of the velocity of each particle in three perpendicular directions. If $\rho(v_x)$ is the probability density function of v_x, then the fraction of particles for which this velocity lies between v_x and $v_x + \mathrm{d}v_x$ is $\rho(v_x)\,\mathrm{d}v_x$. The x-direction is special in no way and therefore the same function applies to v_y and v_z also.

These three velocities must not affect each other in any way because they are at right angles and statistically independent, and so the fraction of particles with velocities between v_x and $v_x+\mathrm{d}v_x$, and between v_y and $v_y + \mathrm{d}v_y$, and between v_z and $v_z + \mathrm{d}v_z$, is $\rho(v_x)\,\mathrm{d}v_x\rho(v_y)\,\mathrm{d}v_y\rho(v_z)\,\mathrm{d}v_z = \rho(v_x)\rho(v_y)\rho(v_z)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$.

However, the choice of direction of the axes we're using is purely arbitrary, and so this fraction must depend only on the speed of the particle $v^2 = v_x^2 + v_y^2 + v_z^2$. Therefore,

$$\rho(v_x)\rho(v_y)\rho(v_z)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z = \phi(v_x^2 + v_y^2 + v_z^2)\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$$

for some function ϕ.

We note that a product appears on the left and a sum on the right; thus, the solution to this *must* be an exponential. We let $\rho(x) = A\mathrm{e}^{-Bx^2}$ for some positive constants A and B so that $\phi(v^2) = A^3\mathrm{e}^{-Bv^2}$. We add the negative sign to B since the number of particles with velocities of increasing size must decrease.

So, the fraction of particles with velocity vector in the 'box' of volume $\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$ with its innermost vertex a distance v from the origin is $A^3\mathrm{e}^{-Bv^2}\,\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$. A particle with speed between v and $v + \mathrm{d}v$ will have its velocity vector in the space occupied by the sphere of radius $v + \mathrm{d}v$ with its centre at the origin, and *not* by the similar sphere of radius v. The volume of this space, due to the smallness of $\mathrm{d}v$, is $4\pi v^2\,\mathrm{d}v$ and so, replacing $\mathrm{d}v_x\,\mathrm{d}v_y\,\mathrm{d}v_z$, the fraction of particles with speed between v and $v+\mathrm{d}v$ is $4\pi A^3 v^2\mathrm{e}^{-Bv^2}\,\mathrm{d}v$. If the probability density function for particle speed is $f(v)$ then it follows that

$$f(v) = 4\pi A^3 v^2\mathrm{e}^{-Bv^2}.$$

As this must be normalised,

$$4\pi A^3\int_0^\infty v^2\mathrm{e}^{-Bv^2}\,\mathrm{d}v = 1$$

which implies that since $\int_0^\infty x^2 e^{-x^2}\,\mathrm{d}x = \frac{\sqrt{\pi}}{4}$,

$$\begin{aligned} 4\pi A^3 \frac{\sqrt{\pi}}{4B^{3/2}} &= 1 \\ \implies A &= \sqrt{\frac{B}{\pi}} \end{aligned}$$

and thus $f(v) = 4\pi \left(\frac{B}{\pi}\right)^{3/2} v^2 \mathrm{e}^{-Bv^2}$.

The mean square speed is given by

$$\langle v^2 \rangle = \int_0^\infty v^2 f(v)\,\mathrm{d}v = 4\pi \left(\frac{B}{\pi}\right)^{3/2} \int_0^\infty v^4 \mathrm{e}^{-Bv^2}\,\mathrm{d}v$$

and so as $\int_0^\infty x^4 \mathrm{e}^{-x^2}\,\mathrm{d}x = \frac{3\sqrt{\pi}}{8}$, we come to $\langle v^2 \rangle = 4\pi \left(\frac{B}{\pi}\right)^{3/2} \frac{3\sqrt{\pi}}{8B^{5/2}} = \frac{3}{2B}$.

However, from kinetic theory we know that

$$\frac{1}{2} m \langle v^2 \rangle = \frac{3}{2} k_B T$$

where m is the mass of each particle, T is the thermodynamic temperature of the gas and k_B is Boltzmann's constant. Therefore,

$$\begin{aligned} \frac{1}{2} m \frac{3}{2B} &= \frac{3}{2} k_B T \\ \implies B &= \frac{m}{2k_B T}. \end{aligned}$$

This leads us to our final expression for the Maxwell-Boltzmann distribution,

$$f(v) = 4\pi \left(\frac{m}{2\pi k_B T}\right)^{3/2} v^2 \mathrm{e}^{-\frac{mv^2}{2k_B T}}.$$

Chapter 4

An investigation into electric fields around charged spheres

This was the result of a couple of weeks of Dr Cheung's lessons on electric fields in September 2017. I was very interested in the topic and wanted another chance to combine mathematics with a bit of programming.

AN INVESTIGATION INTO ELECTRIC FIELDS AROUND CHARGED SPHERES

Damon Falck

June 30, 2018

I A HOLLOW CHARGED SPHERE

Consider a thin, spherical shell of radius r and surface charge density σ. To find the charge at a point h from its centre, we note that by symmetry all tangential fields cancel and so we need only consider the radial component of the electric field at this point.

Suppose the shell is centred at the origin in $\mathbb{R}^3$ and that our test point is on the z-axis at position

$$\mathbf{h} = \begin{pmatrix} 0 \\ 0 \\ h \end{pmatrix}$$

so that any point on the shell has position vector

$$\mathbf{r} = \begin{pmatrix} r\cos\phi\cos\theta \\ r\cos\phi\sin\theta \\ r\sin\phi \end{pmatrix}$$

where θ and ϕ are the two angles shown in fig. 1.

By Coulomb's law, the contribution of a small amount of charge at point $\mathbf{r}$ to the field at $\mathbf{h}$ is

$$\mathrm{d}E = \frac{\mathrm{d}Q}{4\pi\varepsilon_0|\mathbf{h}-\mathbf{r}|^2}.$$

By definition of surface charge density,

$$\sigma = \frac{\mathrm{d}Q}{\mathrm{d}A}$$

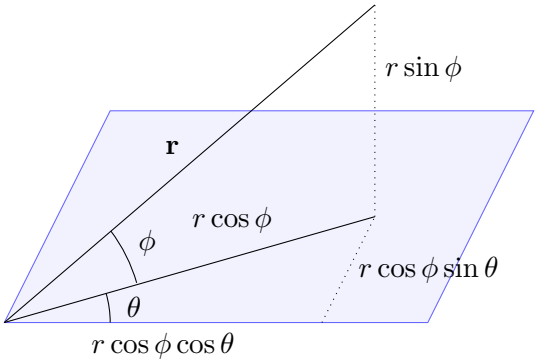


Figure 1: A point r on a thin spherical shell.

where $\mathrm{d}A$ is a small area of the shell at $\mathbf{r}$, and it can be seen that this small area has 'height' $r \cdot \mathrm{d}\phi$ and 'width' $r\cos\phi \cdot \mathrm{d}\theta$ (both arc lengths), so that

$$\mathrm{d}E = \frac{\sigma r^2 \cos\phi \,\mathrm{d}\theta \,\mathrm{d}\phi}{4\pi\varepsilon_0 |\mathbf{h}-\mathbf{r}|^2}.$$

As explained, we only want the radial (vertical) component of this field, and so we multiply by $\cos\psi$ where ψ is the angle $\mathbf{h}-\mathbf{r}$ makes to the vertical:

$$\mathrm{d}E_z = \frac{\sigma r^2 \cos\phi \,\mathrm{d}\theta \,\mathrm{d}\phi}{4\pi\varepsilon_0 |\mathbf{h}-\mathbf{r}|^2} \cos\psi.$$

Therefore by the principle of superposition, the total field at point $\mathbf{h}$ is

$$E = E_z = \frac{\sigma r^2}{4\pi\varepsilon_0} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} \frac{\cos\phi\cos\psi}{|\mathbf{h}-\mathbf{r}|^2} \,\mathrm{d}\theta \,\mathrm{d}\phi.$$

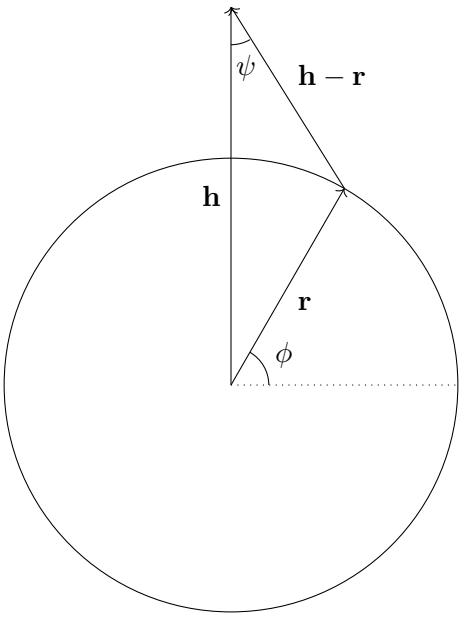


Figure 2: A vertical slice through the sphere (the situation is of course rotationally symmetric around the z-axis).

From fig. 2 it can be seen that

$$\cos\psi = \frac{\mathbf{h}_z - \mathbf{r}_z}{|\mathbf{h}-\mathbf{r}|} = \frac{h - r\sin\phi}{|\mathbf{h}-\mathbf{r}|}$$

and so our value for the total electric field becomes

$$E = \frac{\sigma r^2}{4\pi\varepsilon_0} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} \frac{\cos\phi(h - r\sin\phi)}{|\mathbf{h} - \mathbf{r}|^3} \,\mathrm{d}\theta \,\mathrm{d}\phi.$$

The situation is rotationally symmetric around the z-axis, which means nothing can be a function of θ. Therefore performing the inner integral is a trivial matter of multiplying by 2π:

$$E = \frac{\sigma r^2}{2\varepsilon_0} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\cos\phi(h - r\sin\phi)}{|\mathbf{h} - \mathbf{r}|^3} \,\mathrm{d}\phi.$$

Next, we want to find the denominator:

$$\begin{aligned}
|\mathbf{h} - \mathbf{r}| &= \left| \begin{pmatrix} 0 \\ 0 \\ h \end{pmatrix} - \begin{pmatrix} r\cos\phi\cos\theta \\ r\cos\phi\sin\theta \\ r\sin\phi \end{pmatrix} \right| \\
&= \left| \begin{pmatrix} -r\cos\phi\cos\theta \\ -r\cos\phi\sin\theta \\ h - r\sin\phi \end{pmatrix} \right| \\
&= \sqrt{r^2\cos^2\phi\cos^2\theta + r^2\cos^2\phi\sin^2\theta + h^2 - 2hr\sin\phi + r^2\sin^2\phi} \\
&= \sqrt{r^2\cos^2\phi + h^2 - 2hr\sin\phi + r^2\sin^2\phi} \\
&= \sqrt{r^2 + h^2 - 2hr\sin\phi}
\end{aligned}$$

which, as expected, is not a function of θ. So, the integral we need to evaluate is

$$E = \frac{\sigma r^2}{2\varepsilon_0} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\cos\phi(h - r\sin\phi)}{\left(h^2 + r^2 - 2hr\sin\phi\right)^{3/2}} \,\mathrm{d}\phi.$$

Substituting $u = \sin\phi$ so that $\mathrm{d}u = \cos\phi\,\mathrm{d}\phi$, this becomes

$$\begin{aligned}
E &= \frac{\sigma r^2}{2\varepsilon_0} \int_{-1}^{1} \frac{h - ru}{\left(h^2 + r^2 - 2hru\right)^{3/2}} \,\mathrm{d}u \\
&= \frac{\sigma r^2 h}{2\varepsilon_0} \int_{-1}^{1} \frac{1}{\left(h^2 + r^2 - 2hru\right)^{3/2}} \,\mathrm{d}u - \frac{\sigma r^3}{2\varepsilon_0} \int_{-1}^{1} \frac{u}{\left(h^2 + r^2 - 2hru\right)^{3/2}} \,\mathrm{d}u. \qquad (1)
\end{aligned}$$

The first integral evaluates to

$$\begin{aligned}
\int_{-1}^{1} \frac{1}{\left(h^2 + r^2 - 2hru\right)^{3/2}} \,\mathrm{d}u &= \left[\frac{-2}{-2hr\sqrt{h^2 + r^2 - 2hru}} \right]_{-1}^{1} \\
&= \frac{1}{hr} \left(\frac{1}{\sqrt{h^2 + r^2 - 2hr}} - \frac{1}{\sqrt{h^2 + r^2 + 2hr}} \right) \\
&= \frac{1}{hr} \left(\frac{1}{|h - r|} - \frac{1}{h + r} \right). \qquad (2)
\end{aligned}$$

For the second integral we'll use parts, relying on the result we just showed:

$$\begin{aligned}
\int_{-1}^{1} \frac{u}{\left(h^2+r^2-2hru\right)^{3/2}}\,\mathrm{d}u &= \left[\frac{u}{hr\sqrt{h^2+r^2-2hru}}\right]_{-1}^{1} - \int_{-1}^{1} \frac{1}{hr\sqrt{h^2+r^2-2hru}}\,\mathrm{d}u \\
&= \left[\frac{u}{hr\sqrt{h^2+r^2-2hru}}\right]_{-1}^{1} - \frac{1}{hr}\left[\frac{2}{-2hr}\sqrt{h^2+r^2-2hru}\right]_{-1}^{1} \\
&= \frac{1}{hr}\left[\frac{u}{\sqrt{h^2+r^2-2hru}} + \frac{\sqrt{h^2+r^2-2hru}}{hr}\right]_{-1}^{1} \\
&= \frac{1}{hr}\left(\frac{1}{|h-r|} + \frac{1}{h+r} + \frac{|h-r|-(h+r)}{hr}\right) \qquad (3)
\end{aligned}$$

Putting the expressions from eqs. (2) and (3) together into our main expression for E in eq. (1), we come to

$$\begin{aligned}
E &= \frac{\sigma r}{2\varepsilon_0}\left(\frac{1}{|h-r|} - \frac{1}{h+r}\right) - \frac{\sigma r^2}{2\varepsilon_0 h}\left(\frac{1}{|h-r|} + \frac{1}{h+r} + \frac{|h-r|-(h+r)}{hr}\right) \\
&= \frac{\sigma}{2\varepsilon_0}\left(\frac{r}{|h-r|} - \frac{r}{h+r} - \frac{r^2}{h|h-r|} - \frac{r^2}{h(h+r)} - \frac{r|h-r|}{h^2} + \frac{r(h+r)}{h^2}\right) \\
&= \frac{\sigma}{2\varepsilon_0}\left(\frac{rh^2(h+r) - rh^2|h-r| - r^2h(h+r) - r^2h|h-r| - r|h-r||h-r|\,(h+r) + r(h+r)(h+r)|h-r|}{h^2(h+r)|h-r|}\right) \\
&= \frac{\sigma r}{2\varepsilon_0 h^2}\left(\frac{h^3 + rh^2 - h^2|h-r| - rh^2 - r^2h - rh|h-r| - |h-r|\left|h^2-r^2\right| + (h+r)\left|h^2-r^2\right|}{\left|h^2-r^2\right|}\right) \qquad (4)
\end{aligned}$$

From this rather complicated-looking expression there arise two simple cases: that where $h \leqslant r$ and that where $h > r$. If $h \leqslant r$, meaning our test point is *inside* the spherical shell, then

$$|h-r| = r-h$$

and

$$\left|h^2-r^2\right| = r^2-h^2.$$

Then, eq. (4) simplifies to

$$\begin{aligned}
E &= \frac{\sigma r}{2\varepsilon_0 h^2(r^2-h^2)}\left(h^3 + rh^2 - h^2(r-h) - rh^2 - r^2h - rh(r-h) - (r-h)(r^2-h^2) + (h+r)(r^2-h^2)\right) \\
&= \frac{\sigma r}{2\varepsilon_0 h^2(r^2-h^2)} \cdot 0 \\
&= 0.
\end{aligned}$$

This is wonderful: at no point within the sphere is there any electric field whatsoever! The second case, where $h > r$, means our test point is outside the sphere, and implies

$$|h-r| = h-r$$

and

$$\left|h^2-r^2\right| = h^2-r^2.$$

Therefore, eq. (4) reduces to:

$$\begin{aligned}
E &= \frac{\sigma r}{2\varepsilon_0 h^2(h^2-r^2)}\left(h^3 + rh^2 - h^2(h-r) - rh^2 - r^2h - rh(h-r) - (h-r)(h^2-r^2) + (h+r)(h^2-r^2)\right) \\
&= \frac{\sigma r}{2\varepsilon_0 h^2(h^2-r^2)}\left(2rh^2 - 2r^3\right) \\
&= \frac{\sigma r^2}{\varepsilon_0 h^2}. \qquad (5)
\end{aligned}$$

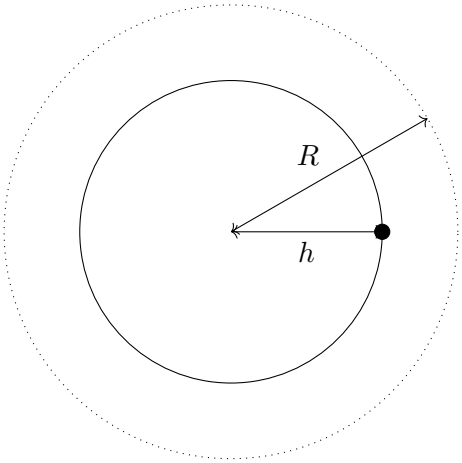


Figure 3: A smaller sphere of radius h within a larger one of radius R.

This is a beautiful result: the electric field depends, as makes sense intuitively, only on the ratio of the distance h to the radius r. It doesn't end there, though: note that the surface area of the sphere is $4\pi r^2$ and so where Q is the total charge,

$$\sigma = \frac{Q}{4\pi r^2}.$$

Then, eq. (5) becomes

$$E = \frac{Q}{4\pi\varepsilon_0 h^2};$$

the spherical shell generates the same electric field as would a point charge at its centre.

II A SOLID CHARGED SPHERE

Now that we know the behaviour of a hollow sphere, the case of a solid sphere is trivial. By splitting the solid sphere up into infinitely many shells, each with its equivalent charge at the centre of the sphere, it follows that the equivalent charge of the entire sphere must be at its centre and must have the value Q where Q is the total charge on the sphere. So, outside the sphere, the same formula applies:

$$E = \frac{Q}{4\pi\varepsilon_0 h^2}$$

at a distance h from the centre of the sphere.

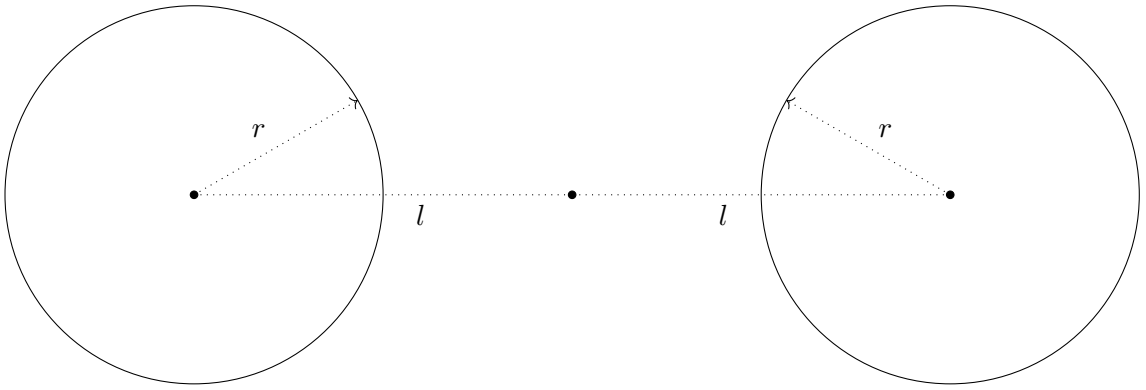
If the test charge is *within* the sphere, then as proven above, no subshell of radius r such that $h < r$ will exert a force on the test charge; we can disregard these shells so that we have a smaller solid sphere of radius h behaving in the same way.

If the original sphere has radius R then the new sphere is a fraction $\frac{h^3}{R^3}$ of the original volume, and so assuming the charge is uniformly distributed, the smaller sphere has a total charge of $\frac{h^3}{R^3}Q$. Consequently the electric field within the sphere is

$$E = \frac{\frac{h^3}{R^3}Q}{4\pi\varepsilon_0 h^2} = \frac{hQ}{4\pi\varepsilon_0 R^3}.$$

III THE ELECTRIC FIELD AROUND TWO SOLID CHARGED SPHERES

Now consider two insulating spheres, each with radius r and volume charge density ρ, with their centres separated by distance $2l$ such that $l > r$. We'll place the spheres along the z-axis in $\mathbb{R}^3$ so that the origin is midway between their centres.



At an arbitrary point $\mathbf{s} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ we wish to determine the electric field $\mathbf{E}$.

There are three cases to consider: outside both spheres, inside the upper sphere, and inside the lower sphere.

Outside both spheres, each sphere can be modelled as a point charge at the centre of the sphere. The total charge of each sphere is $\rho V = \frac{4\pi r^3 \rho}{3}$ and these point charges are located at $\mathbf{z_1} = \begin{pmatrix} 0 \\ 0 \\ l \end{pmatrix}$ and $\mathbf{z_2} = \begin{pmatrix} 0 \\ 0 \\ -l \end{pmatrix}$. Hence, by the principle of superposition, the total field at a point $\mathbf{s}$ outside both spheres is

$$\begin{aligned}
\mathbf{E} &= \frac{\frac{4\pi r^3 \rho}{3}}{4\pi\varepsilon_0 |\mathbf{s} - \mathbf{z_1}|^3}(\mathbf{s} - \mathbf{z_1}) + \frac{\frac{4\pi r^3 \rho}{3}}{4\pi\varepsilon_0 |\mathbf{s} - \mathbf{z_2}|^3}(\mathbf{s} - \mathbf{z_2}) \qquad (6) \\
&= \frac{r^3\rho}{3\varepsilon_0 \left| \begin{pmatrix} x \\ y \\ z \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ l \end{pmatrix} \right|^3} \left[\begin{pmatrix} x \\ y \\ z \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ l \end{pmatrix} \right] + \frac{r^3\rho}{3\varepsilon_0 \left| \begin{pmatrix} x \\ y \\ z \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ -l \end{pmatrix} \right|^3} \left[\begin{pmatrix} x \\ y \\ z \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ -l \end{pmatrix} \right] \\
&= \frac{r^3\rho}{3\varepsilon_0} \left[\frac{1}{\sqrt{x^2 + y^2 + (z-l)^2}^3} \begin{pmatrix} x \\ y \\ z - l \end{pmatrix} + \frac{1}{\sqrt{x^2 + y^2 + (z+l)^2}^3} \begin{pmatrix} x \\ y \\ z + l \end{pmatrix} \right] \\
&= \frac{r^3\rho}{3\varepsilon_0} \left[\frac{1}{(x^2 + y^2 + z^2 - 2zl + l^2)^{3/2}} \begin{pmatrix} x \\ y \\ z - l \end{pmatrix} + \frac{1}{(x^2 + y^2 + z^2 + 2zl + l^2)^{3/2}} \begin{pmatrix} x \\ y \\ z + l \end{pmatrix} \right]. \qquad (7)
\end{aligned}$$

When inside the upper sphere, the contribution from the lower sphere remains unchanged, but the electric field of the upper sphere decreases in magnitude as described in section II. The total equivalent charge of the upper

sphere changes from $\frac{4}{3}\pi r^3\rho$ to $\frac{4}{3}\pi|\mathbf{s}-\mathbf{z_1}|^3\rho$, so the new electric field function (based on eq. (6)) is

$$\begin{aligned}\mathbf{E} &= \frac{\frac{4}{3}\pi|\mathbf{s}-\mathbf{z_1}|^3\rho}{4\pi\varepsilon_0|\mathbf{s}-\mathbf{z_1}|^3}(\mathbf{s}-\mathbf{z_1}) + \frac{\frac{4\pi r^3\rho}{3}}{4\pi\varepsilon_0|\mathbf{s}-\mathbf{z_2}|^3}(\mathbf{s}-\mathbf{z_2}) \\ &= \frac{\rho}{3\varepsilon_0}\left[\begin{pmatrix}x\\y\\z-l\end{pmatrix} + \frac{r^3}{(x^2+y^2+z^2+2zl+l^2)^{3/2}}\begin{pmatrix}x\\y\\z+l\end{pmatrix}\right].\end{aligned} \tag{8}$$

When our test charge is inside the lower sphere instead, we just switch the sign of l wherever it occurs:

$$E = \frac{\rho}{3\varepsilon_0}\left[\begin{pmatrix}x\\y\\z+l\end{pmatrix} + \frac{r^3}{(x^2+y^2+z^2-2zl+l^2)^{3/2}}\begin{pmatrix}x\\y\\z-l\end{pmatrix}\right]. \tag{9}$$

When to use each of these three formulae (eqs. (7) to (9)) can be determined by the following table:

Position	Condition 1	Condition 2	Equation to use
Outside both	$\lvert\mathbf{s}-\mathbf{z_1}\rvert > r$	$\lvert\mathbf{s}-\mathbf{z_2}\rvert > r$	eq. (7)
Inside upper	$\lvert\mathbf{s}-\mathbf{z_1}\rvert \leqslant r$	$\lvert\mathbf{s}-\mathbf{z_2}\rvert > r$	eq. (8)
Inside lower	$\lvert\mathbf{s}-\mathbf{z_1}\rvert > r$	$\lvert\mathbf{s}-\mathbf{z_2}\rvert \leqslant r$	eq. (9)

Writing these conditions in terms of scalars, if $|\mathbf{s}-\mathbf{z_1}| > r$ then

$$\sqrt{x^2+y^2+z^2-2zl+l^2} > r$$

and likewise if $|\mathbf{s}-\mathbf{z_2}| > r$ then

$$\sqrt{x^2+y^2+z^2+2zl+l^2} > r.$$

We can express the electric field as a piecewise function in Mathematica:

```
In[32]:= Field[x_, y_, z_] :=
 Piecewise[{{r^3 ρ/(3 ϵ) (1/(x^2 + y^2 + z^2 - 2 z 1 + 1^2)^(3/2) {x, y, z - 1} + 1/(x^2 + y^2 + z^2 + 2 z 1 + 1^2)^(3/2) {x, y, z + 1}),
   Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] > r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] > r},
  {ρ/(3 ϵ) ({x, y, z - 1} + r^3/(x^2 + y^2 + z^2 + 2 z 1 + 1^2)^(3/2) {x, y, z + 1}),
   Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] ≤ r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] > r},
  {ρ/(3 ϵ) ({x, y, z + 1} + r^3/(x^2 + y^2 + z^2 - 2 z 1 + 1^2)^(3/2) {x, y, z - 1}),
   Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] > r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] ≤ r}}]
```

This leads to some rather nice visualisations, as shown in figs. 4 and 5. We set l to 2 m and the radius r to 1 m.

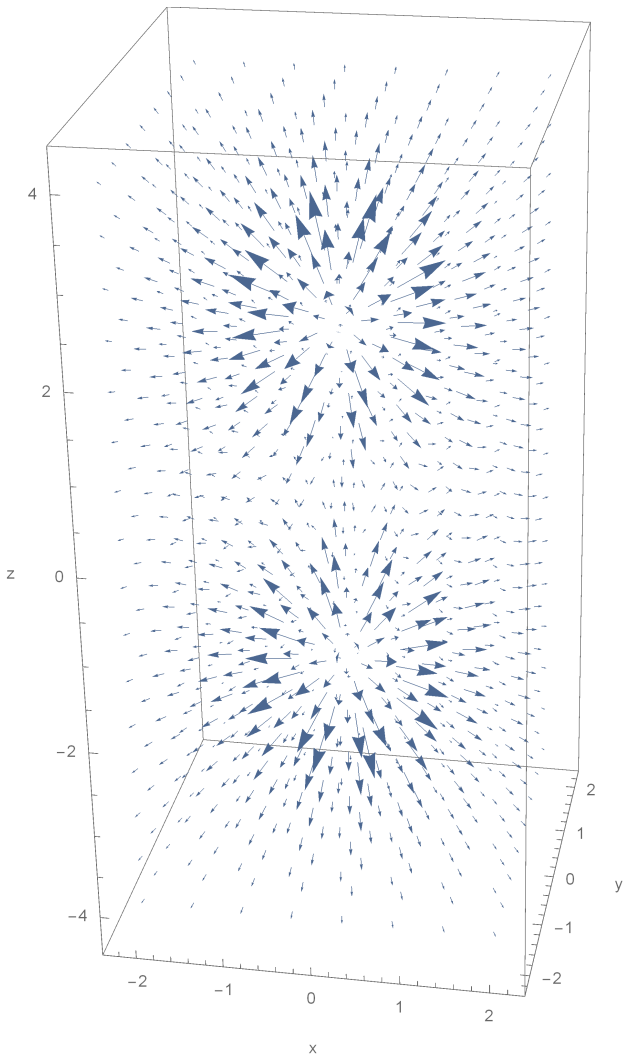


Figure 4: A 3D vector field visualisation of the electric field $\vec{E}$ for the spheres with like charges.

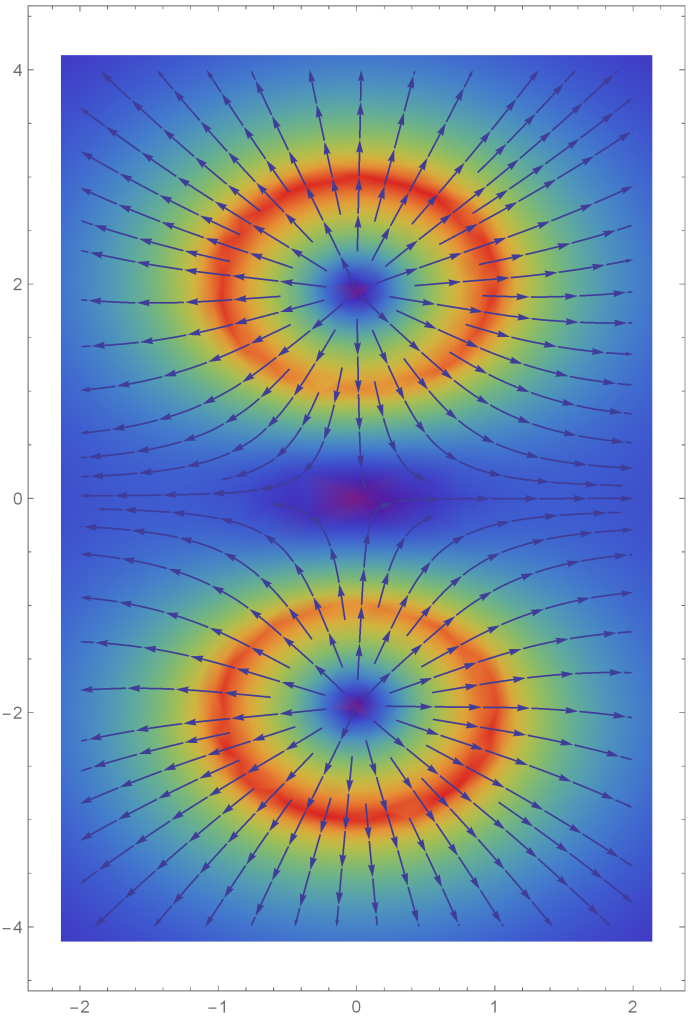


Figure 5: A stream plot overlayed on a density plot in the plane $y = 0$ (for the spheres with like charges).

So far we have assumed the spheres are either both positively charged, or both negatively charged. If we set them to have opposite charges, the visualisation changes. Our piecewise function becomes:

```
In[85]:= Field'[x_, y_, z_] :=
  Piecewise[{{r^3 ρ/(3 ϵ) (1/(x^2 + y^2 + z^2 - 2 z 1 + 1^2)^(3/2) {x, y, z - 1} - 1/(x^2 + y^2 + z^2 + 2 z 1 + 1^2)^(3/2) {x, y, z + 1}),
    Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] > r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] > r},
   {ρ/(3 ϵ) ({x, y, z - 1} - r^3/(x^2 + y^2 + z^2 + 2 z 1 + 1^2)^(3/2) {x, y, z + 1}),
    Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] ≤ r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] > r},
   {ρ/(3 ϵ) (-{x, y, z + 1} + r^3/(x^2 + y^2 + z^2 - 2 z 1 + 1^2)^(3/2) {x, y, z - 1}),
    Sqrt[x^2 + y^2 + z^2 - 2 z 1 + 1^2] > r ∧ Sqrt[x^2 + y^2 + z^2 + 2 z 1 + 1^2] ≤ r}}]
```

Figures 6 to 9 show the field when the two spheres have opposite charges.

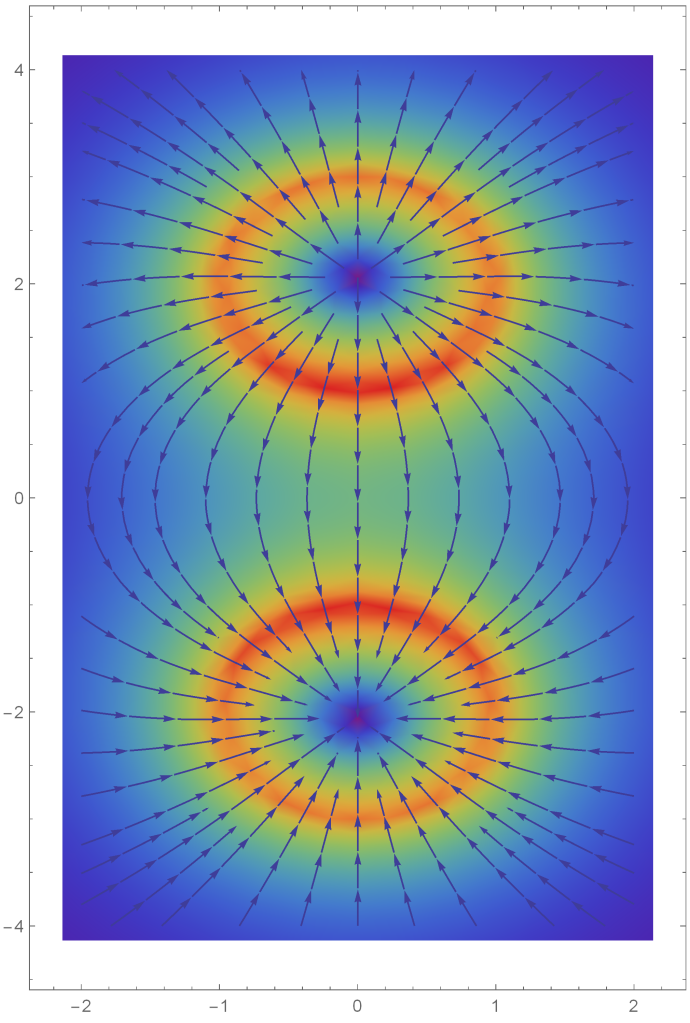


Figure 6: A stream plot and density plot at $y = 0$ for the spheres with opposite charges.

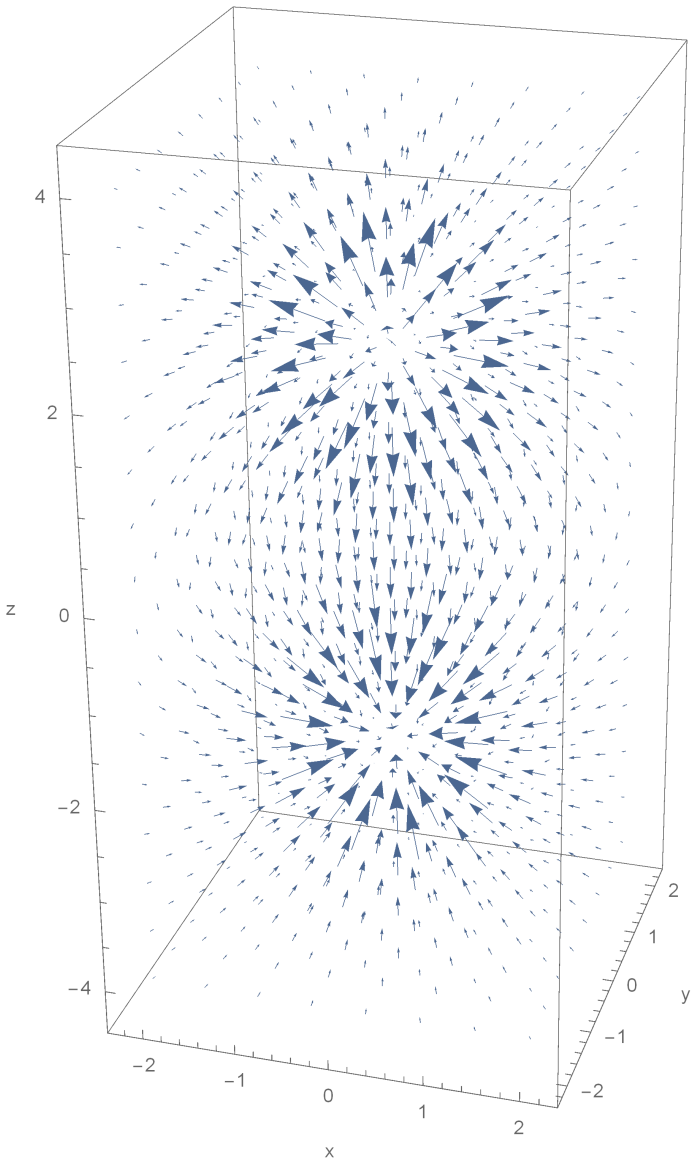


Figure 7: A 3D vector field visualisation of the electric field $\vec{E}$ for the spheres with opposite charges.

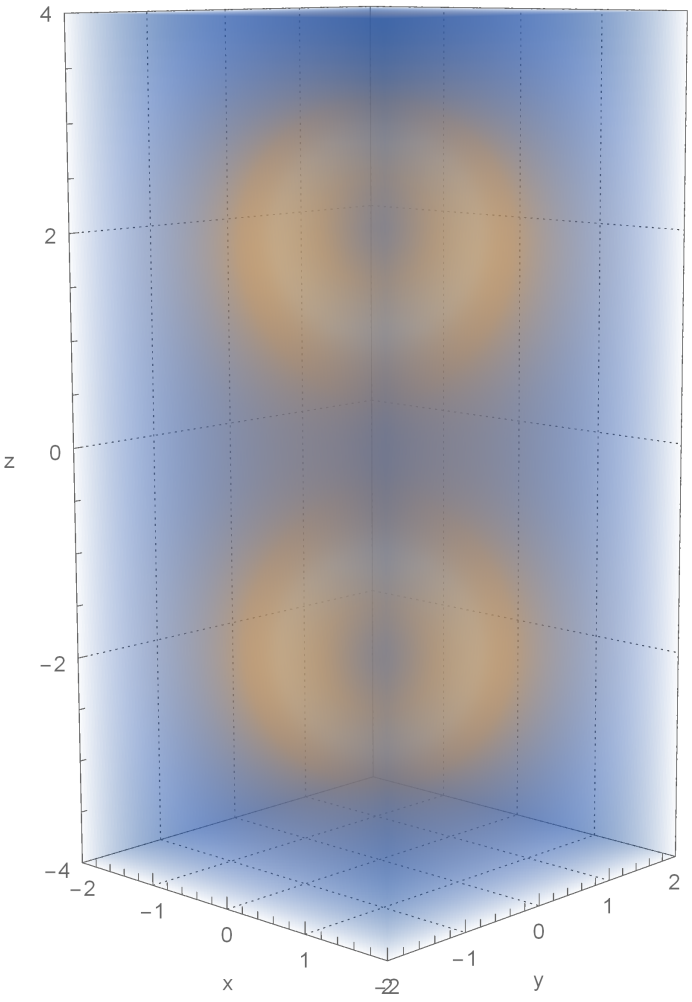


Figure 8: A 3D plot of the electric field magnitude E for the spheres with opposite charges.

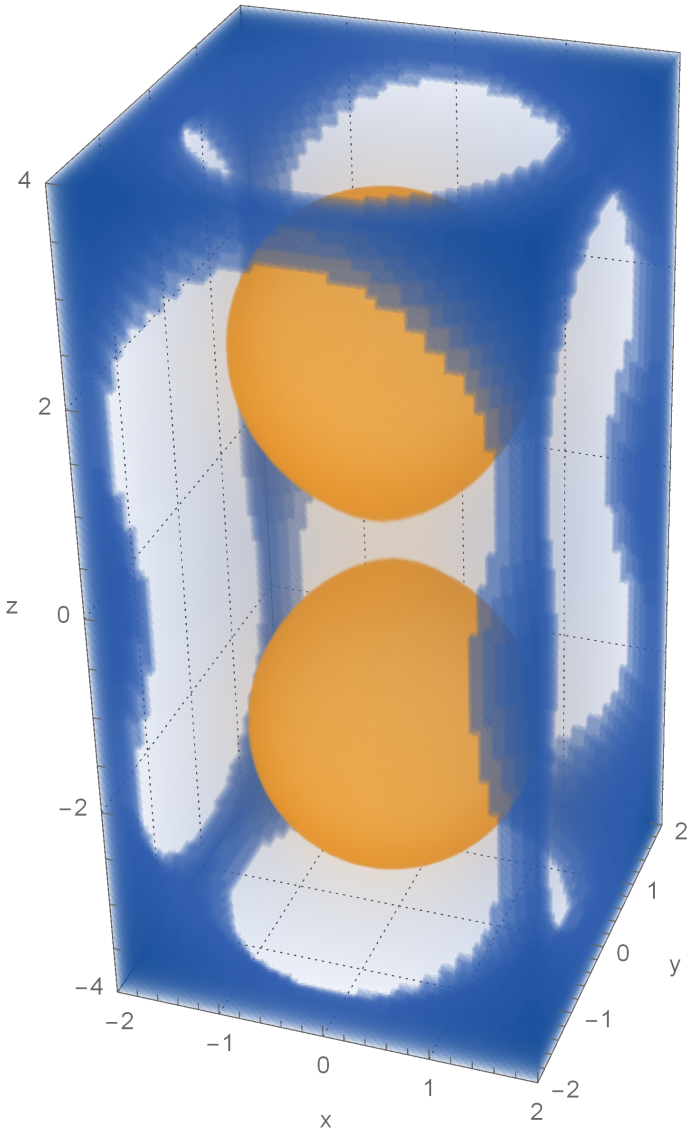


Figure 9: Another 3D plot of the electric field magnitude E for the spheres with opposite charges, this time with the opacity function adjusted to make the spheres more visible.

IV THE FORCE BETWEEN TWO SPHERES OF DIFFERENT VOLUMES

We'll now consider two solid spheres of radius r and R centred at arbitrary positions in Cartesian space $\mathbf{a}$ and $\mathbf{b}$ respectively, such that $|\mathbf{a}-\mathbf{b}| > r + R$. We wish to find the net force each sphere exerts on the other due to their electric fields; we will assume the first sphere is positively charged and the second is negatively charged, both with volume charge density ρ_c.

As shown previously, we can model each sphere as a point charge. The first sphere will have charge

$$q = \rho_c V = \frac{4}{3}\pi r^3 \rho_c \tag{10}$$

and the second will have charge

$$Q = \frac{4}{3}\pi R^3 \rho_c. \tag{11}$$

Therefore, by direct application of Coulomb's law, the force that the first sphere exerts on the second is

$$\begin{aligned}\mathbf{F} &= \frac{\frac{4}{3}\pi r^3 \rho_c \cdot \frac{4}{3}\pi R^3 \rho_c}{4\pi\varepsilon_0 |\mathbf{a}-\mathbf{b}|^3}(\mathbf{a}-\mathbf{b}) \\ &= \frac{4\pi\rho_c^2 r^3 R^3}{9\varepsilon_0 |\mathbf{a}-\mathbf{b}|^3}(\mathbf{a}-\mathbf{b})\end{aligned}$$

and by Newton's third law, the corresponding force from the second sphere on the first is $-\mathbf{F}$.

V GENERALISED MOTION OF TWO FREE-MOVING CHARGED SPHERES

In order to use equations of motion, we must have mass. Let both spheres have volume mass density ρ_m. Then, by analogy with eqs. (10) and (11), the mass of the first sphere is

$$m = \frac{4}{3}\pi r^3 \rho_m$$

and the mass of the second sphere is

$$M = \frac{4}{3}\pi R^3 \rho_m.$$

Therefore, at any given moment, the acceleration of the first sphere is

$$\ddot{\mathbf{a}} = \frac{\mathbf{F}}{m} = \frac{\frac{4\pi\rho_c^2 r^3 R^3}{9\varepsilon_0|\mathbf{a}-\mathbf{b}|^3}(\mathbf{b}-\mathbf{a})}{\frac{4}{3}\pi r^3 \rho_m} = \frac{\rho_c^2 R^3}{3\varepsilon_0 \rho_m |\mathbf{a}-\mathbf{b}|^3}(\mathbf{b}-\mathbf{a})$$

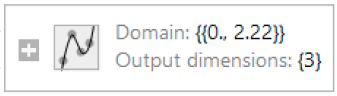
and the acceleration of the second sphere is

$$\ddot{\mathbf{b}} = -\frac{\mathbf{F}}{M} = -\frac{\frac{4\pi\rho_c^2 r^3 R^3}{9\varepsilon_0|\mathbf{a}-\mathbf{b}|^3}(\mathbf{a}-\mathbf{b})}{\frac{4}{3}\pi R^3 \rho_m} = \frac{\rho_c^2 r^3}{3\varepsilon_0 \rho_m |\mathbf{a}-\mathbf{b}|^3}(\mathbf{a}-\mathbf{b}).$$

Thus our problem reduces to a pair of coupled second-order non-linear ordinary differential equations. We won't try to solve these, but rather we'll numerically solve them in Mathematica.

Stripping away all constants for now, the following code numerically generates functions for $\mathbf{a}$ and $\mathbf{b}$ (we set their initial velocities to zero and their initial positions to $\begin{pmatrix}0\\0\\-1\end{pmatrix}$ and $\begin{pmatrix}0\\0\\1\end{pmatrix}$ respectively):

```
In[221]:= Positions =
  NDSolve[{a''[t] == 1/Norm[a[t] - b[t]]^3 (b[t] - a[t]),
    b''[t] == 1/Norm[a[t] - b[t]]^3 (a[t] - b[t]), a[0] == {0, 0, -1},
    b[0] == {0, 0, 1}, a'[0] == {0, 0, 0}, b'[0] == {0, 0, 0}}, {a, b}, {t, 0, 100}]
```

```
Out[221]= {{a → InterpolatingFunction[   ],
  b → InterpolatingFunction[   ]}}
```

Now we'll add the constants back in (with $\rho_c = \lambda$ and $\rho_m = \mu$). Setting $\lambda = \mu = 1$ and the constant of permittivity $\varepsilon = \frac{1}{3}$ for convenience, we allow control of the radii r and R as well as all of the initial positions and velocities through a 'Manipulate' function. For now we'll plot the z-positions of the two spheres from $t = 0$ to t_max. The code is as follows:

```
In[247]:= Manipulate[
  PositionFunc =
   NDSolve[{a''[t] == λ^2 R^3/(3 ϵ μ Norm[a[t] - b[t]]^3) (b[t] - a[t]),
     b''[t] == λ^2 r^3/(3 ϵ μ Norm[a[t] - b[t]]^3) (a[t] - b[t]), a[0] == {sax, say, saz},
     b[0] == {sbx, sby, sbz}, a'[0] == {vax, vay, vaz}, b'[0] == {vbx, vby, vbz}},
    {a, b}, {t, 0, 100}];
  Plot[{Evaluate[a[time][[3]] /. PositionFunc],
    Evaluate[b[time][[3]] /. PositionFunc]}, {time, 0, tmax}], {{r, 1}, 0.1, 10},
  {{R, 1}, 0.1, 10}, {{tmax, 50}, 0.001, 100}, {{sax, 0}, -1, 1}, {{say, 0}, -1, 1},
  {{saz, -1}, -1, 1}, {{sbx, 0}, -1, 1}, {{sby, 0}, -1, 1}, {{sbz, 1}, -1, 1},
  {{vax, 0}, -1, 1}, {{vay, 0}, -1, 1}, {{vaz, 0}, -1, 1}, {{vbx, 0}, -1, 1},
  {{vby, 1}, -1, 1}, {{vbz, 0}, -1, 1}]
```

The resulting manipulate applet is shown in fig. 10.

VI ORBITAL MOTION

We'll now set up a small test charge in orbit around a larger one. Let's make a widget to animate the spheres' motion in three dimensions to aid visualisation.

Based on the previous part, the following code contains an 'Animate' function within a 'Manipulate' function, which allows customised animations based on different starting conditions. I've also included a little plot similar to in fig. 10 to make things easier.

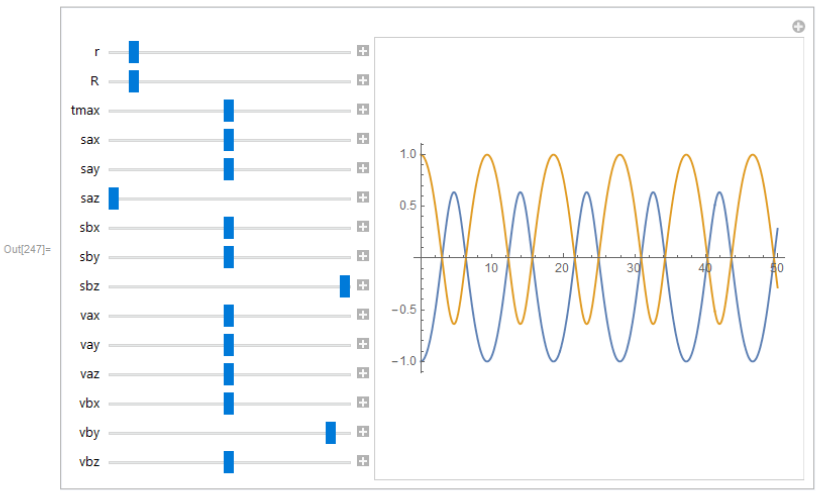


Figure 10: A manipulate function with control over both spheres' radii and starting positions. We already see some nice patterns developing.

```
In[274]:= Manipulate[
 PositionFunc =
  NDSolve[{a''[t] == λ^2 R^3 / (3 ∈ μ Norm[a[t] - b[t]]^3) (b[t] - a[t]), b''[t] == λ^2 r^3 / (3 ∈ μ Norm[a[t] - b[t]]^3) (a[t] - b[t]),
    a[0] == {sax, say, saz}, b[0] == {sbx, sby, sbz}, a'[0] == {vax, vay, vaz}, b'[0] == {vbx, vby, vbz}}, {a, b}, {t, 0, 100}];
 Grid[
  {{Plot[{Evaluate[a[time][[3]] /. PositionFunc], Evaluate[b[time][[3]] /. PositionFunc]}, {time, 0, tmax},
     ImageSize → Small],
    Animate[Graphics3D[{Sphere[Evaluate[a[time] /. PositionFunc], r], Sphere[Evaluate[b[time] /. PositionFunc], R],
       Opacity[0], Sphere[{0, 0, 0}, 2]}, ViewPoint → {1.3, -2.4, 2.}, Boxed → True, FaceGrids → All, ImageSize → Large],
     {time, 0, tmax}]}}], {{r, .5}, 0.01, 5}, {{R, .5}, 0.01, 5}, {{tmax, 50}, 0.001, 100}, {{sax, 0}, -5, 5},
 {{say, 0}, -5, 5}, {{saz, -1}, -5, 5}, {{sbx, 0}, -5, 5}, {{sby, 0}, -5, 5}, {{sbz, 1}, -5, 5}, {{vax, 0}, -5, 5},
 {{vay, 0}, -5, 5}, {{vaz, 0}, -5, 5}, {{vbx, 0}, -5, 5}, {{vby, 1}, -5, 5}, {{vbz, 0}, -5, 5}]
```

We can experiment with all sorts of different radii and sizes and come up with lots of different orbital patterns. I recommend just trying out the attached applet to see the animation working.

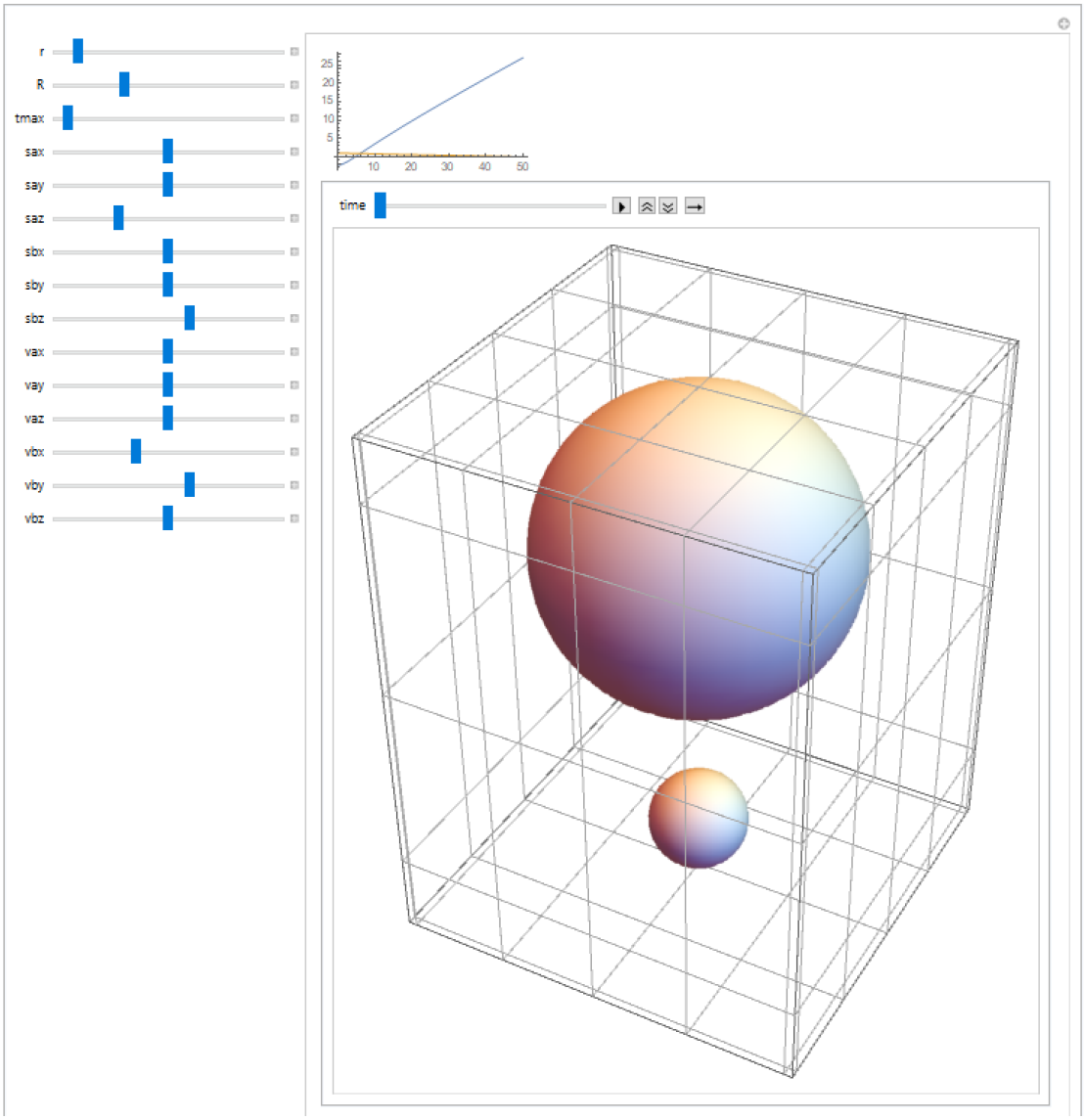


Figure 11: A screenshot of the animation of orbital motion.

Chapter 5

Charged planes and capacitors

Again, while we were studying electric fields with Dr Cheung in September 2018 he set the problem of thinking about charged infinite planes and their relation to capacitors, and this was my response.

Charged planes and capacitors

Damon Falck

June 30, 2018

1 ELECTRIC FIELD AROUND AN INFINITE PLANE OF UNIFORM CHARGE DENSITY

We will consider the electric field $\vec{E}$ at a perpendicular distance h above an infinite horizontal plane of area charge density σ.

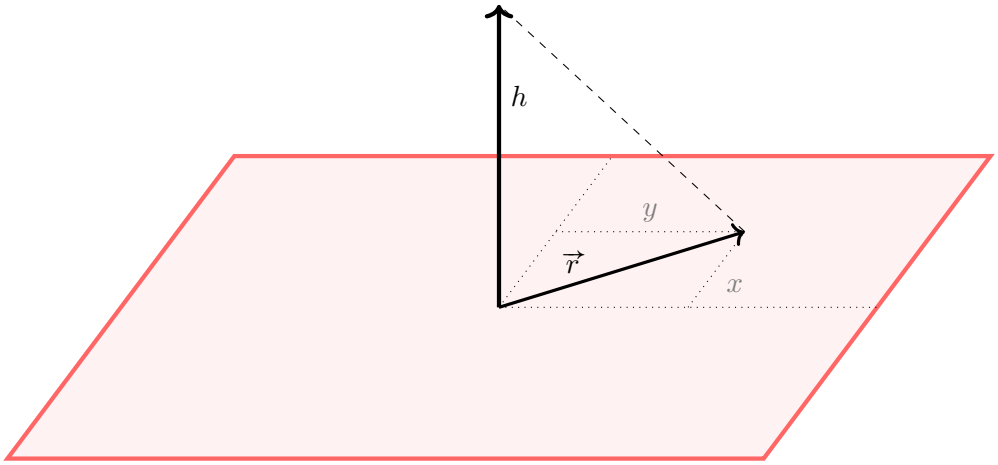


Figure 1: A plane of charge density σ

Noting that all horizontal field components will cancel (as the plane is infinite), we need only consider the vertical component of the field; and assuming the plane is positively charged, the vector $\vec{E}$ will always be in the direction away from the plane.

At a height h directly above the origin, the contribution to the field of a small area $\mathrm{d}x\,\mathrm{d}y$ of the plane at a position $\vec{r} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}$ is given by Coulomb's law as

$$\mathrm{d}E = \frac{\sigma}{4\pi\epsilon_0(|\vec{r}|^2 + h^2)}\,\mathrm{d}x\,\mathrm{d}y. \tag{1}$$

To find the vertical component, we multiply by $\cos\theta$, where θ is the angle shown in fig. 2.

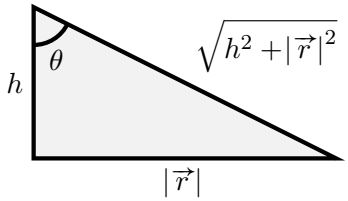


Figure 2: A right triangle showing θ

Clearly this angle is given by

$$\cos\theta = \frac{h}{\sqrt{|\vec{r}^2| + h^2}} \tag{2}$$

and so the vertical component of the field is

$$\mathrm{d}E_z = \frac{\sigma}{4\pi\epsilon_0(|\vec{r}|^2 + h^2)}\,\mathrm{d}x\,\mathrm{d}y \cdot \frac{h}{\sqrt{|\vec{r}^2| + h^2}} \tag{3}$$

$$= \frac{\sigma h}{4\pi\epsilon_0\left(x^2 + y^2 + h^2\right)^{3/2}}\,\mathrm{d}x\,\mathrm{d}y. \tag{4}$$

Using the principle of superposition, the total field at this point is the sum of the field contributions of every small area $\mathrm{d}x\,\mathrm{d}y$ in the plane, and so the total electric field strength is

$$E = \frac{\sigma h}{4\pi\epsilon_0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\mathrm{d}x\,\mathrm{d}y}{\left(x^2 + y^2 + h^2\right)^{3/2}}, \tag{5}$$

which, splitting the plane into four, we can rewrite for simplicity as

$$E = \frac{\sigma h}{\pi\epsilon_0}\int_{0}^{\infty}\int_{0}^{\infty}\frac{\mathrm{d}x\,\mathrm{d}y}{\left(x^2 + y^2 + h^2\right)^{3/2}}. \tag{6}$$

Letting $\tan\alpha = \dfrac{x}{\sqrt{y^2+h^2}}$ so that $\dfrac{\mathrm{d}x}{\mathrm{d}\alpha} = \sqrt{y^2+h^2}\sec^2\alpha$, this becomes

$$E = \frac{\sigma h}{\pi\epsilon_0}\int_0^{\infty}\frac{1}{\left(y^2+h^2\right)^{3/2}}\int_0^{\frac{\pi}{2}}\frac{\mathrm{d}\alpha}{\left(\tan^2\alpha+1\right)^{3/2}}\cdot\sqrt{y^2+h^2}\sec^2\alpha\,\mathrm{d}y \tag{7}$$

$$= \frac{\sigma h}{\pi\epsilon_0}\int_0^{\infty}\frac{1}{y^2+h^2}\int_0^{\frac{\pi}{2}}\cos\alpha\,\mathrm{d}\alpha\,\mathrm{d}y \tag{8}$$

$$= \frac{\sigma h}{\pi\epsilon_0}\int_0^{\infty}\frac{1}{y^2+h^2}\left[\sin\alpha\right]_0^{\frac{\pi}{2}}\mathrm{d}y = \frac{\sigma h}{\pi\epsilon_0}\int_0^{\infty}\frac{\mathrm{d}y}{y^2+h^2}. \tag{9}$$

Similarly for this outer integral, we let $\tan\beta = \dfrac{y}{h}$ so that $\dfrac{\mathrm{d}y}{\mathrm{d}\beta} = h\sec^2\beta$. Then,

$$E = \frac{\sigma h}{\pi\epsilon_0 h^2}\int_0^{\frac{\pi}{2}}\frac{\mathrm{d}\beta}{\tan^2\beta+1}\cdot h\sec^2\beta \tag{10}$$

$$= \frac{\sigma}{\pi\epsilon_0 h}\int_0^{\frac{\pi}{2}} h\,\mathrm{d}\beta \tag{11}$$

$$= \frac{\sigma}{2\epsilon_0}. \tag{12}$$

This is a remarkable result: an infinite charged plane will produce a uniform electric field throughout all space, regardless of a test charge's distance from the plane.

2 ELECTRIC FIELD BETWEEN TWO CHARGED PLANES

As a corollary of the previous result, if we have two parallel infinite planes of *opposite* charge but of the same charge density σ, then the resulting field in between the planes is simply

$$E = \frac{\sigma}{\epsilon_0} \tag{13}$$

by the principle of superposition. On either side of the two planes the fields cancel completely and there is no electric field.

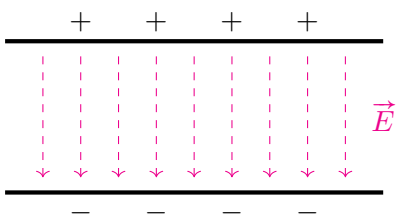


Figure 3: The field between two infinite parallel planes

If the two parallel planes in the previous section are made finite, then as long as the distance d between them is small compared to their area, the electric field between them can still be approximated by $E = \frac{\sigma}{\epsilon_0}$. This is the case in a standard parallel plate capacitor.

Let each plate have area A and store a total charge Q. So, the charge density is

$$\sigma = \frac{Q}{A} \tag{14}$$

and therefore the electric field between the two plates is

$$E = \frac{Q}{A\epsilon_0}. \tag{15}$$

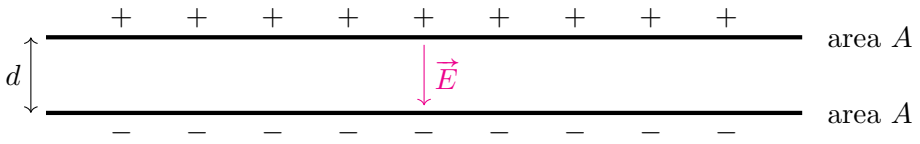


Figure 4: A parallel plate capacitor, each with total charge Q

The potential difference V across the plates is defined as the work done per unit charge against the electric field to move a test charge from one plate to the other. So, with a small test charge q,

$$V = \frac{Fd}{q}, \tag{16}$$

where F is the force exerted by the electric field on the test charge. However, by the definition of an electric field of strength E,

$$E = \frac{F}{q} \tag{17}$$

and so our potential difference becomes

$$V = Ed. \tag{18}$$

Now substituting in our electric field value from eq. (15),

$$V = \frac{Qd}{A\epsilon_0}. \tag{19}$$

Capacitance C is defined as the charge stored per unit potential difference, and so

$$C = \frac{Q}{V} = \frac{QA\epsilon_0}{Qd} = \frac{A\epsilon_0}{d}. \tag{20}$$

Finally, the use of a dielectric (rather than a vacuum) between the plates requires the adjustment of the constant of permittivity, and so following convention we write

$$C = \frac{A\epsilon_0\epsilon_r}{d} \tag{21}$$

where ϵ_r is the relative permittivity of the material used. This is a well-known equation for the capacitance of a parallel plate capacitor.

Chapter 6

C3 coursework

This is pretty self-explanatory; my numerical methods coursework for the C3 module.

An Investigation into Methods for the Numerical Solution of Equations

(C3 Coursework)

Damon Falck

October 2017

Contents

1 Introduction

In this coursework I will investigate the numerical solution of equations by means of:

- Systematic search for a change of sign using interval bisection
- The Newton-Raphson fixed-point iteration method
- Rearrangement of the equation $f(x) = 0$ into the form $x = g(x)$

For each method I will demonstrate a successful application and then discuss under what circumstances the method will fail to find a root. I will also directly compare the three methods' advantages and disadvantages by using them all to solve one equation.

All equations that I use in this coursework will be analytically insoluble quintic polynomials.

2 The Change of Sign Method

2.1 Application of the Change of Sign Method

The *intermediate value theorem* says that if a continuous function takes a positive value at one end of a given interval and a negative value at the other end, then at some point in that interval the function will have a root.

As a consequence of this, we can search for a root of a function by looking for a smaller and smaller interval over which the function changes sign.

Consider the function

$$f(x) = 7x^5 - 5x^4 + 3x^3 + 15x^2 - 10x + 32 = 0.$$

Testing reveals $f(-10) = -751,368 < 0$ and $f(10) = 654,432 > 0$ and so $f(x)$ must have a root on the interval $(-10, 10)$. We then bisect this interval and test $f(0)$, which evaluates to $32 > 0$, so we have narrowed our root down to the interval $(-10, 0)$. Repeatedly applying this procedure generates a table of values as in table 1.

Lower bound a	Upper bound b	Interval midpoint $\frac{a+b}{2}$	$f\left(\frac{a+b}{2}\right)$	Change of sign over
-10	10	0	32	Lower interval
-10	0	-5	$-24,918$	Upper interval
-5	0	-2.5	-775.0	Upper interval
-2.5	0	-1.25	28.5	Lower interval
-2.5	-1.25	-1.875	-140.3	Upper interval
-1.875	-1.25	-1.5625	-22.2	Upper interval
-1.5625	-1.25	-1.40625	9.3	Lower interval
-1.5625	-1.40625	-1.484375	-4.6	Upper interval
-1.484375	-1.40625	-1.44531	2.7	Lower interval
-1.484375	-1.44531	-1.46484	-0.83	Upper interval
-1.46484	-1.44531	-1.45508	1.0	Lower interval
-1.46484	-1.45508	-1.45996	0.08	Lower interval
-1.46484	-1.45996	-1.4624	-0.37	Upper interval
-1.4624	-1.45996	-1.46118	-0.13	Upper interval
-1.46118	-1.45996	-1.46057	-0.02	Upper interval
-1.46057	-1.45996	-1.46027	0.03	Lower interval
-1.46057	-1.46027	-1.46042	0.004	Lower interval
-1.46057	-1.46042	-1.46049	-0.009	Upper interval
-1.46049	-1.46042	-1.460455	-0.002	Upper interval
-1.460455	-1.46042	-1.460438	0.0003	Lower interval
-1.460455	-1.460438	-1.460447	-0.001	Upper interval
-1.460447	-1.460438	-1.460443	-0.0005	Upper interval

Table 1: *Repeated interval bisection to reduce the error of our root.*

Thus if our root is x_0 we know $x_0 \in (-1.460443, -1.460438)$. The midpoint of this interval is -1.4604405 and thus we can say

$$x_0 = -1.4604405 \pm 0.0000025.$$

In other words, $f(x)$ has a root at $x = 1.46044$ to 5 decimal places.

The procedure can be visualised graphically as shown in fig. 1.

2.2 Failure of the Change of Sign Method

The change of sign method does not always successfully reveal a root. Consider the function

$$f(x) = 25x^5 - 20x^4 - 494x^3 + 512x^2 + 1504x - 1792,$$

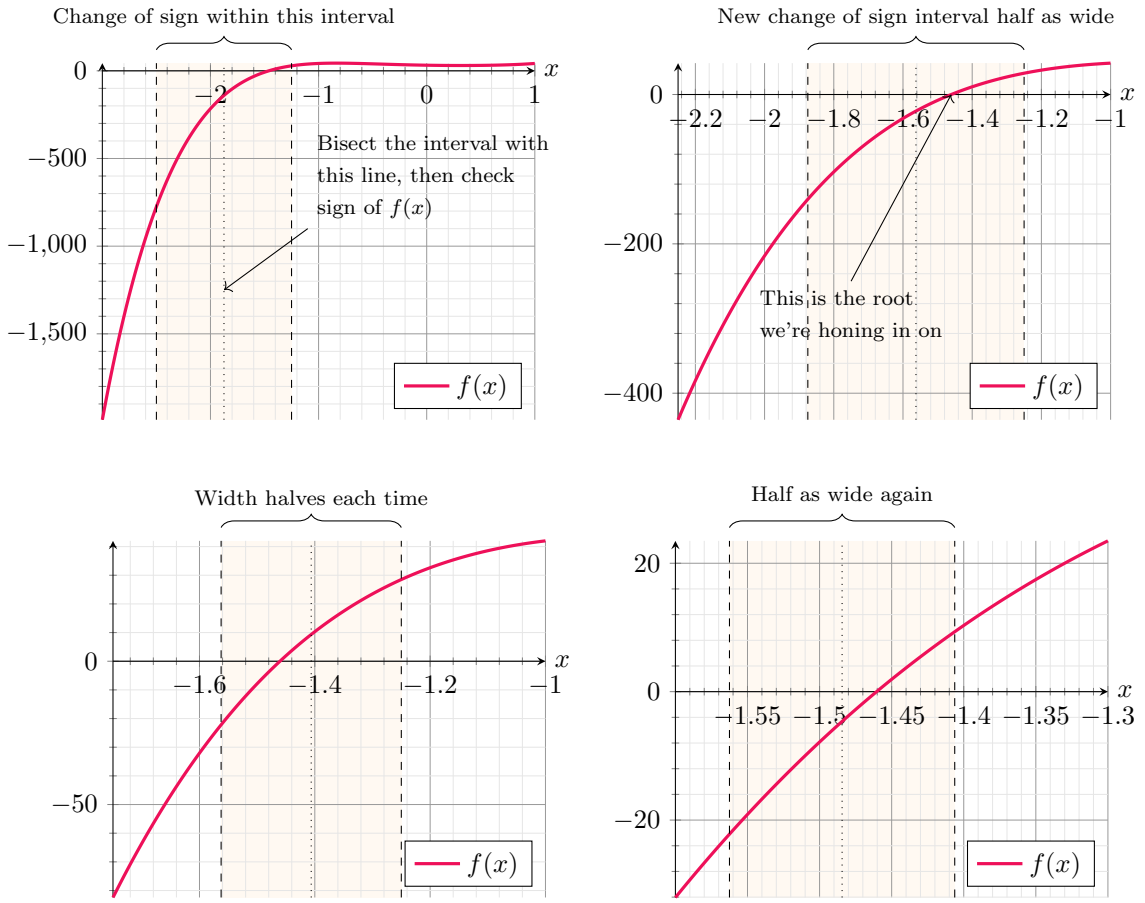


Figure 1: *This shows four levels of 'zoom' when applying the interval bisection method, corresponding to the fifth through eighth rows of table 1.*

for instance. Looking at the interval $(1, 2)$, we see that $f(1) = -265 < 0$ and $f(2) = -208 < 0$. There is no change of sign, suggesting that there is no root on this interval.

However, plotting the function reveals that there are actually two roots here, as can be seen in fig. 2. The curve crosses the x-axis *twice* between $x = 1$ and $x = 2$, so that there is no overall change of sign. This is one of the situations in which the change of sign method will fail to find a root of an equation.

Change of sign may also fail when there is a discontinuity in the function or when the function is tangent to the x-axis (at a repeated root).

3 The Newton-Raphson Method

3.1 Application of the Newton-Raphson Method

The next two methods will use fixed-point iterations — techniques that involve finding a single estimate for the root as opposed to an interval within which it must lie.

For the Newton-Raphson method we start with an estimate for the root, and we draw a tangent to the curve at this point. Where the tangent intersects the x-axis we take the tangent again and repeat this process, converging towards the root. This iteration takes the form

$$x_{n+1} \coloneqq x_n - \frac{f(x_n)}{f'(x_n)}. \tag{1}$$

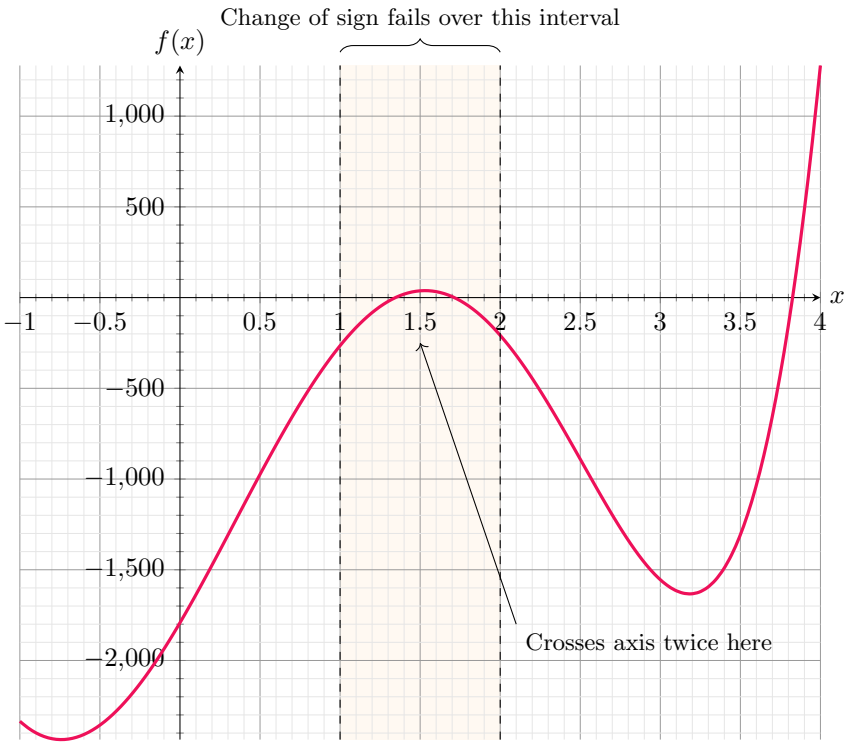


Figure 2: *The new function has two roots between the neighbouring integers $x = 1$ and $x = 2$, so there is no change of sign over this interval.*

Consider the function

$$f(x) = 5x^5 - 36x^4 + 50x^3 - 5x^2 + 35x - 50. \tag{2}$$

To apply the iteration we will need to know the derivative of the function, which in this case is

$$f'(x) = 25x^4 - 144x^3 + 150x^2 - 10x + 35.$$

Therefore, for this specific function the Newton-Raphson iteration given by eq. (1) is

$$x_{n+1} \coloneqq x_n - \frac{5x_n^5 - 36x_n^4 + 50x_n^3 - 5x_n^2 + 35x_n - 50}{25x_n^4 - 144x_n^3 + 150x_n^2 - 10x_n + 35}. \tag{3}$$

We will now try and find a root of this function. After experimenting a little with the function's behaviour, we decide to use $x_0 = 3$ as a starting value. Therefore, repeatedly applying the iteration eq. (3), we start to generate increasingly accurate estimates of a root:

$$\begin{aligned}
x_1 &\coloneqq x_0 - \frac{5x_0^5 - 36x_0^4 + 50x_0^3 - 5x_0^2 + 35x_0 - 50}{25x_0^4 - 144x_0^3 + 150x_0^2 - 10x_0 + 35} \\
&= 3 - \frac{5(3)^5 - 36(3)^4 + 50(3)^3 - 5(3)^2 + 35(3) - 50}{25(3)^4 - 144(3)^3 + 150(3)^2 - 10(3) + 35} \\
&= 2.329 \\
\implies x_2 &\coloneqq (2.329) - \frac{5(2.329)^5 - 36(2.329)^4 + 50(2.329)^3 - 5(2.329)^2 + 35(2.329) - 50}{25(2.329)^4 - 144(2.329)^3 + 150(2.329)^2 - 10(2.329) + 35} \\
&= 2.017 \\
\implies x_3 &= 1.888
\end{aligned}$$

and so on. The process is continued until the difference $x_{n+1} - x_n$ is acceptably small. The table of values given by repeatedly applying the Newton-Raphson iteration starting from $x_0 = 3$ is shown in table 2.

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	3	-341	-508	2.32874
1	2.32874	-80.4686	-258.153	2.01703
2	2.01703	-18.3824	-142.789	1.88829
3	1.88829	-2.74990	-100.740	1.86100
4	1.86100	-0.114816	-92.3603	1.85975
5	1.85975	-0.000234292	-91.9834	1.85975

Table 2: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (2) using $x_0 = 3$ as the starting value.*

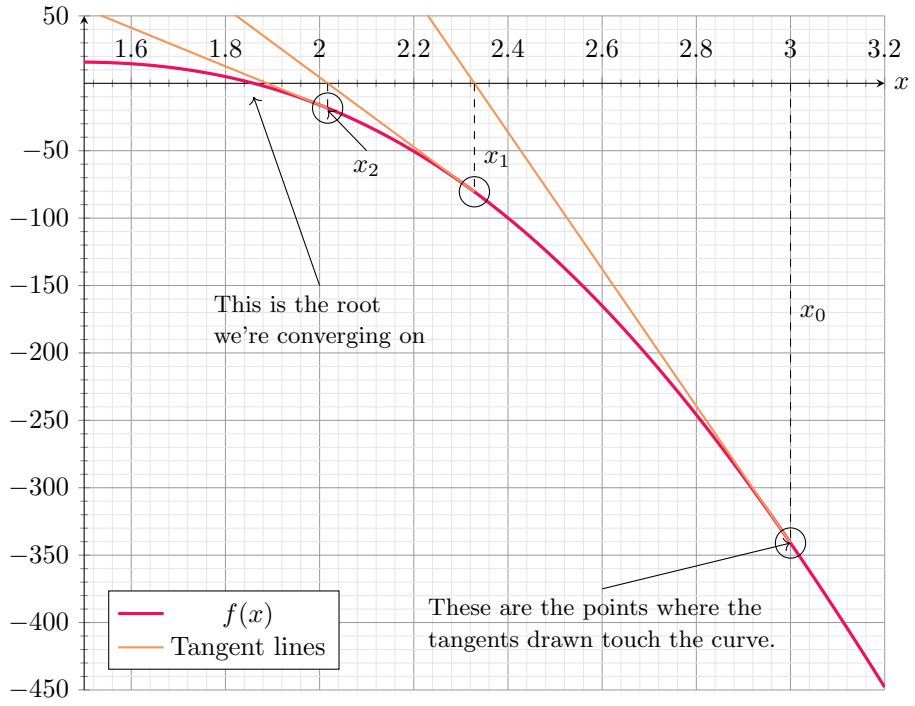


Figure 3: *Visually, the Newton-Raphson method works by repeatedly finding the intersection between the tangent to the curve and the x-axis.*

As can be seen, we converge to a value of 1.85975 after only six iterations. The location of this root using the Newton-Raphson method is illustrated graphically in fig. 3.

To establish error bounds on this root, we must demonstrate a change of sign. If this root is $r = 1.85975$, we can test the function's value just under and just above this value to confirm the existence of a root between the two.

Testing reveals that $f(1.859745) = 0.00045 > 0$ and $f(1.859755) = -0.00047 < 0$, hence there *is* a change of sign and so we can confidently say that $r \in (1.859745, 1.859755)$. That is,

$$r = 1.859750 \pm 0.000005$$

or in other words, $r = 1.85975$ accurate to 5 decimal places.

Now we'll apply the Newton-Raphson method to also find the other two roots of this function. Starting at $x = 1$, we generate a table of values as given in table 3 and starting at $x = 6$, we generate a table of values as given in table 4.

Therefore, our other two roots are to five decimal places $x = 1.01798$ and $x = 5.32253$.

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	1	-1	56	1.01786
1	1.01786	-0.00694809	55.2018	1.01798
2	1.01798	$-3.70375 \cdot 10^{-7}$	55.2108	1.01798

Table 3: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (2) using $x_0 = 1$ as the starting value.*

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	6	3004	6671	5.54969
1	5.54969	709.243	3700.69	5.35804
2	5.35804	94.5182	2741.96	5.32357
3	5.32357	2.69050	2586.70	5.32253
4	5.32253	0.00239622	2582.09	5.32253

Table 4: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (2) using $x_0 = 6$ as the starting value.*

3.2 Failure of the Newton-Raphson Method

There are some situations in which the Newton-Raphson method will fail to find a root of an equation. For instance, consider the function

$$f(x) = 500x^5 + 2600x^4 + 5520x^3 + 6364x^2 + 4265x + 1344. \tag{4}$$

Plotting this function as in fig. 4 reveals that there are three roots on the interval $(-2, -1)$. However, without having seen such a graph, no reasonable integer starting values will converge to the middle root under the Newton-Raphson iteration.

Figure 4: *A plot of the quintic function $f(x) = 500x^5 + 2600x^4 + 5520x^3 + 6364x^2 + 4265x + 1344$.*

The derivative of $f(x)$ is

$$f'(x) = 2500x^4 + 10400x^3 + 16560x^2 + 12728x + 4265,$$

which will allow us to apply the Newton-Raphson iteration. Take $x = -1$ as a starting value. Table 5 shows that from this starting value, the iteration converges to -1.17906.

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	-1	23	197	-1.11675
1	-1.11675	5.32363	107.400	-1.16632
2	-1.16632	0.871372	72.6449	-1.17831
3	-1.17831	0.0482528	64.6281	-1.17906
4	-1.17906	0.000184170	64.1349	-1.17906

Table 5: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (4) using $x_0 = -1$ as the starting value.*

Now take $x = -2$ as a starting value. Table 6 shows that for this starting value, the iteration converges to -1.57476.

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	-2	-290	1849	-1.84316
1	-1.84316	-90.3486	795.396	-1.72957
2	-1.72957	-27.3038	352.075	-1.65202
3	-1.65202	-7.80633	164.173	-1.60447
4	-1.60447	-1.96846	85.5427	-1.58146
5	-1.58147	-0.349589	56.0277	-1.57522
6	-1.57522	-0.0224098	48.9054	-1.57476
7	-1.57476	-0.000116634	48.3966	-1.57476

Table 6: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (4) using $x_0 = -2$ as the starting value.*

These are the only two reasonable integer starting values from which one might expect to converge to the middle root; however, neither do: starting from -1 converges to the upper of the three roots and starting from -2 converges to the lower of the three roots. Any other integer starting values in this area (e.g. $0, -3$) will also converge to one of the outer roots. Figure 5 illustrates why this occurs graphically — it can be seen that in order to converge to the middle root (the one near -1.45), a non-integer starting value *between the two turning points of the curve* must be chosen. This would only be done if the existence of the middle root was already known.

This is one example of a situation where the Newton-Raphson method will fail. Failure may also occur if the function is discontinuous or not defined over the whole of $\mathbb{R}$, or if a starting point is chosen at a stationary point or in some other inconvenient place.

4 The $x = g(x)$ Rearrangement Method

4.1 Application of the $x = g(x)$ Rearrangement Method

A second method of using fixed-point iterations to find roots of an equation is to manually rearrange the equation into a different form which can then be iterated upon. Consider the function

$$f(x) = 2x^5 - 10x^4 + 4x^3 + 18x^2 - 13x + 16. \tag{5}$$

To solve the equation $f(x) = 0$, we wish to somehow rearrange it into the form $x = g(x)$. We can then perform the iteration $x_{n+1} \coloneqq g(x_n)$. One possibly way to make this rearrangement is

$$x = \left(\frac{-2x^5 + 10x^4 - 18x^2 + 13x - 16}{4} \right)^{\frac{1}{3}} \tag{6}$$

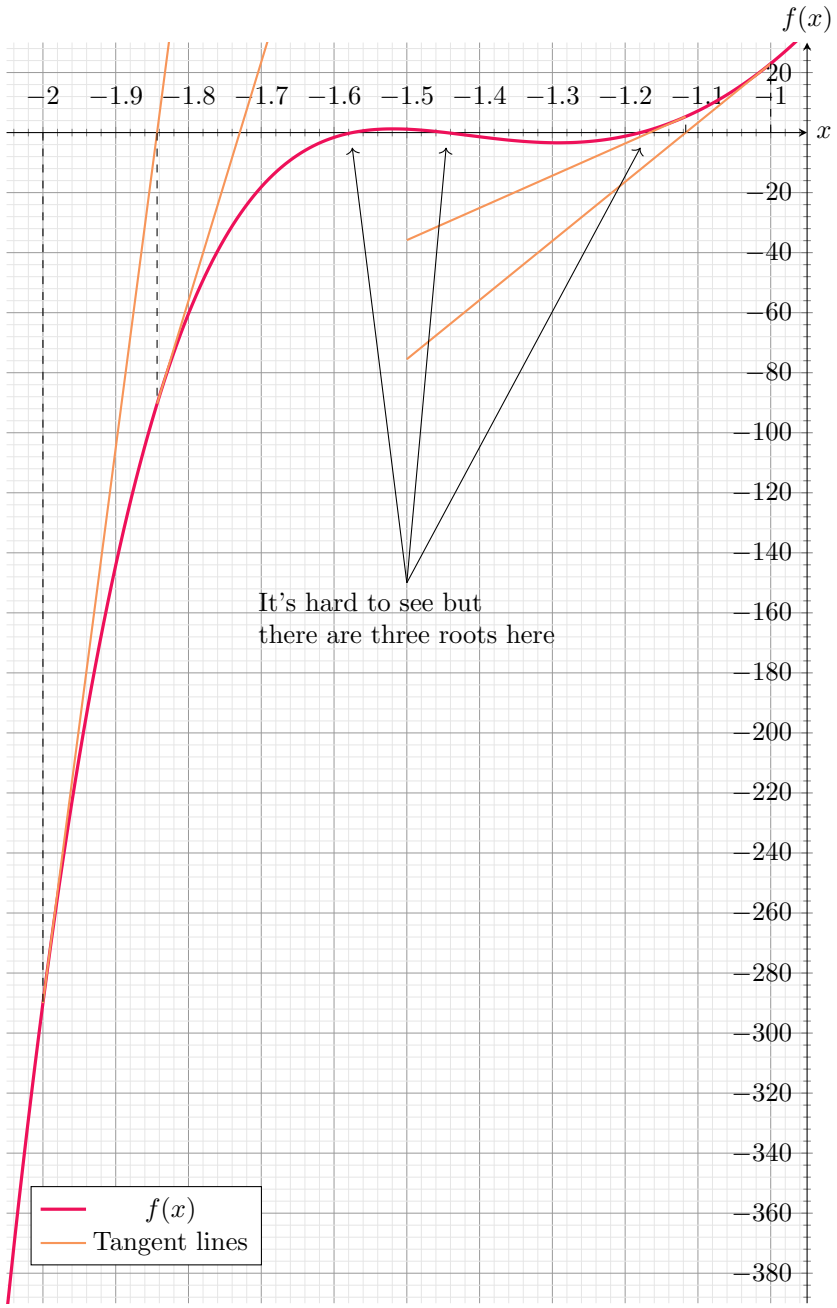


Figure 5: *No integer starting value will converge to the middle root of this function, as there are three roots on the interval* $(-2, -1)$.

where $x^{\frac{1}{3}}$ is the real-valued cube root of x. For this rearrangement, we can therefore create the iteration

$$x_{n+1} = \left(\frac{-2x_n^5 + 10x_n^4 - 18x_n^2 + 13x_n - 16}{4} \right)^{\frac{1}{3}}. \tag{7}$$

Using $x_0 = 3$ as a starting value, repeatedly applying this iteration generates the table of values in table 7.

This iteration was quite slow to converge, but did succeed in locating a fixed point — a root — at $x = 4.01221$ to five decimal places.

This process can also be visualised rather nicely. Figures 6 and 7 show a plot of both $y = x$ and $y = g(x)$, where $g(x)$ is the right hand side of eq. (6). Any roots (when $f(x) = 0$) will of course occur

n	x_n
0	3
1	3.589527
2	4.056795
3	3.976107
4	4.035820
5	3.994088
6	4.024654
7	4.002928
8	4.018733
9	4.007415
10	4.015616
11	4.009723
12	4.013984
13	4.010916
14	4.013131
15	4.011535
16	4.012687
17	4.011857
18	4.012456
19	4.012024
20	4.012336
21	4.012111
22	4.012273
23	4.012156
24	4.012240
25	4.012179
26	4.012223
27	4.012192
28	4.012215
29	4.012198
30	4.012210
31	4.012201
32	4.012208
33	4.012203
34	4.012206
35	4.012204
36	4.012206
37	4.012204
38	4.012205
39	4.012205
40	4.012205

Table 7: *Applying the $x = g(x)$ iteration in eq. (7) with a starting value of 3.*

at the intersection of these two curves (when $x = g(x)$). To repeatedly apply the iteration, we start at the first value of x and find $g(x)$ here. Then we make this the new x-value and repeat. This leads to a staircase/cobweb-like diagram as shown.

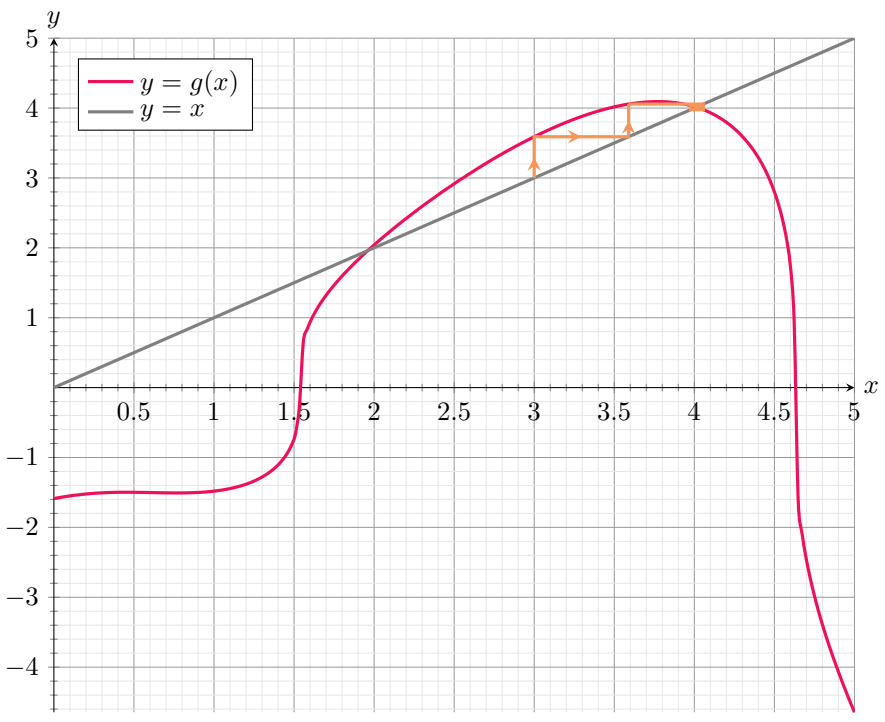


Figure 6: *The $x = g(x)$ iteration in eq. (7) is applied, where $g(x) = \left(\frac{-2x^5+10x^4-18x^2+13x-16}{4}\right)^{\frac{1}{3}}$.*

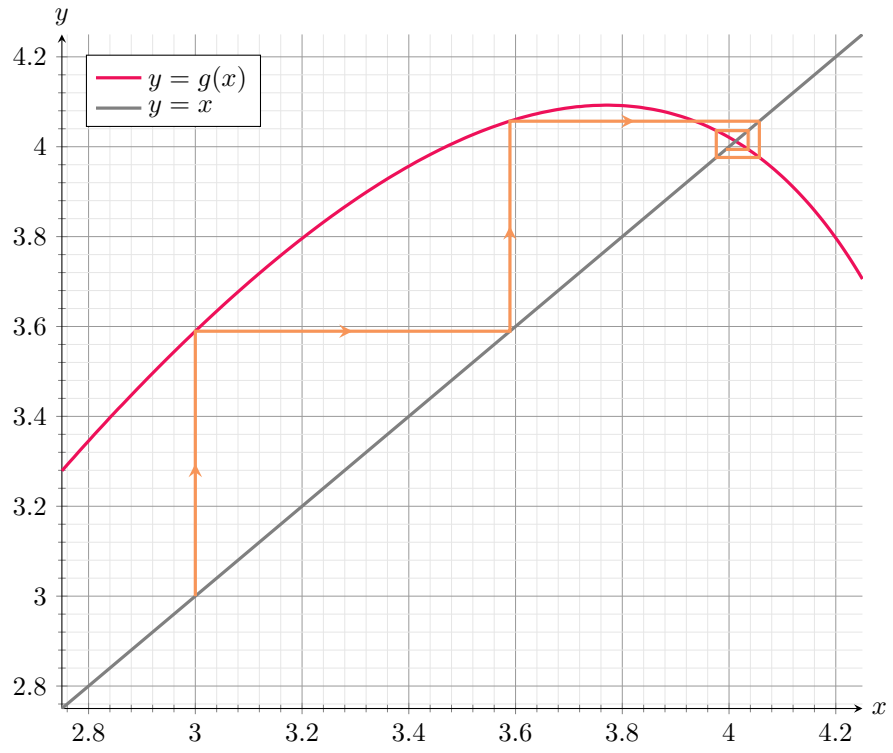


Figure 7: *A closer look at what's going on.*

In this case, it can be seen that the gradient of $g(x)$ is just a little more than -1 at the root. That is,

$$g'(4.01221) = -0.721.$$

Slightly nearer to our starting point the gradient is actually positive, as there is a turning point:

$$g'(3.5) = 0.524.$$

This is why in fig. 7 the orange line switches from a 'staircase' pattern to a 'cobweb' pattern. In both of these cases, however, $-1 < g'(x) < 1$. That is,

$$\left|g'(x)\right| < 1.$$

4.2 Failure of the $x = g(x)$ Rearrangement Method

Some rearrangements will fail to find particular roots of an equation. For instance, using the same function as in eq. (5), another rearrangement of $f(x) = 0$ into the form $x = g(x)$ is

$$x = \frac{2x^5 - 10x^4 + 4x^3 + 18x^2 + 16}{13}, \tag{8}$$

giving rise to the iteration

$$x_{n+1} \coloneqq \frac{2x_n^5 - 10x_n^4 + 4x_n^3 + 18x_n^2 + 16}{13}. \tag{9}$$

Using the same starting value of $x_0 = 3$ as before, this iteration diverges to negative infinity as shown in table 8.

We know the root we're trying to find has a value of 4.01221 and so let's try other starting values close to this root. Using $x_0 = 4$ as a starting value diverges too, as does $x_0 = 5$, as shown in tables 9 and 10 respectively.

Figures 8 and 9 illustrate this failure for starting values of both 3 and 5.

n	x_n
0	3
1	-2.92
2	-83.6
3	$-6.67 \cdot 10^{8}$
4	$-2.02 \cdot 10^{43}$
5	$-5.22 \cdot 10^{215}$

Table 8: *Applying the $x = g(x)$ iteration in eq. (9) with a starting value of 3 fails to find a fixed point.*

n	x_n
0	4
1	3.69
2	-1.80
3	-6.98
4	$-4.40 \cdot 10^{3}$
5	$-2.53 \cdot 10^{17}$
6	$-1.60 \cdot 10^{86}$

Table 9: *The same thing happens when applying the $x = g(x)$ iteration in eq. (9) with a starting value of 4.*

n	x_n
0	5
1	74.3
2	$3.25 \cdot 10^{8}$
3	$5.60 \cdot 10^{41}$
4	$8.45 \cdot 10^{207}$

Table 10: *Applying the $x = g(x)$ iteration in eq. (9) with a starting value of 5 causes a divergence again, this time towards positive infinity.*

As there are no other roots in this immediate neighbourhood, it is becoming apparent that the iteration in eq. (9) will not find this root from *any* starting point. Why is this?

Testing reveals that at the root in question ($x = 4.01221$), the function $g(x)$ (the right hand side of eq. (8)) has a gradient of

$$g'(4.0122) = 26.58.$$

This is *far* outside our established limit of $\left|g'(x)\right| < 1$, and it explains why this condition is necessary: if $g'(x) > 1$ near the root then our orange line will just be deflected off to infinity, and if $g'(x) < -1$, it will be deflected off towards negative infinity. At every step of the iteration the distance from the root is multiplied by approximately $\left|g'(x)\right|$ and so if $\left|g'(x)\right| > 1$ we will just get further and further from the root; we will diverge.

In fact, a root that *can* be found by a particular iteration is called a 'stable' fixed point, and one that *cannot* be found by that iteration is called an 'unstable' fixed point.

5 Comparison of Numerical Methods

5.1 Application of each method to find the same root

Now that the use of each of the three methods has been covered, we'll apply them all to find the same root of an equation in order to compare their efficiency.

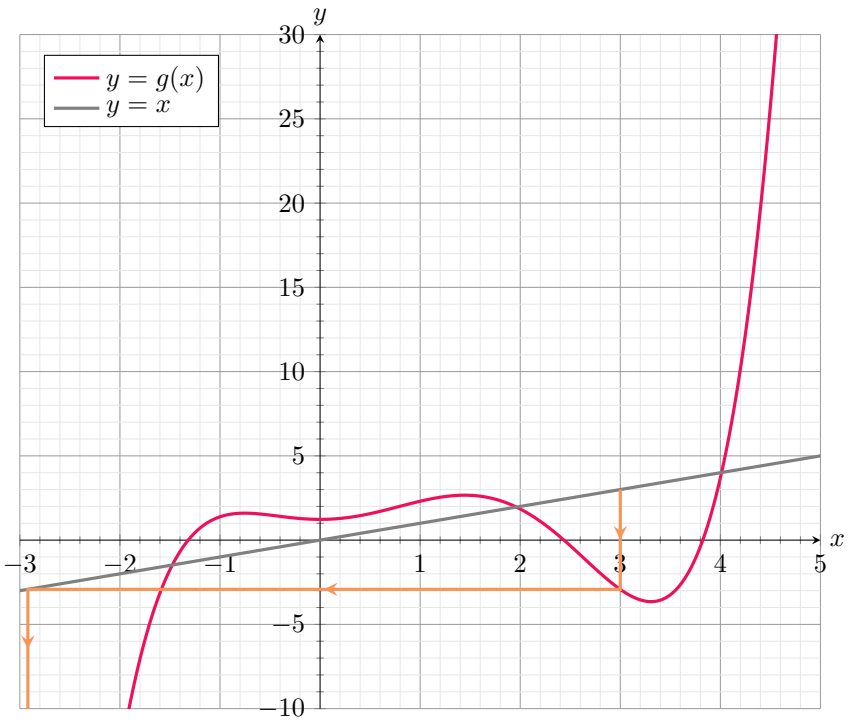


Figure 8: *The $x = g(x)$ iteration in eq. (9) is applied with a starting value of $x = 3$, where $g(x) = \frac{2x^5-10x^4+4x^3+18x^2+16}{13}$.*

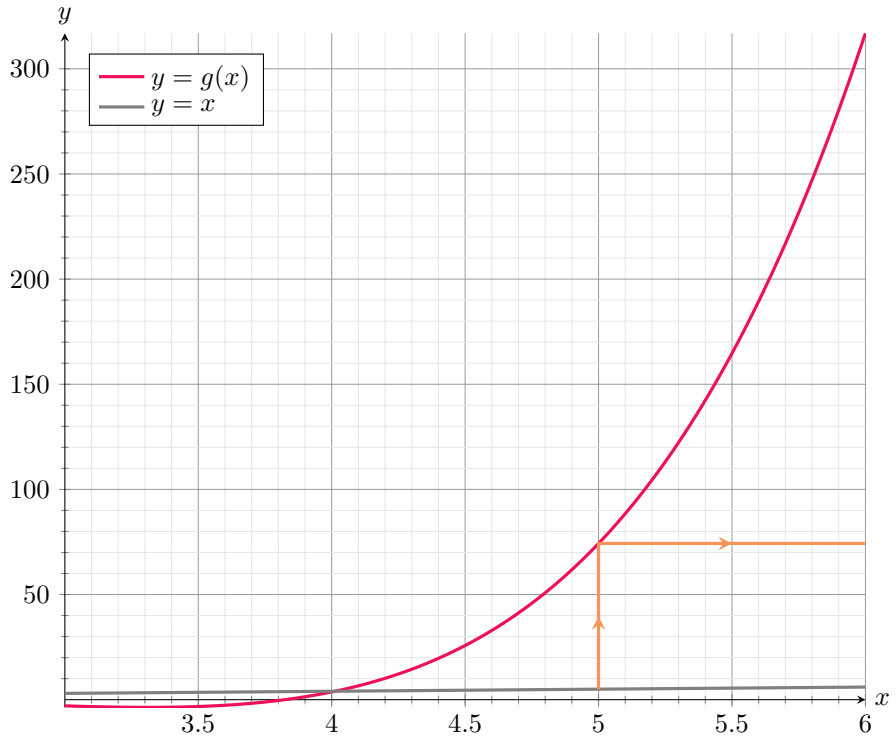


Figure 9: *The $x = g(x)$ iteration in eq. (9) is applied with a starting value of $x = 5$, where $g(x) = \frac{2x^5-10x^4+4x^3+18x^2+16}{13}$.*

The function that we'll use will be the same as in section 4:

$$f(x) = 2x^5 - 10x^4 + 4x^3 + 18x^2 - 13x + 16.. \tag{10}$$

We have already found the root at $x = 4.01221$ using the iteration

$$x_{n+1} = \left(\frac{-2x_n^5 + 10x_n^4 - 18x_n^2 + 13x_n - 16}{4} \right)^{\frac{1}{3}} \tag{11}$$

as explained in the previous section. Since we will have to start to the right of the turning point for the Newton-Raphson iteration, we quickly use the same rearrangement but a starting value of 3.5 to find this root again. The table of values produced is shown in table 11.

n	x_n
0	3.5
1	4.016533
2	4.009047
3	4.014463
4	4.010566
5	4.013382
6	4.011354
7	4.012818
8	4.011762
9	4.012524
10	4.011975
11	4.012371
12	4.012085
13	4.012291
14	4.012143
15	4.012250
16	4.012173
17	4.012228
18	4.012188
19	4.012217
20	4.012196
21	4.012211
22	4.012200
23	4.012208
24	4.012203
25	4.012207
26	4.012204
27	4.012206
28	4.012204
29	4.012205
30	4.012205

Table 11: *Applying the $x = g(x)$ iteration in eq. (11) with a starting value of 3.5.*

Let's now find the same root using the Newton-Raphson method. The derivative of our function is

$$f'(x) = 10x^4 - 40x^3 + 12x^2 + 36x - 13.$$

So, our Newton-Raphson iteration is

$$\begin{aligned} x_{n+1} &\coloneqq x_n - \frac{f(x_n)}{f'(x_n)} \\ &= x_n - \frac{2x_n^5 - 10x_n^4 + 4x_n^3 + 18x_n^2 - 13x_n + 16}{10x_n^4 - 40x_n^3 + 12x_n^2 + 36x_n - 13}. \end{aligned}$$

Applying this recursively given the same starting value of 3.5 yields the table of values given in table 12.

n	x_n	$f(x_n)$	$f'(x_n)$	x_{n+1}
0	3.5	-87.6875	45.625	5.421918
1	5.421918	1841.460	2801.338	4.764568
2	4.764568	552.6690	1257.902	4.325210
3	4.325210	147.8503	630.3359	4.090652
4	4.090652	28.57686	397.1165	4.018691
5	4.018691	2.173143	337.6022	4.012254
6	4.012254	0.01633323	332.5343	4.012205
7	4.012205	$9.5 \cdot 10^{-7}$	332.4957	4.012205

Table 12: *Applying the Newton-Raphson iteration to the function $f(x)$ given in eq. (10) using $x_0 = 3.5$ as the starting value.*

We arrive at the same value of 4.01221 to five decimal places, after only seven iterations!

Finally, we will locate the root using the change of sign method with interval bisection. Starting with an interval of $(3, 5)$, we see that $f(3) = -77 < 0$ and $f(5) = 901 > 0$, implying that there is definitely a root within this interval. Now, repeatedly bisecting the interval checking the value of $f(x)$ leads to table 13.

Lower bound a	Upper bound b	Interval midpoint $\frac{a+b}{2}$	$f\left(\frac{a+b}{2}\right)$	Change of sign over
3	5	4	-4	Upper interval
4	5	4.5	276.4	Lower interval
4	4.5	4.25	103.6	Lower interval
4	4.25	4.125	42.7	Lower interval
4	4.125	4.0625	17.7	Lower interval
4	4.0625	4.03125	6.48	Lower interval
4	4.03125	4.015625	1.14	Lower interval
4	4.015625	4.007813	-1.45	Upper interval
4.007813	4.015625	4.011719	-0.161	Upper interval
4.011719	4.015625	4.013672	0.489	Lower interval
4.011719	4.013672	4.012696	0.163	Lower interval
4.011719	4.012696	4.012208	0.001011	Lower interval
4.011719	4.012208	4.011964	-0.080	Upper interval
4.011964	4.012208	4.012086	-0.039	Upper interval
4.012086	4.012208	4.012147	-0.019	Upper interval
4.012147	4.012208	4.012178	-0.0089	Upper interval
4.012178	4.012208	4.012193	-0.0040	Upper interval
4.012193	4.012208	4.012201	-0.0013	Upper interval
4.012201	4.012208	4.0122045	-0.000153	Upper interval
4.0122045	4.012208	4.0122063	0.00045	Lower interval
4.0122045	4.0122063	4.0122054	0.000146	Lower interval
4.0122045	4.0122054	4.01220495	$-3.4 \cdot 10^{-6}$	Upper interval

Table 13: *Repeated interval bisection to find the root of $f(x)$ in the interval $(3, 5)$.*

The repeated bisection tells us that the root is on the interval $(4.0122050, 4.0122054)$ and therefore the root is 4.01221 accurate to five decimal places.

5.2 Discussion of relative efficiency

We have now succeeded in finding a root of the function in eq. (10) at 4.01221 to five decimal places using each of the three methods covered.

5.2.1 Speed of convergence

The iteration used for the $x = g(x)$ rearrangement method took 31 iterations to converge to the root with the desired accuracy, whereas the change of sign method starting on an approximately two-integer wide interval around the root took 22 iterations. The Newton-Raphson method, in contrast, took only 8 iterations to accurately find the root. In this respect, the Newton-Raphson method is far more efficient than either of the other two in computing a root to a desired level of accuracy in the least time.

Of course, some calculations are more computationally difficult than others and I haven't taken this into consideration.

5.2.2 Ease of use

The change of sign method is the simplest method conceptually, however the computation of each step requires an element of 'if-then' logic which the other two methods don't. In this way, actually automating the task was more difficult than the other two, and combined with the slow convergence this makes the method more difficult to use than the other two in my opinion. Using a different algorithm such as decimal search or linear interpolation may have alleviated this problem.

The primary difficulty in applying the rearrangement method was choosing a suitable $x = g(x)$ rearrangement such that the modulus of $g'(x)$ near the root is less than 1. This took considerable trial and error, but once a suitable iteration had been decided upon the actual implementation was very easy.

The Newton-Raphson method, in contrast, was incredibly quick to apply. While it does require the initial computation of the derivative of the function, which for non-polynomials may be more difficult, the application of the iterative formula is very simple and efficient.

Both of the fixed-point iteration methods require choosing a starting value, while the change of sign method requires choosing a starting interval. The difficulty in making these choices is roughly similar.

6 Conclusion

In conclusion, for most day-to-day numerical methods needs, the Newton-Raphson method is the fastest and easiest to use out of these three. There is however room for improvement on each algorithm, and the nature of the equation being solved may make a significant difference to this result.

Chapter 7

Mock DE coursework

This was the mock differential equations coursework set for us by Mr Dales.

Modelling the velocity of an aircraft after landing

(Mock DE Coursework)

Damon Falck

January 2018

Contents

1 Introduction

We are given the following situation to model:

> *An aeroplane of mass* 120 000 kg *comes in to land at a runway. After touchdown air resistance slows it initially, and then when it is slow enough this is augmented by a constant force from the wheel brakes.*

We are to predict the aeroplane's velocity v $\mathrm{m\,s^{-1}}$ as a function of time t s after touchdown. We will first set up an initial model of air resistance and use this to make preliminary predictions, and then based on comparison with the given data we will create a second, improved model.

2 Setting up an initial model

2.1 Assumptions

We will begin by making the following simplifying assumptions about the situation:

- The magnitude of the force due to air resistance is proportional to the velocity of the aeroplane.
- Any frictional forces other than air resistance acting horizontally on the aeroplane are negligible.
- The aeroplane's mass is constant (we assume no fuel is used up during its landing).
- The runway is horizontal so that there is no vertical component to the velocity.
- The structure of the aeroplane does not change throughout the landing (e.g. flaps could extend) so that the air resistance constant of proportionality does not change.

These assumptions are listed in approximate order of importance. While most of the assumptions are necessary just for us to get started with a model, the first assumption is extremely important and not necessarily true.

If the force due to air resistance had magnitude F_R N, then the first assumption implies that $F_R = kv$ for some positive constant k. This will mean that our differential equations are linear.

Our assumption about constant mass also has a direct impact on our differential equations; were mass variable we would have to use the full differential statement of Newton II.

2.2 Differential equations for our first model

We will split our model into two parts:

- Part A: before the aircraft applies its brakes.
- Part B: after the aircraft applies its brakes.

Let m kg be the mass of the aeroplane and let F_B N be the magnitude of the constant braking force applied in part B. So, applying Newton II horizontally in the direction of travel, for part A we have

$$\begin{aligned} -F_R &= m\dot{v} \\ \implies -kv &= m\dot{v} \end{aligned} \tag{1}$$

and for part B we have

$$\begin{aligned} -F_R - F_B &= m\dot{v} \\ \implies -kv - F_B &= m\dot{v}. \end{aligned} \tag{2}$$

3 Manipulating our initial model

3.1 Solving the differential equations

3.1.1 Part A

We will solve eq. (1) using separation of variables:

$$\begin{aligned} -kv &= m\frac{\mathrm{d}v}{\mathrm{d}t} \\ \implies \int \frac{1}{v}\,\mathrm{d}v &= -\frac{k}{m}\int \mathrm{d}t \\ \implies \ln v &= -\frac{k}{m}t + C \end{aligned}$$

for some constant C. Exponentiating,

$$v = A\mathrm{e}^{-\frac{k}{m}t} \tag{3}$$

for some constant $A = \mathrm{e}^{C}$.

3.1.2 Part B

We will solve eq. (2) using an integrating factor:

$$\begin{aligned} -kv - F_B &= m\frac{\mathrm{d}v}{\mathrm{d}t} \\ \implies \frac{\mathrm{d}v}{\mathrm{d}t} + \frac{k}{m}v &= -\frac{F_B}{m} \end{aligned}$$

so multiplying by $\mathrm{e}^{\frac{k}{m}t}$,

$$\begin{aligned} \frac{\mathrm{d}v}{\mathrm{d}t}\mathrm{e}^{\frac{k}{m}t} + \frac{k}{m}v\mathrm{e}^{\frac{k}{m}t} &= -\frac{F_B}{m}\mathrm{e}^{\frac{k}{m}t} \\ \implies \frac{\mathrm{d}}{\mathrm{d}t}\left[v\mathrm{e}^{\frac{k}{m}t}\right] &= -\frac{F_B}{m}\mathrm{e}^{\frac{k}{m}t} \\ \implies v\mathrm{e}^{\frac{k}{m}t} &= -\frac{F_B}{m}\int \mathrm{e}^{\frac{k}{m}t}\,\mathrm{d}t \\ &= -\frac{F_B}{k}\mathrm{e}^{\frac{k}{m}t} + B \end{aligned}$$

for some constant B. Exponentiating,

$$v = B\mathrm{e}^{-\frac{k}{m}t} - \frac{F_B}{k}. \tag{4}$$

3.2 Parameter choice

There are four parameters to adjust: the parameters A and B determine the importance of the exponential term in eqs. (3) and (4) respectively; the constant k determines the strength of the air resistance in both cases; and the constant F_B determines the strength of the braking force applied in part B. (We are given that $m = 120\,000$.)

Looking at a graph of the data (shown in fig. 1) it seems that the brakes are applied at $t = 9$ so we will consider part A to be valid for $t \in [0, 9)$ and part B to be valid for $t \in [9, 26]$.

Hence we would like both parts of the model to predict the same velocity at $t = 9$. We also want part A to predict a velocity of $96\,\mathrm{m\,s^{-1}}$ at $t = 0$ (as this is the touchdown velocity) and for part B to predict a velocity of zero at $t = 26$ (as this is when the aircraft comes to rest).

Subject to these constraints, we will choose the parameters A, B, k, F_B which minimise the sum of square residuals between the predicted velocities and the measured velocities from our data. This leads to the following parameter choices:

$$\begin{aligned} A &= 96, \\ B &= 148.413, \\ k &= 7\ 347.13, \\ F_B &= 221\ 953. \end{aligned}$$

Therefore $\frac{k}{m} = 0.0612242$ and $\frac{F_B}{k} = 30.2105$.

An increase in k will cause a steeper exponential decay in both parts, whereas an increase in F_B will cause a steeper gradient in part B only. Varying A will vary the predicted touchdown velocity and varying B will vary the velocity predicted by part B at $t = 9$.

3.3 Predictions

Using the optimised parameters as found above, our first model predicts the following:

For $t \in [0, 9)$,

$$v = 96 \cdot \mathrm{e}^{-0.0612242t} \tag{5}$$

and for $t \in [9, 26]$,

$$v = 148.413 \cdot \mathrm{e}^{-0.0612242t} - 30.2105. \tag{6}$$

4 Data collection

4.1 Measuring the velocity

For this task we have been given a set of measured data already. This data could have been recorded in the following ways:

- Use an ultrasonic sensor directed at the aeroplane to track its velocity as it slows down.
- Use the flight recorder ('black box') on the aeroplane to recover the velocities recorded by the internal instruments on the aeroplane.
- Film the aeroplane landing and play back in slow motion, estimating the distance of the aircraft along the runway at regular intervals in time.
- Use the radar system in a nearby air traffic control tower to track the aeroplane's velocity as it lands.

Although any of these methods would work, the ultrasonic sensor or the flight recorder method would probably be the most accurate.

4.2 Our data

The data we are given is summarised in table 1. The velocities are plotted on a scatter graph in fig. 1.

v	96	89	82	77	72	68	64	61	58	55	50	46	41	38
t	0	1	2	3	4	5	6	7	8	9	10	11	12	13

v	34	31	27	24	21	18	16	13	10	8	5	3	0
t	14	15	16	17	18	19	20	21	22	23	24	25	26

Table 1: *Measured velocity v $\mathrm{m\,s^{-1}}$ at t seconds after landing.*

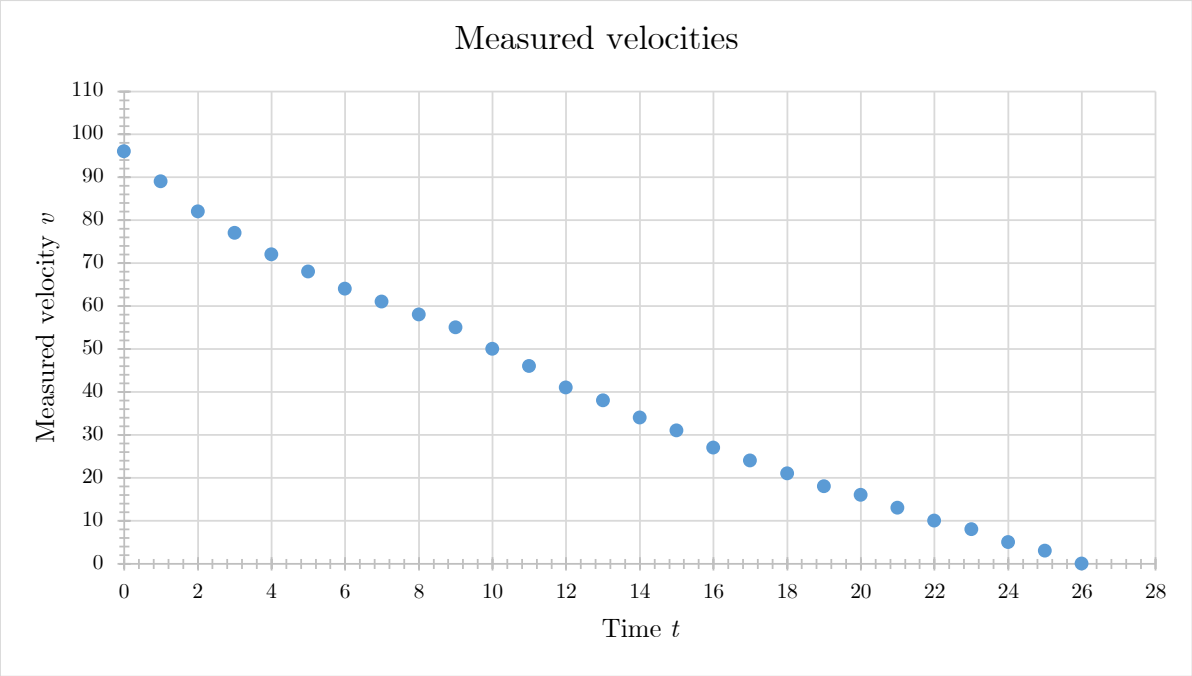


Figure 1: *A scatter plot of measured velocity against time.*

5 Efficacy of our initial model

Now we will compare the predictions made by eqs. (5) and (6) with the measured values in table 1.

Computing the predicted velocity at each point in time leads to the values shown in table 2, where the square residuals are also shown. The sum of the square residuals for this model is

$$\sum_i (v_i' - v_i)^2 = 84.88.$$

Time t	Measured velocity v	Predicted velocity v'	Square residual $(v'-v)^2$
0	96	96.00	0.00
1	89	90.30	1.69
2	82	84.94	8.62
3	77	79.89	8.36
4	72	75.15	9.91
5	68	70.68	7.21
6	64	66.49	6.18
7	61	62.54	2.37
8	58	58.82	0.68
9	55	55.33	0.11
10	50	50.25	0.06
11	46	45.47	0.28
12	41	40.98	0.00
13	38	36.75	1.56
14	34	32.77	1.50
15	31	29.03	3.87
16	27	25.51	2.21
17	24	22.21	3.22
18	21	19.09	3.64
19	18	16.16	3.37
20	16	13.41	6.71
21	13	10.82	4.75
22	10	8.38	2.61
23	8	6.09	3.64
24	5	3.94	1.13
25	3	1.91	1.19
26	0	0.00	0.00

Table 2: *Actual values, predicted values and the square residuals.*

To analyse exactly how the model performs, we plot the predicted velocities and measured velocities together in fig. 2 and we plot the residuals as a function of time in fig. 3.

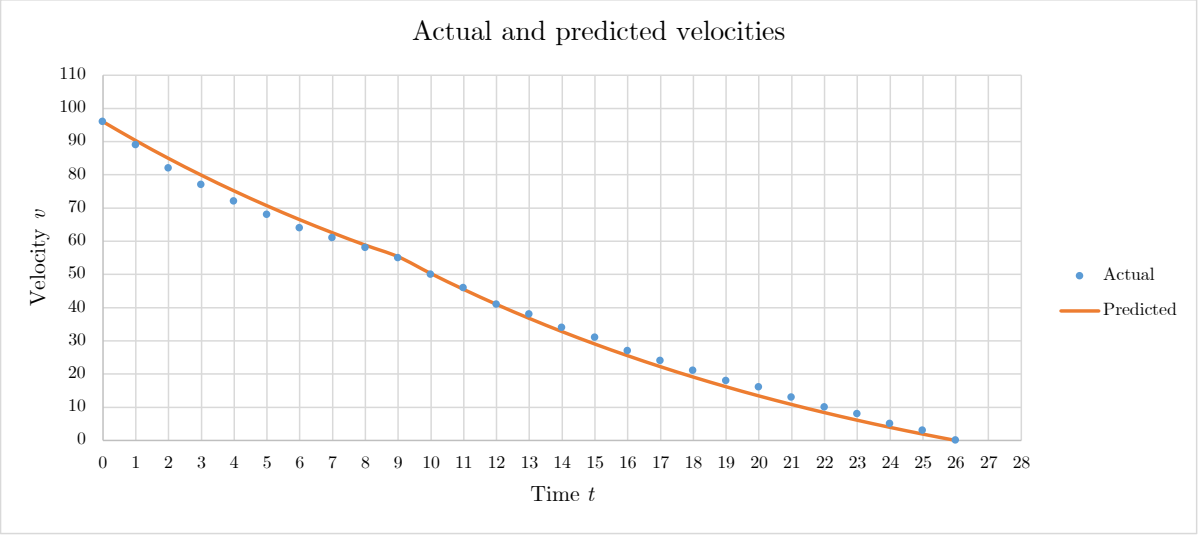


Figure 2: *Measured velocity and predicted velocity plotted together against time.*

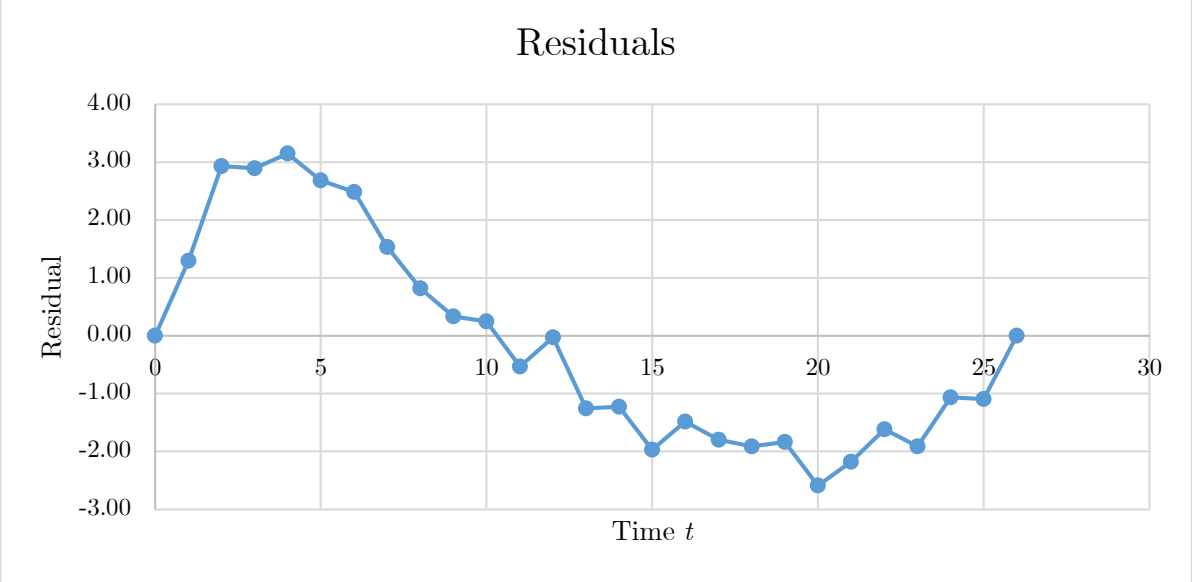


Figure 3: *Residual (difference between predicted and measured velocity) as a function of time.*

From these graphs it is clear that the current model produces an overestimate for most of part A and an underestimate for most of part B. Indeed, the exponential decay constant is too slow near the beginning and too quick near the end.

This implies that we have underestimated the air resistance $F_R = kv$ for high velocities and overestimated it for low velocities.

6 An improved model

6.1 New assumptions

In light of this comparison, one might suggest a refined model of air resistance as $F_R = kv^2$ rather than kv. This will make the gradient of the predicted velocity curve steeper for high velocities and shallower for lower velocities (due to the parabolic shape of the v^2 dependency); this should decrease our predicted velocities for part A and increase them for part B, as desired.

Hence, our new assumptions are the same as presented previously but with air resistance instead proportional to the square of velocity.

Indeed, research online shows that this is universally accepted to be a much better model for air resistance.

6.2 New differential equations

Now with $F_R = kv^2$, applying Newton II horizontally in the direction of travel for part A gives

$$\begin{aligned} -F_R &= m\dot{v} \\ \implies -kv^2 &= m\dot{v} \end{aligned} \tag{7}$$

and for part B gives

$$\begin{aligned} -F_R - F_B &= m\dot{v} \\ \implies -kv^2 - F_B &= m\dot{v}. \end{aligned} \tag{8}$$

7 Efficacy of our improved model

7.1 Solving our new differential equations

7.1.1 Part A

We will solve eq. (7) using separation of variables:

$$\begin{aligned} -kv^2 &= m\frac{\mathrm{d}v}{\mathrm{d}t} \\ \implies \int \frac{1}{v^2}\,\mathrm{d}v &= -\frac{k}{m}\int \mathrm{d}t \\ \implies -\frac{1}{v} &= -\frac{k}{m}t - A \end{aligned}$$

for some constant A. Therefore,

$$v = \frac{1}{\frac{k}{m}t + A}. \tag{9}$$

7.1.2 Part B

We will solve eq. (8) using separation of variables as well:

$$\begin{aligned} -kv^2 - F_B &= m\frac{\mathrm{d}v}{\mathrm{d}t} \\ \implies \int \frac{1}{kv^2 + F_B}\,\mathrm{d}v &= -\frac{1}{m}\int \mathrm{d}t \\ \implies \frac{1}{F_B}\int \frac{1}{\frac{k}{F_B}v^2 + 1}\,\mathrm{d}v &= -\frac{1}{m}t. \end{aligned}$$

Substituting $v = \sqrt{\frac{F_B}{k}}\tan\theta$, we see that

$$\begin{aligned} \int \frac{1}{\frac{k}{F_B}v^2 + 1}\,\mathrm{d}v &= \int \frac{1}{\tan^2\theta + 1}\cdot\frac{\mathrm{d}v}{\mathrm{d}\theta}\,\mathrm{d}\theta \\ &= \int \frac{1}{\sec^2\theta}\cdot\sqrt{\frac{F_B}{k}}\sec^2\theta\,\mathrm{d}\theta \\ &= \sqrt{\frac{F_B}{k}}\int \mathrm{d}\theta \end{aligned}$$

and so becomes

$$\begin{aligned} \frac{1}{F_B}\sqrt{\frac{F_B}{k}}\int \mathrm{d}\theta &= -\frac{1}{m}t \\ \implies \frac{1}{\sqrt{F_B k}}\theta &= -\frac{1}{m}t + C \end{aligned}$$

for some constant C. Hence,

$$\begin{aligned} \arctan\left(\sqrt{\frac{k}{F_B}}v\right) &= -\frac{\sqrt{F_B k}}{m}t + C\sqrt{F_B k} \\ \implies v &= \sqrt{\frac{F_B}{k}}\tan\left(B - \frac{\sqrt{F_B k}}{m}t\right) \end{aligned} \tag{10}$$

for some constant $B = C\sqrt{F_B k}$.

7.2 Parameter choice and predictions

As before, we have four parameters A, B, k, F_B to adjust. We will require the same constraints as previously: eqs. (9) and (10) must predict the same velocity at $t = 9$; eq. (9) must predict $v = 96$ at $t = 0$; and eq. (10) must predict $v = 0$ at $t = 26$.

Under these constraints, optimising the sum of square residuals with the measured data leads to the following parameter values:

$$\begin{aligned} A &= \frac{1}{96} = 0.0104167, \\ B &= 10.6378, \\ k &= 103.250, \\ F_B &= 303558. \end{aligned}$$

These values lead to the following specific solutions to our differential equations:

For $t \in [0, 9)$,

$$v = (0.000860417t + 0.0104167)^{-1}$$

and for $t \in [9, 26]$,

$$v = 54.2220 \cdot \tan\left(10.6378 - 0.0466535t\right).$$

7.3 Comparison with the data

The predicted velocities and square residuals from this new model are shown in table 3; the sum of square residuals is now just 2.05. This is an enormous decrease from the first model!

Time t	Measured velocity v	Predicted velocity v'	Square residual $(v' - v)^2$
0	96	96.00	0.00
1	89	88.68	0.11
2	82	82.39	0.15
3	77	76.94	0.00
4	72	72.16	0.03
5	68	67.94	0.00
6	64	64.19	0.04
7	61	60.83	0.03
8	58	57.80	0.04
9	55	55.06	0.00
10	50	50.16	0.02
11	46	45.65	0.12
12	41	41.49	0.24
13	38	37.61	0.15
14	34	33.98	0.00
15	31	30.56	0.20
16	27	27.31	0.09
17	24	24.21	0.04
18	21	21.23	0.05
19	18	18.37	0.13
20	16	15.59	0.17
21	13	12.88	0.01
22	10	10.24	0.06
23	8	7.64	0.13
24	5	5.07	0.01
25	3	2.53	0.22
26	0	0.00	0.00

Table 3: *Predicted velocities, measured velocities and square residuals as calculated using the new model.*

The predicted and measured velocities are plotted together in fig. 4, and the residuals as a function of time are shown in fig. 5. Both graphs show that the fit of the new line is remarkable, and that the residuals are sufficiently random to indicate that this model cannot be improved upon.

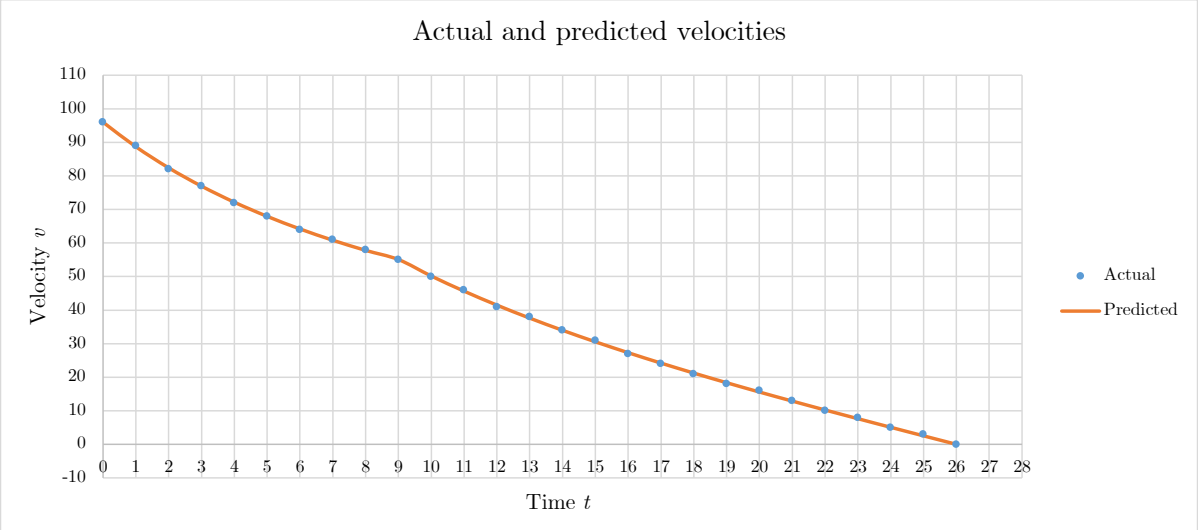


Figure 4: *A plot of predicted and measured velocities as a function of time.*

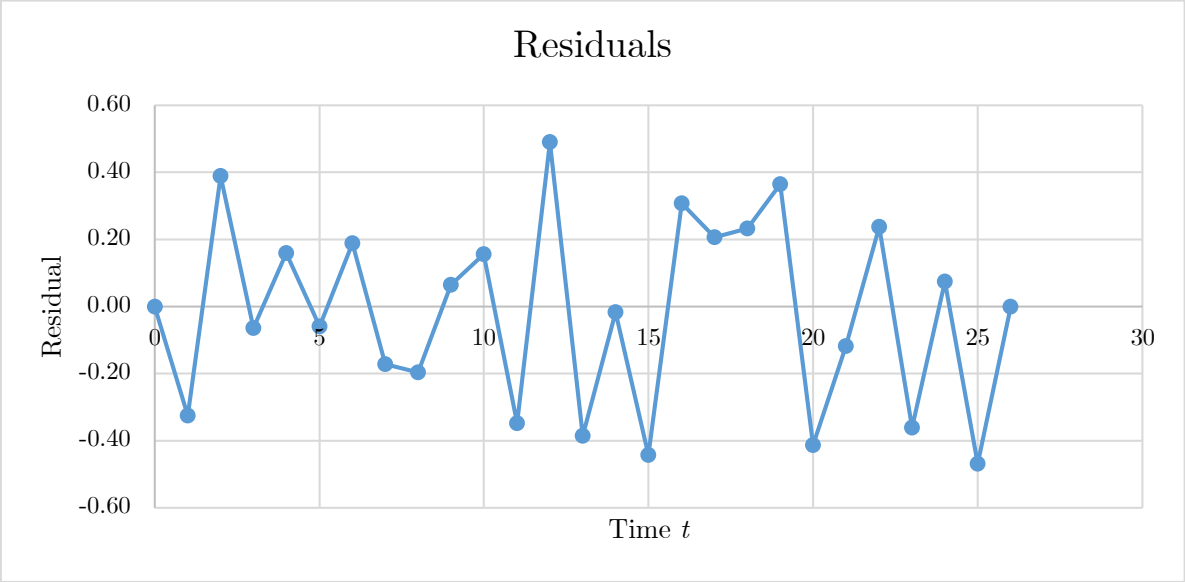


Figure 5: *Residuals as a function of time for the new model.*

8 Conclusion

We have tried two different models for the situation; the first with air resistance proportional to velocity, and the second with air resistance proportional to the square of velocity. It is clear that the second model is far superior to the first, and indeed the superb accuracy of the second model indicates that our other simplifying assumptions are good assumptions (that making the assumptions hardly changes the situation).

Using the second model, we can find an estimate for the length of runway the aircraft would need. If the $x(t)$ is the horizontal displacement (in metres) of the aeroplane from its touchdown location at a time t, then for part A eq. (9) predicts

$$\begin{aligned} x(t) &= \int_0^t v \,\mathrm{d}\tau \\ &= \int_0^t \frac{1}{\frac{k}{m}\tau + A} \,\mathrm{d}\tau \\ &= \left[\frac{m}{k}\ln(\frac{k}{m}\tau + A)\right]_0^t \\ &= \frac{m}{k}\ln(\frac{k}{m}t + A) - \frac{m}{k}\ln A \\ &= \frac{m}{k}\ln\left(\frac{k}{mA}t + 1\right) \end{aligned}$$

and for part B eq. (10) predicts

$$\begin{aligned}
x(t) &= x(9) + \int_9^t v\,\mathrm{d}\tau \\
&= x(9) + \int_9^t \sqrt{\frac{F_B}{k}} \tan\left(B - \frac{\sqrt{F_B k}}{m}\tau\right) \mathrm{d}\tau \\
&= x(9) + \left[\sqrt{\frac{F_B}{k}} \frac{m}{\sqrt{F_B k}} \ln\left\{\cos\left(B - \frac{\sqrt{F_B k}}{m}\tau\right)\right\}\right]_9^t \\
&= \frac{m}{k}\ln\left(\frac{9k}{mA} + 1\right) + \frac{m}{k}\ln\left[\frac{\cos\left(B - \frac{\sqrt{F_B k}}{m}t\right)}{\cos\left(B - \frac{9\sqrt{F_B k}}{m}\right)}\right] \\
&= \frac{m}{k}\ln\left[\left(\frac{9k}{mA} + 1\right) \cdot \frac{\cos\left(B - \frac{\sqrt{F_B k}}{m}t\right)}{\cos\left(B - \frac{9\sqrt{F_B k}}{m}\right)}\right].
\end{aligned}$$

Hence the total distance travelled before the plane comes to rest at $t = 26$ is

$$\begin{aligned}
x_{\text{tot}} &= \frac{m}{k}\ln\left[\left(\frac{9k}{mA} + 1\right) \cdot \frac{\cos\left(B - \frac{26\sqrt{F_B k}}{m}\right)}{\cos\left(B - \frac{9\sqrt{F_B k}}{m}\right)}\right] \\
&= 1162.23 \cdot \ln\left[1.74340 \cdot \frac{\cos(9.42478)}{\cos(10.2179)}\right] \\
&= 1057.86.
\end{aligned}$$

We can therefore recommend a minimum runway length of about 1.2 km for the plane to land safely (accounting for the plane having a length of about 60 m and an extra 80 m for safety).

Chapter 8

DE coursework

This is the real coursework for the differential equations module. I actually found this problem quite fascinating and enjoyed the data collection process too.

Investigating the rate of flow of water through a small hole

(DE Coursework)

Damon Falck

March 2018

Contents

1 Introduction

The situation which we will investigate is as follows. A bucket with a small (temporarily covered) hole in the side is filled up with water and positioned above a second, identical bucket, also with a small hole in the side. The hole in the first bucket is then uncovered and water allowed to flow between the two buckets and out of the second bucket. (See fig. 1 for a diagram of this.)

We are to investigate the height of water in each bucket as a function of time. We will first set up an initial model relating the rate of flow out of a bucket to the height of water in the bucket, and use this to make preliminary predictions. Then, after collecting experimental data and comparing it to our predictions, we will set up a second, improved model.

2 Setting up an initial model

2.1 Assumptions

We will begin by making the following simplifying assumptions about the situation, in rough order of importance:

- The rate of flow of water out of each bucket is proportional to the height of the water in the bucket. This reasonable as the pressure (and therefore the force) acting down on the hole is proportional to the height since $p = \rho g h$.
- The buckets have a constant cross-sectional area, so that the volume of water is proportional to the height of the water.
- The buckets are identical.
- The position and size of the hole is the same in each bucket. This and the previous assumption mean that the constants of proportionality k for flow rate will be the same for both buckets.
- Both buckets are level, so that the height of water measured is the height above the hole.
- The time the water takes to flow from the first bucket to the second is negligible. This is an approximation but simplifies the situation (otherwise we would have to calculate the time taken using constant acceleration formulae and then model the water height in the second bucket with a time offset from the first).

These assumptions are listed in approximate order of importance. While most of the assumptions are necessary for our initial model, the first assumption (that the flow rate is proportional to the height of the water) is extremely important and not necessarily true: this is the main assumption we may need to alter later. This assumption determines the form of our initial differential equation.

2.2 Differential equations for our first model

Let h_1 and h_2 be the height, in centimetres, of the water in the first and second buckets respectively, at t seconds after the hole in the first bucket is uncovered.

Since we are assuming that the rate of flow *out* of each bucket is proportional to the water height, we can say

$$\frac{\mathrm{d}h_1}{\mathrm{d}t} = -kh_1 \tag{1}$$

and

$$\frac{\mathrm{d}h_2}{\mathrm{d}t} = -\frac{\mathrm{d}h_1}{\mathrm{d}t} - kh_2 \tag{2}$$

for some positive real constant k. We use the same constant of proportionality k for both equations due to our second assumption that the buckets are identical.

3 Manipulating our initial model

3.1 Solving the differential equations

The first equation, eq. (1), can be solved by separating the variables:

$$\begin{aligned} \frac{\mathrm{d}h_1}{\mathrm{d}t} &= -kh_1 \\ \implies \int \frac{1}{h_1}\,\mathrm{d}h_1 &= \int -k\,\mathrm{d}t \\ \implies \ln h_1 &= -kt + \ln A \\ \implies h_1 &= A\mathrm{e}^{-kt} \end{aligned} \tag{3}$$

for some constant A.

This is our general solution for h_1. Differentiating,

$$\frac{\mathrm{d}h_1}{\mathrm{d}t} = -kA\mathrm{e}^{-kt},$$

and substituting this into eq. (2) gives

$$\frac{\mathrm{d}h_2}{\mathrm{d}t} = kA\mathrm{e}^{-kt} - kh_2.$$

We can solve this differential equation using an integrating factor:

$$\begin{aligned}
\frac{\mathrm{d}h_2}{\mathrm{d}t} + kh_2 &= kA\mathrm{e}^{-kt} \\
\implies \mathrm{e}^{kt}\frac{\mathrm{d}h_2}{\mathrm{d}t} + k\mathrm{e}^{kt}h_2 &= kA\mathrm{e}^{0} \\
\implies \frac{\mathrm{d}}{\mathrm{d}t}\left[h_2\mathrm{e}^{kt}\right] &= kA \\
\implies h_2\mathrm{e}^{kt} &= \int kA\,\mathrm{d}t \\
&= kAt + B \\
\implies h_2 &= kAt\mathrm{e}^{-kt} + B\mathrm{e}^{-kt} \qquad (4)
\end{aligned}$$

for some constant B.

3.2 Parameter choice

Our solutions, eqs. (3) and (4), are dependent on three parameters: k, A and B. Of these, A and B are dependent on boundary conditions only, whereas k is a constant which is intrinsic to the geometry of the buckets.

For our experiment (as described later), the first bucket had an initial water height of $h_1 = 13.4$ (and the second bucket was initially empty). Therefore at $t = 0$, eq. (3) gives

$$\begin{aligned}
13.4 &= A\mathrm{e}^{-0} \\
\implies A &= 13.4
\end{aligned}$$

and eq. (4) gives

$$\begin{aligned}
0 &= 0 + B \cdot 0 \\
\implies B &= 0.
\end{aligned}$$

Therefore our particular solutions, dependent only on the parameter k, are

$$h_1 = 13.4\mathrm{e}^{-kt}$$

and

$$h_2 = 13.4kt\mathrm{e}^{-kt}.$$

We will now choose the parameter k which minimises the sum of square residuals between the predicted velocities and the measured velocities from our data. This optimisation leads to a value of $k = 0.01008$, producing a sum of square residuals of 93.4.

3.3 Predictions

Using the optimised value of $k = 0.01008$ as found above, our first model predicts the following:

At time t, the height in centimetres of water in the first bucket is

$$h_1 = 13.4\mathrm{e}^{-0.01008t} \tag{5}$$

and in the second bucket is

$$h_2 = 0.1351t\mathrm{e}^{-0.01008t}. \tag{6}$$

4 Conducting the experiment

4.1 Experimental setup

In order to design a worthwhile experiment, we must consider some of the assumptions we made when setting up our first model. We said that the two buckets are identical, so we must find two identical buckets with the same size and position of hole. We also said that the buckets are perfectly level, so we must ensure that we set up the system on flat, level surfaces.

Additionally, we assumed that the time taken for water to flow between the two buckets is negligible, so we should try to minimise the vertical separation of the buckets so as to make this assumption valid.

We started by selecting two rectangular buckets, each with a small hole (of roughly 8 mm diameter) near the bottom of one side. We then inserted a small plastic spout into the hole in each bucket and attached a ruler vertically to the inside of both, and then positioned the two buckets such that water would flow between them. (See fig. 1 for a diagram of our setup.)

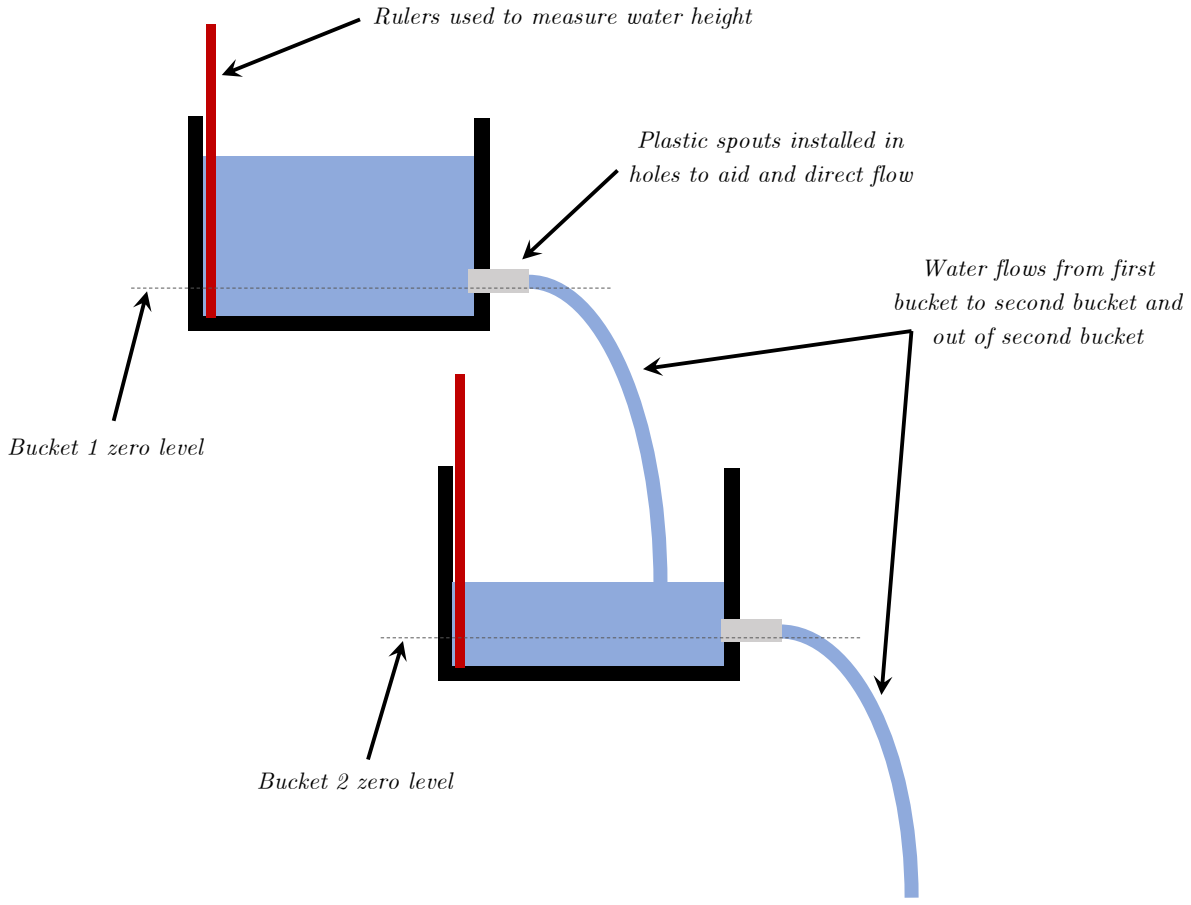


Figure 1: *The experimental setup of the system of two buckets which we are modelling.*

4.2 Experimental method

After setting up the buckets as shown in fig. 1, we filled each bucket up with water until the water just started flowing out of the spout. At this point, we measured the water height in each bucket: this was the height of the spout above the bottom of the bucket, and served as our 'zero level'.

We then covered the first bucket's spout with blu tack, and filled this bucket up to a total height of 14.9 cm of water.

At the same time as starting a stopwatch, the blu tack covering the first bucket's spout was removed, allowing water to flow between the buckets. At five second intervals the height of water in each bucket was measured and noted down. We stopped taking measurements once the water stopped flowing between the buckets.

Once we finished taking readings, we reset the experiment and repeated it once, with different students measuring the height. This repetition was so as to reduce the overall random error in our experiment, and we had different people take the measurements so as to reduce any type of human error particular to the individual taking the readings.

4.3 Our data

Our data is summarised below. Values measured for each bucket on both the first and second reading are shown, and then the 'zeroed average'. This value was calculated by first finding the average height between the two readings at every point in time, and then subtracting the zero level as measured before the blu tack was removed. This zero level was a height of 1.5 cm for the first bucket and 0.9 cm for the second bucket, on both readings.

	First reading		Second reading		Zeroed average	
t / s	h_1 / cm	h_2 / cm	h_1 / cm	h_2 / cm	h_1 / cm	h_2 / cm
0	14.9	0.9	14.9	0.9	13.4	0.0
5	14.4	1.3	14.5	1.2	13.0	0.4
10	14.0	1.8	14.1	1.6	12.6	0.8
15	13.6	2.1	13.6	2.0	12.1	1.2
20	13.2	2.4	13.3	2.3	11.8	1.5
25	12.8	2.7	12.8	2.6	11.3	1.8
30	12.4	2.9	12.5	2.8	11.0	2.0
35	12.0	3.2	12.0	3.1	10.5	2.3
40	11.8	3.4	11.6	3.3	10.2	2.5
45	11.3	3.6	11.3	3.5	9.8	2.7
50	10.9	3.8	11.0	3.7	9.5	2.9
55	10.6	3.9	10.5	3.9	9.1	3.0
60	10.2	4.0	10.2	4.1	8.7	3.2
65	9.8	4.2	9.9	4.3	8.4	3.4
70	9.5	4.3	9.5	4.4	8.0	3.5
75	9.2	4.5	9.1	4.5	7.7	3.6
80	8.8	4.6	8.7	4.6	7.3	3.7
85	8.5	4.7	8.5	4.6	7.0	3.8
90	8.2	4.8	8.2	4.7	6.7	3.9
95	7.9	4.9	7.8	4.8	6.4	4.0
100	7.5	4.9	7.5	4.9	6.0	4.0
105	7.3	4.9	7.2	4.9	5.8	4.0
110	6.9	5.0	6.9	4.9	5.4	4.1
115	6.7	5.0	6.6	5.0	5.2	4.1
120	6.4	5.0	6.4	5.0	4.9	4.1
125	6.1	5.0	6.0	5.0	4.6	4.1
130	5.9	5.0	5.8	5.0	4.4	4.1
135	5.7	5.0	5.5	5.0	4.1	4.1
140	5.4	5.0	5.3	5.0	3.9	4.1

145	5.2	5.0	5.1	5.0	3.7	4.1
150	4.9	5.0	4.8	5.0	3.4	4.1
155	4.7	5.0	4.6	5.0	3.2	4.1
160	4.5	5.0	4.4	5.0	3.0	4.1
165	4.3	5.0	4.2	5.0	2.8	4.1
170	4.0	5.0	4.0	5.0	2.5	4.1
175	3.8	4.9	3.8	4.9	2.3	4.0
180	3.7	4.9	3.6	4.8	2.2	4.0
185	3.5	4.7	3.4	4.7	2.0	3.8
190	3.4	4.6	3.2	4.6	1.8	3.7
195	3.2	4.5	3.1	4.6	1.7	3.7
200	3.0	4.4	3.0	4.5	1.5	3.6
205	2.9	4.4	2.8	4.4	1.4	3.5
210	2.7	4.2	2.6	4.3	1.2	3.4
215	2.5	4.1	2.5	4.2	1.0	3.3
220	2.5	4.0	2.3	4.0	0.9	3.1
225	2.4	3.8	2.2	3.9	0.8	3.0
230	2.3	3.7	2.1	3.8	0.7	2.9
235	2.2	3.6	2.1	3.7	0.7	2.8
240	2.1	3.5	2.0	3.6	0.6	2.7
245	2.1	3.5	1.9	3.5	0.5	2.6
250	2.0	3.4	1.8	3.4	0.4	2.5
255	2.0	3.1	1.8	3.3	0.4	2.3
260	2.0	3.0	1.7	3.1	0.4	2.2
265	1.9	2.9	1.7	3.0	0.3	2.1
270	1.8	2.8	1.6	2.9	0.2	2.0
275	1.8	2.7	1.5	2.8	0.2	1.9
280	1.8	2.6	1.5	2.6	0.2	1.7
285	1.8	2.5	1.5	2.5	0.2	1.6
290	1.7	2.3	1.5	2.3	0.1	1.4
295	1.6	2.1	1.5	2.2	0.1	1.3
300	1.6	2.0	1.5	2.1	0.1	1.2
305	1.6	1.9	1.5	2.0	0.1	1.1
310	1.6	1.8	1.5	2.0	0.1	1.0
315	1.6	1.6	1.5	1.9	0.1	0.9
320	1.6	1.6	1.5	1.7	0.1	0.8
325	1.6	1.6	1.5	1.6	0.1	0.7
330	1.6	1.6	1.5	1.5	0.1	0.7
335	1.6	1.5	1.5	1.5	0.1	0.6
340	1.6	1.4	1.5	1.5	0.1	0.6
345	1.6	1.3	1.5	1.4	0.1	0.5
350	1.6	1.3	1.5	1.3	0.1	0.4
355	1.6	1.2	1.5	1.3	0.1	0.4
360	1.6	1.2	1.5	1.2	0.1	0.3
365	1.6	1.1	1.5	1.2	0.1	0.3
370	1.6	1.1	1.5	1.1	0.1	0.2
375	1.6	1.1	1.5	1.0	0.1	0.2
380	1.6	1.0	1.5	1.0	0.1	0.1
385	1.6	1.0	1.5	1.0	0.1	0.1
390	1.6	1.0	1.5	1.0	0.1	0.1
395	1.6	1.0	1.5	1.0	0.1	0.1
400	1.6	1.0	1.5	0.9	0.1	0.0
405	1.6	1.0	1.5	0.9	0.1	0.0
410	1.5	1.0	1.5	0.9	0.0	0.0
415	1.5	0.9	1.5	0.9	0.0	0.0

The zeroed average water heights, which we will use as our values from now on, are plotted on a scatter graph in fig. 2.

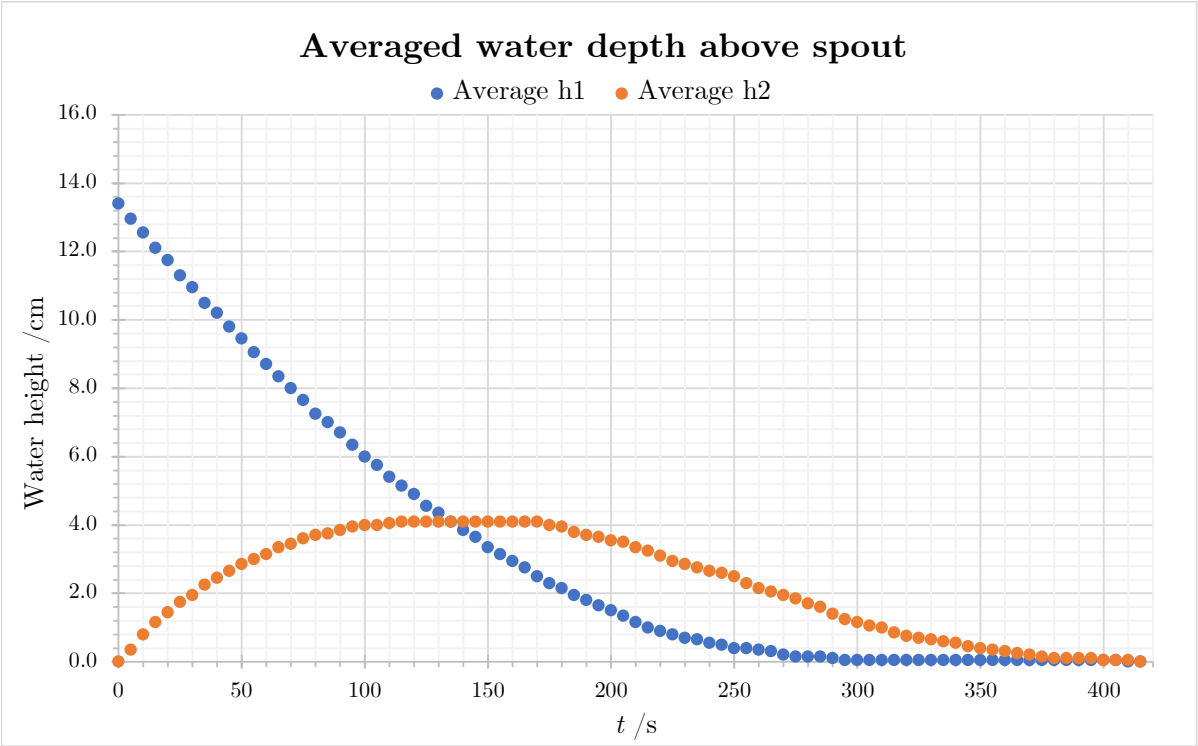


Figure 2: *A scatter plot of measured heights against time.*

In general, the variability of our data is quite low: each reading has an error of at most $\pm 0.2\,\text{cm}$ which is partly due to systematic errors due to, for instance, parallax effects while reading the ruler. This type of error is reduced greatly by taking two readings and taking the average, as we did.

The data from our two readings was very similar, with only a couple of slightly anomalous points in either which were clearly due to human error when reading from the ruler.

Overall, therefore, this dataset is both reliable and precise, due to the low variability between trials and due to the low experimental error.

5 Efficacy of our initial model

Now I will compare the predictions made by eqs. (5) and (6) with the values measured in my experiment.

Computing the predicted heights at each point in time leads to the values shown below, where the square residuals are also shown. The sum of the square residuals for this model is

$$\sum_i \left[(h_1')_i - (h_1)_i\right]^2 + \sum_i \left[(h_2')_i - (h_2)_i\right]^2 = 93.44.$$

	Measured values		Predicted values		Residual	
t / s	h_1 / cm	h_2 / cm	h_1' / cm	h_2' / cm	$h_1' - h_1$ / cm	$h_2' - h_2$ / cm
0	13.4	0.0	13.40	0.00	0.0	0.0
5	13.0	0.4	12.74	0.64	-0.2	0.3
10	12.6	0.8	12.11	1.22	-0.4	0.4
15	12.1	1.2	11.52	1.74	-0.6	0.6
20	11.8	1.5	10.95	2.21	-0.8	0.8

25	11.3	1.8	10.41	2.63	-0.9	0.9
30	11.0	2.0	9.90	3.00	-1.0	1.0
35	10.5	2.3	9.42	3.32	-1.1	1.1
40	10.2	2.5	8.95	3.61	-1.2	1.2
45	9.8	2.7	8.51	3.86	-1.3	1.2
50	9.5	2.9	8.09	4.08	-1.4	1.2
55	9.1	3.0	7.70	4.27	-1.4	1.3
60	8.7	3.2	7.32	4.43	-1.4	1.3
65	8.4	3.4	6.96	4.56	-1.4	1.2
70	8.0	3.5	6.62	4.67	-1.4	1.2
75	7.7	3.6	6.29	4.76	-1.4	1.2
80	7.3	3.7	5.98	4.82	-1.3	1.1
85	7.0	3.8	5.69	4.87	-1.3	1.1
90	6.7	3.9	5.41	4.91	-1.3	1.1
95	6.4	4.0	5.14	4.93	-1.2	1.0
100	6.0	4.0	4.89	4.93	-1.1	0.9
105	5.8	4.0	4.65	4.92	-1.1	0.9
110	5.4	4.1	4.42	4.90	-1.0	0.9
115	5.2	4.1	4.20	4.87	-0.9	0.8
120	4.9	4.1	4.00	4.84	-0.9	0.7
125	4.6	4.1	3.80	4.79	-0.8	0.7
130	4.4	4.1	3.61	4.74	-0.7	0.6
135	4.1	4.1	3.44	4.68	-0.7	0.6
140	3.9	4.1	3.27	4.61	-0.6	0.5
145	3.7	4.1	3.11	4.54	-0.5	0.4
150	3.4	4.1	2.95	4.47	-0.4	0.4
155	3.2	4.1	2.81	4.39	-0.3	0.3
160	3.0	4.1	2.67	4.31	-0.3	0.2
165	2.8	4.1	2.54	4.22	-0.2	0.1
170	2.5	4.1	2.41	4.14	-0.1	0.0
175	2.3	4.0	2.30	4.05	0.0	0.0
180	2.2	4.0	2.18	3.96	0.0	0.0
185	2.0	3.8	2.08	3.87	0.1	0.1
190	1.8	3.7	1.97	3.78	0.2	0.1
195	1.7	3.7	1.88	3.69	0.2	0.0
200	1.5	3.6	1.78	3.60	0.3	0.0
205	1.4	3.5	1.70	3.51	0.3	0.0
210	1.2	3.4	1.61	3.41	0.5	0.1
215	1.0	3.3	1.53	3.32	0.5	0.1
220	0.9	3.1	1.46	3.23	0.6	0.1
225	0.8	3.0	1.39	3.15	0.6	0.2
230	0.7	2.9	1.32	3.06	0.6	0.2
235	0.7	2.8	1.25	2.97	0.6	0.2
240	0.6	2.7	1.19	2.88	0.6	0.2
245	0.5	2.6	1.13	2.80	0.6	0.2
250	0.4	2.5	1.08	2.72	0.7	0.2
255	0.4	2.3	1.02	2.63	0.6	0.3
260	0.4	2.2	0.97	2.55	0.6	0.4
265	0.3	2.1	0.93	2.47	0.6	0.4
270	0.2	2.0	0.88	2.40	0.7	0.4
275	0.2	1.9	0.84	2.32	0.7	0.5
280	0.2	1.7	0.80	2.25	0.6	0.5
285	0.2	1.6	0.76	2.18	0.6	0.6
290	0.1	1.4	0.72	2.10	0.6	0.7
295	0.1	1.3	0.68	2.04	0.6	0.8

300	0.1	1.2	0.65	1.97	0.6	0.8
305	0.1	1.1	0.62	1.90	0.6	0.9
310	0.1	1.0	0.59	1.84	0.5	0.8
315	0.1	0.9	0.56	1.78	0.5	0.9
320	0.1	0.8	0.53	1.72	0.5	1.0
325	0.1	0.7	0.51	1.66	0.5	1.0
330	0.1	0.7	0.48	1.60	0.4	1.0
335	0.1	0.6	0.46	1.54	0.4	0.9
340	0.1	0.6	0.43	1.49	0.4	0.9
345	0.1	0.5	0.41	1.44	0.4	1.0
350	0.1	0.4	0.39	1.39	0.3	1.0
355	0.1	0.4	0.37	1.34	0.3	1.0
360	0.1	0.3	0.36	1.29	0.3	1.0
365	0.1	0.3	0.34	1.24	0.3	1.0
370	0.1	0.2	0.32	1.20	0.3	1.0
375	0.1	0.2	0.31	1.16	0.3	1.0
380	0.1	0.1	0.29	1.11	0.2	1.0
385	0.1	0.1	0.28	1.07	0.2	1.0
390	0.1	0.1	0.26	1.03	0.2	0.9
395	0.1	0.1	0.25	0.99	0.2	0.9
400	0.1	0.0	0.24	0.96	0.2	0.9
405	0.1	0.0	0.23	0.92	0.2	0.9
410	0.0	0.0	0.21	0.89	0.2	0.8
415	0.0	0.0	0.20	0.85	0.2	0.9

It is worth taking a look at how changing the value of the parameter k influences the accuracy of the predictions. Based on looking at the graph, a reasonable upper bound is $k = 0.013$ whereas a reasonable lower bound is $k = 0.008$. Using the value $k = 0.013$ generates a sum of square residuals of 197.2, whereas using the value $k = 0.008$ generates a sum of square residuals of 191.3. Both of these sets of predictions therefore have extremely high error when compared to a sum of square residuals of 93.4 when using the optimised value of $k = 0.01008$.

The variation in our measurements (we can estimate a height error of $\pm 0.2\,\text{cm}$ and a temporal error of $\pm 1\,\text{s}$) is very small in comparison to most of our residuals, and so we can safely say that our high sum of square residuals is not due to experimental error but due to an inaccurate model.

To analyse exactly how the model performs using the value $k = 0.01008$, we plot the predicted heights and actual heights together in fig. 3 and we plot the residuals as a function of time in fig. 4.

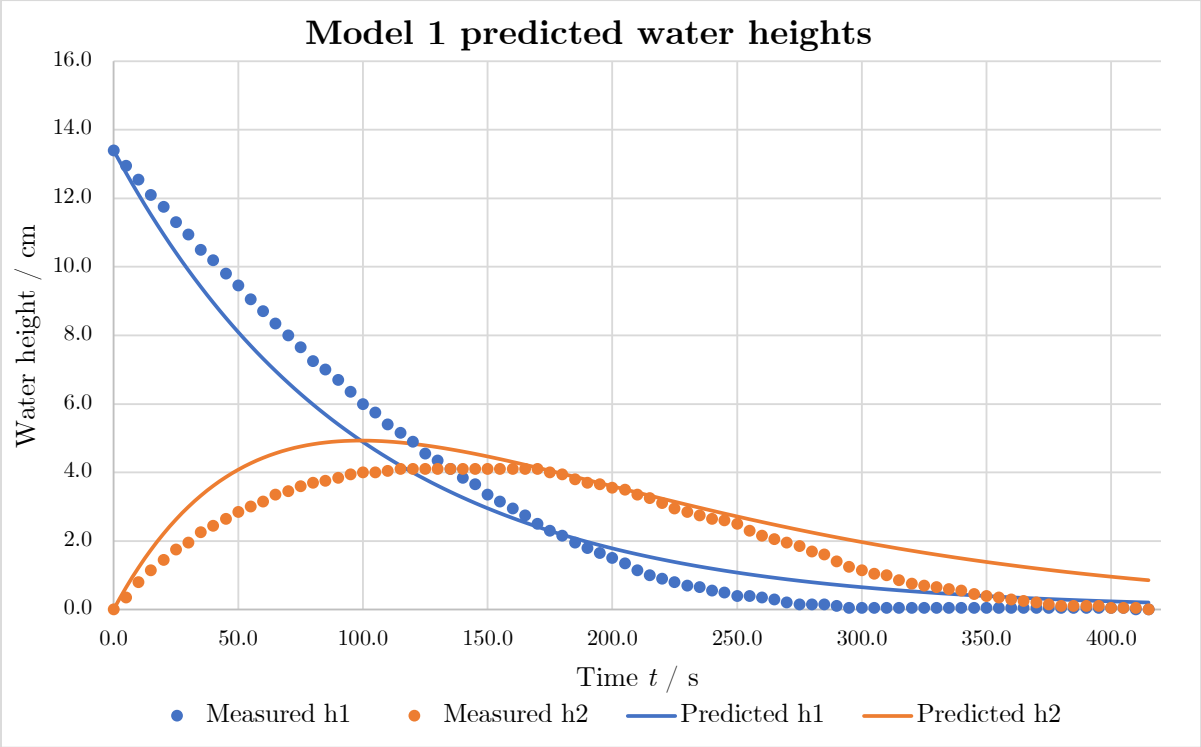


Figure 3: *Measured heights and predicted heights plotted together against time.*

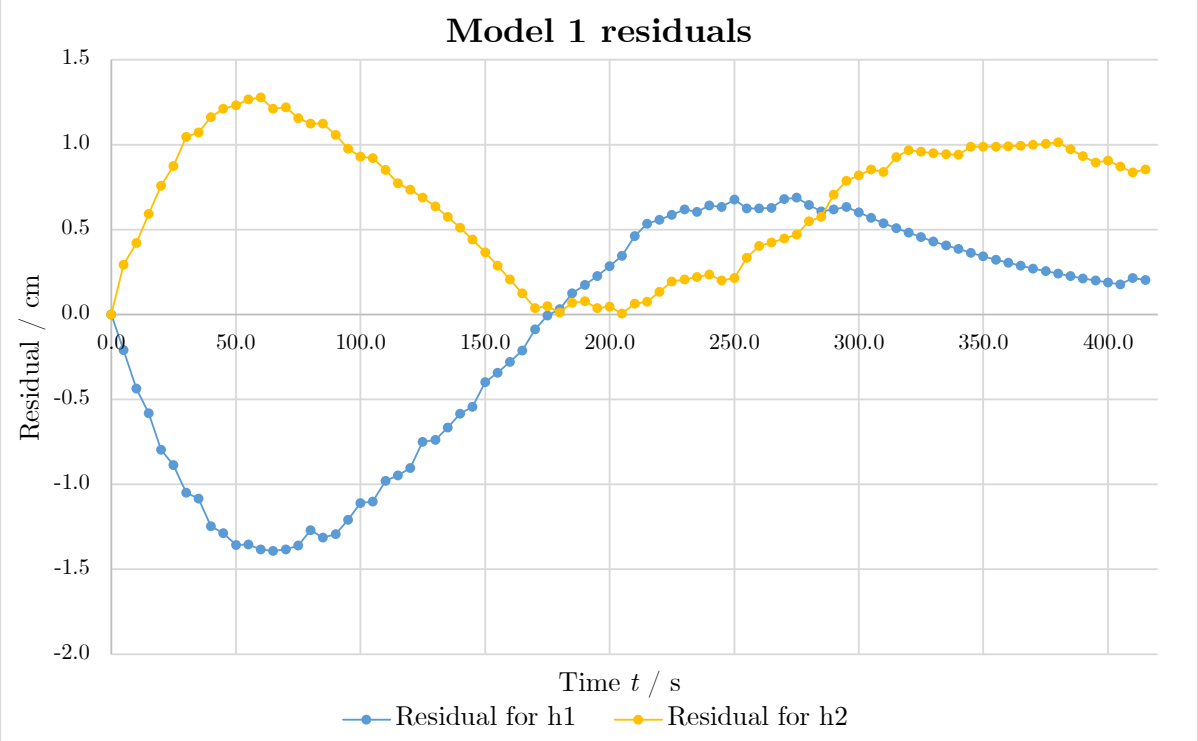


Figure 4: *Residuals (difference between predicted and measured heights) as a function of time.*

From these graphs it is clear that while the current model behaves approximately similarly to real life, the actual curve is not a good fit. The highly patterned nature of the graph of residuals against time suggests that our model can be improved significantly. We must therefore reexamine our initial assumptions to make an improvement.

6 An improved model

In light of this comparison, we will need to make a revision to the modelling process but not the conduct of the experiment. This is because our residuals were high but not at all randomly distributed, and the experimental error was very small in comparison, suggesting that the discrepancy is due to an insufficient model rather than a poorly conducted experiment.

6.1 New assumptions

Our main assumption made for the first model was that the rate of flow of water out of each bucket is proportional to the height of the water in that bucket. However, Torricelli's law, which is a particular case of Bernoulli's principle, implies that the rate of flow of water out of a bucket through a small hole is actually proportional to the *square root* of the height of the water in that bucket.

We will therefore make this change for our second model (keeping all of our other assumptions).

6.2 New differential equations

Now with the rate of flow proportional to $\sqrt{h}$, our new differential equations are

$$\frac{\mathrm{d}h_1}{\mathrm{d}t} = -k\sqrt{h_1} \tag{7}$$

and

$$\frac{\mathrm{d}h_2}{\mathrm{d}t} = -\frac{\mathrm{d}h_1}{\mathrm{d}t} - k\sqrt{h_2}, \tag{8}$$

where h_1 and h_2 are the water heights in the first and second buckets as before, and as before k is a positive constant intrinsic to the geometry of each bucket.

7 Efficacy of our improved model

7.1 Solving our new differential equations

Our first differential equation, eq. (7), can be solved by separating the variables:

$$\begin{aligned} \frac{\mathrm{d}h_1}{\mathrm{d}t} &= -k\sqrt{h_1} \\ \implies \int \frac{1}{\sqrt{h_1}}\,\mathrm{d}h_1 &= \int -k\,\mathrm{d}t \\ \implies 2\sqrt{h_1} &= -kt + A \\ \implies h_1 &= \frac{k^2}{4}t^2 - \frac{Ak}{2}t + \frac{A^2}{4} \end{aligned} \tag{9}$$

for some constant A.

This gives $\frac{\mathrm{d}h_1}{\mathrm{d}t} = \frac{k^2}{2}t - \frac{Ak}{2}$, and so substituting into our second differential equation, eq. (8), we have

$$\frac{\mathrm{d}h_2}{\mathrm{d}t} = -\frac{k^2}{2}t + \frac{Ak}{2} - k\sqrt{h_2}. \tag{10}$$

Unfortunately, this differential equation is analytically insoluble — we will have to find an approximate numerical solution.

7.2 Parameter choice and predictions for h_1

We will start by finding the constant A which depends on boundary conditions for the first bucket only. Since at $h_1 = 13.4$ at $t = 0$ (from the data), eq. (9) gives

$$13.4 = 0 - 0 + \frac{A^2}{4}$$
$$\implies A = 2\sqrt{13.4} = 7.321.$$

Therefore our particular solution for h_1 as a function of time is

$$h_1 = \frac{k^2}{4}t^2 - 3.661kt + 13.4.. \tag{11}$$

As before, we now only have the constant k to choose. We only have an analytic solution for h_1 but if we select the value of k which optimises the sum of square residuals for h_1 as predicted by eq. (11), we can then use that value to numerically solve the differential equation for h_2.

As an important note, the function given in *eq.* (11) is a parabola and so predicts that h_1 will eventually increase again (to infinity!) after it reaches zero. Of course, this is not physical, and so we will simply disregard predictions after about $t = 300$ seconds, which is when the first bucket becomes empty.

Optimising the sum of square residuals for h_1 on the interval $t \in [0, 300]$ leads to a value of $k = 0.02421$. This value gives a sum of square residuals of 0.232, which is absolutely tiny.

With this value of k, our predicted water height in the first bucket as a function of t is:

$$h_1 = \begin{cases} 0.0001465t^2 - 0.08862t + 13.4 & \text{for } 0 \leqslant t \leqslant 300 \\ 0 & \text{for } t > 300. \end{cases}$$

7.3 Numerical solution for h_2

Now we will use Euler's method to numerically solve the differential equation

$$\frac{\mathrm{d}h_2}{\mathrm{d}t} = 0.08862 - 0.0002930t - 0.02421\sqrt{h_2}$$

as given by eq. (10) and our optimised value $k = 0.02421$.

Using a step size of 0.2 s, a standard Euler method iteration was applied, and the result plotted alongside the data in the next section.

7.4 Comparison with the data

The predicted water heights and square residuals from this new model are shown below; the sum of square residuals is now just 2.39. This is an enormous decrease from the first model!

	Measured values		Predicted values		Residual	
t / s	h_1 / cm	h_2 / cm	h_1' / cm	h_2' / cm	$h_1' - h_1$ / cm	$h_2' - h_2$ / cm
0	13.4	0.0	13.40	0.00	0.0	0.0
5	13.0	0.4	12.96	0.38	0.0	0.0
10	12.6	0.8	12.53	0.73	0.0	-0.1
15	12.1	1.2	12.10	1.04	0.0	-0.1
20	11.8	1.5	11.69	1.33	-0.1	-0.1
25	11.3	1.8	11.28	1.59	0.0	-0.2
30	11.0	2.0	10.87	1.84	-0.1	-0.1
35	10.5	2.3	10.48	2.06	0.0	-0.2
40	10.2	2.5	10.09	2.27	-0.1	-0.2

45	9.8	2.7	9.71	2.46	-0.1	-0.2
50	9.5	2.9	9.33	2.64	-0.1	-0.2
55	9.1	3.0	8.97	2.81	-0.1	-0.2
60	8.7	3.2	8.61	2.96	-0.1	-0.2
65	8.4	3.4	8.26	3.10	-0.1	-0.2
70	8.0	3.5	7.91	3.23	-0.1	-0.2
75	7.7	3.6	7.58	3.35	-0.1	-0.2
80	7.3	3.7	7.25	3.46	0.0	-0.2
85	7.0	3.8	6.92	3.55	-0.1	-0.2
90	6.7	3.9	6.61	3.64	-0.1	-0.2
95	6.4	4.0	6.30	3.72	0.0	-0.2
100	6.0	4.0	6.00	3.78	0.0	-0.2
105	5.8	4.0	5.71	3.84	0.0	-0.2
110	5.4	4.1	5.42	3.89	0.0	-0.2
115	5.2	4.1	5.15	3.93	0.0	-0.2
120	4.9	4.1	4.87	3.96	0.0	-0.1
125	4.6	4.1	4.61	3.98	0.1	-0.1
130	4.4	4.1	4.35	4.00	0.0	-0.1
135	4.1	4.1	4.11	4.00	0.0	-0.1
140	3.9	4.1	3.86	4.00	0.0	-0.1
145	3.7	4.1	3.63	4.00	0.0	-0.1
150	3.4	4.1	3.40	3.98	0.1	-0.1
155	3.2	4.1	3.18	3.96	0.0	-0.1
160	3.0	4.1	2.97	3.93	0.0	-0.2
165	2.8	4.1	2.76	3.90	0.0	-0.2
170	2.5	4.1	2.57	3.86	0.1	-0.2
175	2.3	4.0	2.38	3.81	0.1	-0.2
180	2.2	4.0	2.19	3.76	0.0	-0.2
185	2.0	3.8	2.02	3.70	0.1	-0.1
190	1.8	3.7	1.85	3.64	0.0	-0.1
195	1.7	3.7	1.69	3.57	0.0	-0.1
200	1.5	3.6	1.53	3.49	0.0	-0.1
205	1.4	3.5	1.39	3.42	0.0	-0.1
210	1.2	3.4	1.25	3.33	0.1	0.0
215	1.0	3.3	1.12	3.24	0.1	0.0
220	0.9	3.1	0.99	3.15	0.1	0.1
225	0.8	3.0	0.88	3.06	0.1	0.1
230	0.7	2.9	0.77	2.96	0.1	0.1
235	0.7	2.8	0.66	2.85	0.0	0.1
240	0.6	2.7	0.57	2.75	0.0	0.1
245	0.5	2.6	0.48	2.63	0.0	0.0
250	0.4	2.5	0.40	2.52	0.0	0.0
255	0.4	2.3	0.33	2.40	-0.1	0.1
260	0.4	2.2	0.26	2.28	-0.1	0.1
265	0.3	2.1	0.20	2.16	-0.1	0.1
270	0.2	2.0	0.15	2.04	0.0	0.1
275	0.2	1.9	0.11	1.91	0.0	0.1
280	0.2	1.7	0.07	1.78	-0.1	0.1
285	0.2	1.6	0.04	1.65	-0.1	0.1
290	0.1	1.4	0.02	1.52	-0.1	0.1
295	0.1	1.3	0.01	1.39	0.0	0.1
300	0.1	1.2	0.00	1.26	-0.1	0.1
305	0.1	1.1	0.00	1.13	-0.1	0.1
310	0.1	1.0	0.00	1.00	-0.1	0.0
315	0.1	0.9	0.00	0.86	-0.1	0.0

320	0.1	0.8	0.00	0.73	-0.1	0.0
325	0.1	0.7	0.00	0.60	-0.1	-0.1
330	0.1	0.7	0.00	0.48	-0.1	-0.2
335	0.1	0.6	0.00	0.36	-0.1	-0.2
340	0.1	0.6	0.00	0.24	-0.1	-0.3
345	0.1	0.5	0.00	0.13	-0.1	-0.3
350	0.1	0.4	0.00	0.03	-0.1	-0.4
355	0.1	0.4	0.00	0.00	-0.1	-0.4
360	0.1	0.3	0.00	0.00	-0.1	-0.3
365	0.1	0.3	0.00	0.00	-0.1	-0.3
370	0.1	0.2	0.00	0.00	-0.1	-0.2
375	0.1	0.2	0.00	0.00	-0.1	-0.2
380	0.1	0.1	0.00	0.00	-0.1	-0.1
385	0.1	0.1	0.00	0.00	-0.1	-0.1
390	0.1	0.1	0.00	0.00	-0.1	-0.1
395	0.1	0.1	0.00	0.00	-0.1	-0.1
400	0.1	0.0	0.00	0.00	-0.1	0.0
405	0.1	0.0	0.00	0.00	-0.1	0.0
410	0.0	0.0	0.00	0.00	0.0	0.0
415	0.0	0.0	0.00	0.00	0.0	0.0

The predicted and measured heights are plotted together in fig. 5, and the residuals as a function of time are shown in fig. 6.

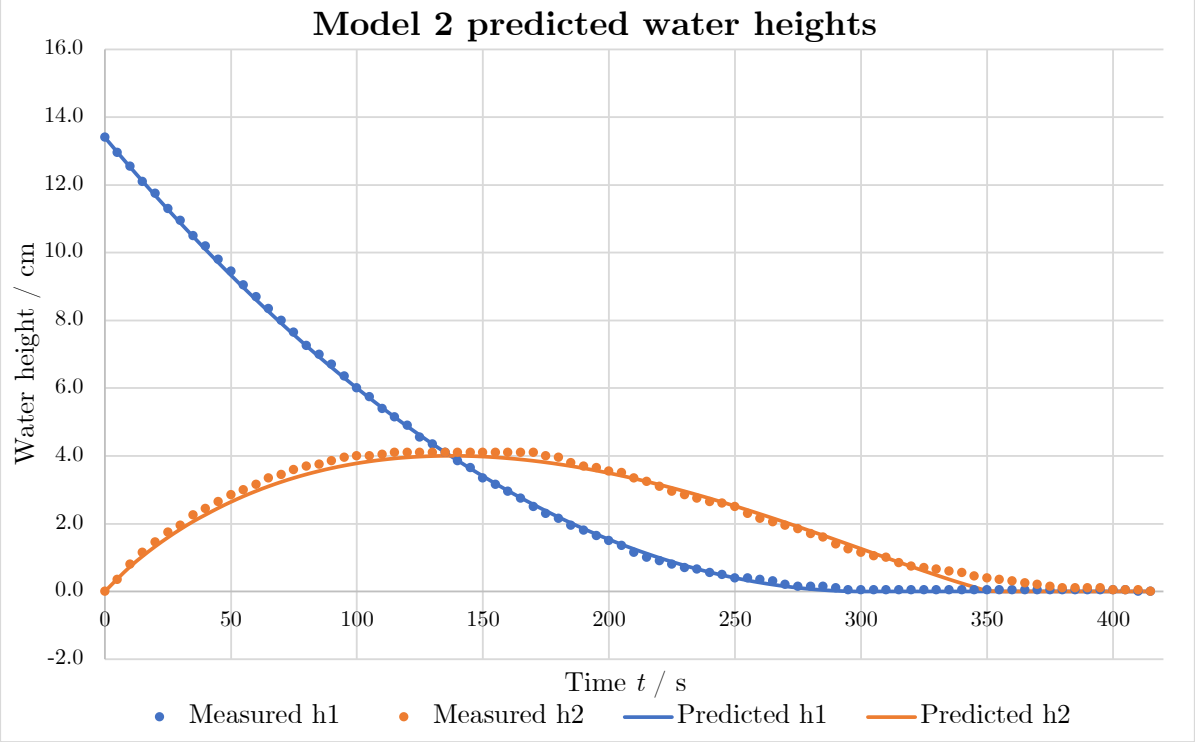


Figure 5: *A plot of predicted and measured heights as a function of time.*

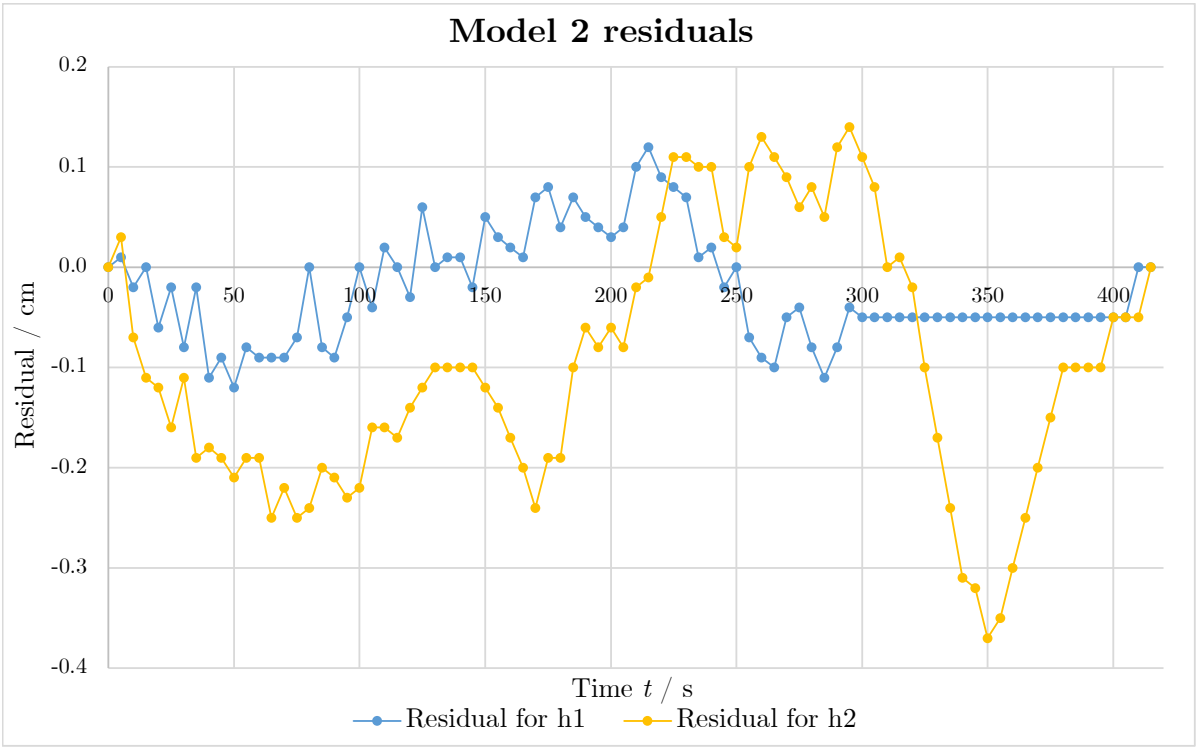


Figure 6: *Residuals as a function of time for the new model.*

Both graphs show that the fit of the new line is remarkable. The distribution of residuals is random (if you consider the levels of precision are working to, any patterns in fig. 6 are not significant). More importantly, however, almost all of the residuals are within our experimental error that we estimated as $\pm 0.2\,\text{cm}$.

The only slight deviation noticeable in fig. 5 is at the very end, where the height is predicted to drop to zero slightly earlier than it actually does. Because of the small amounts of water flowing here and the slow rates, it is quite possible that there are factors coming into play which we did not account for (such as the nature of our spout, the surface tension of the water, the angle of the buckets, etc.).

8 Conclusion

We have tried two different models for the situation; the first with rate of flow proportional to height, and the second with rate of flow proportional to the square root of height. It is clear that the second model is far superior to the first, and indeed the superb accuracy of the second model indicates that our other simplifying assumptions are good assumptions (that making the assumptions hardly changes the situation).

Using the second model, the maximum height of water in the second bucket is predicted to be 4.0 cm, whereas the actual measured value was 4.1 cm. This difference is within experimental error.

In conclusion, we have demonstrated an effective model for the situation.

Chapter 9

Radioactivity and mass-energy equivalence

After studying the radioactive decay with Dr Cheung in Lent 2017, he set us one of his infamous homeworks. These were my solutions (which may be pretty uninteresting without the problems themselves) but I have included them here because of the first section about radioactive decay chains, which I thought was a very beautiful piece of maths and was my first real encounter with differential equations.

Radioactivity and Mass-Energy Equivalence

Damon Falck

April 25, 2017

Solution of a 4-isotope decay chain

Consider four isotopes A, B, C and D, where A decays to B with decay constant λ_1, B decays to C with decay constant λ_2, C decays to D with decay constant λ_3, and D is stable.

We can model the rates of change of numbers N_A, N_B, N_C and N_D of nuclei of isotopes A, B, C and D respectively with the following four differential equations:

$$\frac{\mathrm{d}N_A}{\mathrm{d}t} = -\lambda_1 N_A, \tag{1}$$
$$\frac{\mathrm{d}N_B}{\mathrm{d}t} = \lambda_1 N_A - \lambda_2 N_B, \tag{2}$$
$$\frac{\mathrm{d}N_C}{\mathrm{d}t} = \lambda_2 N_B - \lambda_3 N_C, \tag{3}$$
$$\frac{\mathrm{d}N_D}{\mathrm{d}t} = \lambda_3 N_C. \tag{4}$$

We want to find N_A, N_B, N_C and N_D explicitly as functions of time. At $t = 0$, let the number of particles of N_A be N_0. There are initially no particles of the other three isotopes.

First, we'll solve eq. (1). Rearranging slightly, we get

$$\frac{\mathrm{d}N_A}{N_A} = -\lambda_1 \,\mathrm{d}t$$

and integrating both sides,

$$\begin{aligned} \int \frac{\mathrm{d}N_A}{N_A} &= -\lambda_1 \int \mathrm{d}t \\ \implies \ln N_A &= -\lambda_1 t + c \\ \implies N_A &= \mathrm{e}^{-\lambda_1 t + c} \\ &= \mathrm{e}^{c} \cdot \mathrm{e}^{-\lambda_1 t}. \end{aligned}$$

Setting $\mathrm{e}^c = N_0$ (which must be true to satisfy $N_A = N_0$ at $t = 0$), we have our solution

$$N_A = N_0 \mathrm{e}^{-\lambda_1 t}. \tag{5}$$

Now we move to solving eq. (2). Rewriting it as

$$\frac{\mathrm{d}N_B}{\mathrm{d}t} + \lambda_2 N_B = \lambda_1 N_A,$$

we see that the left hand side resembles the product rule. This works if we multiply through $e^{\lambda_2 t}$:

$$\begin{aligned} e^{\lambda_2 t}\frac{dN_B}{dt} + e^{\lambda_2 t}\lambda_2 N_B &= e^{\lambda_2 t}\lambda_1 N_A \\ \implies \frac{d}{dt}\left(e^{\lambda_2 t}N_B\right) &= e^{\lambda_2 t}\lambda_1 N_A. \end{aligned}$$

Now substituting in our solution for N_A in eq. (5) and integrating,

$$\begin{aligned} e^{\lambda_2 t}N_B &= \int e^{\lambda_2 t}\lambda_1 N_0 e^{-\lambda_1 t}\,dt \\ &= \lambda_1 N_0 \int e^{(\lambda_2-\lambda_1)t}\,dt \\ &= \lambda_1 N_0 \cdot \frac{e^{(\lambda_2-\lambda_1)t}}{\lambda_2-\lambda_1} + c \end{aligned}$$

and so

$$N_B = \frac{\lambda_1 N_0}{\lambda_2-\lambda_1}e^{-\lambda_1 t} + ce^{-\lambda_2 t}. \tag{6}$$

We know that at $t = 0$, $N_B = 0$ and so

$$\begin{aligned} 0 &= \frac{\lambda_1 N_0}{\lambda_2-\lambda_1}e^0 + ce^0 \\ \implies c &= -\frac{\lambda_1 N_0}{\lambda_2-\lambda_1}. \end{aligned}$$

Substituting this value for c back into eq. (6), we finally come to

$$N_B = \frac{\lambda_1 N_0}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right). \tag{7}$$

Now we must solve eq. (3). Although the algebra is longer, we can do this using a very similar method. Rewriting it as

$$\frac{dN_C}{dt} + \lambda_3 N_C = \lambda_2 N_B,$$

we again notice the similarity to the product rule and multiply by $e^{\lambda_3 t}$ to let this simplify:

$$\begin{aligned} e^{\lambda_3 t}\frac{dN_C}{dt} + e^{\lambda_3 t}\lambda_3 N_C &= e^{\lambda_3 t}\lambda_2 N_B \\ \implies \frac{d}{dt}\left(e^{\lambda_3 t}N_C\right) &= e^{\lambda_3 t}\lambda_2 N_B. \end{aligned}$$

Now integrating and substituting and factorising constants, we have

$$e^{\lambda_3 t}N_C = \lambda_2 \int e^{\lambda_3 t}N_B\,dt$$

and so substituting in our solution for N_B from eq. (7), we come to

$$\begin{aligned} e^{\lambda_3 t}N_C &= \lambda_2 \int e^{\lambda_3 t}\frac{\lambda_1 N_0}{\lambda_2-\lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right)dt \\ &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\int \left(e^{(\lambda_3-\lambda_1)t} - e^{(\lambda_3-\lambda_2)t}\right)dt \\ &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{e^{(\lambda_3-\lambda_1)t}}{\lambda_3-\lambda_1} - \frac{e^{(\lambda_3-\lambda_2)t}}{\lambda_3-\lambda_2}\right) + c. \end{aligned}$$

Therefore,

$$N_C = \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}}{\lambda_3-\lambda_2}\right)+c\mathrm{e}^{-\lambda_3 t}. \tag{8}$$

At $t=0$, $N_C=0$, so

$$\begin{aligned} 0 &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{\mathrm{e}^0}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^0}{\lambda_3-\lambda_2}\right)+c\mathrm{e}^0 \\ \implies c &= -\frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{1}{\lambda_3-\lambda_1}-\frac{1}{\lambda_3-\lambda_2}\right). \end{aligned}$$

Substituting this value of c back into eq. (8), we come to our solution

$$\begin{aligned} N_C &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left[\left(\frac{\mathrm{e}^{-\lambda_1 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}}{\lambda_3-\lambda_2}\right)-\left(\frac{1}{\lambda_3-\lambda_1}-\frac{1}{\lambda_3-\lambda_2}\right)\mathrm{e}^{-\lambda_3 t}\right] \\ \implies N_C &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_2}\right). \end{aligned} \tag{9}$$

Finally, we have eq. (4) to solve. Directly integrating, we have

$$N_D = \int \lambda_3 N_C \,\mathrm{d}t$$

and we can now substitute in our solution for N_C from eq. (9) and simplify, coming to

$$\begin{aligned} N_D &= \int \lambda_3 \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_2}\right)\mathrm{d}t \\ &= \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\int\left(\frac{\mathrm{e}^{-\lambda_1 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_2}\right)\mathrm{d}t \\ &= \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\left[\frac{1}{\lambda_3-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}}{-\lambda_1}-\frac{\mathrm{e}^{-\lambda_3 t}}{-\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_2}\left(\frac{\mathrm{e}^{-\lambda_2 t}}{-\lambda_2}-\frac{\mathrm{e}^{-\lambda_3 t}}{-\lambda_3}\right)\right]+c \\ &= \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\left[\frac{1}{\lambda_3-\lambda_2}\left(\frac{\mathrm{e}^{-\lambda_2 t}}{\lambda_2}-\frac{\mathrm{e}^{-\lambda_3 t}}{\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}}{\lambda_1}-\frac{\mathrm{e}^{-\lambda_3 t}}{\lambda_3}\right)\right]+c. \end{aligned} \tag{10}$$

Now again using the fact that $N_D=0$ at $t=0$,

$$c = -\frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\left[\frac{1}{\lambda_3-\lambda_2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_1}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_3}\right)\right].$$

Substituting this expression back into eq. (10), we get

$$\begin{aligned} N_D = \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\Bigg[&\frac{1}{\lambda_3-\lambda_2}\left(\frac{\mathrm{e}^{-\lambda_2 t}}{\lambda_2}-\frac{\mathrm{e}^{-\lambda_3 t}}{\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}}{\lambda_1}-\frac{\mathrm{e}^{-\lambda_3 t}}{\lambda_3}\right) \\ &-\frac{1}{\lambda_3-\lambda_2}\left(\frac{1}{\lambda_2}-\frac{1}{\lambda_3}\right)+\frac{1}{\lambda_3-\lambda_1}\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_3}\right)\Bigg] \end{aligned}$$

$$\implies N_D = \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\left[\frac{1}{\lambda_3-\lambda_2}\left(\frac{\mathrm{e}^{-\lambda_2 t}-1}{\lambda_2}-\frac{\mathrm{e}^{-\lambda_3 t}-1}{\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}-1}{\lambda_1}-\frac{\mathrm{e}^{-\lambda_3 t}-1}{\lambda_3}\right)\right].$$

This is our final solution for N_D.

Hence, our solutions to the original differential equations are:

$$\begin{aligned}
N_A &= N_0\mathrm{e}^{-\lambda_1 t},\\
N_B &= \frac{\lambda_1 N_0}{\lambda_2-\lambda_1}\left(\mathrm{e}^{-\lambda_1 t}-\mathrm{e}^{-\lambda_2 t}\right),\\
N_C &= \frac{\lambda_1\lambda_2 N_0}{\lambda_2-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_1}-\frac{\mathrm{e}^{-\lambda_2 t}-\mathrm{e}^{-\lambda_3 t}}{\lambda_3-\lambda_2}\right),\\
N_D &= \frac{\lambda_1\lambda_2\lambda_3 N_0}{\lambda_2-\lambda_1}\left[\frac{1}{\lambda_3-\lambda_2}\left(\frac{\mathrm{e}^{-\lambda_2 t}-1}{\lambda_2}-\frac{\mathrm{e}^{-\lambda_3 t}-1}{\lambda_3}\right)-\frac{1}{\lambda_3-\lambda_1}\left(\frac{\mathrm{e}^{-\lambda_1 t}-1}{\lambda_1}-\frac{\mathrm{e}^{-\lambda_3 t}-1}{\lambda_3}\right)\right].
\end{aligned}$$

Chapter 64

1. An increase in temperature causes a higher average vibrational kinetic energy of the molecules of a substance — but does not change the motion of subatomic particles within each molecule with respect to one another. Therefore, with higher temperature the rate of chemical reaction increases as there are more frequent successful collisions (those with an energy greater or equal to the activation energy of the reaction) between molecules; but the relative energies of the particles within each molecule are unchanged and so the rate of nuclear reaction remains constant.

5. Let a nucleus have initial velocity 0. So, when it disintegrates into an α-particle of mass m_α and velocity v_α and a smaller nucleus of mass m' and velocity v', by conservation of momentum
$$m'v' = m_\alpha v_\alpha.$$
Therefore, by measuring the recoil velocity v' of the nucleus (which is much easier to measure), we can find the speed
$$v_\alpha = \frac{m'v'}{m_\alpha} = \frac{m'v'}{4m_p}$$
of the alpha particle and hence find its kinetic energy
$$\frac{1}{2}m_\alpha v_\alpha^2 = 2m_p v_\alpha^2 = \frac{(m'v')^2}{8m_p}.$$
(This is using non-relativistic methods only.)

13. Most rocks contain between 1 and 3 ppm of uranium, which has a half life of about 4.5 billion years (roughly equal to the age of the Earth). Uranium undergoes alpha decay to create thorium, which in turn emits an alpha particle to form radium. (This can happen with several different isotopes of uranium and radium.) Therefore, while radium has a very short half life, it is present as the result of the natural decay of uranium.

15. (a) The number of particles of any radioactive isotope after a time t is

$$N(t) = N_0 \mathrm{e}^{-\lambda t} \tag{11}$$

where λ is the decay constant of the isotope and N_0 is the initial number of particles. Using this equation we also come to find that the half-life $T_{1/2}$ of the isotope is

$$T_{1/2} = \frac{\ln 2}{\lambda}. \tag{12}$$

Therefore we can find the number of particles in terms of the half-life to be

$$N(t) = N_0 \cdot 2^{-t/T_{1/2}},$$

a result that makes intuitive sense. So, if $T_{1/2} = 3.00\,\mathrm{s}$ and $N_0 = 5.12 \times 10^{20}$ as in the question, then:

i. At $t = 3.00\,\mathrm{s}$,

$$N(3.00\,\mathrm{s}) = 5.12 \times 10^{20} \cdot 2^{-\frac{3.00\,\mathrm{s}}{3.00\,\mathrm{s}}} = 2.56 \times 10^{20}.$$

ii. At $t = 6.00\,\mathrm{s}$,

$$N(6.00\,\mathrm{s}) = 5.12 \times 10^{20} \cdot 2^{-\frac{6.00\,\mathrm{s}}{3.00\,\mathrm{s}}} = 1.28 \times 10^{20}.$$

iii. At $t = 12.00\,\mathrm{s}$,

$$N(12.00\,\mathrm{s}) = 5.12 \times 10^{20} \cdot 2^{-\frac{12.00\,\mathrm{s}}{3.00\,\mathrm{s}}} = 3.20 \times 10^{19}.$$

iv. At $t = 1\,\mathrm{min} = 60.0\,\mathrm{s}$,

$$N(60.0\,\mathrm{s}) = 5.12 \times 10^{20} \cdot 2^{-\frac{60.0\,\mathrm{s}}{3.00\,\mathrm{s}}} = 4.88 \times 10^{14}.$$

These answers are fairly exact because of the large number of particles in question.

(b) By dividing eq. (11) by time, we come to

$$A(t) = A_0 \mathrm{e}^{-\lambda t} = A_0 \cdot 2^{-t/T_{1/2}} \tag{13}$$

where $A(t)$ is the activity at time t and A_0 is the initial activity. So,

$$\begin{aligned} \frac{A(t)}{A_0} &= 2^{-t/T_{1/2}} \\ \implies -\frac{t}{T_{1/2}} \log 2 &= \log\left(\frac{A(t)}{A_0}\right) \\ \implies t &= -\frac{T_{1/2} \log\left(\frac{A(t)}{A_0}\right)}{\log 2}. \end{aligned}$$

Therefore, if as given in the question $\frac{A(t)}{A_0} = 2^{-40}$ and as before $T_{1/2} = 3.00\,\mathrm{s}$, then

$$\begin{aligned} t &= -\frac{3.00\,\mathrm{s} \cdot \log_2 2^{-40}}{\log_2 2} \\ &= -\frac{3.00\,\mathrm{s} \cdot (-40)}{1} = 120\,\mathrm{s}. \end{aligned}$$

16. (a) Using eq. (12) above with $T_{1/2} = 51 \times 10^9\,\mathrm{s}$, our decay constant is

$$\lambda = \frac{\ln 2}{51 \times 10^9\,\mathrm{s}} = 1.4 \times 10^{-11}\,\mathrm{s}^{-1}.$$

(b) Using the same equation with $\lambda = 6.9 \times 10^{-4}\,\mathrm{s}^{-1}$,

$$T_{1/2} = \frac{\ln 2}{6.9 \times 10^{-4}\,\mathrm{s}^{-1}} = 1.0 \times 10^{3}\,\mathrm{s}.$$

17. If the large sample has a decay constant λ an initial number of nuclei N_0, then after a time period of observation τ_1 we have

$$N(\tau_1) = N_0 \cdot 2^{-\tau_1/T_{1/2}}$$

nuclei remaining. Therefore after a further interval τ_2 there are

$$N(\tau_1 + \tau_2) = N_0 \cdot 2^{-(\tau_1+\tau_2)/T_{1/2}}$$

nuclei and so the proportion of nuclei not decayed during this interval τ_2 is

$$\begin{aligned}\frac{N(\tau_1+\tau_2)}{N(\tau_1)} &= \frac{N_0 \cdot 2^{-(\tau_1+\tau_2)/T_{1/2}}}{N_0 \cdot 2^{-\tau_1/T_{1/2}}} \\ &= \frac{2^{\tau_1/T_{1/2}}}{2^{(\tau_1+\tau_2)/T_{1/2}}} \\ &= \frac{1}{2^{\tau_2/T_{1/2}}} = 2^{-\tau_2/T_{1/2}}.\end{aligned}$$

This is the same as the probability that a randomly selected nucleus *will* decay during this second time interval. We see that this is independent of the initial observation interval τ_1, a sensible result as the probability of an individual particle nucleus decaying during a given interval should depend only on the half-life of the isotope (or its decay constant).

The probability of a randomly selected nucleus *not* decaying during time τ_2 is therefore $P = 1 - 2^{-\tau_2/T_{1/2}}$. (We could have found this also by working from the definition of the half-life.) Hence,

(a) If $\tau_2 = T_{1/2}$ then

$$P = 1 - 2^{-T_{1/2}/T_{1/2}} = 1 - 2^{-1} = 0.5.$$

(b) If $\tau_2 = 3T_{1/2}$ then

$$P = 1 - 2^{-3T_{1/2}/T_{1/2}} = 1 - 2^{-3} = 0.875.$$

19. The activity A_{Rn} of the radon-222 in equilibrium with a sample of radium-226 (with decay constant λ_{Ra}) is

$$A_{\mathrm{Rn}} = A_{\mathrm{Ra}} = -\frac{\mathrm{d}N_{\mathrm{Ra}}}{\mathrm{d}t} = N_{\mathrm{Ra}}\lambda_{\mathrm{Ra}}$$

where N_{Ra} is the number of radium-226 nuclei present, and so if we have mass m of radium then

$$A_{\mathrm{Rn}} = \frac{mN_A\lambda_{\mathrm{Ra}}}{226\,\mathrm{g\,mol}^{-1}}$$

where N_A is Avogadro's constant. Therefore, using $\lambda_{\mathrm{Ra}} = 1.4 \times 10^{-11}\,\mathrm{s}^{-1}$ and $N_A = 6.02 \times 10^{-23}\,\mathrm{mol}^{-1}$, if $m = 1.0\,\mathrm{g}$ then

$$A_{\mathrm{Rn}} = \frac{1.0\,\mathrm{g} \cdot 6.02 \times 10^{-23}\,\mathrm{mol}^{-1} \cdot 1.4 \times 10^{-11}\,\mathrm{s}^{-1}}{226\,\mathrm{g\,mol}^{-1}} = 3.7 \times 10^{10}\,\mathrm{Bq}.$$

21. The corrected count rate of a substance is proportional to its activity, and so

$$\frac{C_1 - C_B}{C_0 - C_B} = \frac{A_1}{A_0}$$

where C_B is the background count rate, C_0 and C_1 are the initial and final measured count rates respectively and A_0 and A_1 are the initial and final activities respectively. So,

$$\frac{C_1 - C_B}{C_0 - C_B} = \frac{A_0 \cdot 2^{-t/T_{1/2}}}{A_0} = 2^{-t/T_{1/2}}$$

which implies

$$\frac{t}{T_{1/2}} = -\log_2\left(\frac{C_1 - C_B}{C_0 - C_B}\right)$$

and so

$$T_{1/2} = \frac{t}{\log_2(C_0 - C_B) - \log_2(C_1 - C_B)}.$$

Therefore, if as in the question $t = 210\,\mathrm{s}$, $C_0 = 82\,\mathrm{s}^{-1}$, $C_1 = 19\,\mathrm{s}^{-1}$ and $C_B = 10\,\mathrm{s}^{-1}$, then

$$T_{1/2} = \frac{210\,\mathrm{s}}{\log_2(82\,\mathrm{s}^{-1} - 10\,\mathrm{s}^{-1}) - \log_2(19\,\mathrm{s}^{-1} - 10\,\mathrm{s}^{-1})} = 70\,\mathrm{s}.$$

22. The activity (rate of disintegration) of a substance with N particles and decay constant λ is

$$-\frac{\mathrm{d}N}{\mathrm{d}t} = N\lambda$$

and so using the ideal gas law $PV = Nk_BT$,

$$-\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{PV\lambda}{k_BT}.$$

Therefore for a sample of volume $V = 1.0\,\mathrm{mm}^3$ and decay constant $\lambda = 2.1 \times 10^{-6}\,\mathrm{s}^{-1}$ at s.t.p. ($P = 101.325\,\mathrm{kPa}$, $T = 273.15\,\mathrm{K}$), using the Boltzmann constant $k_B = 1.38 \times 10^{-23}\,\mathrm{J\,K}^{-1}$, our activity is

$$-\frac{\mathrm{d}N}{\mathrm{d}t} = \frac{101.325\,\mathrm{kPa} \cdot 1.0\,\mathrm{mm}^3 \cdot 2.1 \times 10^{-6}\,\mathrm{s}^{-1}}{1.38 \times 10^{-23}\,\mathrm{J\,K}^{-1} \cdot 273.15\,\mathrm{K}} = 5.6 \times 10^{10}\,\mathrm{Bq}.$$

(This could also have been done using the nucleon number of radon-222 and the molar gas volume at s.t.p.)

25. (a) Graph in fig. 1.

(b) Since by eq. (13)

$$A = A_0\mathrm{e}^{-\lambda t},$$

it follows that

$$\ln A = \ln(A_0) - \lambda t.$$

Hence, the gradient of a graph of $\ln A$ against t is $-\lambda$. The gradient of the line of regression in fig. 1 is $-\lambda = -0.0128\,\mathrm{s}^{-1}$ and so

$$T_{1/2} = \frac{\ln 2}{\lambda} = \frac{\ln 2}{0.0128\,\mathrm{s}^{-1}} = 54\,\mathrm{s}.$$

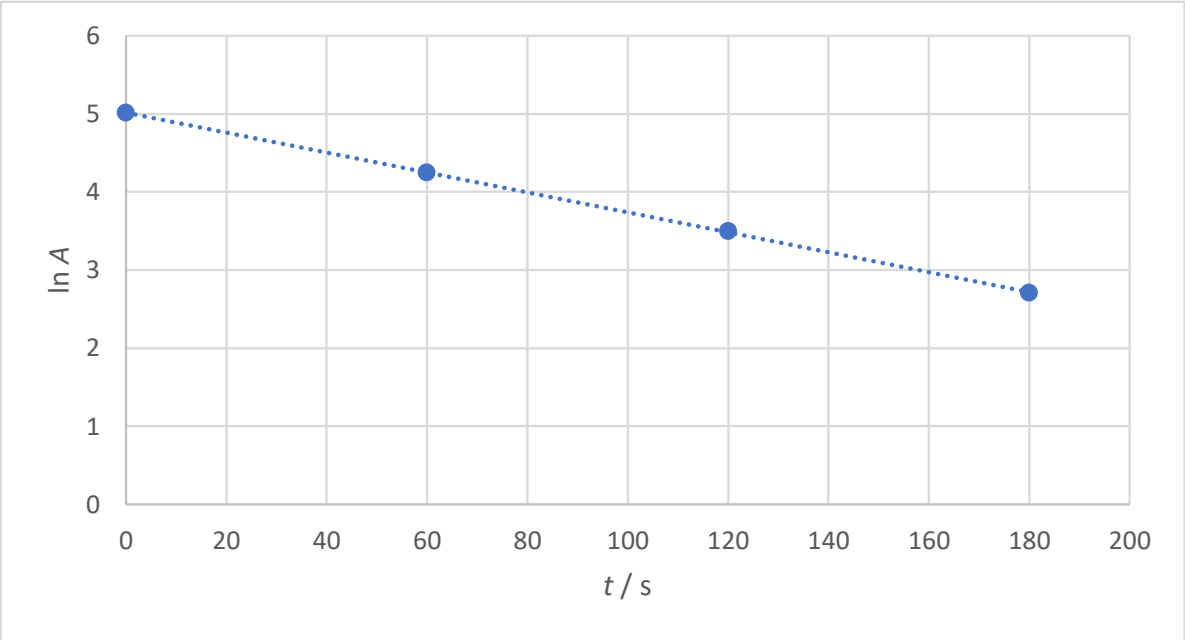


Figure 1: Graph for question 25.

26. (a) In a sample containing a mass m of uranium-238 and a mass ηm of lead-206, we know that the number of atoms N_{U} of uranium-238 is

$$N_{\mathrm{U}} = \frac{mN_A}{A_{\mathrm{U}}}$$

where $A_{\mathrm{U}} = 238\,\mathrm{g\,mol^{-1}}$ is the atomic mass of uranium-238 and $N_A = 6.02 \times 10^{23}\,\mathrm{mol^{-1}}$ is Avogadro's constant. Similarly,

$$N_{\mathrm{Pb}} = \frac{\eta m N_A}{A_{\mathrm{Pb}}}$$

where $A_{\mathrm{Pb}} = 206\,\mathrm{g\,mol^{-1}}$ is the atomic mass of lead-206. Therefore, if $m = 1.0\,\mathrm{g}$ and $\eta = \frac{1}{5}$ as in the question, then

$$N_{\mathrm{U}} = \frac{1.0\,\mathrm{g} \cdot 6.02 \times 10^{23}\,\mathrm{mol^{-1}}}{238\,\mathrm{g\,mol^{-1}}} = 2.5 \times 10^{21}$$

and

$$N_{\mathrm{Pb}} = \frac{\frac{1}{5} \cdot 1.0\,\mathrm{g} \cdot 6.02 \times 10^{23}\,\mathrm{mol^{-1}}}{206\,\mathrm{g\,mol^{-1}}} = 5.8 \times 10^{20}.$$

(b) Assuming all of the lead-206 in this sample was produced by the decay of uranium-238, that all uranium-238 nuclei that decay become lead-206, and that lead-206 is stable, it's apparent that the initial number of uranium-238 atoms $N_{0,\mathrm{U}}$ is equal to the sum of the current number of lead-206 and uranium-238 atoms:

$$\begin{aligned} N_{0,\mathrm{U}} &= N_{\mathrm{Pb}} + N_{\mathrm{U}} \\ &= 5.8 \times 10^{20} + 2.5 \times 10^{21} = 3.1 \times 10^{21}. \end{aligned}$$

(c) Therefore, we can find the age of the rock. We know

$$N_{\mathrm{U}} = N_{0,\mathrm{U}} \cdot 2^{-t/T_{1/2}}$$

and so using $T_{1/2} = 1.4 \times 10^{17}\,\mathrm{s}$,

$$\begin{aligned} t &= -T_{1/2} \log_2 \left(\frac{N_\mathrm{U}}{N_{0,\mathrm{U}}} \right) \\ &= -1.4 \times 10^{17}\,\mathrm{s} \cdot \log_2 \left(\frac{2.5 \times 10^{21}}{3.1 \times 10^{21}} \right) = 4.2 \times 10^{16}\,\mathrm{s}. \end{aligned}$$

27. (a) i. The rate of decay of the uranium is its activity $N_0\lambda_0$, which is clearly equal to the rate of formation of radium atoms.

ii. The rate of decay of the radium atoms is $N\lambda$. However, by the definition of equilibrium, the number of radium atoms is constant so

$$\begin{aligned} \frac{\mathrm{d}N}{\mathrm{d}t} &= N_0\lambda_0 - N\lambda = 0 \\ \implies N\lambda &= N_0\lambda_0. \end{aligned} \tag{14}$$

(b) By eq. (14),

$$\begin{aligned} \frac{N}{N_0} &= \frac{\lambda_0}{\lambda} \\ &= \frac{\left(\frac{\ln 2}{T_{0,1/2}}\right)}{\left(\frac{\ln 2}{T_{1/2}}\right)} \\ \implies \frac{N}{N_0} &= \frac{T_{1/2}}{T_{0,1/2}}. \end{aligned} \tag{15}$$

(c) By eq. (15), if $N_0 = 1.0 \times 10^{23}$, $T_{0,1/2} = 1.4 \times 10^{17}\,\mathrm{s}$ and $T_{1/2} = 51 \times 10^9\,\mathrm{s}$, we have

$$N = \frac{T_{1/2} N_0}{T_{0,1/2}} = \frac{51 \times 10^9\,\mathrm{s} \cdot 1.0 \times 10^{23}}{1.4 \times 10^{17}\,\mathrm{s}} = 3.6 \times 10^{16}$$

radium nuclei.

(d) Radium-226 has atomic mass $226\,\mathrm{g\,mol^{-1}}$ and so the mass of radium present is

$$m = \frac{N \cdot 226\,\mathrm{g\,mol^{-1}}}{N_A} = \frac{3.6 \times 10^{16} \cdot 226\,\mathrm{g\,mol^{-1}}}{6.02 \times 10^{23}\,\mathrm{mol^{-1}}} = 1.4 \times 10^{-8}\,\mathrm{kg}.$$

28. (a) We'll derive the result $T_\mathrm{av} = \frac{1}{\lambda}$ in two ways from the given equations

$$T_\mathrm{av} = \frac{T_\mathrm{tot}}{N_0} \tag{16}$$

and

$$T_\mathrm{tot} = \int_{t=0}^{t=\infty} t\,\mathrm{d}N \tag{17}$$

as well as the decay equation

$$N = N_0 \mathrm{e}^{-\lambda t}. \tag{18}$$

i. First, we'll directly evaluate the integral given. We start by noting that eq. (18) implies that at $t = \infty$,

$$N = \lim_{t\to\infty} N_0 \mathrm{e}^{-\lambda t} = 0$$

and that at $t = 0$,

$$N = \lim_{t \to 0^+} N_0 \mathrm{e}^{-\lambda t} = N_0.$$

So, using these new bounds, our integral from eq. (17) becomes

$$T_{\text{tot}} = \int_0^{N_0} t \, \mathrm{d}N.$$

Now finding t in terms of N, eq. (18) gives us that

$$\begin{aligned} -\lambda t &= \ln\left(\frac{N}{N_0}\right) \\ \implies t &= \frac{\ln N_0 - \ln N}{\lambda} \end{aligned}$$

and so our integral is now

$$\begin{aligned} T_{\text{tot}} &= \int_0^{N_0} \left(\frac{\ln N_0 - \ln N}{\lambda}\right) \mathrm{d}N \\ &= -\frac{1}{\lambda} \int_0^{N_0} \ln N \, \mathrm{d}N + \int_0^{N_0} \frac{\ln N_0}{\lambda}. \end{aligned}$$

Making use of the identity $\int \ln x = x(\ln x - 1) + c$, we come to

$$\begin{aligned} T_{\text{tot}} &= -\frac{1}{\lambda} \left[N(\ln N - 1)\right]_0^{N_0} + \left[\frac{\ln N_0 N}{\lambda}\right]_0^{N_0} \\ &= \left[N\left(\frac{\ln N_0 - \ln N_0 + 1}{\lambda}\right)\right]_0^{N_0} \\ &= \left[\frac{N}{\lambda}\right]_0^{N_0} \end{aligned}$$

and so,

$$T_{\text{tot}} = \frac{N_0}{\lambda} - \frac{0}{\lambda} = \frac{N_0}{\lambda}.$$

Therefore, by eq. (16),

$$T_{\text{av}} = \frac{T_{\text{tot}}}{N_0} = \frac{\frac{N_0}{\lambda}}{N_0} = \frac{1}{\lambda}$$

as desired.

ii. Secondly, we can use the intuitive fact that due to the positive exponential nature of the relation between N and t,

$$\int_{t=0}^{t=\infty} t \, \mathrm{d}N = \int_{t=0}^{t=\infty} N \, \mathrm{d}t.$$

(This is intuitively obvious because the total area under the curve in the upper right quadrant will not change if the axes are flipped.) So,

$$\begin{aligned} T_{\text{tot}} &= \int_0^{\infty} N \, \mathrm{d}t \\ &= \int_0^{\infty} N_0 \mathrm{e}^{-\lambda t} \, \mathrm{d}t. \end{aligned}$$

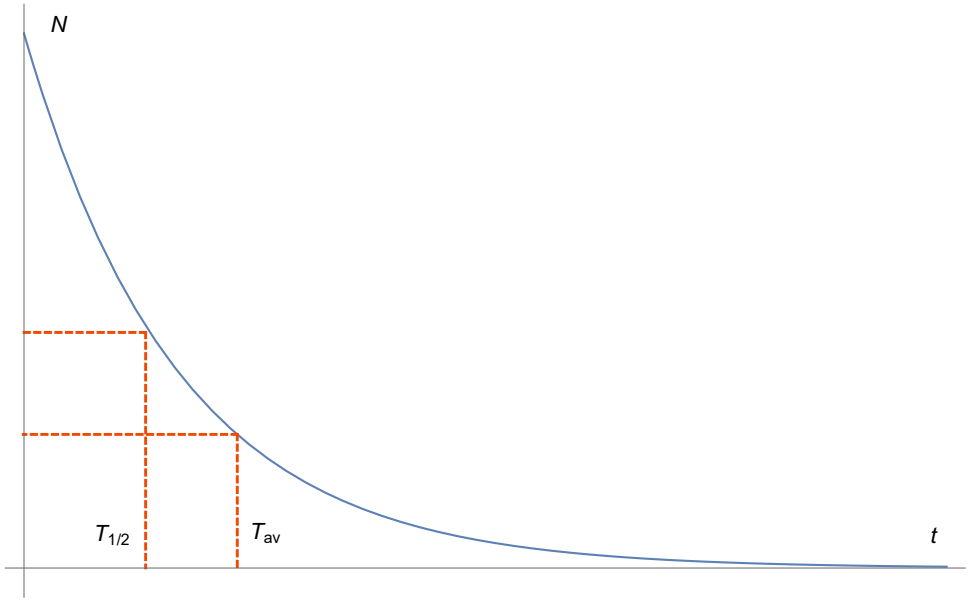


Figure 2: A graph of N against t with the positions of $T_{1/2}$ and T_{av} marked, for question 28.

Evaluating this simple integral,

$$\begin{aligned} T_{\text{tot}} &= \left[-\frac{N_0}{\lambda}\mathrm{e}^{-\lambda t}\right]_0^\infty \\ &= \lim_{t\to\infty}\left(-\frac{N_0}{\lambda}\mathrm{e}^{-\lambda t}\right) - \left(-\frac{N_0}{\lambda}\mathrm{e}^{-\lambda\cdot 0}\right) \\ &= 0 - \left(-\frac{N_0}{\lambda}\cdot 1\right) = \frac{N_0}{\lambda} \end{aligned}$$

and so as previously, using eq. (16) gives

$$T_{\text{av}} = \frac{1}{\lambda}$$

as required.

(b) A graph of N against t is shown in fig. 2. The average lifetime is related to the half-life by

$$T_{1/2} = \ln 2 \cdot T_{\text{av}}.$$

(c) Using $T_{1/2} = 51 \times 10^9\,\text{s}$,

$$T_{\text{av}} = \frac{1}{\lambda} = \frac{T_{1/2}}{\ln 2} = 51 \times 10^9\,\text{s}\ln 2 = 7.4 \times 10^{10}\,\text{s}.$$

30. (a) Absorbed dose is defined as

$$D = \frac{E}{m}$$

where E is the energy absorbed and m is the mass of the absorbing material (air in this case). Using the definition of density ρ,

$$D = \frac{E}{\rho V}$$

where V is the volume of air. Now let E_I be the energy required to create one ion pair and let N_I be the total number of ion pairs created. So, $E = E_I N_I$ and hence

$$D = \frac{E_I N_I}{\rho V}.$$

Therefore the dose received D_G by $V = 1.0 \times 10^{-6}\,\mathrm{m}^3$ of air at s.t.p. ($\rho = 1.3\,\mathrm{kg\,m^{-3}}$), in which $N_I = 2.1 \times 10^9$ ion pairs were created and it takes $E_I = 5.1 \times 10^{-18}\,\mathrm{J}$ of energy to create one pair, is

$$D_G = \frac{5.1 \times 10^{-18}\,\mathrm{J} \cdot 2.1 \times 10^9}{1.3\,\mathrm{kg\,m^{-3}} \cdot 1.0 \times 10^{-6}\,\mathrm{m}^3} = 8.2 \times 10^{-3}\,\mathrm{Gy}.$$

This is equal to 1 röntgen.

(b) Consequently, 1 milliröntgen per week is equal to

$$\frac{D_G}{1000 \cdot 7 \cdot 24 \cdot 3600\,\mathrm{s}} = \frac{8.2 \times 10^{-3}\,\mathrm{Gy}}{60.48 \times 10^7\,\mathrm{s}} = 1.4 \times 10^{-11}\,\mathrm{Gy\,s^{-1}}.$$

Chapter 63

40. Let the body have mass m and fall through height h. The change in gravitational potential energy of the body is

$$\Delta E = mgh$$

and the equivalent change in mass is

$$\Delta m = \frac{\Delta E}{c^2}.$$

(We're neglecting the fact that the force due to gravity will be changing very slowly due to the continuous change in the object's height and mass.) So, the fractional change in mass of the object is

$$\frac{\Delta m}{m} = \frac{\frac{\Delta E}{c^2}}{\frac{\Delta E}{gh}} = \frac{gh}{c^2}.$$

Therefore with $g = 9.8\,\mathrm{m\,s^{-2}}$ and $c = 3.0 \times 10^8\,\mathrm{m\,s^{-1}}$ for a body falling through $h = 6.0\,\mathrm{m}$,

$$\frac{\Delta m}{m} = \frac{9.8\,\mathrm{m\,s^{-2}} \cdot 6.0\,\mathrm{m}}{(3.0 \times 10^8\,\mathrm{m/s})^2} = 6.5 \times 10^{-16}.$$

41. Consider an electron of momentum p_1 and a positron of momentum p_2 from their centre-of-momentum frame (which is equivalent to all other inertial frames). Taking the electron-positron system as one body, the total energy is

$$E = \sqrt{(2m_e)^2c^4 + (p_1 + p_2)^2c^2},$$

however since in this frame the total initial momentum $p_1 + p_2$ is zero, this simplifies down to

$$E = 2m_e c^2.$$

If the particles now collide, let the total momentum of the photons released be p. So, as photons are massless,

$$E = pc$$

and therefore

$$p = 2m_e c.$$

However, by conservation of momentum we require $p_1 + p_2 = p = 0$ which clearly we cannot satisfy with one photon which can never have zero momentum. Therefore we need at least two photons to be released such that their momenta sum to zero.

43. (a) If the γ-ray has wavelength λ and frequency $f = 2.6 \times 10^{20}\,\mathrm{Hz}$, then the photon has linear momentum

$$p = \frac{h}{\lambda} = \frac{hf}{c} = \frac{6.63 \times 10^{-34}\,\mathrm{J\,s} \cdot 2.6 \times 10^{20}\,\mathrm{Hz}}{3.00 \times 10^{8}\,\mathrm{m\,s^{-1}}} = 5.7 \times 10^{-22}\,\mathrm{N\,s}$$

where $h = 6.63 \times 10^{-34}\,\mathrm{J\,s}$ is Planck's constant.

(b) The initial energy of the photon is pc, and the final energy of the electron-positron pair is (as they are both massive) $2\gamma m_e c^2$, where

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{19}$$

and v is the velocity of each β-particle. Therefore, assuming energy is conserved (since the k.e. imparted to the nucleus is negligible),

$$pc = 2\gamma m_e c^2$$

and so

$$p = 2\gamma m_e c.$$

Using eq. (19),

$$p = \frac{2m_e c}{\sqrt{1 - \frac{v^2}{c^2}}}$$

and now rearranging to find v,

$$\begin{aligned}
\sqrt{1 - \frac{v^2}{c^2}} &= \frac{2m_e c}{p} \\
\implies \frac{v^2}{c^2} &= 1 - \frac{4m_e^2 c^2}{p^2} \\
\implies v &= c\sqrt{1 - \frac{4m_e^2 c^2}{p^2}}.
\end{aligned}$$

Hence using the value $p = 5.7 \times 10^{-22}\,\mathrm{N\,s}$ with the electron rest mass $m_e = 9.11 \times 10^{-31}\,\mathrm{kg}$,

$$\begin{aligned}
v &= 3.00 \times 10^{8}\,\mathrm{m\,s^{-1}} \cdot \sqrt{1 - \frac{4 \cdot (9.11 \times 10^{-31}\,\mathrm{kg})^2 \cdot (3.00 \times 10^{8}\,\mathrm{m\,s^{-1}})^2}{(5.7 \times 10^{-22}\,\mathrm{N\,s})^2}} \\
&= 9.3 \times 10^{7}\,\mathrm{m\,s^{-1}}.
\end{aligned}$$

(c) The maximum total linear momentum would be if both particles are emitted in the same direction and so the total linear momentum is equal to the sum of the linear momenta of each β-particle. Since the β-particles are massive, their momentum is given by

$$p_e = \gamma m_e v$$

and so the maximum total linear momentum is

$$\begin{aligned} 2p_e &= 2\gamma m_e v \\ &= \frac{2m_e v}{\sqrt{1-\frac{v^2}{c^2}}}. \end{aligned}$$

So, if $v = 9.3 \times 10^7\,\mathrm{m\,s^{-1}}$ as found previously,

$$2p_e = \frac{2 \cdot 9.11 \times 10^{-31}\,\mathrm{kg} \cdot 9.3 \times 10^7\,\mathrm{m\,s^{-1}}}{\sqrt{1-\frac{(9.3 \times 10^7\,\mathrm{m\,s^{-1}})^2}{(3.00 \times 10^8\,\mathrm{m\,s^{-1}})^2}}} = 1.8 \times 10^{-22}\,\mathrm{N\,s}.$$

44. (a) Because there is only one electron and one proton in the hydrogen atom, the binding energy of the system is equal to the ionization energy of a hydrogen atom, which is given to be $E = 2.2\,\mathrm{aJ}$.

(b) The mass defect of the system is

$$\Delta m = \frac{E}{c^2} = 2.44 \times 10^{-35}\,\mathrm{kg}.$$

(c) The fractional mass decrease is

$$\frac{\Delta m}{m} = \frac{2.44 \times 10^{-35}\,\mathrm{kg}}{1.67 \times 10^{-27}\,\mathrm{kg}} = 1.5 \times 10^{-8}$$

where $m = 1.67 \times 10^{-27}\,\mathrm{kg}$ is the total mass of the hydrogen atom.

The mass of protons, electrons and hydrogen atoms are normally measured using mass spectrometry, a very precise technique that measures each particle's mass-to-charge ratio by measuring its deflection through an electromagnetic field. Modern mass spectrometry can have a resolving power of up to $2,000,000$, meaning it can detect a minimum fractional change in mass of $\frac{1}{2,000,000} = 50 \times 10^{-8}$.

However, the fractional mass decrease from above, 1.5×10^{-8}, is much smaller than this threshold and so we cannot directly measure this mass defect.

45. (a) The mass defect Δm of one deuterium atom is

$$\Delta m = m_\mathrm{H} + m_n - m$$

where $m = 2.014\,10\,\mathrm{u}$ is the measured mass of the deuterium atom, $m_\mathrm{H} = 1.007\,82\,\mathrm{u}$ is the mass of a hydrogen atom and $m_n = 1.008\,66\,\mathrm{u}$ is the mass of a neutron. So,

$$\Delta m = 1.007\,82\,\mathrm{u} + 1.008\,66\,\mathrm{u} - 2.014\,10\,\mathrm{u} = 2.38 \times 10^{-3}\,\mathrm{u} = 3.95 \times 10^{-30}\,\mathrm{kg}.$$

(b) The binding energy of one atom is therefore given by

$$E = \Delta m c^2 = 3.95 \times 10^{-30}\,\mathrm{kg} \cdot (3.00 \times 10^8\,\mathrm{m\,s^{-1}})^2 = 3.56 \times 10^{-13}\,\mathrm{J}.$$

(c) The binding energy per nucleon is therefore $\frac{3.56 \times 10^{-13}\,\mathrm{J}}{2} = 1.78 \times 10^{-13}\,\mathrm{J}$.

This is about $\frac{1}{7}$ of the binding energy per nucleon of iron; this indicates that deuterium will very readily fuse with other nuclei.

46. (a) The mass m_T of 9 hydrogen atoms plus 8 neutrons (at infinite separation at rest) is

$$m_T = 8m_\mathrm{H} + 8m_n = 8 \cdot 1.0078\,\mathrm{u} + 8 \cdot 1.0087\,\mathrm{u} = 16.132\,\mathrm{u}$$

where m_H is the mass of a hydrogen atom and m_n is the mass of a neutron.

(b) The mass defect Δm when they come together to form an oxygen atom is

$$\Delta m = m_T - m_{\mathrm{O}} = 16.132\,\mathrm{u} - 15.995\,\mathrm{u} = 0.137\,\mathrm{u} = 2.27 \times 10^{-28}\,\mathrm{kg}$$

where m_{O} is the mass of an oxygen atom.

(c) As a fraction of the mass of an oxygen atom, this is

$$\frac{\Delta m}{m_{\mathrm{O}}} = \frac{0.137\,\mathrm{u}}{15.995\,\mathrm{u}} = 8.57 \times 10^{-3}.$$

A mass spectrometer with a sensitivity of 10^{-5} will indeed be able to detect this change, as $8.57 \times 10^{-3} > 10^{-5}$.

(d) The average binding energy per nucleon is

$$\frac{E}{16} = \frac{\Delta m c^2}{16} = \frac{2.27 \times 10^{-28}\,\mathrm{kg} \cdot (3.00 \times 10^{8}\,\mathrm{m\,s^{-1}})^2}{16} = 1.28 \times 10^{-12}\,\mathrm{J}.$$

47. Let N be the number of particles undergone fission and let C be the number of chain links (the number of 'stages' of the chain reaction). Clearly, since at every stage twice the number of particles undergo fission as at the previous stage, we can say that

$$N = 2^C.$$

Hence,

$$C = \log_2 N$$

and so if we have 1 mole ($N = 6 \times 10^{23}$) of atoms having undergone fission, we must have had

$$C = \log_2(6 \times 10^{23}) \approx 79$$

chain links.

Chapter 10

Mathematics and Computer Science interview questions

This is different to most of the things in here; I wanted to remember the interview questions I got asked at Oxford (because they were interesting questions) so I wrote them down here.

Mathematics and Computer Science interview questions

Damon Falck

June 30, 2018

I thought I'd make a quick note of what I got asked in my interviews at Oxford (applying for Mathematics & Computer Science) in case it's useful to anyone, as I really enjoyed the questions. I had one interview for Mathematics and one for Computer Science at each of Worcester College and New College.

These questions are difficult without any hints and so for effective preparation they should definitely be used in a mock interview format.

(None of the tutors asked me anything about my personal statement.)

1 Problems

1.1 Maths at Worcester

They asked me two printed questions in the interview and guided me through solving them.

1. You are given a function $f : \mathbb{N} \to \mathbb{N}$ defined by the following rules:

$$f(mn) \equiv f(m) + f(n); \tag{1}$$
$$f(n) = 0 \text{ if the last digit of } n \text{ is a } 3; \tag{2}$$
$$f(10) = 1. \tag{3}$$

   (a) Find $f(17)$.

   (b) What are the possible values of $f(500)$?

   (c) In general, for which values of n does $f(n)$ have only one possible value?

   As an aside, does the behaviour of this function under rule (1) remind you of any other functions?

2. For some $a, b, c \in \mathbb{R}$ with $a < b < c$, you are told that

$$a + b + c = 6 \tag{4}$$

   and

$$ab + ac + bc = 9. \tag{5}$$

   Show that $0 < a < 1 < b < 3 < c < 4$.

1.2 Maths at New

This interview consisted of one long informal discussion about differential inequalities. This is my best attempt to replicate what I was given. Each part relies closely on the previous.

As is conventional, $\mathbb{R}_0^+$ denotes the set of nonnegative real numbers.

1. If the derivative of some function $y : \mathbb{R}_0^+ \to \mathbb{R}$ has the property $\frac{\mathrm{d}y}{\mathrm{d}x} \leqslant 0$ for all x, what can we say about the function $y(x)$ graphically? Therefore how can we relate $y(x)$ to $y(0)$?

2. Given that $\frac{\mathrm{d}y}{\mathrm{d}x} \leqslant ky$ for some constant $k \in \mathbb{R}$, make a conclusion about how $y(x)$ compares to $y(0)$. You may wish to consider the function

$$u(x) = \mathrm{e}^{-kx} y(x).$$

3. Given instead that $y\frac{\mathrm{d}y}{\mathrm{d}x} \leqslant ky^2$, can we say something similar about $y(x)$ and $y(0)$? Remember we can't blindly divide both sides of an inequality by a variable.

4. Now what about if $\frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x} \leqslant k$?

5. Functions $y : \mathbb{R}_0^+ \to \mathbb{R}$ and $z : \mathbb{R}_0^+ \to \mathbb{R}$ satisfy the differential equations

$$\frac{\mathrm{d}y}{\mathrm{d}x} = g(y(x)) \tag{6}$$

and

$$\frac{\mathrm{d}z}{\mathrm{d}x} = g(z(x)), \tag{7}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a general function such that

$$\frac{g(a) - g(b)}{a - b} \leqslant k$$

for all values of a and b. The values of $y(0)$ and $z(0)$ are known.

   (a) What can you tell about $y(x) - z(x)$?

   (b) Therefore, given that $y(0) = z(0) = 0$, what can you say about the functions $y(x)$ and $z(x)$?

   (c) Hence deduce the number of functions f that solve the differential equation

$$f'(x) = g(f(x))?.$$

   (d) Find two solutions to the differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}x} = 2\sqrt{y(x)}.$$

   (e) The answers to the last two parts contradict each other! Where have we gone wrong?

1.3 Compsci at Worcester

I was given a computer science example sheet to complete when I arrived on Sunday evening — **see separate document**. Here are some extensions they gave me in the interview to the problems I completed the night before. These are labelled by question number on the example sheet.

1. (b) iii. Maximise the worst-case efficiency of this algorithm. You should be able to make two comparisons per number. Then:

      A. Santa now gives you another four storage locations and we want to find the 7 largest values on the tape (rather than the 3 largest). What algorithm would you use and how many comparisons will it need?

      B. Can we generalise to finding the m highest values from a tape of length n, using a total of $m + 1$ storage locations? For convenience you may assume that $m + 1$ is a power of 2. Find the total number of comparisons necessary for this task.

      C. Now what if we let $m = n$? What task have we accomplished?

   (c) Once you have found the only such number between 10 and 20:

      i. Why is it this value? Try to find a general expression for all numbers that *can* be written as the sum of consecutive numbers.

      ii. Look at the factorisation of your expression. What can you tell about odd and even factors?

      iii. Therefore, which numbers in general cannot be written as the sum of consecutive integers?

1.4 Compsci at New

This interview was much more off-the-cuff than my others. It was split roughly into the following parts:

1. A brief discussion of my programming experience.
2. A discussion of my experience with sorting algorithms. I was asked to give two examples of sorting algorithms and describe them. When I mentioned insertion sort, he went on to derive with me a formula for exactly how many comparisons must be made to sort a list of length n using insertion sort.
3. A short problem to test my physical intuition. The tutor described a solid cube with its vertex pointing downwards being slowly lowered into a liquid. I was asked to draw a series of nine diagrams of what shape the cube would make in the surface of the water (were it removed instantaneously) at each point.
4. The tutor got out some dominoes and began to demonstrate how one might stack them overhanging the edge of the table. I commented that I had seen this problem before (google the block-stacking problem) and so he moved on.
5. A discussion of the basic geometry of hyperdimensional cubes. After the tutor explained what a 4-dimensional, 5-dimensional and then n-dimensional cube meant, he asked me to find the angle between the diagonal and one edge of first a 4-dimensional cube and then an n-dimensional cube.

2 Solutions

2.1 Maths at Worcester

1. (a) Since 17 is prime, we cannot break down $f(17)$ usefully using rule (1). However, $17 \cdot 9 = 153$ ends in a 3, and so

$$\begin{aligned} & f(17 \cdot 9) = 0 \\ \implies & f(17) + f(9) = 0 \\ \implies & f(17) = -f(9). \end{aligned}$$

   But $f(9) = f(3^2) = 2f(3) = 2 \cdot 0 = 0$ so $f(17) = 0$.

   (b) Since $500 = 5 \cdot 10^2$, $f(500) = f(5) + 2f(10) = f(5) + 2$. So we must determine the possible values of $f(5)$.

   We know $f(10) = f(5) + f(2)$ and $f(10) = 1$, so since f is a natural-valued function, either $f(5) = 1$ and $f(2) = 0$ *or* $f(2) = 1$ and $f(5) = 0$.

   Hence $f(500)$ either takes the value 2 or 3.

   (c) We know already that $f(n) = 0$ if n ends in a 3.

   If n ends in a 7 then $f(n) = 0$ by the argument presented in part i.

   If n ends in a 1 then $f(3n) = f(n)+f(3) = f(n)+0$ but $3n$ must end in a 3 so $f(n) = f(3n) = 0$. If n ends in a 9 then similarly $f(n) = f(7n) = 0$ since $7n$ must end in a 3.

   So $f(n) = 0$ if n ends in a 1, 3, 7 or 9.

   If on the other hand n is a multiple of 2 or 5 then in general $f(n)$ has two possible values (as explained in part ii).

   The exception to this rule is if the prime factorisation of n has the same powers of 2 and 5 — that is, $n = 2^k \cdot 5^k \cdot q$ where q is coprime with 2 and 5. In this case, $f(n) = kf(2)+kf(5)+f(q) = kf(10)+f(q) = k+f(q)$ and since $f(q)$ must end in a 1, 3, 7 or 9, $f(n)$ is uniquely determined.

   The behaviour of the function f should remind you of logarithmic functions.

2. Let a, b and c be the roots of a monic cubic equation. This cubic equation is

$$\begin{aligned}(x-a)(x-b)(x-c) &= 0\\ \implies x^3-(a+b+c)x^2+(ab+ac+bc)x-abc &= 0\\ \implies x^3-6x^2+9x-abc &= 0.\end{aligned}$$

Thus the values a, b and c are the x-intercepts of the graph $y=x^3-6x^2+9x-abc$.

In order to sketch this graph, we will find the turning points:

$$\begin{aligned}\frac{\mathrm{d}y}{\mathrm{d}x} &= 3x^2-12x+9 &&= 0\\ \implies 3(x-3)(x-1)=0 &\implies x=1,3.\end{aligned}$$

Since our function has three distinct roots $a<b<c$ and turning points at $x=1$ and $x=3$, we must have $a<1$, $1<b<3$, $c>3$.

Looking at a sketch of the graph reveals that, depending on the value of the y-intercept $-abc$, the third root c will be largest when the first turning point is almost tangential to the x-axis. In this situation roots a and b will converge to the turning point $x=1$ and so

$$\begin{aligned}a+b+c &= 6\\ \implies 1+1+c &= 6\\ \implies c &= 4.\end{aligned}$$

Similarly, the first root a will be smallest when the second turning point is almost tangential to the x-axis; roots b and c will converge to $x=3$ so

$$\begin{aligned}a+3+3 &= 6\\ \implies a &= 0.\end{aligned}$$

Thus by looking at the roots' limiting values, we have shown that

$$0<a<1<b<3<c<4$$

as required.

2.2 Maths at New

1. The function is always either decreasing or constant. Hence $y(x)\leqslant y(0)$ for all x.
2. Differentiating the function u gives

$$\frac{\mathrm{d}u}{\mathrm{d}x}=-k\mathrm{e}^{-kx}y+\mathrm{e}^{-kx}\frac{\mathrm{d}y}{\mathrm{d}x}=\mathrm{e}^{-kx}\left(\frac{\mathrm{d}y}{\mathrm{d}x}-ky\right).$$

Since $\frac{\mathrm{d}y}{\mathrm{d}x}\leqslant ky$, we know $\frac{\mathrm{d}y}{\mathrm{d}x}-ky\leqslant 0$ and so $\frac{\mathrm{d}u}{\mathrm{d}x}\leqslant 0$.

Therefore, $u(x)\leqslant u(0)\implies \mathrm{e}^{-kx}y(x)\leqslant \mathrm{e}^{-k(0)}y(0)\implies y(x)\leqslant \mathrm{e}^{kx}y(0)$.

3. Noticing that $\frac{\mathrm{d}}{\mathrm{d}x}\left[y^2(x)\right]=2y\frac{\mathrm{d}y}{\mathrm{d}x}$, we consider the substitution $w(x)=y^2(x)$. Differentiating,

$$\frac{\mathrm{d}w}{\mathrm{d}x}=2y\frac{\mathrm{d}y}{\mathrm{d}x}$$

and so $\frac{1}{2}\frac{\mathrm{d}w}{\mathrm{d}x}\leqslant kw\implies \frac{\mathrm{d}w}{\mathrm{d}x}\leqslant 2kw$. Hence, as shown before, $w(x)\leqslant \mathrm{e}^{2kx}w(0)$.

It follows that $y^2(x)\leqslant \mathrm{e}^{2kx}y^2(0)$ and so taking square roots, $y(x)\leqslant \mathrm{e}^{kx}y(0)$. (We can take the square root of both sides as $\sqrt{x}$ is an increasing function.)

4. Multiplying both sides by y^2 (which must be positive), $y\frac{\mathrm{d}y}{\mathrm{d}x}\leqslant ky^2$ and so as previously shown, $y(x)\leqslant \mathrm{e}^{kx}y(0)$.

5. (a) Substituting $a = y(x)$ and $b = z(x)$

$$\frac{g(y(x)) - g(z(x))}{y(x) - z(x)} \leqslant k$$
$$\implies \frac{\frac{\mathrm{d}y}{\mathrm{d}x} - \frac{\mathrm{d}z}{\mathrm{d}x}}{y - z} \leqslant k.$$

Hence as shown in the previous part, $y(x) - z(x) \leqslant \mathrm{e}^{kx}\big(y(0) - z(0)\big)$.

(b) Substituting $y(0) = z(0) = 0$,

$$y(x) - z(x) \leqslant 0 \implies y(x) \leqslant z(x)$$

but we could just as easily have substituted $a = z(x)$ and $b = y(x)$ initially, so to preserve symmetry we must have $y(x) = z(x)$.

(c) We have shown that if there are any two functions that satisfy the given differential equation, they must be equal. Hence there can be at most one function that solves this differential equation.

(d) The functions $y = 0$ and $y = x^2$ both satisfy the given differential equation.

(e) In the last example, the general function g is $g(x) = \sqrt{x}$. However, this does not satisfy the condition that

$$\frac{g(a) - g(b)}{a - b} \leqslant k$$

for all a, b; set $b = 0$, then we require $\frac{\sqrt{a}}{a} \leqslant k \iff \frac{1}{\sqrt{a}} \leqslant k$ which is clearly a contradiction as

$$\lim_{a \to 0^+} \left(\frac{1}{\sqrt{a}} \right) = \infty.$$

2.3 Compsci at Worcester

See separate document for my solutions to the example sheet given to me the night before.

Below are answers for the extensions I was given in the interview.

1. (b) i. A. Make initial comparisons to load the first seven slots on the tape to storage locations 1 through 7 in ascending order. For each consequent value on the tape, first load the value to slot 8. Then compare this to slot 4 (the middle value). If it is greater, compare to slot 6, and if it is less then compare to slot 2. Depending on the result of this comparison, compare either to slot 1, 3, 5 or 7. If it is less than slot 1 then proceed to the next value on the tape without making any changes; otherwise, insert the value of slot 8 into the relevant position in slots 1 through 7.

This process will require a total of $3 \cdot 14 \times 10^9 = 42 \cdot 10^9$ comparisons.

B. Assuming $m + 1$ is a power of 2, we will make $\log_2(m + 1)$ comparisons for each value on the storage tape. Therefore the total number of comparisons is $n \log_2(m + 1)$.

C. If we let $m = n$ then we have sorted the entire list of values with $n \log_2(n + 1)$ comparisons. So, we have derived an $O(n \log n)$ sorting algorithm (this is insertion sort).

(c) i. In the attached solutions I find a general formula for this: a number $n \in \mathbb{N}$ can be expressed as the sum of consecutive numbers if and only if

$$n = \frac{a(a + 1)}{2} + ba$$

for some $a, b \in \mathbb{N}$.

ii. This can be rewritten as

$$n = \frac{a}{2}(a + 1 + 2b).$$

If a is even then $a+1+2b$ is odd so n has an odd factor, and if a is odd then n clearly still has an odd factor a. Therefore, all numbers that can be written as the sum of consecutive integers must have an odd factor.

iii. We can conclude that the only numbers which cannot be written as such are those with no odd factors, i.e. powers of 2.

2.4 Compsci at New

This interview was very unstructured so there aren't any 'solutions' as such.

The derivation for insertion sort (involving logs and floor functions) is easily found, and for the last part of the interview, the angle between the diagonal and edge of a 4-dimensional cube is $\arctan\left(\frac{1}{\sqrt{3}}\right)$ and in general the angle between the diagonal and edge of an n-dimensional cube is $\arctan\left(\frac{1}{\sqrt{n-1}}\right)$.

Chapter 11

Proving Pick's theorem

After Miss Brownlee mentioned this very cool theorem at the beginning of a lesson in October 2018, she suggested I try to prove it, and this was what I came up with after a bit of thinking.

Proving Pick's theorem

Damon Falck

June 30, 2018

Definition. A *lattice polygon* is a polygon constructed on a grid of points with integer coordinates such that every vertex of the polygon is on a grid point.

Definition. An *interior lattice point* of a polygon is a point with integer coordinates contained by the polygon.

Definition. A *boundary lattice point* of a polygon is a point with integer coordinates on one of the polygon's sides or vertices.

Definition. A lattice polygon is *peculiar* if, where α is its number of interior lattice points and β is its number of boundary lattice points, its area is given by

$$A = \alpha + \frac{\beta}{2} - 1.$$

Lemma 1. *Consider two lattice polygons P and Q with a common edge, and define PQ as the polygon formed by joining P and Q along their common edge. Then:*

(a) If P and Q are peculiar then PQ is peculiar.

(b) If PQ and P are peculiar then Q is peculiar.

(c) If PQ and Q are peculiar then P is peculiar.

Proof. Let P and Q have α_P and α_Q interior lattice points, and β_P and β_Q boundary lattice points, respectively.

Where γ is the number of lattice points on the common edge between P and Q, by joining the polygons we convert these to interior lattice points — except for the two lattice points at either end of the common edge, which remain boundary lattice points.

So, the number of interior lattice points of PQ is

$$\alpha_{PQ} = \alpha_P + \alpha_Q + (\gamma - 2)$$

and the number of boundary lattice points of PQ is

$$\beta_{PQ} = (\beta_P - \gamma) + (\beta_Q - \gamma) + 2.$$

Hence, if and only if PQ is peculiar, the area of PQ is given by

$$\begin{aligned} A_{PQ} &= \alpha_{PQ} + \frac{\beta_{PQ}}{2} - 1 \\ &= \alpha_P + \alpha_Q + (\gamma - 2) \\ &\quad + \frac{1}{2}\left[(\beta_P - \gamma) + (\beta_Q - \gamma) + 2\right] - 1 \\ &= \alpha_P + \alpha_Q + \gamma - 2 + \frac{\beta_P}{2} \\ &\quad - \frac{\gamma}{2} + \frac{\beta_Q}{2} - \frac{\gamma}{2} + 1 - 1 \\ &= \left(\alpha_P + \frac{\beta_P}{2} - 1\right) + \left(\alpha_Q + \frac{\beta_Q}{2} - 1\right). \end{aligned} \tag{1}$$

It is also clear that the areas of the polygons always add: that is,

$$A_{PQ} = A_P + A_Q. \tag{2}$$

Therefore, we can tackle our three cases:

(a) If P and Q are peculiar then

$$A_P = \alpha_P + \frac{\beta_P}{2} - 1$$

and

$$A_Q = \alpha_Q + \frac{\beta_Q}{2} - 1$$

and so by eq. (2),

$$A_{PQ} = \left(\alpha_P + \frac{\beta_P}{2} - 1\right) + \left(\alpha_Q + \frac{\beta_Q}{2} - 1\right)$$

which is identical to eq. (1); thus, PQ must be peculiar.

(b) If PQ and P are peculiar then

$$A_P = \alpha_P + \frac{\beta_P}{2} - 1$$

and also eq. (1) must hold, which therefore gives

$$A_{PQ} = A_P + \left(\alpha_Q + \frac{\beta_Q}{2} - 1\right),$$

but comparing this to eq. (2) implies

$$A_Q = \alpha_Q + \frac{\beta_Q}{2} - 1;$$

thus, Q must be peculiar.

(c) The situation is symmetrical in P and Q and so by the same argument as in part (b), if PQ and Q are peculiar then P must be peculiar.

□

Lemma 2. *Every lattice rectangle with its sides parallel to the axes is peculiar.*

Proof. Consider a lattice rectangle with height h and width w. If the rectangle's sides are parallel to the axes, then the number of boundary lattice points is equal to its perimeter,

$$\beta = 2w + 2h,$$

and the number of interior lattice points is just

$$\alpha = (w-1)(h-1).$$

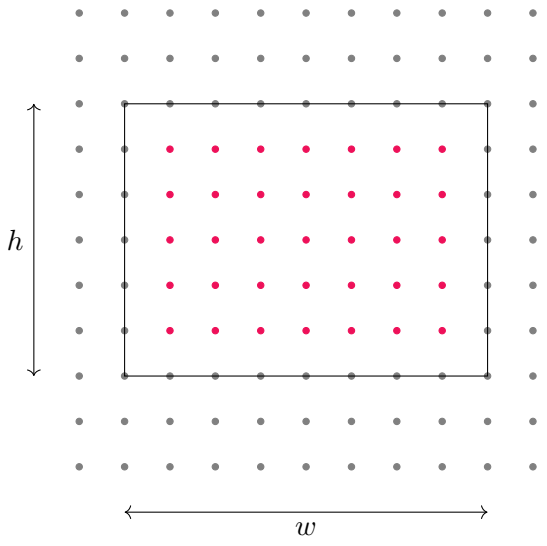


Figure 1: A lattice rectangle with its sides parallel to the axes.

So, if and only if the rectangle is peculiar, its area is

$$\begin{aligned} A &= \alpha + \frac{\beta}{2} - 1 \\ &= (w-1)(h-1) + \frac{2w+2h}{2} - 1 \\ &= wh - w - h + 1 + (w+h) - 1 \\ &= wh \end{aligned}$$

but this is always the area of any rectangle, and so the rectangle must be peculiar. □

Lemma 3. *Every right-angled lattice triangle with its short sides parallel to the axes is peculiar.*

Proof. Consider a right-angled lattice triangle with short sides of length w and h. If the short sides are parallel to the axes, such a triangle can always be formed by cutting the rectangle described in lemma 2 along a diagonal.

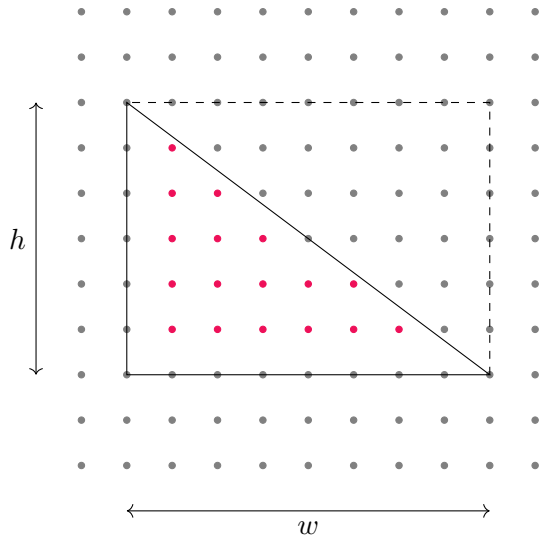


Figure 2: A right-angled lattice triangle formed by cutting a lattice rectangle along its diagonal. In this case $\gamma = 3$.

If the rectangle has α_R interior lattice points and β_R boundary lattice points, then as it must be peculiar its area is

$$wh = \alpha_R + \frac{\beta_R}{2} - 1. \tag{3}$$

Now, suppose the diagonal of the rectangle in question contains γ lattice points. Then $\gamma - 2$ of the rectangle's interior lattice points will be converted to boundary lattice points on the triangle, and exactly half of the remainder will become interior lattice points of the triangle. Therefore, the number of interior lattice points of the triangle is

$$\alpha_T = \frac{\left[\alpha_R - (\gamma - 2)\right]}{2}$$

and the number of boundary lattice points is

$$\beta_T = \frac{\beta_R - 2}{2} + \gamma.$$

Hence, if and only if the new right-angled triangle is peculiar, its area is

$$\begin{aligned} A_T &= \alpha_T + \frac{\beta_T}{2} - 1 \\ &= \frac{\left[\alpha_R - (\gamma - 2)\right]}{2} + \frac{\beta_R - 2}{4} + \frac{\gamma}{2} - 1 \\ &= \frac{\alpha_R}{2} - \frac{\gamma}{2} + 1 + \frac{\beta_R}{4} - \frac{1}{2} + \frac{\gamma}{2} - 1 \\ &= \frac{1}{2}\left(\alpha_R + \frac{\beta_R}{2} - 1\right) \end{aligned}$$

which by eq. (3) is

$$A_T = \frac{wh}{2}.$$

This, however, is clearly the area of any triangle, and so our right-angled triangle must be peculiar. □

Lemma 4. *Every lattice triangle is peculiar.*

Proof. Take any general lattice triangle and enclose it in the smallest possible lattice rectangle with its sides parallel to the axes — there are two cases. This rectangle will consist of either

1. three small right triangles with their sides parallel to the axes, and our triangle under consideration, or
2. the above shapes in addition to one small rectangle with its sides parallel to the axes.

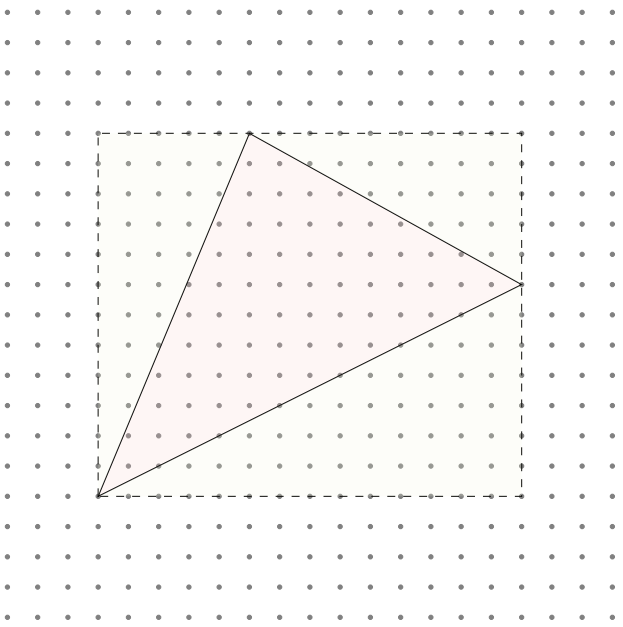


Figure 3: The first case, where the enclosing rectangle is made up of three right triangles and the triangle under consideration.

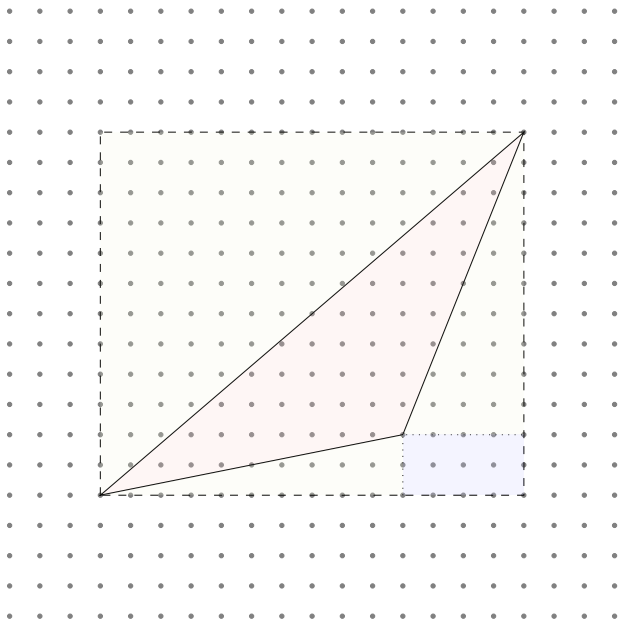


Figure 4: The second case, where the enclosing rectangle is made up of three right triangles, a small rectangle, and the triangle under consideration.

Now, lemma 1 guarantees that the sum or difference of peculiar lattice polygons will also be peculiar. Therefore, as we know that the large enclosing rectangle must be peculiar by lemma 2 and that all of the small right triangles and rectangles must also be peculiar by lemmas 2 and 3, the general lattice triangle we're considering is always the difference of peculiar lattice polygons and so lemma 1 guarantees that it will also be peculiar itself. □

Theorem (Pick's theorem)**.** *Every lattice polygon is peculiar.*

Proof. Every polygon can be decomposed into a number of triangles with their vertices at the vertices of the original polygon, and so any lattice polygon can be decomposed into a number of adjacent lattice triangles.

However, by lemma 4 these lattice triangles are all peculiar, and so lemma 1 guarantees that the polygon formed by joining them all together is also peculiar.

Therefore, every lattice polygon is peculiar. □

Chapter 12

Integrating $\sqrt{\tan x}$

Two friends and I ran into Dr Cheung in a pub in August 2017, just before the start of term, and he gave us this integral to do on the spot. It took each of us about a week to solve the problem and write it up.

$\displaystyle\int \sqrt{\tan x}\,\mathrm{d}x$

Damon Falck

September 2017

Let $I = \displaystyle\int \sqrt{\tan x}\,\mathrm{d}x$. We start by substituting $u^2 = \tan x$ so that, differentiating implicitly,

$$\begin{aligned} 2u &= \sec^2 x \frac{\mathrm{d}x}{\mathrm{d}u} \\ \implies \frac{\mathrm{d}x}{\mathrm{d}u} &= 2u\cos^2 x, \end{aligned}$$

but drawing a right triangle (see fig. 1) reveals $\cos^2 x = \dfrac{1}{1+u^4}$ and so

$$\frac{\mathrm{d}x}{\mathrm{d}u} = \frac{2u}{1+u^4}.$$

Therefore, our integral is equivalent to

$$\begin{aligned} I &= \int u \frac{\mathrm{d}x}{\mathrm{d}u}\,\mathrm{d}u \\ &= \frac{2u^2}{1+u^4}\,\mathrm{d}u. \end{aligned}$$

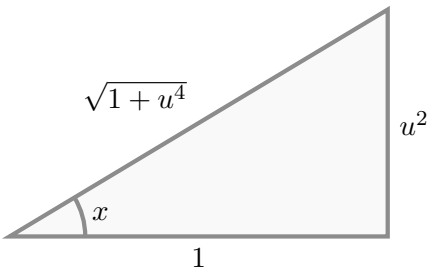


Figure 1: A right triangle showing $u^2 = \tan x$.

Now we want to try to break up this fraction as much as possible. Completing the square,

$$1 + u^4 = (u^2+1)^2 - 2u^2$$

which is just the difference of two squares, so we can factorise further to

$$1 + u^4 = (u^2 + 1 + \sqrt{2}u)(u^2 + 1 - \sqrt{2}u)$$

and hence

$$I = \int \frac{2u^2}{(u^2+\sqrt{2}u+1)(u^2-\sqrt{2}u+1)}\,\mathrm{d}u.$$

Next, using partial fraction decomposition (we require linear numerators as the factors are irreducible quadratics), we write

$$\frac{2u^2}{(u^2+\sqrt{2}u+1)(u^2-\sqrt{2}u+1)} = \frac{A+Bu}{u^2+\sqrt{2}u+1} + \frac{C+Du}{u^2-\sqrt{2}u+1} \tag{1}$$

and now we work through the algebra:

$$\begin{aligned}
\frac{2u^2}{(u^2+\sqrt{2}u+1)(u^2-\sqrt{2}u+1)} &= \frac{(A+Bu)(u^2-\sqrt{2}u+1)+(C+Du)(u^2+\sqrt{2}u+1)}{(u^2+\sqrt{2}u+1)(u^2-\sqrt{2}u+1)} \\
\implies 2u^2 &= (A+Bu)(u^2-\sqrt{2}u+1)+(C+Du)(u^2+\sqrt{2}u+1) \\
&= (B+D)u^3 + (A-\sqrt{2}B+C-\sqrt{2}D)u^2 \\
&\quad + (-\sqrt{2}A+B+\sqrt{2}C+D)u + (A+C).
\end{aligned}$$

Therefore,

$$\begin{aligned}
B+D &= 0, \\
A-\sqrt{2}B+C+\sqrt{2}D &= 2, \\
-\sqrt{2}A+B+\sqrt{2}C+D &= 0, \\
A+C &= 0
\end{aligned}$$

which implies

$$\begin{aligned}
B &= -\frac{2}{\sqrt{2}} = -\sqrt{2}, \\
D &= -B = \sqrt{2}, \\
C &= 0, \\
A &= -C = 0.
\end{aligned}$$

Hence, our fraction from eq. (1) becomes

$$\frac{2u^2}{(u^2+\sqrt{2}u+1)(u^2-\sqrt{2}u+1)} = \frac{-\sqrt{2}u}{u^2+\sqrt{2}u+1} + \frac{\sqrt{2}u}{u^2-\sqrt{2}u+1}$$

and thus

$$I = \sqrt{2}\int \frac{u}{u^2-\sqrt{2}u+1}\,\mathrm{d}u - \sqrt{2}\int \frac{u}{u^2+\sqrt{2}u+1}\,\mathrm{d}u.$$

Completing the square in both integrals, we come to

$$\begin{aligned}
I &= \sqrt{2}\int \frac{u}{\left(u-\frac{\sqrt{2}}{2}\right)^2+\frac{1}{2}}\,\mathrm{d}u - \sqrt{2}\int \frac{u}{\left(u+\frac{\sqrt{2}}{2}\right)^2+\frac{1}{2}}\,\mathrm{d}u \\
&= 2\sqrt{2}\int \frac{u}{\left(\sqrt{2}u-1\right)^2+1}\,\mathrm{d}u - 2\sqrt{2}\int \frac{u}{\left(\sqrt{2}u+1\right)^2+1}\,\mathrm{d}u
\end{aligned}$$

Now substituting $\tan\theta = \sqrt{2}u-1$ and $\tan\phi = \sqrt{2}u+1$,

$$\begin{aligned}
I &= 2\sqrt{2}\int \frac{\frac{\tan\theta+1}{\sqrt{2}}}{\tan^2\theta+1}\frac{\mathrm{d}u}{\mathrm{d}\theta}\,\mathrm{d}\theta - 2\sqrt{2}\int \frac{\frac{\tan\phi-1}{\sqrt{2}}}{\tan^2\phi+1}\frac{\mathrm{d}u}{\mathrm{d}\phi}\,\mathrm{d}\phi \\
&= 2\int \frac{\tan\theta+1}{\sec^2\theta}\frac{\mathrm{d}u}{\mathrm{d}\theta}\,\mathrm{d}\theta - 2\int \frac{\tan\phi-1}{\sec^2\phi}\frac{\mathrm{d}u}{\mathrm{d}\phi}\,\mathrm{d}\phi.
\end{aligned}$$

Differentiating the substitutions gives $\frac{\mathrm{d}u}{\mathrm{d}\theta} = \frac{\sec^2\theta}{\sqrt{2}}$ and $\frac{\mathrm{d}u}{\mathrm{d}\phi} = \frac{\sec^2\phi}{\sqrt{2}}$, and so

$$\begin{aligned} I &= \frac{2}{\sqrt{2}}\int \frac{\tan\theta+1}{\sec^2\theta}\sec^2\theta\,\mathrm{d}\theta - \frac{2}{\sqrt{2}}\int \frac{\tan\phi-1}{\sec^2\phi}\sec^2\phi\,\mathrm{d}\phi \\ &= \frac{2}{\sqrt{2}}\int (\tan\theta+1)\,\mathrm{d}\theta - \frac{2}{\sqrt{2}}\int (\tan\phi-1)\,\mathrm{d}\phi. \end{aligned}$$

Now integrating (finally),

$$\begin{aligned} I &= \sqrt{2}\left(-\ln|\cos\theta| + \theta + \ln|\cos\phi| + \phi\right) + c \\ &= \sqrt{2}\ln\left|\frac{\cos\phi}{\cos\theta}\right| + \sqrt{2}(\theta+\phi) + c. \end{aligned}$$

All there is remaining to do is back-substitute. With the help of another right triangle we see that

$$\begin{aligned} \cos\theta &= \cos\left[\arctan(\sqrt{2}u-1)\right] \\ &= \frac{1}{\sqrt{1+(\sqrt{2}u-1)^2}} \\ &= \frac{1}{\sqrt{2}\sqrt{u^2-\sqrt{2}u+1}} \end{aligned}$$

and by analogy,

$$\cos\phi = \frac{1}{\sqrt{2}\sqrt{u^2+\sqrt{2}u+1}}$$

which leads to

$$\begin{aligned} I &= \sqrt{2}\ln\left|\frac{\frac{1}{\sqrt{2}\sqrt{u^2+\sqrt{2}u+1}}}{\frac{1}{\sqrt{2}\sqrt{u^2-\sqrt{2}u+1}}}\right| + \sqrt{2}\left[\arctan(\sqrt{2}u-1)+\arctan(\sqrt{2}u+1)\right] + c \\ &= \sqrt{2}\ln\left|\frac{\sqrt{u^2-\sqrt{2}u+1}}{\sqrt{u^2+\sqrt{2}u+1}}\right| + \sqrt{2}\left[\arctan(\sqrt{2}u-1)+\arctan(\sqrt{2}u+1)\right] + c \\ &= \frac{\sqrt{2}}{2}\ln\left|\frac{u^2-\sqrt{2}u+1}{u^2+\sqrt{2}u+1}\right| + \sqrt{2}\left[\arctan(\sqrt{2}u-1)+\arctan(\sqrt{2}u+1)\right] + c. \end{aligned}$$

Finally, we can use the arctangent additive identity (which stems directly from the tangent additive identity) $\arctan a + \arctan b = \arctan\left(\frac{a+b}{1-ab}\right)$ to simplify this further to

$$\begin{aligned} I &= \frac{\sqrt{2}}{2}\ln\left|\frac{u^2-\sqrt{2}u+1}{u^2+\sqrt{2}u+1}\right| + \sqrt{2}\arctan\left[\frac{\sqrt{2}u-1+\sqrt{2}u+1}{1-(\sqrt{2}u-1)(\sqrt{2}u+1)}\right] + c \\ &= \frac{\sqrt{2}}{2}\ln\left|\frac{u^2-\sqrt{2}u+1}{u^2+\sqrt{2}u+1}\right| + \sqrt{2}\arctan\left(\frac{\sqrt{2}u}{1-u^2}\right) + c. \end{aligned}$$

At last we can substitute x back in, leading to our final answer:

$$\int \sqrt{\tan x}\,\mathrm{d}x = \frac{\sqrt{2}}{2}\ln\left|\frac{\tan x - \sqrt{2\tan x}+1}{\tan x + \sqrt{2\tan x}+1}\right| + \sqrt{2}\arctan\left(\frac{\sqrt{2\tan x}}{1-\tan x}\right) + c.$$

Chapter 13

An infinite series for π

I honestly cannot remember why I did this... I think I was thinking about talks for π day (March 2017) and came across this way to find an expression for π after researching the Basel problem? Either way, it is a very nice quick bit of maths I think.

We start with $\displaystyle\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x$. Substituting $x = \tan\theta$, so $\dfrac{\mathrm{d}x}{\mathrm{d}\theta} = \sec^2\theta$,

$$\begin{aligned}\int \frac{1}{1+x^2}\,\mathrm{d}x &= \int \frac{1}{1+\sin^2\theta}\,\frac{\mathrm{d}x}{\mathrm{d}\theta}\,\mathrm{d}\theta \\ &= \int \frac{1}{\sec^2\theta}\sec^2\theta\,\mathrm{d}\theta \\ &= \int 1\,\mathrm{d}\theta = \theta + c = \arctan x + c.\end{aligned}$$

$$\text{So, } \int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x = \Big[\arctan x\Big]_0^1 = \arctan 1 = \frac{\pi}{4}.$$

However, $\dfrac{1}{1+x^2} = \dfrac{1}{1-(-x^2)}$ is also the sum to infinity of the geometric series with common ratio $-x^2$. So,

$$\begin{aligned}\frac{1}{1+x^2} &= 1 - x^2 + x^4 - x^6 \pm \cdots \\ \implies \int \frac{1}{1+x^2}\,\mathrm{d}x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} \pm \cdots + c.\end{aligned}$$

Hence, $\displaystyle\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x = \Big[x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} \pm \cdots\Big]_0^1 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} \pm \cdots$. Thus,

$$\begin{aligned}\frac{\pi}{4} &= 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} \pm \cdots \\ \implies \pi &= 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} \pm \cdots.\end{aligned}$$

Chapter 14

Does upturning a cathode-ray television affect the picture?

At the end of the first or second lesson on electric fields in September 2017, Dr Cheung asked this question, so I tried to answer it here. I loved this because it was short and sweet but a satisfying application of a principle we'd just learned.

Does upturning a cathode-ray television affect the picture?

Damon Falck

June 30, 2018

We'll say our vertical deflection plates have length ℓ and generate an electric field of magnitude E, and that each electron is travelling horizontally at speed v before entering this field.

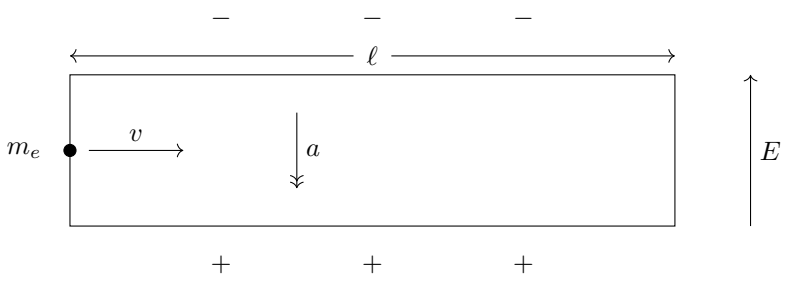


Figure 1: A simplified cathode-ray tube

As there are no horizontal forces, the electron will leave this field at time $t = \frac{\ell}{v}$. If the electron is initially accelerated through a voltage V, then the total kinetic energy acquired is

$$\frac{1}{2}m_e v^2 = eV$$
$$\implies v^2 = 2\frac{eV}{m_e}.$$

and so, if during its passage between the deflection plates the electron has vertical acceleration a, the final deflection is therefore

$$h = \frac{1}{2}at^2 = \frac{a\ell^2 m_e}{4eV}.$$

The Coulomb force on the electron is $EQ = Ee$ and so the electron's vertical acceleration due to the field is $\frac{Ee}{m_e}$. Therefore with gravity acting in the same direction as the field, the total acceleration is $a = \frac{Ee}{m_e} + g$, and with gravity acting against the field, the acceleration is $a = \frac{Ee}{m_e} - g$.

The fractional difference in deflection when reversing the direction of gravity is therefore

$$\frac{\Delta h}{h} = \frac{\left(\frac{Ee}{m_e} + g\right)\frac{\ell^2 m_e}{4eV} - \left(\frac{Ee}{m_e} - g\right)\frac{\ell^2 m_e}{4eV}}{\left(\frac{Ee}{m_e} + g\right)\frac{\ell^2 m_e}{4eV}}$$
$$= \frac{\frac{g\ell^2 m_e}{2eV}}{\frac{E\ell^2}{4V} + \frac{g\ell^2 m_e}{4eV}} = \frac{2gm_e}{eE + gm_e}.$$

So, with a typical field strength of $E = 1.00\,\mathrm{kN\,C^{-1}}$,

$$\frac{\Delta h}{h} = 1.34 \times 10^{-15},$$

a tiny fraction; we can conclude that turning the device upside down does not meaningfully affect the picture. Of course, there may be other effects such as the Earth's magnetic field that we haven't taken into account.

Part II

Interesting notes from lessons

Chapter 15

The three laws of motion (a parody)

Quite some months after the start of Year 12 I did this for fun (probably out of sleep deprivation).

The Three Laws of Motion

Damon Falck

June 30, 2018

The following three laws form the basis for all classical mechanics, and were revolutionary at their time of introduction. It is important to fully understand their consequences in order to facilitate the further study of kinetics.

I

If there is no resultant force acting on a body, then it will eventually come to rest.

II

The resultant force acting on an object of constant mass is proportional to its rate of change of displacement.

$$\sum \vec{F} = m \frac{\mathrm{d}\,\vec{s}}{\mathrm{d}t}$$
$$\sum \vec{F} = m\,\vec{v}$$

In fact, the first law is just a special case of this law; it was, nevertheless, required as a presupposition upon the introduction of these axioms.

III

If body A exerts a force on body B, then body B must exert a force back on body A such that the ratio of the magnitudes of these forces equals the ratio of the bodies' masses.

$$\vec{F}_{A\to B} = -\frac{m_A}{m_B}\vec{F}_{B\to A}$$

For instance, when a car and a truck collide head-on, the truck will exert a much greater force on the car than the car will exert on the truck; therefore the car will undergo much greater damage.

Chapter 16

Some introductory physics from Michaelmas 2016

In December 2017 I tried to write up nice notes from all of the subjects we'd studied that term in physics and mathematics. I only got this far but I want to include these notes because I did put quite a lot of effort into presenting them, despite their simple subject content.

Notes in Physics

Damon Falck

Michaelmas 2016

Contents

1 Kinematics

1.1 Newton's Laws

I

If there is no resultant force acting on a body, then its velocity is constant.

II

The resultant force acting on an object is proportional to its rate of change of momentum.

$$\sum \vec{F} = \frac{\mathrm{d}\,\vec{p}}{\mathrm{d}t} \tag{1}$$

$$\sum \vec{F} = m\,\vec{a} \tag{2}$$

III

If body A exerts a force on body B, then body B must exert a force of equal magnitude and opposite direction on body A of the same type and along the same line of action.

$$\vec{F}_{A\to B} = -\vec{F}_{B\to A} \tag{3}$$

1.1.1 An Aside on Horses and Carts

In Newtonian pairs of forces, the two forces always act on different objects - otherwise no motion would ever be able to occur.

A notorious question is the following:

> A horse is harnessed to a cart. If the horse tries to pull the cart, the horse must exert a force on the cart. By Newton's third law the cart must then exert an equal and opposite force on the horse. Newton's second law tells us that acceleration is equal to the net force divided by the mass of the system. ($F = ma$, so $a = \frac{F}{m}$.) Since the two forces are equal and opposite, they must add to zero, so Newton's second law tells us that the acceleration of the system must be zero. If it doesn't accelerate, and it started it rest, it must remain at rest (by the definition of acceleration), and therefore no matter how hard the horse pulls, it can never move the cart.

There are several things wrong with this statement.

- First, the forces do not sum to zero. The force from horse on the cart is matched by an equal and opposite force from the cart on the horse, but as these are on different objects they do not cancel and so motion is not prevented.
- Secondly, 'the system' is not defined. The cart and the horse should each be considered separately. One will find then that there is a resultant force forwards on both bodies.

The reason that the horse and cart can move is that the force of the horse backwards on the ground causes an equal and opposite force of friction forwards from the ground on the horse. The system will move forwards if the frictional force forwards on the horse is greater than the pull of the wagon backwards.

1.2 Kinematics Definitions

Displacement, a vector quantity denoted $\vec{s}$ or $\vec{r}$, is distance travelled in a particular direction. It is also defined as the shortest distance between an object's initial position and final position.

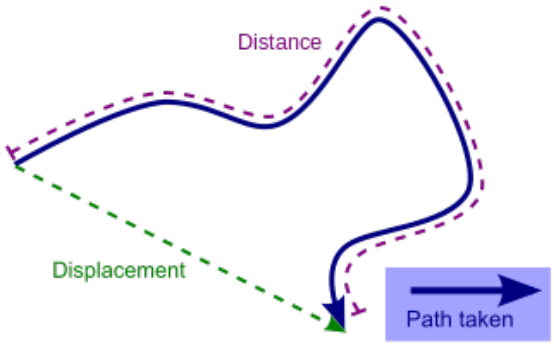


Figure 1: The difference between distance (d) and displacement ($\vec{s}$).

Velocity, a vector quantity denoted $\vec{v}$, is defined as the ratio of change in displacement to time taken, so that

$$\vec{v} = \frac{\Delta \vec{s}}{t}. \tag{4}$$

The speed of an object (usually also denoted by the letter v, is the magnitude of its velocity, $||\vec{v}||$.

N.B. If an object returns to it's starting point,

$$\vec{s} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \mathrm{m} \tag{5}$$

$$\vec{v}_{\mathrm{avg}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \mathrm{m\,s^{-1}}. \tag{6}$$

Acceleration, a vector quantity denoted $\vec{a}$, is the ratio of change in velocity to time taken, so that

$$\vec{a} = \frac{\Delta \vec{v}}{\Delta t} = \frac{\Delta \vec{s}}{\Delta t^2}. \tag{7}$$

It should be noted that velocity is the first derivative of displacement with respect to time and acceleration the second, such that

$$\vec{v} = \dot{\vec{s}} \quad \text{and} \quad \vec{a} = \ddot{\vec{s}}. \tag{8}$$

1.3 One-dimensional Motion: Derivation of the Constant Acceleration Formulae

With the condition of constant acceleration, we can derive five formulae relating s, u, v, a and t where u is the initial velocity and v is the final velocity of the object. Note that we are only considering one dimension, so we can drop vector notation.

We know that

$$v = \frac{\mathrm{d}s}{\mathrm{d}t} \tag{9}$$

and also that

$$a = \frac{\mathrm{d}v}{\mathrm{d}t}\,. \tag{10}$$

Therefore, by the fundamental theorem of calculus

$$v = \int a \,\mathrm{d}t \tag{11}$$

and so where u is the constant of integration, that is the velocity at $t = 0$, then

$$v = u + at. \tag{12}$$

We can also see that

$$s = \int v \,\mathrm{d}t \tag{13}$$

and so where the constant of integration is 0, that is there is no displacement at $t = 0$, then

$$s = \int (u + at)\,\mathrm{d}t = ut + \frac{1}{2}at^2. \tag{14}$$

Integrating u in the same manner gives

$$s = \int u \, \mathrm{d}t = \int (v - at) \, \mathrm{d}t = vt - \frac{1}{2}at^2. \tag{15}$$

Manipulating (14) slightly gives us

$$s = \frac{1}{2}t(2u + at) = \frac{1}{2}t[u + (u + at)] \tag{16}$$

and so substituting in (12) results in

$$s = \frac{1}{2}(v + u)t. \tag{17}$$

Finally, rearranging (12) gives us

$$t = \frac{v - u}{a} \tag{18}$$

and substituting this into (17) gives

$$\begin{aligned} s &= \frac{1}{2}(v + u)\frac{v - u}{a} \\ s &= \frac{v^2 - u^2}{2a} \\ v^2 &= u^2 + 2as. \end{aligned} \tag{19}$$

Thus our five equations governing one-dimensional motion with constant acceleration are:

$$\begin{aligned} v &= u + at \\ s &= ut + \frac{1}{2}at^2 \\ s &= vt - \frac{1}{2}at^2 \\ v^2 &= u^2 + 2as \\ s &= \frac{1}{2}(v + u)t \end{aligned}$$

It should be noted that the last of these equations is the same as saying that the displacement is the the average velocity times the time taken, where because acceleration is constant the average velocity is $\frac{v+u}{2}$.

1.4 Projectiles

It's important to note that experimental work shows that x-axis and y-axis motion is independent — that is, motion in any axes at an angle of $\frac{\pi}{2}$ to each other can be considered separately.

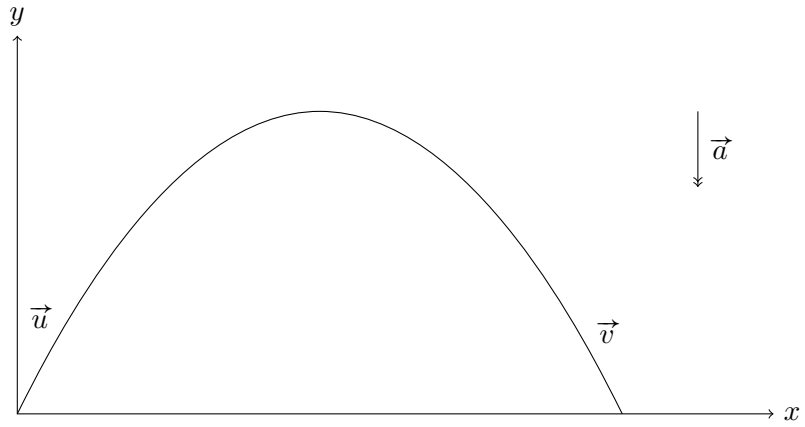


Figure 2: The motion of a projectile in the xy-plane with initial velocity $\vec{u}$ and final velocity $\vec{v}$.

We know from the constant acceleration formulae that

$$\vec{v} = \vec{u} + \vec{a}t.$$

This vector summation can be represented like so:

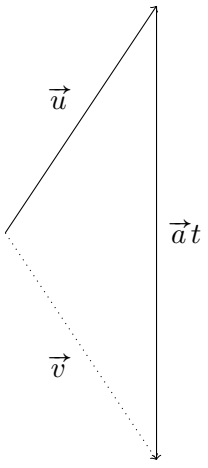


Figure 3: A triangle showing that $\vec{v}$ is the vector sum of $\vec{u}$ and $\vec{a}t$.

If the launch angle is θ, then we can resolve $\vec{u}$ to find the component initial velocities:

$$u_x = ||\vec{u}|| \cos\theta \tag{20}$$
$$u_y = ||\vec{u}|| \sin\theta \tag{21}$$

or, as would more frequently be written,

$$u_x = u\cos\theta \tag{22}$$
$$u_y = u\sin\theta. \tag{23}$$

Now we know that the only force acting on the projectile is its weight, and so in the vertical direction its acceleration will be g downwards. (In the horizontal direction there is no acceleration as there are no forces.) So, applying the constant acceleration formula $s = ut + \frac{1}{2}at^2$, we can see that

$$x = u\cos\theta t \tag{24}$$
$$y = u\sin\theta t - \frac{1}{2}gt^2. \tag{25}$$

where x is its horizontal displacement and y is its vertical displacement.

1.4.1 Deriving the Path Equation

Given the above result, we can easily find an equation for the altitude of the projectile y in terms of its horizontal displacement x. In other words, this is the locus of points in the xy-plane that the projectile passes through.

We know from (24) that

$$t = \frac{x}{u\cos\theta} \tag{26}$$

and so substituting this expression for t into (25) gives us

$$y = u\sin\theta \cdot \frac{x}{u\cos\theta} - \frac{1}{2}g \cdot \frac{x^2}{u^2\cos^2\theta}$$
$$y = \tan\theta x - \frac{gx^2}{2u^2\cos^2\theta}. \tag{27}$$

This is a quadratic in x, a parabola as we would expect.

1.4.2 Finding the Launch Angle for Maximum Range

Let us first find an expression for the range R of the projectile, its total horizontal displacement.

Applying $s = ut + \frac{1}{2}at^2$ vertically upwards where $s = 0$ (the projectile has landed so its vertical displacement is zero) and where T is the time of flight,

$$\begin{aligned} 0 &= u\sin\theta T - \frac{1}{2}gT^2 \\ 0 &= T(u\sin\theta - \frac{1}{2}gT) \end{aligned} \tag{28}$$

and assuming that $T \neq 0$,

$$\begin{aligned} \frac{1}{2}gT &= u\sin\theta \\ T &= \frac{2u\sin\theta}{g}. \end{aligned} \tag{29}$$

Now that we know the projectile's time of flight, we can find its range by applying $s = ut + \frac{1}{2}at^2$ horizontally, which gives us

$$\begin{aligned} R &= u\cos\theta T + \frac{1}{2}aT^2 \\ R &= u\cos\theta \cdot \frac{2u\sin\theta}{g} + 0 \\ R &= \frac{u^2 \cdot 2\sin\theta\cos\theta}{g} \end{aligned} \tag{30}$$

which using the inverse double angle identity we can simplify to

$$R = \frac{u^2\sin 2\theta}{g}. \tag{31}$$

Since u and g are constant, we can maximise R by maximising $\sin 2\theta$. The largest value the sine function can take is 1, so assuming θ is acute,

$$\begin{aligned} \sin 2\theta &= 1 \\ 2\theta &= \frac{\pi}{2} \\ \theta &= \frac{\pi}{4}. \end{aligned} \tag{32}$$

Thus, as we would expect, the launch angle resulting in the longest range of the projectile is 45°.

Therefore, the maximum range of a projectile is

$$R_{\max} = \frac{u^2}{g}. \tag{33}$$

1.5 Relative Velocities

The 'resultant' velocity of an object is the vector sum of its observed velocity and the velocity of the frame of observation.

For example, if an airplane is perceived to be moving at $\vec{v}_1$ from a helicopter moving at $\vec{v}_2$ relative to the ground, then the plane's velocity relative to the ground $\vec{v}_{\text{resultant}}$ will be

$$\vec{v}_{\text{resultant}} = \vec{v}_1 + \vec{v}_2. \tag{34}$$

N.B. It is always a good idea to draw a vector summation triangle in these circumstances. One should be particularly careful with sign conventions when dealing with resolved velocities.

A typical question involving relative velocities will ask what a boat's velocity relative to the riverbank is, given its speed and direction relative to the water and given the speed of the current. One should simply resolve and find the relative velocities both parallel and perpendicular to the banks, and then combine them to find the magnitude and angle of the resultant velocity.

1.6 Pulley Systems

Consider an infinite Atwood machine. A string passes over each pulley, with one end attached to a mass and the other end attached to another pulley. All the masses are equal to m, and all the pulleys and strings are massless. The masses are held fixed and then simultaneously released. What is the acceleration of the top mass?

For the sake of completeness, let us first consider a case with three pulleys. We will then show that the infinite case is, in fact, simpler to solve (at least algebraically).

1.6.1 Three Pulleys

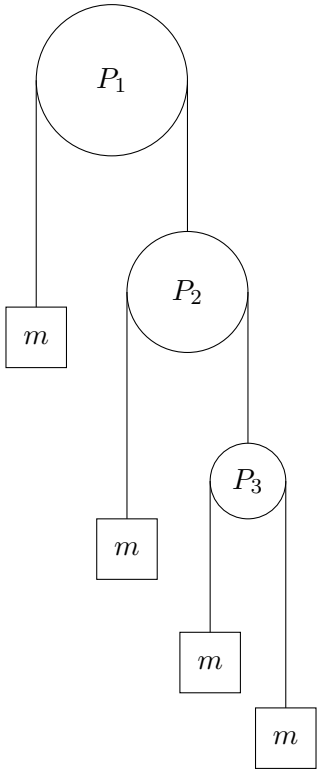


Figure 4: A system of four blocks of mass m and three light smooth pulleys, connected with light inextensible strings.

We will start by giving names to the various tensions and accelerations present. Let a_1, a_2, a_3, a_4 be the accelerations of the first through fourth masses from the top respectively, with upwards positive, and let T_1, T_2, T_3 be the tensions in the strings over pulleys P_1, P_2 and P_3 respectively.

Applying Newton II on the first mass vertically upwards,

$$T_1 - mg = ma_1. \tag{35}$$

Applying the law similarly to the other three masses, we know that

$$T_2 - mg = ma_2 \tag{36}$$
$$T_3 - mg = ma_3 \tag{37}$$
$$T_3 - mg = ma_4. \tag{38}$$

Because the pulleys are massless, we also know that

$$T_1 = 2T_2 \tag{39}$$
$$T_2 = 2T_3. \tag{40}$$

Additionally, we know that because P_2 and the first mass are connected, they must have the same magnitude of acceleration, and so

$$a_1 = -a_{P_2}. \tag{41}$$

The acceleration of P_2 is equal to the average acceleration of the second mass and P_3, so

$$a_1 = -\frac{a_2 + a_{P_3}}{2}. \tag{42}$$

Finally, the acceleration of P_3 is likewise equal to the average acceleration of the last two masses, so

$$a_1 = -\frac{a_2 + \frac{a_3+a_4}{2}}{2}$$
$$a_1 = -\frac{2a_2 + a_3 + a_4}{4}. \tag{43}$$

Substituting (39) and (40) into (35), (36), (37) and (38) gives us

$$a_1 = \frac{T_1 - mg}{m} \tag{44}$$
$$a_2 = \frac{\frac{T1}{2} - mg}{m} \tag{45}$$
$$a_3 = \frac{\frac{T1}{4} - mg}{m} \tag{46}$$
$$a_4 = \frac{\frac{T1}{4} - mg}{m}. \tag{47}$$

Now substituting these expressions for a_2, a_3 and a_4 into (43) gives us

$$a_1 = -\frac{2 \cdot \frac{\frac{T1}{2}-mg}{m} + \frac{\frac{T1}{4}-mg}{m} + \frac{\frac{T1}{4}-mg}{m}}{4}$$
$$a_1 = -\frac{2 \cdot \frac{T_1-2mg}{2m} + 2 \cdot \frac{T_1-4mg}{4m}}{4}$$
$$a_1 = -\frac{2(T_1 - 2mg) + (T_1 - 4mg)}{4 \cdot 2m}$$
$$a_1 = -\frac{3T_1 - 8mg}{8m} \tag{48}$$

Rearranging this for T_1 gives

$$T_1 = \frac{8mg - 8ma_1}{3} \tag{49}$$

and rearranging (44) similarly gives

$$T_1 = ma_1 + mg. \tag{50}$$

Setting these two expressions equal gives us

$$\begin{aligned} \frac{8mg - 8ma_1}{3} &= ma_1 + mg \\ 8mg - 8ma_1 &= 3ma_1 + 3mg \\ a_1 &= \frac{5}{11}g \end{aligned} \tag{51}$$

which is what we wanted to find.

1.6.2 Infinite Pulleys

Now we'll consider the same problem but with an infinite number of pulleys (and all of the masses still with equal mass m.

Let the tension in the string on the first pulley be T. Then the tension in the string on the second pulley is $\frac{T}{2}$ (because the pulley is massless). Let the downwards acceleration of the second pulley be a_2. Then the second pulley effectively lives in a world where gravity has strength $g - a_2$.

We know that if if we were to multiply the strength of gravity by some factor η, the tension in all of the strings in the Atwood machine would also be multiplied by η. This is true because the only way to produce a quantity with the units of tension (force) is to multiply a mass by g.

In other words, the ratio of every tension to the strength of gravity is going to be the same no matter what the strength of gravity is.

Now consider the subsystem of all the pulleys except the top one. This infinite subsystem is identical to the original infinite system of all the pulleys, except the tension of the top string is $\frac{T}{2}$ and the strength of gravity is $g - a_2$. Hence,

$$\frac{T}{g} = \frac{\frac{T}{2}}{g - a_2}. \tag{52}$$

and so

$$\begin{aligned} T(g - a_2) &= T \cdot \frac{g}{2} \\ a_2 &= \frac{g}{2}. \end{aligned} \tag{53}$$

2 Momentum and Energy

2.1 Momentum and Impulse

The definition of momentum, a vector quantity denoted by $\vec{p}$, is

$$\vec{p} = m\vec{v}. \tag{54}$$

The units of momentum are $\mathrm{kg\,m\,s^{-1}}$ or $\mathrm{N\,s}$. The second comes from the fact that impulse, J, is defined as the integral of a force over the time for which it acts. So, for a force acting between times t_1 and t_2

$$\vec{J} = \int_{t_1}^{t_2} \vec{F}\,\mathrm{d}t \tag{55}$$

and so by Newton II,

$$\begin{aligned}
\vec{J} &= \int_{t_1}^{t_2} \frac{\mathrm{d}\vec{p}}{\mathrm{d}t}\,\mathrm{d}t \\
\vec{J} &= \int_{\vec{p}_1}^{\vec{p}_2} \mathrm{d}\vec{p} \\
\vec{J} &= \vec{p}_2 - \vec{p}_1 \\
\vec{J} &= \Delta\vec{p}
\end{aligned} \tag{56}$$

or with constant mass and force,

$$\vec{F}\cdot\Delta t = m\cdot\Delta\vec{v}. \tag{57}$$

2.2 Conservation of Momentum

In a **closed system** with no external forces acting, momentum is conserved, i.e.

$$\sum \vec{p}_{\text{initial}} = \sum \vec{p}_{\text{final}}. \tag{58}$$

2.2.1 Proof of Conservation of Momentum with Two Particles

We will first consider the collision of two particles.

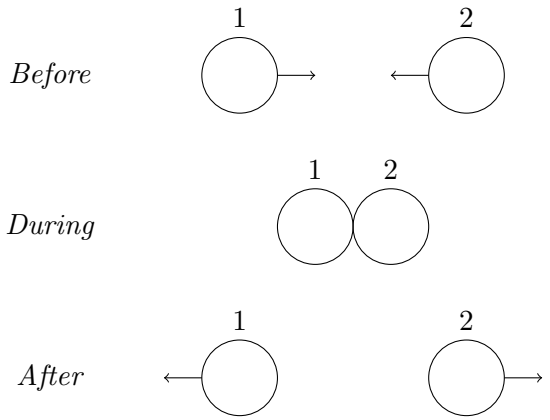


Figure 5: Both particles before, during and after collision.

When these two particles collide, there is a force acting on each. Let $\vec{F}_1$ be the force acting on particle 1, and $\vec{F}_2$ be the force acting on particle 2.

Newton II tells us that

$$\vec{F}_1 = \frac{\mathrm{d}\,\vec{p}_1}{\mathrm{d}t} \tag{59}$$

and

$$\vec{F}_2 = \frac{\mathrm{d}\,\vec{p}_2}{\mathrm{d}t} \tag{60}$$

where $\vec{p}_1$ and $\vec{p}_2$ are the momenta of particles 1 and 2 respectively. We also know from Newton III that

$$\vec{F}_1 = -\vec{F}_2. \tag{61}$$

Therefore,

$$\begin{aligned} \frac{\mathrm{d}\,\vec{p}_1}{\mathrm{d}t} &= -\frac{\mathrm{d}\,\vec{p}_2}{\mathrm{d}t} \\ \frac{\mathrm{d}\,\vec{p}_1}{\mathrm{d}t} + \frac{\mathrm{d}\,\vec{p}_2}{\mathrm{d}t} &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\vec{p}_1 + \vec{p}_2\right) &= 0 \\ \frac{\mathrm{d}\,\vec{p}_{\text{total}}}{\mathrm{d}t} &= 0 \end{aligned} \tag{62}$$

and so for two particles colliding in a closed system, $\vec{p}_{\text{total}}$ is constant — i.e. total momentum is conserved. □

2.2.2 Proof of Conservation of Momentum with Multiple Particles

We can now perform a similar proof for a situation in which there are n particles colliding in some unknown way.

Let us pick any single particle i. Newton II tells us that the change of momentum of i is equal to the sum of all the forces acting on i from any other particles that happen to collide with it. So,

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \vec{F}_{j\to i} = \frac{\mathrm{d}\,\vec{p}_i}{\mathrm{d}t} \tag{63}$$

Note that if $j = i$ then the force will be zero as a particle cannot exert a force on itself, and so we can dispose of the condition $j \neq i$. For some j, if j does not collide with i then $\vec{F}_{j\to i}$ will be zero and so will not contribute to i's change of momentum.

Therefore, now summing through i,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i} = \sum_{i=1}^{n} \frac{\mathrm{d}\,\vec{p}_i}{\mathrm{d}t} \tag{64}$$

and so

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum \vec{p} = \sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i}. \tag{65}$$

Newton III tells us that

$$\begin{aligned} \vec{F}_{j\to i} &= -\vec{F}_{i\to j} \\ \vec{F}_{j\to i} + \vec{F}_{i\to j} &= 0 \end{aligned} \tag{66}$$

for any particular i and j. Therefore if you write out the double sum above, the forces will always come in pairs and they will cancel out. We can show this by using Newton III to write that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i} = -\sum_{j=1}^{n}\sum_{i=1}^{n} \vec{F}_{i\to j} \tag{67}$$

and so by switching the dummy variable names on the right and swapping the order of summation,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i} = -\sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i} \tag{68}$$

which implies that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \vec{F}_{j\to i} = 0 \tag{69}$$

and therefore eq. (65) gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum \vec{p} = 0. \tag{70}$$

That is, total momentum is conserved. □

It's a good idea to try this with four or five particles and write out the double summation to build intuition as to what it actually represents and how the terms cancel.

Chapter 17

Notes on the kinetic theory of gases

When we started learning about ideal gases with Dr Cheung in December 2017, he went through several proofs of the ideal gas law with us and talked about what intuition we could gain from it. Here is a summary of the arguments from one particular lesson.

Notes on the Kinetic Theory of Gases

Damon Falck

Lent 2017

Contents

Derivations of the Ideal Gas Law

In a sphere with one particle

Consider a spherical container of radius R. Suppose there is a particle of mass m and speed v undergoing elastic collisions with the frictionless walls.

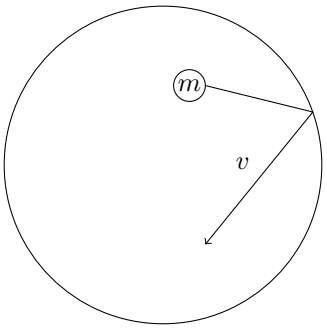


Figure 1: The particle is colliding continuously with the walls of the sphere.

Because the collisions are frictionless, there are no tangential forces acting on the particle.

The collisions are elastic, so kinetic energy is conserved — and thus because the mass is constant, the speed is conserved.

Hence, as tangential velocity doesn't change and radial velocity keeps its magnitude, it can be seen that the angle of approach is equal to the angle of departure.

Let θ be this angle, to the normal.

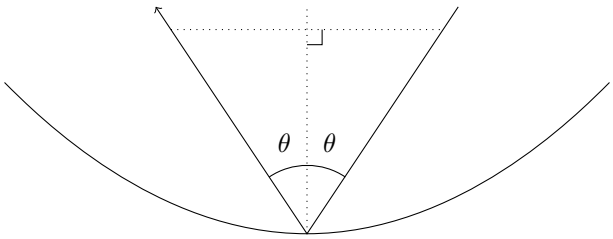


Figure 2: The particle is colliding continuously with the walls of the sphere.

Let us now find the average acceleration between the midpoints of two of the particle's paths, before and after a collision:

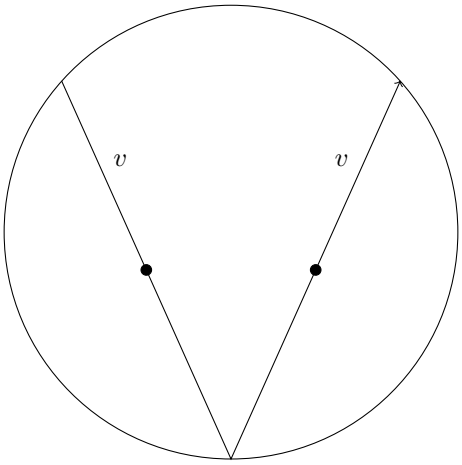


Figure 3: We want the acceleration between the two points marked.

For this, we need the change in velocity and the time passed.

Resolving tangentially,

$$\Delta v_t = 0$$

as there are no tangential forces. Resolving radially,

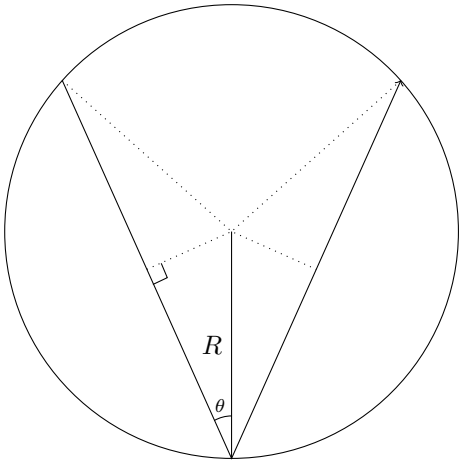
$$\begin{aligned}\Delta v_r &= v_{r,f} - v_{r,i}\\ \Delta v_r &= v\cos\theta - (-v\cos\theta)\\ \Delta v_r &= 2v\cos\theta.\end{aligned}$$

So,

$$\begin{aligned}\Delta v &= \sqrt{(\Delta v_t)^2 + (\Delta v_r)^2}\\ \Delta v &= \sqrt{0^2 + (2v\cos\theta)^2}\\ \Delta v &= 2v\cos\theta.\end{aligned}$$

Now we must find the time passed, for which we need the distance travelled by the particle (since we know its speed v).

We can draw construction lines to create four congruent right triangles:



It can be seen that between each collision with the wall, the displacement of the particle is $2R\cos\theta$, and it's clear due to congruence that this is equal to the distance the particle travels between two midpoints. Hence the particle's distance travelled s is

$$s = 2R\cos\theta$$

and so the time taken is

$$t = \frac{2R\cos\theta}{v}.$$

Thus, the average acceleration between the two midpoints is

$$\begin{aligned}a &= \frac{\Delta v}{t}\\ a &= \frac{2v\cos\theta}{\left(\frac{2R\cos\theta}{v}\right)}\\ a &= \frac{v^2}{R}\end{aligned}$$

radially inwards (as this is the direction of the change in velocity).

Corollary: *As θ tends to $\frac{\pi}{2}$, so this becomes the instantaneous acceleration, as the change in velocity becomes continuous. Thus, a particle travelling along the inside surface of a sphere experiences an acceleration of constant magnitude $\frac{v^2}{R}$ radially inwards.*

Applying Newton II, we get that the average force acting on the particle is

$$F = \frac{mv^2}{R} \tag{1}$$

and so by Newton III there is an equal force exerted radially outwards on the sphere by the particle.

Thus as the surface area of the sphere is $4\pi R^2$, the average pressure radially outwards on the sphere is

$$P = \frac{mv^2}{4\pi R^3}. \tag{2}$$

We know that the relation between thermodynamic temperature and average kinetic energy is

$$\frac{1}{2}\langle mv^2\rangle = \frac{3}{2}k_B T$$

where k_B is Boltzmann's constant and T is thermodynamic temperature. We are considering only one particle, so

$$\frac{1}{2}mv^2 = \frac{3}{2}k_B T. \tag{3}$$

We can substitute eq. (3) into eq. (2) to get

$$P = \frac{3k_B T}{4\pi R^3}$$
$$P = \frac{k_B T}{\frac{4}{3}\pi R^3}$$

and so where V is the volume of the sphere,

$$P = \frac{k_B T}{V}$$
$$PV = k_B T. \quad \square$$

Extension to N non-interacting particles using Dalton's law

Therefore, as pressure is cumulative by Dalton's law of partial pressures, if we have N particles then

$$P = \frac{k_B T}{V} \cdot N$$
$$PV = N k_B T. \tag{4}$$

(This is already a common form of the ideal gas law.)

Hence, if n is the total number of moles then

$$n = \frac{N}{N_A} \tag{5}$$

where N_A is Avogadro's number.

Thus, substituting eq. (5) into eq. (4) gives

$$PV = nN_A k_B T. \tag{6}$$

The universal gas constant R is defined as

$$R \coloneqq N_A k_B \tag{7}$$

and so substituting eq. (7) into eq. (6) gives

$$PV = nRT.$$

Thus is the most common form of the ideal gas law, as required. □

(Proven only for non-interacting particles in a sphere.)

In a sphere with N non-interacting particles

In the same sphere let us re-derive the law with N non-interacting particles present. Let F be the total force acting outwards on the wall of the sphere, such that

$$F = \sum_{i=1}^{N} F_i$$

where F_i is the force exerted by each particle i in colliding with the wall. So, from eq. (1) we have that

$$F = \sum_{i=1}^{N} \frac{mv_i^2}{R}$$

where v_i is the speed of each particle, and so as the surface area of the sphere is $4\pi R^2$,

$$P = \sum_{i=1}^{N} \frac{mv_i^2}{4\pi R^3}$$

$$P = \frac{m}{4\pi R^3} \sum_{i=1}^{N} v_i^2.$$

As $\langle v^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} v_i^2$ by definition of the mean, we can rewrite this equation as

$$P = \frac{m}{4\pi R^3} N \langle v^2 \rangle$$

$$P = 2\frac{N}{4\pi R^3} \cdot \frac{1}{2} m v^2.$$

Thus because $\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_BT$,

$$\begin{aligned} P &= 2\frac{N}{4\pi R^3}\cdot\frac{3}{2}k_BT \\ P &= \frac{3Nk_BT}{4\pi R^3} \\ PV &= Nk_BT. \end{aligned}$$

The molar form of the law can then be derived as in the preceding section. □

In a cube with N non-interacting particles

Let us now consider a similar scenario but within a cube of side length ℓ rather than a sphere. First will consider one particle, mass m, in the x-dimension only.

Between midpoints,

$$\begin{aligned} \Delta v_x &= v_{xf} - v_{xi} \\ &= v_x - (-v_x) \\ &= 2v_x \end{aligned}$$

away from the wall, as the collision is elastic and there are no forces acting parallel to the wall it collides with; so the velocity is reversed.

The distance between midpoints can be seen to be ℓ. Therefore, the time taken between midpoints is

$$\Delta t = \frac{\ell}{v_x}.$$

Thus the acceleration between midpoints is

$$a_x = \frac{\Delta v_x}{\Delta t} = \frac{2v_x}{\frac{\ell}{v_x}} = \frac{2v_x^2}{\ell}.$$

Hence the force acting on the one face in question by this particle is

$$F = \frac{2mv_x^2}{\ell}$$

and so the pressure on this one face is

$$\begin{aligned} PA &= \frac{2mv_x^2}{\ell} \\ P &= \frac{2mv_x^2}{V}. \end{aligned}$$

Now we consider the full N particles. It's important to note that (in the x-direction) only one half of the particles are going to be travelling towards and consequently colliding with the face in question. Thus, the number of particles colliding with this face is actually $\frac{N}{2}$:

$$P = \frac{Nm\langle v_x^2\rangle}{V}.$$

Now, because the kinetic motion of the particles is entirely random, it stands to reason that the average velocity in each direction should be equal. Thus,

$$\langle v_x^2\rangle = \langle v_y^2\rangle = \langle v_z^2\rangle$$

so because Pythagoras tells us that

$$\langle v^2\rangle = \langle v_x^2\rangle + \langle v_y^2\rangle + \langle v_z^2\rangle,$$

it follows that

$$\frac{1}{3}\langle v^2\rangle = \langle v_x^2\rangle = \langle v_y^2\rangle = \langle v_z^2\rangle.$$

So, we can write our previous equation in terms of the 3-dimensional speeds:

$$P = \frac{m\frac{1}{3}\langle v^2\rangle N}{V}.$$

Therefore because $E_k = \frac{1}{2}mv^2 = \frac{3}{2}k_BT$, we now know that

$$P = \frac{\frac{1}{3}N\cdot 2\cdot \frac{3}{2}k_BT}{V}$$
$$\therefore \quad PV = Nk_BT.$$

By symmetry this is true for all face of the cube. The molar form of the law derived earlier follows. □

N.B. The next task is to show what happens when the particles do indeed interact — that is, they collide elastically. Before, we had assumed they pass through one another without collision.

An interesting corollary

Suppose we wish to come to an approximate estimate of the average velocity of gaseous particles in a room at room temperature and atmospheric pressure.

Using the equation

$$\frac{3}{2}k_BT = \frac{1}{2}m\langle v^2\rangle,$$

we find that our root-mean-square velocity[1] $v_{\text{RMS}} = \sqrt{\langle v^2\rangle}$ is

$$v_{\text{RMS}} = \sqrt{\frac{3k_BT}{m}}.$$

The Boltzmann constant k_B is approximately 1.38×10^{-23} and the thermodynamic temerature at RTP is roughly $273.15 + 20 = 293$ K. Finally, let us for this purpose assume that air consists entirely of nitrogen. The mass of one N_2 molecule is $(14\cdot 2)/N_A = 4.65\times10^{-23}$ g. Therefore,

$$v_{\text{RMS}} \approx \sqrt{\frac{3\cdot 1.38\times10^{-23}\cdot 293}{4.65\times10^{-26}}} = 511\,\text{m}\,\text{s}^{-1}.$$

[1]Note that this estimate will always be on the large side due to the RMS-AM-GM-HM inequality.

This is approximately the speed of sound (which in dry air at room temperature is 331 $\mathrm{m\,s^{-1}}$) — the speed of the molecules v_{RMS} will always be higher than the speed of sound, since the sound propagates through the gas by disturbing the motion of the molecules. The disturbance is passed on from molecule to molecule by means of collisions; a sound wave can therefore never travel faster than the average speed of the molecules.

Chapter 18

Some calculus for radioactive decay

In Lent 2017 I missed a lesson where Dr Cheung talked about the problem of decay chains, and this was my attempt to understand the content of the lesson in catching up.

Some calculus for radioactive decay

Damon Falck

June 30, 2018

Introduction

A chain of three isotopes

Suppose we have three isotopes A, B and C. Isotope A decays with decay constant λ_1 to isotope B, and isotope B decays with constant λ_2 to isotope C, which is stable.

The number of nuclei of A is given by the differential equation

$$\frac{\mathrm{d}N_A}{\mathrm{d}t} = -N_A\lambda_1 \tag{1}$$

as shown earlier. Therefore, because isotope B is decaying at rate $N\lambda_2$ but is also increasing at the rate of decay of A, we also know that

$$\frac{\mathrm{d}N_B}{\mathrm{d}t} = N_A\lambda_1 - N_B\lambda_2. \tag{2}$$

Finally, it is clear that the change in the number of nuclei of C over time is

$$\frac{\mathrm{d}N_C}{\mathrm{d}t} = N_B\lambda_2.$$

We assume that at $t = 0$ we have $N_A = N_0$ and $N_B, N_C = 0$. Solving eq. (1) therefore yields

$$N_A = N_0\mathrm{e}^{-\lambda_1 t}. \tag{3}$$

(The methods of arriving at this solution are described earlier.) Now, we come to the more arduous task of solving eq. (2).

Using a little insight, we multiply both sides by $\mathrm{e}^{\lambda_2 t}$, coming to

$$\mathrm{e}^{\lambda_2 t}\frac{\mathrm{d}N_B}{\mathrm{d}t} = \mathrm{e}^{\lambda_2 t}(\lambda_1 N_A - \lambda_2 N_B)$$

and substituting in the solution for N_A found in eq. (3), we get

$$\begin{aligned}\mathrm{e}^{\lambda_2 t}\,\frac{\mathrm{d}N_B}{\mathrm{d}t} &= \mathrm{e}^{\lambda_2 t}(\lambda_1 N_0\mathrm{e}^{-\lambda_1 t} - \lambda_2 N_B)\\ &= \lambda_1 N_0\mathrm{e}^{\lambda_2 t-\lambda_1 t} - \lambda_2 N_B\mathrm{e}^{\lambda_2 t}\end{aligned}$$

and so,

$$\mathrm{e}^{\lambda_2 t}\frac{\mathrm{d}N_B}{\mathrm{d}t} + \lambda_2 N_B\mathrm{e}^{\lambda_2 t} = \lambda_1 N_0\mathrm{e}^{(\lambda_2-\lambda_1)t}.$$

The left hand side of this equation reminds us of the product rule, and we see that we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(N_B \mathrm{e}^{\lambda_2 t}\right) = \lambda_1 N_0 \mathrm{e}^{(\lambda_2-\lambda_1)t}$$

and therefore

$$N_B \mathrm{e}^{\lambda_2 t} = \int \lambda_1 N_0 \mathrm{e}^{(\lambda_2-\lambda_1)t}\,\mathrm{d}t.$$

Factoring out constants, we can evaluate this integral:

$$\begin{aligned} N_B \mathrm{e}^{\lambda_2 t} &= \lambda_1 N_0 \int \mathrm{e}^{(\lambda_2-\lambda_1)t}\,\mathrm{d}t \\ &= \lambda_1 N_0 \cdot \frac{\mathrm{e}^{(\lambda_2-\lambda_1)t}}{\lambda_2-\lambda_1} + c. \end{aligned}$$

Hence, dividing by $\mathrm{e}^{\lambda_2 t}$,

$$\begin{aligned} N_B &= \frac{\lambda_1 N_0 \mathrm{e}^{(\lambda_2-\lambda_1)t}}{(\lambda_2-\lambda_1)\mathrm{e}^{\lambda_2 t}} + c \\ &= \frac{\lambda_1}{\lambda_2-\lambda_1} N_0 \mathrm{e}^{-\lambda_1 t} + c\mathrm{e}^{-\lambda_2 t}. \end{aligned}$$

To find this constant of integration c, we can check the initial state. We know $N_B = 0$ at $t = 0$, and so

$$\begin{aligned} 0 &= \frac{\lambda_1}{\lambda_2-\lambda_1} N_0 + c \\ \implies c &= -\frac{\lambda_1}{\lambda_2-\lambda_1} N_0. \end{aligned}$$

Hence, we now know that

$$N_B = \frac{\lambda_1}{\lambda_2-\lambda_1} N_0 \mathrm{e}^{-\lambda_1 t} - \frac{\lambda_1}{\lambda_2-\lambda_1} N_0 \mathrm{e}^{-\lambda_2 t}.$$

Thus, factorising, we come to our final solution to eq. (2):

$$N_B = \frac{\lambda_1}{\lambda_2-\lambda_1} N_0 \left[\mathrm{e}^{-\lambda_1 t} - \mathrm{e}^{-\lambda_2 t}\right].$$

Chapter 19

Some diffraction notes

After internal exams at the end of Year 12, Dr Cheung did some lessons with us on unrelated but interesting mathematics and physics, and after frantic note-taking in a *lovely* lesson on diffraction gratings and infinite series I wrote these notes.

Some diffraction notes

Damon Falck

June 30, 2018

1 A model for a diffraction grating

Consider an array of n narrow slits, equally spaced with separation d, out of which are emitted monochromatic coherent light waves, all of angular frequency ω and amplitude A_0.

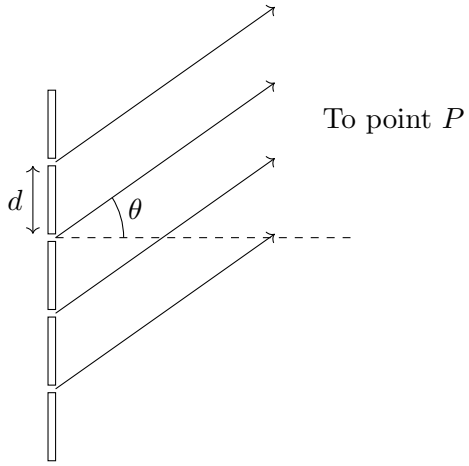


Figure 1: A simple diffraction grating

Now, if we look at these slits at an angle θ, from a point P sufficiently far away that we can consider the paths from each slit to point P to be parallel, then at this point every wave will be out of phase from the next by the same amount, a phase difference we'll call ϕ.

Assuming the light entering the grating is from a single source so that every wave initially has zero phase difference with the next, this phase difference between each will merely be the product of the path difference $d\sin\theta$ and the wavenumber $k = \frac{2\pi}{\lambda}$, so that

$$\phi = \frac{2\pi d \sin\theta}{\lambda}. \tag{1}$$

2 Resultant amplitude as a function of observation angle

Now, we will try to find the resultant amplitude caused by the interference of these waves as a function of the angle θ. Expressing the wave from the first slit as $A_0\sin(\omega t)$, the next will be shifted out of phase by ϕ and so will have form $A_0\sin(\omega t+\phi)$, the next $A_0\sin(\omega t+2\phi)$, and so

on up until the wave from the last slit, which will be given by $A_0 \sin(\omega t + (n-1)\phi)$. So, using the principle of superposition, at this point the sum of these n waves will be

$$\sigma = A_0 \sum_{r=0}^{n-1} \sin(\omega t + r\phi). \tag{2}$$

Now we know from Euler's formula that this is the same as writing

$$\sigma = A_0 \sum_{r=0}^{n-1} \Im\left[\mathrm{e}^{\mathrm{i}(\omega t + r\phi)}\right] = A_0 \cdot \Im\left[\mathrm{e}^{\mathrm{i}\omega t} \sum_{r=0}^{n-1} \mathrm{e}^{\mathrm{i}r\phi}\right] \tag{3}$$

but this is just a geometric series, which we can evaluate as

$$\sigma = A_0 \cdot \Im\left[\mathrm{e}^{\mathrm{i}\omega t} \frac{1-\mathrm{e}^{\mathrm{i}\phi n}}{1-\mathrm{e}^{\mathrm{i}\phi}}\right] = A_0 \sin(\omega t) \cdot \Im\left[\frac{1-\mathrm{e}^{\mathrm{i}\phi n}}{1-\mathrm{e}^{\mathrm{i}\phi}}\right] \tag{4}$$

To create symmetry in the denominator, we bring out a factor of $\mathrm{e}^{\mathrm{i}\frac{\phi}{2}}$, giving us

$$\sigma = A_0 \sin(\omega t) \cdot \Im\left[\frac{\mathrm{e}^{-\mathrm{i}\frac{\phi}{2}} - \mathrm{e}^{\mathrm{i}\phi\left(n-\frac{1}{2}\right)}}{\mathrm{e}^{-\mathrm{i}\frac{\phi}{2}} - \mathrm{e}^{\mathrm{i}\frac{\phi}{2}}}\right]. \tag{5}$$

The denominator of this fraction is clearly $\cos\left(-\frac{\phi}{2}\right) + \mathrm{i}\sin\left(-\frac{\phi}{2}\right) - \cos\left(\frac{\phi}{2}\right) - \mathrm{i}\sin\left(\frac{\phi}{2}\right) = -2\mathrm{i}\sin\left(\frac{\phi}{2}\right)$. Similarly, the numerator must be

$$\cos\left(-\frac{\phi}{2}\right)+\mathrm{i}\sin\left(-\frac{\phi}{2}\right)-\cos(n\phi)\cos\left(\frac{\phi}{2}\right)+\mathrm{i}\cos(n\phi)\sin\left(\frac{\phi}{2}\right)-\mathrm{i}\sin(n\phi)\cos\left(\frac{\phi}{2}\right)-\sin(n\phi)\sin\left(\frac{\phi}{2}\right) \tag{6}$$

and so taking the imaginary part,

$$\sigma = A_0 \sin(\omega t) \cdot \Im\left[\frac{\cos\left(\frac{\phi}{2}\right) - \cos(n\phi)\cos\left(\frac{\phi}{2}\right) - \sin(n\phi)\sin\left(\frac{\phi}{2}\right)}{-2\mathrm{i}\sin\left(\frac{\phi}{2}\right)}\right] \tag{7}$$

$$= A_0 \sin(\omega t) \cdot \frac{\cos\left(\frac{\phi}{2}\right) - \cos(n\phi)\cos\left(\frac{\phi}{2}\right) - \sin(n\phi)\sin\left(\frac{\phi}{2}\right)}{2\sin\left(\frac{\phi}{2}\right)} \tag{8}$$

$$= A_0 \sin(\omega t) \cdot \frac{\cos\left(\frac{\phi}{2}\right) - \cos(n\phi)\cos\left(\frac{\phi}{2}\right) - \sin(n\phi)\sin\left(\frac{\phi}{2}\right)}{2\sin\left(\frac{\phi}{2}\right)} \tag{9}$$

which we can use a couple more compound identities to simplify:

$$\sigma = A_0 \sin(\omega t) \cdot \frac{\cos\left(\frac{\phi}{2}\right)\left(1 - \cos^2\left(\frac{n\phi}{2}\right) + \sin^2\left(\frac{n\phi}{2}\right)\right) - 2\sin\left(\frac{n\phi}{2}\right)\cos\left(\frac{n\phi}{2}\right)\sin\left(\frac{\phi}{2}\right)}{2\sin\left(\frac{\phi}{2}\right)} \tag{10}$$

$$= A_0 \sin(\omega t) \cdot \frac{2\cos\left(\frac{\phi}{2}\right)\sin^2\left(\frac{n\phi}{2}\right) - 2\sin\left(\frac{n\phi}{2}\right)\cos\left(\frac{n\phi}{2}\right)\sin\left(\frac{\phi}{2}\right)}{2\sin\left(\frac{\phi}{2}\right)} \tag{11}$$

$$= A_0 \sin(\omega t) \cdot \frac{\sin\left(\frac{n\phi}{2}\right)}{\sin\left(\frac{\phi}{2}\right)}\left[\cos\left(\frac{\phi}{2}\right)\sin\left(\frac{n\phi}{2}\right) - \cos\left(\frac{n\phi}{2}\right)\sin\left(\frac{\phi}{2}\right)\right] \tag{12}$$

$$= A_0 \frac{\sin\left(\frac{n\phi}{2}\right)}{\sin\left(\frac{\phi}{2}\right)} \sin\left(\frac{n\phi}{2} - \frac{\phi}{2}\right)\sin(\omega t). \tag{13}$$

This is the full form of the resultant wave, and so we can say that at a maximum such that $\sin(\omega t) = 1$, the amplitude (the 'scaling') of the wave $\sin\left(\frac{\phi(n-1)}{2}\right)$ is given by

$$A = A_0 \frac{\sin\left(\frac{n\phi}{2}\right)}{\sin\left(\frac{\phi}{2}\right)} \tag{14}$$

and so it is very easy to calculate the intensity of light here just by squaring, getting

$$I = I_0 \frac{\sin^2\left(\frac{n\phi}{2}\right)}{\sin^2\left(\frac{\phi}{2}\right)} \tag{15}$$

where I_0 is the intensity of the wave from each slit. Finally, we can substitute our expression for ϕ in terms of the viewing angle θ, and so

$$I = I_0 \frac{\sin^2\left(\frac{n\pi d \sin\theta}{\lambda}\right)}{\sin^2\left(\frac{\pi d \sin\theta}{\lambda}\right)}. \tag{16}$$

3 Position of the maxima

Neglecting the denominator (this is an approximation), the overall maxima will occur whenever $\frac{n\pi d \sin\theta}{\lambda}$ is a multiple of π; that is, we must have $\frac{d\sin\theta}{\lambda} = m$ for some natural m. In fact, for every new integer m a new maximum occurs, and so we can say

$$m\lambda = d\sin\theta \tag{17}$$

at every mth order maximum.

Chapter 20

Some notes from Dr Cheung's lessons on electric fields and potential

Some of the most mathematically intense physics lessons at Highgate were Dr Cheung's lessons on electric fields and the shell theorem near the start of Year 12. I thought that three or four lessons in particular formed a nice series of notes so I typed them up here.

Some notes from Dr Cheung's lessons on electric fields and potential

Damon Falck

October 4, 2017

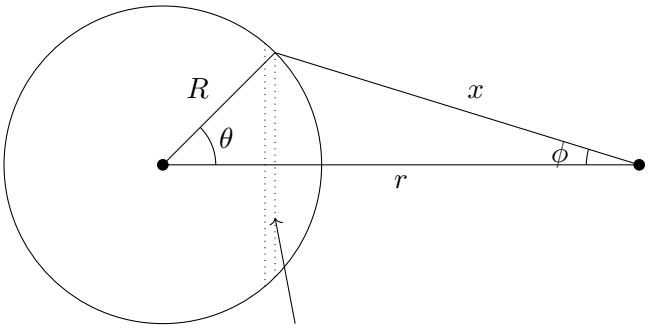
Contents

Miscellaneous wisdom from the lessons is in italics.

1 Several solutions to the shell theorem integral

1.1 Setting up the integral

We're trying to find the electric field near a spherical shell of radius R and surface charge density σ. We'll consider a point of distance r from the centre of the sphere. Look at this diagram:



Ring of charge shown dotted

All tangential forces cancel so the field strength is just the radial component, for which we just multiply by $\cos\phi$. The contribution of the dotted ring to this field is therefore

$$\mathrm{d}E = \frac{\mathrm{d}Q}{4\pi\varepsilon_0 x^2}\cos\phi \tag{1}$$

where $\mathrm{d}Q$ is the charge of the ring. Clearly the ring has radius $R\sin\theta$, hence circumference $2\pi R\sin\theta$, and so its area is $\mathrm{d}A = 2\pi R\sin\theta\cdot R\,\mathrm{d}\theta$. By definition of charge density σ,

$$\sigma = \frac{\mathrm{d}Q}{\mathrm{d}A}$$

and so

$$\mathrm{d}Q = \sigma\,\mathrm{d}A = 2\pi\sigma R^2\sin\theta\,\mathrm{d}\theta.$$

Substituting this into eq. (1) and integrating both sides yields

$$\begin{aligned} E &= \int_0^\pi \frac{2\pi\sigma R^2\sin\theta\cos\phi\,\mathrm{d}\theta}{4\pi\varepsilon_0 x^2} \\ &= \frac{\sigma R^2}{2\varepsilon_0}\int_0^\pi \frac{\sin\theta\cos\phi}{x^2}\,\mathrm{d}\theta. \end{aligned} \tag{2}$$

This gives us the total field strength at the point we're interested in, and this is what we want to integrate.

1.2 Integrating with respect to θ

We have a choice of integration variable here. This one is the long, hardcore, most obvious way to do it. It's never a good idea, however, to try to solve a problem without knowing the answer at the start.

The cosine rule tells us

$$x^2 = r^2 + R^2 - 2rR\cos\theta$$

and also

$$\begin{aligned} \cos\phi &= \frac{x^2 + r^2 - R^2}{2xr} \\ &= \frac{r^2 + R^2 - 2rR\cos\theta + r^2 - R^2}{2xr} \\ &= \frac{r - R\cos\theta}{x} \end{aligned}$$

and so our shell theorem integral eq. (2) becomes

$$E = \frac{\sigma R^2}{2\varepsilon_0}\int_0^\pi \frac{\sin\theta(r - R\cos\theta)}{(r^2 + R^2 - 2rR\cos\theta)^{3/2}}\,\mathrm{d}\theta.$$

Now we have everything in terms of θ, we'll make a substitution of $u = \cos\theta$ (and so $\mathrm{d}u = -\sin\theta\,\mathrm{d}\theta$):

$$E = \frac{\sigma R^2}{2\varepsilon_0}\int_{-1}^1 \frac{r - Ru}{(r^2 + R^2 - 2rRu)^{3/2}}\,\mathrm{d}u. \tag{3}$$

At this point, we just need to plough through and integrate. Separate it into two integrals,

$$\int \frac{r}{(r^2 + R^2 - 2rRu)^{3/2}}\,\mathrm{d}u$$

and

$$\int \frac{Ru}{(r^2 + R^2 - 2rRu)^{3/2}}\,\mathrm{d}u.$$

The first is just a matter of chain rule/substitution, and the second can be done by parts. Combine these and we find that outside the sphere, the field is

$$E = \frac{\sigma R^2}{\varepsilon_0 r^2}. \tag{4}$$

Using the total charge $Q = 4\pi R^2\sigma$ instead,

$$E = \frac{Q}{4\pi\varepsilon_0 r^2}$$

as would be a point charge at the shell's centre; this is the shell theorem.

1.3 Integrating with respect to x

A slightly more sneaky way to do it.

Looking again at eq. (2), this time we'll get everything in terms of x. (Integrating with respect to ϕ seems pointless as the algebra would be very similar to when using θ.)

As before, the cosine rule gives

$$x^2 = r^2 + R^2 - 2rR\cos\theta$$

but this time let's differentiate with respect to θ, yielding

$$2x\frac{\mathrm{d}x}{\mathrm{d}\theta} = 2rR\sin\theta,$$

or,

$$\frac{x}{rR}\,\mathrm{d}x = \sin\theta\,\mathrm{d}\theta.$$

Substituting this into the integral in eq. (2) immediately simplifies it to

$$E = \frac{\sigma R}{2\varepsilon_0 r}\int_{r-R}^{r+R}\frac{\cos\phi}{x}\,\mathrm{d}x$$

and the cosine rule also tells us

$$\cos\phi = \frac{x^2 + r^2 - R^2}{2xr}$$

which means

$$E = \frac{\sigma R}{4\varepsilon_0 r^2}\int_{r-R}^{r+R}\frac{x^2 + r^2 - R^2}{x^2}\,\mathrm{d}x.$$

This is easily integrable:

$$\begin{aligned}
E &= \frac{\sigma R}{4\varepsilon_0 r^2}\int_{r-R}^{r+R}\left(1 + \frac{r^2 - R^2}{x^2}\right)\mathrm{d}x \\
&= \frac{\sigma R}{4\varepsilon_0 r^2}\left[x - \frac{r^2 - R^2}{x}\right]_{r-R}^{r+R} \\
&= \frac{\sigma R}{4\varepsilon_0 r^2}\left(r + R - r + R - \frac{r^2 - R^2}{r+R} + \frac{r^2 - R^2}{r - R}\right) \\
&= \frac{\sigma R}{4\varepsilon_0 r^2}\left(2R - (r - R) + (r + R)\right) \\
&= \frac{\sigma R}{4\varepsilon_0 r^2}(4R) \\
&= \frac{\sigma R^2}{\varepsilon_0 r^2}
\end{aligned}$$

which is the correct result as in eq. (4).

1.4 Chain rule trickery

Return to our expression in eq. (3):

$$E = \frac{\sigma R^2}{2\varepsilon_0}\int_{-1}^{1}\frac{r - Ru}{(r^2 + R^2 - 2rRu)^{3/2}}\,\mathrm{d}u.$$

Last time we split it up and used parts — however, this fraction strikingly resembles an application of the chain rule, and we in fact notice that

$$\frac{\partial}{\partial r}\left[-\frac{1}{\sqrt{r^2 + R^2 - 2rRu}}\right] = \frac{r - Ru}{(r^2 + R^2 - 2rRu)^{3/2}}.$$

This is a rather peculiar fact, but it means we can write

$$E = \frac{\sigma R^2}{2\varepsilon_0}\int_{-1}^{1}\frac{\partial}{\partial r}\left[-\frac{1}{\sqrt{r^2 + R^2 - 2rRu}}\right]\mathrm{d}u.$$

However, differentiation and integration are commutative with one another (I'll leave this unproven for now) and so

$$\begin{aligned}
E &= -\frac{\sigma R^2}{2\varepsilon_0}\frac{\partial}{\partial r}\int_{-1}^{1}\frac{1}{\sqrt{r^2 + R^2 - 2rRu}}\,\mathrm{d}u \\
&= -\frac{\sigma R^2}{2\varepsilon_0}\frac{\partial}{\partial r}\left[\frac{1}{rR}\sqrt{r^2 + R^2 - 2rRu}\right]_{-1}^{1} \\
&= -\frac{\sigma R}{2\varepsilon_0}\frac{\partial}{\partial r}\frac{1}{r}\left(\sqrt{r^2 + R^2 - 2rR} - \sqrt{r^2 + R^2 + 2rR}\right) \\
&= -\frac{\sigma R}{2\varepsilon_0}\frac{\partial}{\partial r}\frac{1}{r}\left(|r - R| - (r + R)\right).
\end{aligned}$$

Therefore, given $r > R$,

$$\begin{aligned}
E &= -\frac{\sigma R}{2\varepsilon_0}\frac{\partial}{\partial r}\left(\frac{-2R}{r}\right) \\
&= \frac{\sigma R^2}{\varepsilon_0 r^2} \\
&= \frac{Q}{4\pi\varepsilon_0 r^2}
\end{aligned}$$

which is just what we were hoping for.

By the way, this can actually be done in one line. Such a solution is left as an exercise to the reader.

2 Introducing potential

2.1 Force as a gradient

First, consider a simple example of an object with kinetic energy and gravitational potential energy, moving in one dimension. So, the total energy is

$$E = \frac{1}{2}m\dot{x}^2 + mgx.$$

Differentiating both sides gives

$$\frac{\partial E}{\partial t} = m\dot{x}\ddot{x} + mg\dot{x}$$

but by energy conservation the total energy is constant, so $\dfrac{\partial E}{\partial t} = 0$:

$$\begin{aligned} m\dot{x}\ddot{x} + mg\dot{x} &= 0 \\ \implies \ddot{x} &= -g; \end{aligned}$$

this is just a statement of Newton II applied on the object: we've proven the law in this case using only the assumption of conservation of energy conservation.

To find out more about how one might prove something such as energy conservation, read 'Emmy Noether's Wonderful Theorem' by Neuenschwander.

Now let's generalise to an object at position $\mathbf{r}$ in n-dimensional space and total potential energy $U(\mathbf{r})$. Then, the total energy of the object is

$$E = \frac{1}{2}m(\dot{\mathbf{r}} \cdot \dot{\mathbf{r}}) + U(\mathbf{r}).$$

Differentiating as before,

$$\begin{aligned} \frac{\partial E}{\partial t} = m(\dot{\mathbf{r}} \cdot \ddot{\mathbf{r}}) + \frac{\partial U}{\partial \mathbf{r}} \cdot \dot{\mathbf{r}} &= 0 \\ \implies m\ddot{\mathbf{r}} &= -\frac{\partial U}{\partial \mathbf{r}}. \end{aligned}$$

But, by Newton II the left hand side is just the force $\mathbf{F}$ acting on the object, and so we come to a new definition of force as

$$\mathbf{F} = -\frac{\partial U}{\partial \mathbf{r}}.$$

This is unusual notation however (it's only really used by Russians), and we define the *gradient* ∇ of a function of a vector as

$$\nabla f(x, y, z) = \begin{pmatrix} \dfrac{\partial f}{\partial x} \\ \dfrac{\partial f}{\partial y} \\ \dfrac{\partial f}{\partial z} \end{pmatrix}.$$

(The symbol ∇ is called 'del' or 'nabla'.) For an n-dimensional vector, this is

$$\nabla f(x_1, x_2, \ldots, x_n) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{pmatrix}.$$

So, we can rewrite our new-found definition as

$$\mathbf{F} = -\nabla U(\mathbf{r}). \tag{5}$$

2.2 The definition of potential

The potential at a point in an electric field is the work done by an external agent in bringing a unit point positive test charge from infinity to that point.

Note well that usually when we say potential we mean *potential energy per unit charge.*

The above is the A-level definition, where the zero of potential is defined as being at infinity. For non-A-Level physics this may change, and we must also remember to take the limit as the magnitude of the test charge tends to zero.

Electric potential is usually given the symbol V; potential difference, or voltage, is the difference in potential between two points.

2.3 Potential and electric fields

We know that the force due to an electric field $\mathbf{E}$ is $\mathbf{F} = q\mathbf{E}$ where q is the test charge magnitude. So, in this case our new definition in eq. (5) becomes

$$\begin{aligned} q\mathbf{E} &= -\nabla U(\mathbf{r}) \\ \implies \mathbf{E} &= -\nabla \frac{U(\mathbf{r})}{q} \end{aligned}$$

but $\frac{U(\mathbf{r})}{q}$ is the potential energy per unit charge — that is, the potential V:

$$\mathbf{E} = -\nabla V(\mathbf{r}) \tag{6}$$

So, potential is just the gradient of an electric field!

2.4 The potential near a point charge

Take a point charge (or a sphere by the shell theorem!) of total charge Q positioned at the origin. At a distance r from the charge, therefore, the electric field strength is by Coulomb's law

$$E = \frac{Q}{4\pi\varepsilon_0 r^2}.$$

Hence, we know by our definition of potential in eq. (6) that

$$\frac{\mathrm{d}V}{\mathrm{d}r} = -\frac{Q}{4\pi\varepsilon_0 r^2}.$$

(This is a special one-dimensional case.) Integrating both sides, and remembering that potential must be zero at $r = \infty$,

$$\begin{aligned} V &= \int_{-\infty}^{r} -\frac{Q}{4\pi\varepsilon_0 r^2}\,\mathrm{d}r \\ &= \frac{Q}{4\pi\varepsilon_0 r}. \end{aligned} \tag{7}$$

Therefore, for a point charge, potential decreases linearly with distance.

2.5 Capacitors

Consider two parallel charged plates of surface charge density $+\sigma$ and $-\sigma$ separated by distance l. We know the electric field strength between the two is

$$E = \frac{\sigma}{\varepsilon_0}$$

and the potential difference is given by

$$E = \frac{\mathrm{d}V}{\mathrm{d}x} = \frac{V}{l}.$$

So,

$$\begin{aligned} \frac{\sigma}{\varepsilon_0} &= \frac{V}{l} \\ \implies \frac{Q}{\varepsilon_0 A} &= \frac{V}{l} \\ \implies \frac{Q}{V} &= C = \frac{\varepsilon_0 A}{l} \end{aligned}$$

where C is the capacitance of the two plates, Q is the total charge and A is the area of each plate. This is a formula we've come across before for the capacitance of a parallel plate capacitor!

3 The *actual* integral of $1/x$

At A-level, $\displaystyle\int \frac{1}{x}\,\mathrm{d}x$ is usually given as $\ln|x| + c$. The modulus signs are not necessary though, as we'll show. We all know that for $n \neq -1$,

$$\int x^n\,\mathrm{d}x = \frac{x^{n+1}}{n+1} + c$$

and so where $\varepsilon = n + 1$,

$$\int x^{\varepsilon-1}\,\mathrm{d}x = \frac{x^\varepsilon}{\varepsilon} + c$$

for $\varepsilon \neq 0$. We'll run with this to see what happens if ε does equal zero. Rewriting the exponent,

$$\int x^{\varepsilon-1}\,\mathrm{d}x = \frac{1}{\varepsilon}\mathrm{e}^{\varepsilon \ln x} + c$$

and now using the infinite expansion for $\mathrm{e}^{f(x)}$,

$$\begin{aligned} \int x^{\varepsilon-1}\,\mathrm{d}x &= \frac{1}{\varepsilon}\left(1 + \varepsilon\ln x + \frac{\varepsilon^2 \ln^2 x}{2} + O(\varepsilon^3)\right) + c \\ &= \frac{1}{\varepsilon} + \ln x + \frac{\varepsilon \ln^2 x}{2} + O(\varepsilon^2) + c. \end{aligned}$$

Now as $\varepsilon \to 0$, the $O(\varepsilon^2)$ terms become negligible and we end up having an infinite constant $1/\varepsilon$. The rest, however, behaves very nicely:

$$\begin{aligned}\int \frac{1}{x}\,\mathrm{d}x &= \lim_{\varepsilon\to 0}\left[\frac{1}{\varepsilon} + \ln x + \frac{\varepsilon \ln^2 x}{2} + O(\varepsilon^2) + c\right] \\ &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon} + c + \ln x.\end{aligned}$$

We can just 'absorb' the infinite constant into our pre-existing constant of integration, resulting in

$$\int \frac{1}{x}\,\mathrm{d}x = \ln x + c.$$

This is true for all real x.

4 Making integrable the potential near a line of charge

If you're trying to do this yourself, don't spend more than 10 minutes — it takes Dr Cheung 10 seconds.

Consider an infinite line of charge density σ. At a distance r from the line, the potential is clearly

$$V = \frac{\lambda}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{\sqrt{x^2+r^2}}, \tag{8}$$

as $\sqrt{x^2+r^2}$ is the distance to some point on the line. If we try to evaluate this integral, however, we end up with

$$V = \frac{\lambda}{2\pi\varepsilon_0}\int 0^{\frac{\pi}{2}}\sec\theta\,\mathrm{d}\theta$$

where $x = r\tan\theta$. This is a problem, as

$$\int \sec\theta\,\mathrm{d}\theta = \ln\sec\theta + \tan\theta + c$$

but $\sec\frac{\pi}{2}$ and $\tan\frac{\pi}{2}$ are both undefined.

Physically, this is due to our definition of potential: to derive eq. (7) we took $V = 0$ at a distance of infinity. Clearly doing so here results in a potential of infinity wherever we are, and so we must somehow get rid of this infinite offset.

4.1 Differentiating to find the field strength

The most obvious thing to do is find the electric field strength by differentiating immediately (as we know $E = -\frac{\partial V}{\partial r}$), and then we can use that to find the potential. So,

$$\begin{aligned}E &= -\frac{\partial}{\partial r}\left[\frac{\lambda}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{\sqrt{x^2+r^2}}\right] \\ &= \frac{\lambda r}{4\pi\varepsilon_0}\int_{-\infty}^{\infty}\frac{\mathrm{d}x}{(x^2+r^2)^{3/2}}.\end{aligned}$$

This integral we can solve using a substitution of $x = r\tan\theta$, and we come to the result

$$E = \frac{\lambda r}{4\pi\varepsilon_0}\left[\frac{x}{r^2\sqrt{x^2+r^2}}\right]_{-\infty}^{\infty}$$

which can be checked easily by differentiating. Now to evaluate this, clearly

$$\lim_{x\to\infty}\frac{x}{\sqrt{x^2+r^2}} = \frac{x}{x} = 1$$

and

$$\lim_{x\to-\infty}\frac{x}{\sqrt{x^2+r^2}} = \frac{x}{-x} = -1,$$

so

$$E = \frac{\lambda}{4\pi\varepsilon_0 r}\big(1-(-1)\big) = \frac{\lambda}{2\pi\varepsilon_0 r}$$

which we know to be the correct expression for the field strength! Trying to integrate this to find the potential immediately reveals our earlier problem from before: $\ln(\infty)$ and $\ln(0)$ are both undefined. If we want a nice expression for potential we'd have to pick some other zero point. We do know now, though, that the potential V must go like $\ln r$.

4.2 Taking the limit as the line length tends to infinity

The other way we can approach this problem is to start with a line of finite length and then let it go to infinity. Let the line have length $2L$, and to simplify the situation we'll assume our point of measurement is a perpendicular distance r from the *midpoint* of the line. So, eq. (8) becomes

$$V_L = \frac{\lambda}{4\pi\varepsilon_0}\int_{-L}^{L}\frac{\mathrm{d}x}{\sqrt{x^2+r^2}}.$$

We'll solve this integral as described before, by substituting $x = r\tan\theta$:

$$\begin{aligned}V_L &= \frac{\lambda}{2\pi\varepsilon_0 r}\int_0^{\arctan\frac{L}{r}}\frac{r\sec^2\theta\,\mathrm{d}\theta}{\sqrt{\tan^2\theta+1}} \\ &= \frac{\lambda}{2\pi\varepsilon_0}\int_0^{\arctan\frac{L}{r}}\sec\theta\,\mathrm{d}\theta.\end{aligned}$$

Multiplying by $\frac{\tan\theta+\sec\theta}{\tan\theta+\sec\theta}$ gives

$$\begin{aligned} V_L &= \frac{\lambda}{2\pi\varepsilon_0}\Big[\ln(\tan\theta+\sec\theta)\Big]_0^{\arctan\frac{L}{r}} \\ &= \frac{\lambda}{2\pi\varepsilon_0}\left[\ln\left(\frac{L}{r}+\frac{\sqrt{L^2+r^2}}{r}\right)-\ln(0+1)\right] \\ &= \frac{\lambda}{2\pi\varepsilon_0}\left[\ln(L+\sqrt{L^2+r^2})-\ln r\right]. \end{aligned}$$

Now we want to find the asymptotic behaviour of this function. Looking at the expression inside the first logarithm, we can use the generalised binomial expansion to expand the first few terms:

$$\begin{aligned} L+\sqrt{L^2+r^2} &= L\left[1+\sqrt{1+\frac{r^2}{L^2}}\right] \\ &= L\left[1+1+\frac{r^2}{2L^2}+O\left(\frac{r^4}{L^4}\right)\right] \\ &= 2L\left[1+\frac{r}{4L^2}+O\left(\frac{r^4}{L^4}\right)\right]. \end{aligned}$$

So,

$$V=\frac{\lambda}{2\pi\varepsilon_0}\left[\ln 2L-\ln r+\ln\left(1+\frac{r^2}{4L^2}+O\left(\frac{r^4}{L^4}\right)\right)\right].$$

Now, as the length of the line $L\to\infty$, the terms $\frac{r^2}{4L^2}$ and $O\left(\frac{r^4}{L^4}\right)$ become negligible, and so our potential is

$$\begin{aligned} V &= \lim_{L\to\infty} V_L \\ &= \frac{\lambda}{2\pi\varepsilon_0}\lim_{L\to\infty}\left[\ln 2L-\ln r\right. \\ &\qquad \left.+\ln\left(1+\frac{r^2}{4L^2}+O\left(\frac{r^4}{L^4}\right)\right)\right] \\ &= \frac{\lambda}{2\pi\varepsilon_0}(\ln 2L-\ln r+\ln 1) \\ &= \frac{\lambda}{2\pi\varepsilon_0}(\ln 2L-\ln r). \end{aligned}$$

Here's our infinite constant: $\ln 2L$. Ignore that (which constitutes a redefinition of the zero point of potential as before), and we have

$$V=-\frac{\lambda\ln r}{2\pi\varepsilon_0}$$

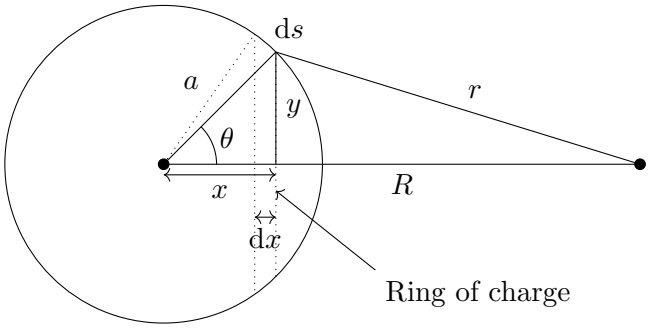
just as we were expecting.

It's worth noting that we had to assume our point was midway along the line; this method does not preserve translation invariance. It might be worthwhile verifying the result is still reached if our point is an arbitrary distance from the end of the line.

5 Proving the shell theorem without nasty integration

So far our proofs of the shell theorem have revolved around setting up integrals of trigonometric functions, the evaluation of which have posed the main difficulty. Here is a way to solve this problem without resorting to such intricate calculus.

We start by setting up a similar diagram to before:



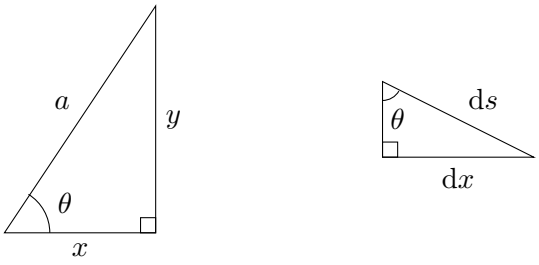
The ring of charge we're considering now has radius y and width $\mathrm{d}s$, so area $2\pi y\,\mathrm{d}s$. Therefore, the charge on this ring is

$$\mathrm{d}Q=2\pi\sigma y\,\mathrm{d}s$$

where σ is the surface charge density of the shell. Hence, at our point of measurement the contribution of potential from this ring is

$$\mathrm{d}V=\frac{\mathrm{d}Q}{4\pi\varepsilon_0 r}=\frac{2\pi\sigma y\,\mathrm{d}s}{4\pi\varepsilon_0 r.} \tag{9}$$

Now look at the following two triangles:



By approximating the arc length $\mathrm{d}s$ with a straight edge, we create a triangle (shown right) which angle

tracing shows is similar to the triangle shown to the left: hence, by comparing the two, we see that

$$\frac{\mathrm{d}s}{\mathrm{d}x} = \frac{a}{y}$$

and so eq. (9) becomes

$$\mathrm{d}V = \frac{\sigma a\,\mathrm{d}x}{2\varepsilon_0 r}. \tag{10}$$

Now applying the Pythagorean theorem gives both

$$r^2 = y^2 + (R-x)^2 = y^2 + x^2 - 2Rx + R^2$$

and

$$y^2 + x^2 = a^2$$

and so

$$r^2 = a^2 - 2Rx + R^2,$$

differentiating which with respect to x leads to

$$2r\frac{\mathrm{d}r}{\mathrm{d}x} = -2R$$
$$\therefore \frac{\mathrm{d}r}{R} = -\frac{\mathrm{d}x}{r}.$$

Substituting this into eq. (10) gives us

$$\mathrm{d}V = -\frac{\sigma a\,\mathrm{d}r}{2\varepsilon_0 R}$$

and now we just need to integrate both sides:

$$\begin{aligned} V &= -\int_{R-a}^{R+a} \frac{\sigma a\,\mathrm{d}r}{2\varepsilon_0 R} \\ &= -\frac{\sigma a}{2\varepsilon_0 R}\int_{R-a}^{R+a} \mathrm{d}r \\ &= -\frac{\sigma a}{2\varepsilon_0 R}\left(R + a - R + a\right) \\ &= -\frac{\sigma a}{2\varepsilon_0 R}(2a) \\ &= -\frac{\sigma a^2}{\varepsilon_0 R} \end{aligned}$$

but as $\sigma = \frac{Q}{4\pi a^2}$,

$$V = -\frac{Q}{4\pi\varepsilon_0 R}$$

as predicted; finally, differentiating gives our electric field strength as

$$E = -\frac{\partial V}{\partial R} = -\frac{Q}{4\pi\varepsilon_0 R^2}.$$

This is a beautiful method and demonstrates how with a little bit of thinking much difficult maths can be avoided.

Part III

Team submissions to competitions

Chapter 21

Princeton University Physics Competition 2016

In November 2016 three of us in Year 12 along with three Year 13s took part in this competition. It was an amazing week and this is what we came up with. Most of it wasn't me but I did do most of the formatting and I was responsible for the majority of the simulation part of the plasma physics section.

Princeton University Physics Competition - 2016 Online Exam

Team: Absolute Noethers

Highgate School
United Kingdom

November 19, 2016

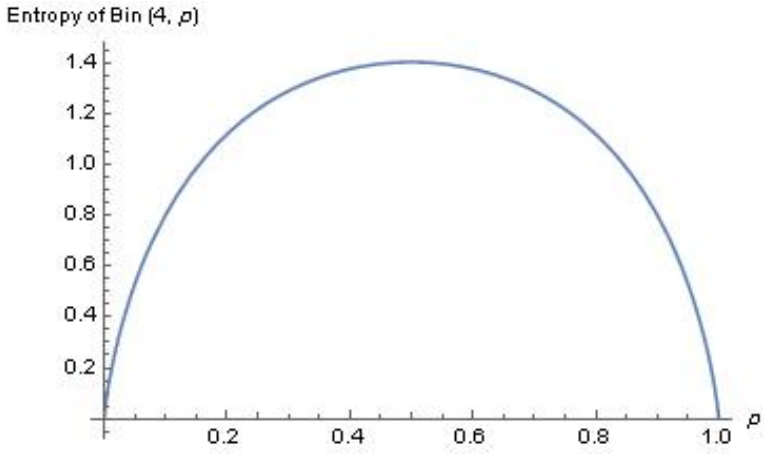
Contents

Entropy and Statistical Mechanics

1 Entropy as information

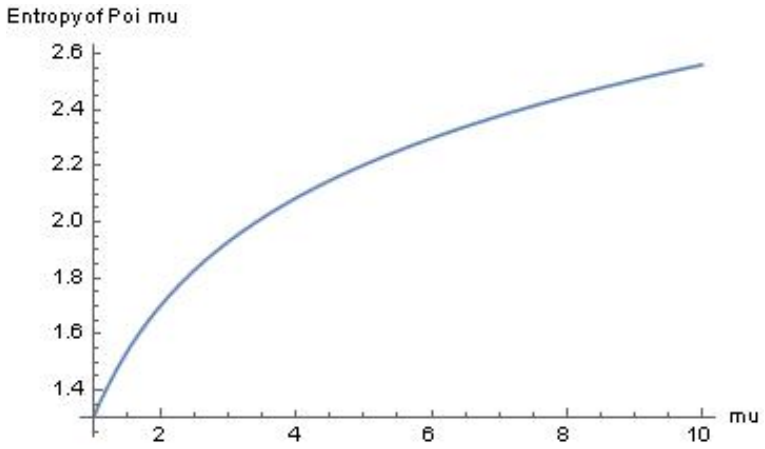
1.1 Quantifying the amount of information in the answer to a question

To get some intuition for our new definition of entropy, we will first apply it to two common discrete probability distributions: the binomial and the Poisson. We first plot the entropy of the binomial distribution for 4 trials, $\sim Bin(4, p)$, against the trial success probability p:



Clearly the entropy is highest when p is around 0.5, or when we are most uncertain about what the result of our 4 trials might be.

Next we plot the entropy of the Poisson distribution $Poi(\mu)$ as a function of its mean μ:



The entropy is an increasing function of μ. This seems reasonable since the variance of the Poisson distribution is equal to μ, so when the mean is higher the "spread" of the distribution is greater, and we are more uncertain about the outcome of a trial.

Now that we have some understanding of the meaning of high or low entropy, let us explore more quantitatively the range of values that I, the information entropy, can take.

We're given the definition

$$I \coloneqq -\sum_{i=1}^{N} p_i \log(p_i). \tag{1}$$

Since the p_i are probabilities, $0 \leq p_i \leq 1$. Thus $\log(p_i) \leq 0$, and so $p_i \log(p_i) \leq 0$ for all i. Hence, $I = -\sum p_i \log(p_i) \geq 0$.

We now ask what combination of p_i's (satisfying $\sum_{i=1}^{N} p_i = 1$) maximise I for a given N.

Jensen's inequality states that if f is a convex function, a_i are real constants satisfying $\sum a_i = 1$, and x_i are arbitrary real numbers, then

$$f\left(\sum (a_i x_i)\right) \leq \sum a_i f(x_i). \tag{2}$$

Noting that $-\log(x)$ is convex, we let $f(x) = -\log(x)$, we let $a_i = p_i$, and we let $x_i = \frac{1}{p_i}$. Then, by Jensen's inequality,

$$\begin{aligned} -\log\left(\sum_{i=1}^{N} p_i \frac{1}{p_i}\right) &\leq \sum_{i=1}^{N} p_i \left(-\log \frac{1}{p_i}\right) \\ -\log \sum_{i=1}^{N} (1) &\leq \sum_{i=1}^{N} p_i \log(p_i) \\ -\sum_{i=1}^{N} p_i \log(p_i) &\leq \log(N) \\ I &\leq \log(N). \end{aligned} \tag{3}$$

Equality is achieved only when each $p_i \log(p_i) = 0$; hence, either $p_i = 0$ or $\log p_i = 0$ for each i. Thus we put $p_k = 1$ and $p_{i \neq k} = 0$, and use the well known result

$$\lim_{x \to 0} x \log x = 0 \tag{4}$$

to give $p_i \log(p_i) \to 0$ when $i \neq k$, as well as $p_k \log p_k = 1 \log 1 = 0$. Therefore,

$$I = 0. \tag{5}$$

In other words, if a system can only be in one state, we gain no information from asking what state it is in. However, $I \geq 0$ and we cannot "lose" information through such a question.

Equality is achieved by setting all p_i equal to $\frac{1}{N}$. Then,

$$\begin{aligned} I &= -\sum_{i=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) \\ &= \sum_{i=1}^{N} \frac{\log(N)}{N} \\ &= N \cdot \frac{\log(N)}{N} \\ &= \log(N). \end{aligned} \tag{6}$$

In other words, we learn the most information by asking about a system whose states occur with equal probability In some sense, this represents a system in which we are "maximally confused" about to begin with.

When dealing with logarithms, we convert between bases using

$$\log_b x = \frac{\log_e x}{\log_e b}. \tag{7}$$

Thus the logarithms withing the formula for entropy are all simply scaled by the same constant, $\frac{1}{\log b}$, if a base b is used. This has no meaningful effect on the value of the entropy calculated and is equivalent to using grams instead of kilograms as a unit, for example.

Let us discuss our definition of entropy with regards to the three conditions listed on page 5 of the question paper:

Condition 1:

Consider two such independent systems, 1 and 2. The first has N_1 possible states ω_i^1, each with probability p_i. The second has N_2 possible states ω_j^2, each with probability q_j.

We now set out to show that the entropy of the joint distribution I_{tot} (where the states ω_i^1 and ω_j^2 are observed with probability $p_i q_j$) is equal to the sum of the individual entropies I_1 and I_2

of system 1 and 2:

$$
\begin{aligned}
I_{\text{tot}} &= -\sum_{1\le i\le N_1, 1\le j\le N_2} p_i q_j log(p_i q_j) \\
&= -\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_i q_j \log(p_i q_j) \\
&= -\sum_{i=1}^{N_1} p_i \sum_{j=1}^{N_2} q_j \log(p_i q_j) \qquad \text{since } p_i \text{ is independent of } j \\
&= -\sum_{i=1}^{N_1} p_i \sum_{j=1}^{N_2} q_j (\log p_i + \log q_j) \\
&= -\sum_{i=1}^{N_1} p_i \left(\log p_i \sum_{j=1}^{N_2} q_j + \sum_{j=1}^{N_2} q_j \log q_j \right) \qquad \text{since } \log(p_i) \text{ is independent of } j \\
&= -\sum_{i=1}^{N_1} \left(p_i \log(p_i) \sum_{j=1}^{N_2} q_j \log q_j \right) \qquad \text{since } \sum_{j=1}^{N_2} q_j = 1 \\
&= -\sum_{i=1}^{N_1} p_i \log(p_i) - \sum_{i=1}^{N_1} \left(p_i \sum_{j=1}^{N_2} q_j \log q_j \right) \\
&= -\sum_{i=1}^{N_1} p_i \log(p_i) - \left(\sum_{i=1}^{N_1} p_i \right) \left(\sum_{j=1}^{N_2} q_j \log q_j \right) \qquad \text{since } \sum_{j=1}^{N_2} q_j \log q_j \text{ is independent of } i \\
&= -\sum_{i=1}^{N_1} p_i \log(p_i) - \sum_{j=1}^{N_2} q_j \log(q_j) \qquad \text{since } \sum_{i=1}^{N_1} p_i = 1 \\
&= I_1 + I_2 \qquad (8)
\end{aligned}
$$

as desired.

Condition 2 is already satisfied, since I is defined only using the set of probabilities p_i.

Condition 3:

We have already shown that $I \le \log(N)$, and that $I = \log(N)$ is a maximum when all p_i are equal to $\frac{1}{N}$. Hence, this condition is also satisfied.

2 Where does entropy show up in Statistical Mechanics?

2.1 An example: rigid chain and the "force" that entropy causes

2.1.2 Finding the probability of certain positions

The free end of the chain may be in any positions n with $-\frac{N}{2} \le n \le \frac{N}{2}$. Each chain link has two possible states: pointing left or pointing right. Since there are N links, there are 2^N possible states of the whole chain.
Since each link has length $\frac{1}{2}$, in order for the end of the chain to lie at a position n there must be $2n$ more links pointing to the right than to the left. There are N links in total, so we find that

there must be $\frac{N}{2}+n$ pointing to the right and $\frac{N}{2}-n$ to the left. The number of configurations which achieve this is

$$\binom{N}{\frac{N}{2}+n} = \frac{N!}{(\frac{N}{2}+n)!(\frac{N}{2}-n)!}. \tag{9}$$

We divide by the total number of configurations, 2^N, to find the probability the end of the chain is at a position n as

$$p(x=n) = \frac{N!}{(\frac{N}{2}+n)!(\frac{N}{2}-n)!}2^{-N}. \tag{10}$$

Since each state with the end of the chain at a position n is equally likely, the entropy at a position n is simply the logarithm of the number of possible states with the end at n. This number is just the probability of finding the end at n, multiplied by the total number of states. Thus,

$$\sigma_{\text{Sys}}(x=n) = \log(2^N p(x=n)). \tag{11}$$

Hence the entropy increases as the probability increases.

2.1.3 Stirling's approximation: a way to simplify the formula

We previously established that

$$\begin{aligned} p(x=n) &= \binom{N}{\frac{N}{2}+n}2^{-N} \\ &= \frac{N!}{(\frac{N}{2}+n)!(\frac{N}{2}-n)!}2^{-N}. \end{aligned} \tag{12}$$

Using Stirling's formula we have

$$\begin{aligned} p &\approx \frac{1}{\sqrt{2\pi}} \cdot \frac{N^N\sqrt{N}}{(\frac{N}{2}+n)^{\frac{N}{2}}(\frac{N}{2}-n)^{\frac{N}{2}}} \cdot \frac{2^{-N}}{(\frac{N^2}{4}-n^2)^{\frac{1}{2}}} \cdot \left(\frac{\frac{N}{2}-n}{\frac{N}{2}+n}\right)^n \\ &\approx \frac{2}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{N}} \left(1-\frac{4n^2}{N^2}\right)^{-\frac{1}{2}} \left(1-\frac{4n^2}{N^2}\right)^{-\frac{N}{2}} \left(1+2\frac{2n}{N}\right)^{-n}. \end{aligned} \tag{13}$$

Taking the logarithm of both sides gives

$$\log p \approx \log\left(\frac{2}{\sqrt{2\pi}}\right) - \frac{1}{2}\log(N) - \frac{1}{2}\log\left(1-\frac{4n^2}{N^2}\right) - \frac{N}{2}\log\left(1-\frac{4n^2}{N^2}\right) - n\log\left(1+2\frac{2n}{N}\right). \tag{14}$$

Using the first order approximation $log(1-u) \approx -u$ simplifies our expression to

$$\log p \approx \log\left(\frac{2}{\sqrt{2\pi}}\right) - \frac{1}{2}\log(N) + \frac{2n^2}{N^2} + \frac{2n^2}{N} - \frac{4n^2}{N}. \tag{15}$$

The third term is smaller than the last two due to the N^2 in the denominator, so it can be dropped to give

$$\log p \approx \log\sqrt{\frac{2}{\pi}} - \frac{1}{2}\log(N) - \frac{2n^2}{N}. \tag{16}$$

Finally, we exponentiate both sides to achieve the desired result:

$$p(x=n) \approx \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{N}} e^{-\frac{2n^2}{N}}, \tag{17}$$

and since $\sigma_{\text{Sys}} = \log(2^N p(x=n))$,

$$\sigma_{\text{Sys}} \approx N\log(2) + \log\sqrt{\frac{2}{\pi}} - \frac{1}{2}\log(N) - \frac{2n^2}{N}. \tag{18}$$

2.2 An abstract derivation of entropy in statistical mechanics

2.2.2 A derivation

Thus:

$$\begin{aligned} p(E_{\text{Sys}} = E_k) &\propto [\text{number of system states with } E_{\text{Sys}} = E_k] \\ &\quad \times [\text{number of reservoir states with } E_{\text{Sys}} = E_k] \\ &\propto [\text{number of system states with } E_{\text{Sys}} = E_k] \\ &\quad \times [\text{number of reservoir states with } E_{\text{Res}} = E_{\text{tot}} - E_k] \\ &\propto \mathrm{e}^{\sigma_{\text{Sys}}(E_k)} \times \mathrm{e}^{\sigma_{\text{Res}}(E_{\text{tot}} - E_k)} \\ &\propto \mathrm{e}^{\sigma_{\text{Sys}}(E_k)} \times \mathrm{e}^{\sigma_{\text{Res}}(E_{\text{tot}} - E_k)} \times e^{-\sigma_{\text{Res}}(E_{\text{tot}})}. \end{aligned} \tag{19}$$

since $-\sigma_{\text{Res}}(E_{\text{tot}})$ is just a constant. So

$$p(E_{\text{Sys}} = E_k) \propto e^{\sigma_{\text{Sys}} + \sigma_{\text{Res}}(E_{\text{tot}} - E_k) - \sigma_{\text{Res}}(E_{\text{tot}})}. \tag{20}$$

But on page 11, after β is introduced in equation (4), we are told that

$$\sigma_{\text{Res}}(E_{\text{tot}} - E_k) = \sigma_{\text{Res}}(E_{\text{tot}}) - \beta E_k \tag{21}$$

to a good approximation when $\frac{E_k}{E_{\text{tot}}} \ll 1$. Therefore

$$\sigma_{\text{Res}}(E_{\text{tot}} - E_k) - \sigma_{\text{Res}}(E_{\text{tot}}) = -\beta E_k. \tag{22}$$

Substituting this equality into our above expression for $p(E_{\text{Sys}} = E_k)$ gives

$$\begin{aligned} p(E_{\text{Sys}} = E_k) &\propto \mathrm{e}^{\sigma_{\text{Sys}} - \beta E_k} \\ &\propto \mathrm{e}^{-\beta[E_k - \frac{1}{\beta}\sigma_{\text{Sys}}(E_k)]} \\ &\propto \mathrm{e}^{-\beta[E_k - \tau\sigma_{\text{Sys}}(E_k)]}. \end{aligned} \tag{23}$$

So for some normalisation constant ζ_3, we have

$$\begin{aligned} p(E_{\text{Sys}} = E_k) &= \frac{1}{\zeta_3}\mathrm{e}^{-\beta[E_k - \tau\sigma_{\text{Sys}}(E_k)]} \\ &= \frac{1}{\zeta_3}\mathrm{e}^{-\beta\mathcal{F}(E_k)} \end{aligned} \tag{24}$$

as given.

The two "competing" effects which determine the most likely state are essentially the terms "number of system states with $E_{\text{Sys}} = E_k$" and "number of reservoir states with $E_{\text{Sys}} = E_k$". Multiplying these two terms gives the total number of possible system-reservoir states at $E_{\text{Sys}} = E_k$, and so the most likely energy maximises their product. Increasing E_{Sys} might cause, for instance, an increase in the number of possible reservoir states, but in causing E_{Res} to decrease it might also reduce the number of reservoir states, thus potentially lowering their product.

2.2.4 A pause for reflection

In our previous work, we made the following assumptions:

- In order to compute well-defined entropies, we assumed that our system and reservoir could only exist in a finite number of states. In order to make this true, we must somehow allow states which are "sufficiently similar" to be deemed the same. For instance, when considering the average speed of a collection of particles, we might round observed speeds to the nearest integer, say, in order to discretise an otherwise continuous quantity and reduce the number of states to a finite value.
- We assumed that *a priori* we have no reason to believe any one state is more likely than another, so long as both have the correct energy. This allows us to write $\sigma_{\text{Sys}} = \log N$ where N is the number of possible states and was vital in our analysis. To explain this, we use the model that the system changes state very quickly - much more frequently than our observations - so that we can sensibly talk about probabilities of finding each state and eventually visit every possible state.
- So far we have assumed that the system and reservoir are isolated from the rest of the universe, so that the energy of the system is conserved.
- In order to use the approximation $p(\omega_k^S) \propto e^{-\beta E_k}$ we assumed that $E_k \ll E_{tot}$, so that the first order approximation found by truncating the Taylor series (shown on page 12 of the question paper) is accurate. In other words, the reservoir contains most of the energy, a realistic assumption.

2.3 Why is our definition of temperature reasonable?

When $E = N\eta$ or $-N\eta$, there is only one state the system can be in: all spin up or all spin down, respectively. Hence $\sigma_{\text{Sys}}(\pm N\eta) = 0$; nothing is learnt from the answer to the question "What state is the system in?".

There are the same number of states with energy E as there are with energy $-E$, since for every arrangement with an energy E we can change each spin up particle to spin down, and vice versa, to create an arrangement with energy $-E$. Hence $\sigma_{\text{Sys}}(E) = \sigma_{\text{Sys}}(-E)$.

The number of possible states with a given energy follows a binomial distribution with a mean E, so the most common energy is near $E = 0$ (it may not be exactly 0 if N is odd). Hence $\sigma_{\text{Sys}}(E)$ is maximised near $E = 0$.

To have an energy $E = n\eta$, the system must have n more spin up particles than spin down ones. This means there must be $\frac{N}{2} + \frac{n}{2}$ spin up particles and $\frac{N}{2} - \frac{n}{2}$ spin down ones, so that $\left(\frac{N}{2} + \frac{n}{2}\right) + \left(\frac{N}{2} - \frac{n}{2}\right) = N$ and $\left(\frac{N}{2} + \frac{n}{2}\right) - \left(\frac{N}{2} - \frac{n}{2}\right) = n$ as desired. Thus the number of states with energy $E = n\eta$ is given by

$$\binom{N}{\frac{N}{2} + \frac{n}{2}} = \frac{N!}{(\frac{N}{2} + \frac{n}{2})!(\frac{N}{2} - \frac{n}{2})!}. \tag{25}$$

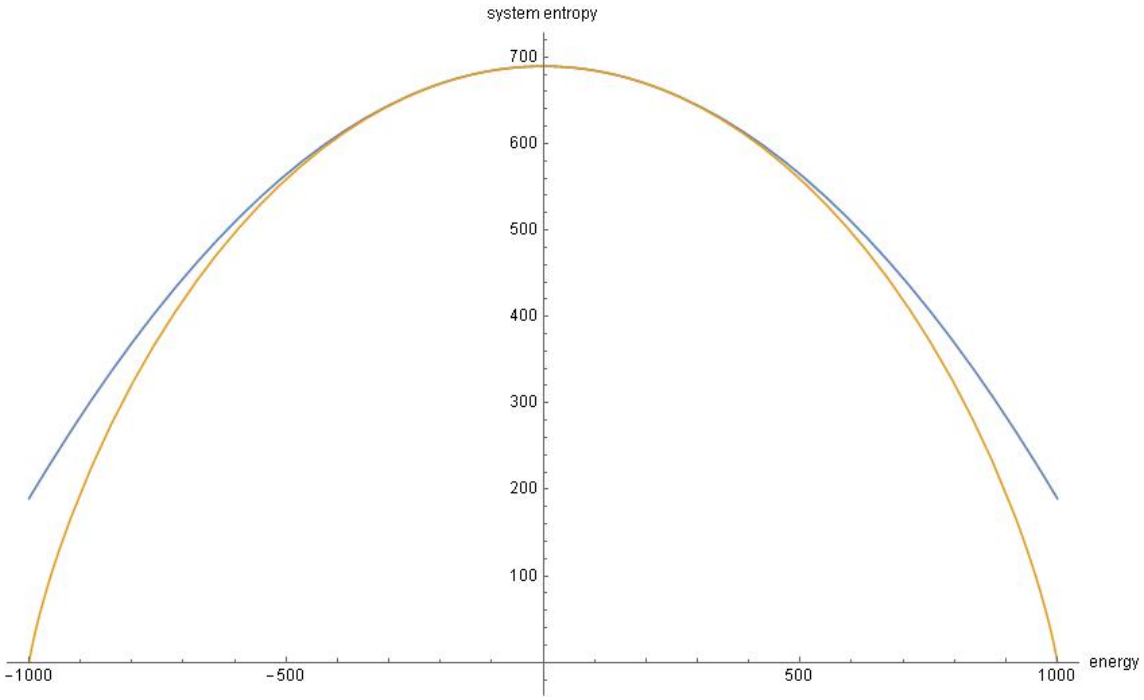
From our work on Stirling's approximation in 2.1.3 (noting that the n in the binomial from before has now become $\frac{n}{2}$), we can write

$$\begin{aligned}\binom{N}{\frac{N}{2} + \frac{n}{2}} &\approx 2^N \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{N}} \mathrm{e}^{-\frac{n^2}{2N}} \\ &\approx 2^N \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{N}} \mathrm{e}^{-\frac{E^2}{2\eta^2 N}}\end{aligned} \tag{26}$$

All spin states are equally likely, so $\sigma_{\text{Sys}}(E)$ is simply the logarithm of number of states at energy E:

$$\begin{aligned}\sigma_{\text{Sys}}(E) &\approx \log(\frac{2^N}{\sqrt{N}}e^{-\frac{E^2}{2\eta^2 N}}) + \log(\frac{2}{\pi}) \\ &\approx \log(\frac{2^N}{\sqrt{N}}) - \frac{E^2}{2\eta^2 N}\end{aligned} \tag{27}$$

Below is a plot of the true value of the entropy, shown in orange, and our approximation, shown in blue. We have chosen the values $N = 1000$ and $\eta = 1$ here. Note that the approximation is very good when $|E|$ is not too close to $N\eta = 1000$, but is an overestimate when E gets larger.



$$\frac{1}{\tau} \coloneqq \frac{\partial \sigma_{\text{Res}}(E)}{\partial E} \tag{28}$$

We have found that for a reservoir of spins, $\sigma_{\text{Sys}}(E) = \log(\frac{2^N}{\sqrt{N}}) - \frac{E^2}{2\eta^2 N}$.

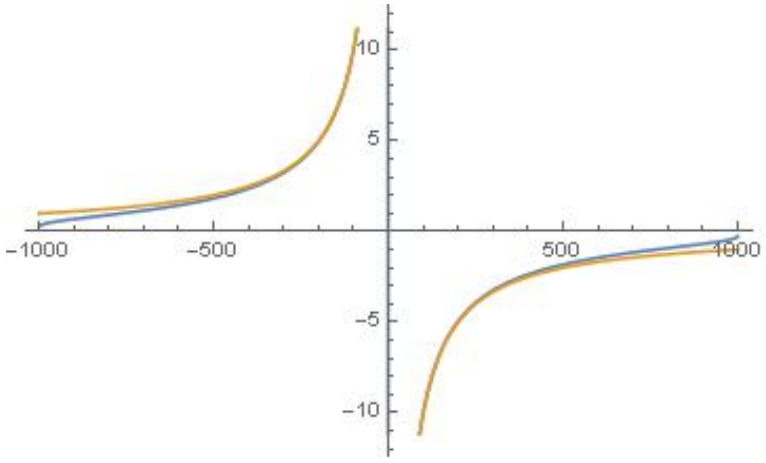
Thus in this case:

$$\begin{aligned}\frac{1}{\tau} &= \frac{\partial}{\partial E}(\log(\frac{2^N}{\sqrt{N}}) - \frac{E^2}{2\eta^2 N}) \\ &= \frac{-E}{\eta^2 N})\end{aligned}$$

So:

$$\tau = \eta^2 N \frac{-1}{E} \tag{29}$$

Below we plot τ against E for $N = 1000$, $\eta = 1$. Our approximation is shown in orange, and the true value is shown in blue.



Again note that this approximation is best when $|E|$ is significantly smaller than ηN.
τ is an increasing function on $[-\infty, 0)$ and $(0, \infty]$, but due to its asymptote at $E = 0$, we still have $\tau(E_1) < \tau(E_2)$ for all $E_1 > 0$ and $E_2 < 0$. Due to the asymptote at $E = 0$, τ has no well defined maximum or minimum: its range is $[-\infty, \infty]$.

Let E'_A and E'_B be the most probable values for E_A and E_B, with $E'_A + E'_B = E_{\text{tot}}$. Then we wish for $\sigma_{\text{tot}} = \sigma_{\text{A}}(E'_A) + \sigma_{\text{B}}(E'_B)$ to be maximised.

Due to our energy conservation relation, E_A depends explicitly on E_B , and so to find the maximum we may partial differentiate with respect to E_{A} and set to 0:

$$\begin{aligned} &\frac{\partial}{\partial E_{\text{A}}}(\sigma_{\text{A}}(E_{\text{A}}) + \sigma_{\text{B}}(E_{\text{B}}))|_{E_{\text{A}}=E'_A, E_B=E'_B} = 0 \\ &\Rightarrow \frac{\partial \sigma_A(E_A)}{\partial E_A}|_{E_A=E'_A} + \frac{\partial \sigma_B(E_B)}{\partial E_{\text{B}}} \times \frac{\partial E_B}{\partial E_A}|_{E_B=E'_B} = 0 \end{aligned} \tag{30}$$

due to the chain rule for differentiation. We now use our definition of temperature in equation (4), and substitute in the energy conservation relation $E_B = E_{tot} - E_A$ to give:

$$\begin{aligned} &\frac{1}{\tau_A} + \frac{1}{\tau_B} \times \frac{\partial}{\partial E_A}(E_{\text{tot}} - E_A) = 0 \\ &\Rightarrow \frac{1}{\tau_A} - \frac{1}{\tau_B} = 0 \\ &\Rightarrow \tau_A = \tau_B \end{aligned} \tag{31}$$

as desired.

If our equilibrium condition $\tau_A = \tau_B$ is satisfied, then $E_A = E_B$. Since $E_A + E_B$ is constant, the energies of both systems must tend towards the average of their starting values: $\frac{E_{A0}+E_{B0}}{2}$. If $E_{A0} < E_{B0}$, then E_A will increase and E_B will decrease until $E_A = E_B = \frac{E_{A0}+E_{B0}}{2}$. If, in addition, E_{A0} and E_{B0} are both positive or both negative, then $E_{A0} < EB0 \Rightarrow \tau_{A0} < \tau_{A0}$. Hence energy flows from higher temperature (system B) to lower temperature (system A), as we might expect.

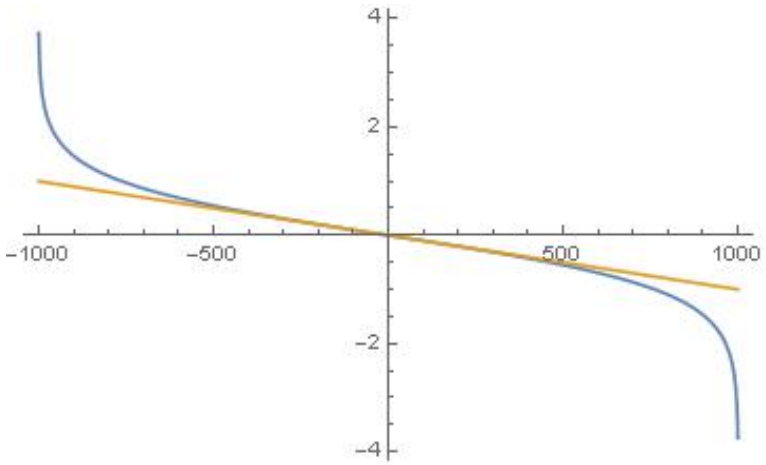
However, in the case that $E_{A0} < 0 < E_{B0}$, then we have $\tau_{A0} > 0 > \tau_{B0}$. Just as before, E_A increases and E_B decreases until equilibrium. Thus the energy flows from lower temperature

(system B) to higher temperature (system A). Note that at some point during relaxation, at least one of the systems will have a temperature which is undefined, as it goes to positive or negative infinity when crossing $E = 0$. If $E_{A0} = -E_{B0}$, then the equilibrium energy is 0 and so the equilibrium temperatures are $\pm\infty$.

We briefly consider the scenario where the energy of one system increases indefinitely whilst the other decreases so as to maintain constant total energy. It may appear that the two systems are heading towards equilibrium; however in nature there is no such system whose energy could increase or decrease without bound - in this case the limiting factor is the number of particles present. What's more, such a scenario would only go to minimise the total entropy, as there would be very few, if any, possible states that each system could be in at such energies. The fact that τ does indeed go to 0 at very high or low energies is in fact indicative of tending towards a global minimum - when we set the derivative to 0 in equation (6) we potentially admitted solutions for E'_A and E'_B which corresponded to minima rather than maxima of the total entropy. Only in the case described above is this fact potentially misleading.

Lastly, we find that we can avoid having to consider these separate cases by looking at $\beta := \frac{1}{\tau}$. Then $\beta(E) = -\eta^2 NE$ is a decreasing function for all E, and lacks the asymptote suffered by $\tau(E)$ at $E = 0$). Now, energy *always* flows from low β (system B) to high β (system A). Hence β is arguably a more convenient and physical quantity to work with than τ.

Below is a plot of β against E, again with $N = 1000$, $\eta = 1$. Our approximation, which clearly now has a smaller range of validity, is shown in orange and the true value is shown in blue. Both curves are clearly monotone decreasing, which is really the important feature.



3 A more precise analysis of free energy and entropy

3.1 A system coupled to an energy reservoir

3.1.1 A new definition of free energy

Question 1

We begin by finding a convenient expression for σ_{Sys}:

$$\begin{aligned}
\sigma_{Sys} &= -\sum p_k \log p_k \\
&= -\sum p(w_k) \log(p(w_k)) \\
&= -\sum \frac{\mathrm{e}^{-\beta E_k}}{\mathcal{Z}} \log\left(\frac{\mathrm{e}^{-\beta E_k}}{\mathcal{Z}}\right) \\
&= -\sum \frac{\mathrm{e}^{-\beta E_k}}{z}(-\beta E_k - \log z) \\
&= -\sum \frac{(\log \mathcal{Z} + \beta E_k)\mathrm{e}^{-\beta E_k}}{\mathcal{Z}} \\
&= \frac{1}{\mathcal{Z}} \log \mathcal{Z} \sum \mathrm{e}^{-\beta E_k} + \frac{1}{\mathcal{Z}} \beta \sum E_k \mathrm{e}^{-\beta E_k} \\
&= \frac{1}{\mathcal{Z}} (\log \mathcal{Z}) \mathcal{Z} + \frac{1}{\mathcal{Z}} \beta \sum -\frac{\partial}{\partial \beta}(\mathrm{e}^{-\beta E_k}) \\
&= \log \mathcal{Z} - \frac{1}{\mathcal{Z}} \beta \frac{\partial}{\partial \beta} \sum e^{-\beta E_k} \\
&= \log \mathcal{Z} - \beta \frac{\left(\frac{\partial \mathcal{Z}}{\partial \beta}\right)}{\mathcal{Z}} \\
&= \log \mathcal{Z} - \beta \frac{\partial}{\partial \beta} \log \mathcal{Z} \qquad (32)
\end{aligned}$$

due to the chain rule used to differentiate $\log \mathcal{Z}$.

Next, we will find expressions for $\beta^2 \frac{\partial}{\partial \beta}\mathcal{F}$, $-\frac{\partial}{\partial \tau}\mathcal{F}$ and $\frac{\partial}{\partial \tau}\tau \log \mathcal{Z}$ and show that they are all equal to expression (32), and thus equal to σ_{Sys}. Using (9) from the question paper:

$$\begin{aligned}
\beta^2 \frac{\partial}{\partial \beta}\mathcal{F} &= \beta^2 \frac{\partial}{\partial \beta}(-\tau \log \mathcal{Z}) \\
&= -\beta^2 \frac{\partial}{\partial \beta}\left(\frac{1}{\beta} \log \mathcal{Z}\right) \\
&= -\beta^2 \left(\frac{1}{\beta^2} \log \mathcal{Z} + \frac{1}{\beta} \frac{\partial}{\partial \beta} \log \mathcal{Z}\right)
\end{aligned}$$

where we have used the product rule for differentiation. Thus

$$\beta^2 \frac{\partial}{\partial \beta}\mathcal{F} = \log \mathcal{Z} - \beta \frac{\partial}{\partial \beta} \log \mathcal{Z}$$

and because of (32),

$$\beta^2 \frac{\partial}{\partial \beta}\mathcal{F} = \sigma_{\text{Sys}}. \qquad (33)$$

Now,

$$-\frac{\partial}{\partial\tau}\mathcal{F} = -\frac{\partial}{\partial\tau}(-\tau\log\mathcal{Z})$$

due to (9) from the question paper, so

$$-\frac{\partial}{\partial\tau}\mathcal{F} = \tau\frac{\partial}{\partial\tau}\log\mathcal{Z} + \frac{\partial\tau}{\partial\tau}\log\mathcal{Z}$$

by the product rule for differentiation. We apply the chain rule and simplify:

$$\begin{aligned}
&= \tau\frac{\partial}{\partial\beta}(\log\mathcal{Z}) \times \frac{\partial\beta}{\partial\tau} + \log\mathcal{Z} \\
&= \tau\frac{\partial}{\partial\beta}(\log\mathcal{Z}) \times \left(\frac{-1}{\tau^2}\right) + \log\mathcal{Z} \\
&= \frac{-1}{\tau}\frac{\partial}{\partial\beta}\log\mathcal{Z} + \log\mathcal{Z} \\
&= \log\mathcal{Z} - \beta\frac{\partial}{\partial\beta}\log\mathcal{Z}.
\end{aligned}$$

Hence, due to (32),

$$-\frac{\partial}{\partial\tau}\mathcal{F} = \sigma_{\text{Sys}}. \tag{34}$$

Finally,

$$\begin{aligned}
\frac{\partial}{\partial\tau}(\tau\log\mathcal{Z}) &= -\frac{\partial}{\partial\tau}(-\tau\log\mathcal{Z}) \\
&= -\frac{\partial}{\partial\tau}\mathcal{F}
\end{aligned}$$

so, as we showed above,

$$\frac{\partial}{\partial\tau}(\tau\log\mathcal{Z}) = \sigma_{\text{Sys}}. \tag{35}$$

Thus

$$\sigma_{\text{Sys}} = \beta^2\frac{\partial}{\partial\beta}\mathcal{F} = -\frac{\partial}{\partial\tau}\mathcal{F} = \frac{\partial}{\partial\tau}(\tau\log\mathcal{Z}). \tag{36}$$

Question 2

$$\begin{aligned}
\langle E\rangle &\coloneqq \sum E_k p(w_k) \\
&= \sum E_k\frac{\mathrm{e}^{-\beta E_k}}{\mathcal{Z}} \\
&= -\frac{1}{\mathcal{Z}}\sum -E_k e^{-\beta E_k} \\
&= -\frac{1}{\mathcal{Z}}\sum\frac{\partial}{\partial\beta}(\mathrm{e}^{-\beta E_k})
\end{aligned}$$

due to the chain rule. So

$$\begin{aligned}
\langle E\rangle &= -\frac{1}{\mathcal{Z}}\frac{\partial}{\partial\beta}\sum\mathrm{e}^{-\beta E_k} \\
&= -\frac{1}{\mathcal{Z}}\frac{\partial}{\partial\beta}\mathcal{Z} \\
\langle E\rangle &= -\frac{\partial}{\partial\beta}\log\mathcal{Z}
\end{aligned} \tag{37}$$

using the fact that $\dfrac{\partial}{\partial\beta}\log\mathcal{Z} = \dfrac{\left(\frac{\partial\mathcal{Z}}{\partial\beta}\right)}{\mathcal{Z}}$ due to the chain rule.

Question 3

$$\begin{aligned}
\langle E\rangle - \tau\sigma_{\text{Sys}} &= \sum E_k p(w_k) + \frac{1}{\beta}\sum p(w_k)\log(p(w_k)) \\
&= \sum E_k p(w_k) + \frac{1}{\beta}\sum p(w_k)\log\left(\frac{\mathrm{e}^{-\beta E_k}}{\mathcal{Z}}\right) \\
&= \sum E_k p(w_k) + \frac{1}{\beta}\sum p(w_k)(-\beta E_k - \log\mathcal{Z}) \\
&= \sum E_k p(w_k) - \frac{1}{\beta}\sum \beta E_k p(w_k) - \frac{1}{\beta}\sum p(w_k)\log\mathcal{Z} \\
&= \sum E_k p(w_k) - \sum E_k p(w_k) - \tau\log\mathcal{Z}\sum p(w_k)
\end{aligned}$$

since $\log\mathcal{Z}$ is constant. Thus,

$$\langle E\rangle - \tau\sigma_{\text{Sys}} = 0 - \tau\log\mathcal{Z}$$

since $\sum p(w_k) = 1$. Hence,

$$\begin{aligned}
\langle E\rangle - \tau\sigma_{\text{Sys}} &= -\tau\log\mathcal{Z} \\
&= \mathcal{F}
\end{aligned} \tag{38}$$

due to equation (9).

Question 4

By equation (10) in the question paper:

$$\begin{aligned}
\mathcal{F} &= \langle E\rangle - \tau\sigma_{\text{Sys}} \\
\Rightarrow \quad \frac{\partial}{\partial\tau}\mathcal{F} &= \frac{\partial\langle E\rangle}{\partial\tau} - \frac{\partial}{\partial\tau}(\tau\sigma_{\text{Sys}}) \\
&= \frac{\partial\langle E\rangle}{\partial\tau} - \sigma_{\text{Sys}} - \tau\frac{\partial}{\partial\tau}\sigma_{\text{Sys}}.
\end{aligned} \tag{39}$$

We found earlier that $\sigma_{\text{Sys}} = -\frac{\partial}{\partial\tau}\mathcal{F}$. Hence:

$$\begin{aligned}
\frac{\partial}{\partial\tau}\mathcal{F} &= \frac{\partial\langle E\rangle}{\partial\tau} + \frac{\partial}{\partial\tau}\mathcal{F} - \tau\frac{\partial\sigma_{\text{Sys}}}{\partial\tau} \\
\Rightarrow \quad \frac{\partial\langle E\rangle}{\partial\tau} &= \tau\frac{\partial\sigma_{\text{Sys}}}{\partial\tau}
\end{aligned} \tag{40}$$

as desired.

Now, using $\mathcal{F} = -\tau\log\mathcal{Z}$ gives:

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial E_i} &= \frac{\partial}{\partial E_i}(\tau \log(z)) \\ &= -\tau \frac{(\frac{\partial z}{\partial E_i})}{z} \\ &= -\frac{1}{z}\tau \frac{\partial}{\partial E_i} \sum_k \mathrm{e}^{-\beta E_k}\end{aligned}$$

Since, $e^{-\beta E_k}$ does not depend on E_i for $k \neq i$, and thus the derivatives of all such terms with respect to E_i are 0:

$$\begin{aligned}&= -\frac{1}{z}\tau \frac{\partial}{\partial E_i} \mathrm{e}^{-\beta E_k} \\ &= -\frac{1}{z}\tau \cdot \beta \mathrm{e}^{-\beta E_k} \\ &= \frac{\mathrm{e}^{-\beta E_k}}{z} \\ &= p(\omega_i)\end{aligned}$$

By equation (10) in the question paper:

$$\frac{\partial \mathcal{F}}{\partial E_i} = \frac{\partial}{\partial E_i}(\langle E \rangle - \tau \sigma_{\mathrm{Sys}}) \tag{41}$$

$$= \frac{\partial \langle E \rangle}{\partial E_i} - \tau \frac{\partial \sigma_{textSys}}{\partial E_i} \tag{42}$$

Now, equating our two expressions for $\frac{\partial \mathcal{F}}{\partial E_i}$ gives:

$$\frac{\partial \mathcal{F}}{\partial E_i} = p(\omega_i) = \frac{\partial \langle E \rangle}{\partial E_i} - \tau \frac{\partial \sigma_{\mathrm{sys}}}{\partial E_i} \tag{43}$$

Hence $\frac{\partial \langle E \rangle}{\partial E_i} = \tau \frac{\partial \sigma_{\mathrm{sys}}}{\partial E_i} + p(\omega_i)$ as desired.

Now we use the fact that, for small changes δx_i to the arguments of a function f of n variables:

$$f(x_1 + \delta x_1, x_2 + \delta x_2, ...x_n + \delta x_n) \approx f(x_1, x_2, ...x_n) + \delta x_1 \frac{\partial f}{\partial x_1} + \delta x_2 \frac{\partial f}{\partial x_2} + ...\delta x_n \frac{\partial f}{\partial x_n}$$

In words, the total change in value of a function is equal to the sum of the change in value of each argument multiplied by the rate of change of the function with respect to that argument (to a good approximation for small changes):

$$\delta f \approx \sum_{i=1}^{n} \delta x_i \frac{\partial f}{\partial x_i} \tag{44}$$

Thus we can conclude:

$$\delta \langle E \rangle = \sum \delta x_i \frac{\partial \langle E \rangle}{\partial x_i}$$

where x_i are the arguments of $\langle E \rangle$, which in this case are E_i and τ.

$$\begin{aligned} &= \sum \delta E_i \frac{\partial \langle E \rangle}{\partial E_i} + \delta\tau \frac{\partial \langle E \rangle}{\partial \tau} \\ &= \sum \delta E_i p(\omega_i) + \delta\tau \cdot \tau \frac{\partial \sigma_{\text{Sys}}}{\delta \tau} \end{aligned} \tag{45}$$

by our previous expressions for $\frac{\partial \langle E \rangle}{\partial E_i}$ and $\frac{\partial \langle E \rangle}{\partial \tau}$.

$$= \sum p(\omega_i)\delta E_i + \tau \delta \sigma_{\text{sys}}$$

since $\frac{\partial \sigma_{\text{sys}}}{\partial \tau} \cdot \delta\tau \approx \delta\sigma_{\text{sys}}$ for small $\delta\tau$.

So $\delta\langle E \rangle = \sum p(\omega_i)\delta E_i + \tau\delta\sigma_{\text{sys}}$ as desired.

3.1.2 Defining energy that is extractable from a system in terms of free energy

4 Applications: non-equilibrium changes in biological and computational systems

4.1 Jarzynski's equality: a non-equilibrium work relation

4.1.1 Statement of Jarzynski's equality

Jensen's inequality states that if f is a convex function and x a random variable, then:

$$\langle f(x) \rangle \geq f(\langle x \rangle) \tag{46}$$

Noting that the function $e^{-\beta x} =$ is convex, we let $f(x) = \mathrm{e}^{-\beta x}$ and $x = W$ where we write W to mean $W_{\text{actual, ext on sys}}$ for brevity. Then Jensen's inequality implies:

$$\langle \mathrm{e}^{-\beta W} \rangle \geq e^{-\beta \langle W \rangle} \tag{47}$$

By Jensen's equality we can rewrite the LHS to give:

$$\mathrm{e}^{-\beta \Delta G} \geq \mathrm{e}^{-\beta \langle W \rangle}$$

$\mathrm{e}^{-\beta x}$ is a decreasing function with x, so:

$$\begin{aligned} & \beta \Delta G \geq \beta \langle W \rangle \\ \Rightarrow \quad & \langle W \rangle \geq \Delta G \\ & \geq \Delta F \quad \text{as desired.} \end{aligned} \tag{48}$$

The results from eq.(29) in the paper are familiar results, when viewed simply from the perspective of conservation of energy. We would not expect the average dissipated work of a process to

be negative, since from the definition, it would imply that more energy is obtained from undoing a process than is used in doing it. Extending this argument, we could conceivably generate any amount of energy we wanted by repeating these steps. It also makes sense from an entropic point of view. The laws of thermodynamics tell us - to paraphrase - that work done in a useful way is inevitably syphoned off during a reaction into a form which is unhelpful (i.e. heat). If we consider for example a computer performing binary operations, the state of each bit is rapidly changed, so it frequently undergoes a reaction which is quickly reversed, returning to its initial state, and yet RAM heats up very quickly in computers.

4.1.2 Experimental test of Jarzynski's equality

1. **Experimental Setup**

   Broadly speaking, J. Liphardt *et al.* set out to test various theoretical measures of energy lost during irreversible reactions. The reaction chosen was the stretching and eventual unfolding of a molecule of RNA. The molecule is formed of a long chain of amino acids looped and bonded to itself. The reaction investigated breaking the looped, self bonded chain, and turning it into a straightened, relatively flat chain. This was achieved by attaching molecular handles to each end of the chain, and attaching each handle to a relatively large polystyrene bead. One bead is held in place by a laser arrangement, and the other held at the end of a micro-pipette. The micro-pipette was moved and the force on the molecule was measured by sensitive photodetectors, which could analyse the changes in deflection of the laser beams. This force was recorded along with the total length of the molecule, though since the molecule was 341 nm when looped this was set as the zero point, and the force recorded for relative lengths $0 - 30\,\text{nm}$. The work in reaching some length z was obtained computationally as the integral of force with respect to distance. Two separate investigations took place: one where the reaction happened slowly, and thus could be modelled as reversible, and one where the reaction happened quickly, not giving the system time to stay close to equilibrium, and making the reaction more apparently irreversible.

2. **Reservoir**

   "Before external perturbation, the molecule is at equilibrium with the thermal bath" - Due to the microscopic nature of the experiment (to reduce the standard deviation of the work values), we may assume the solution, of mainly water and buffer, in which the experiment takes place, acts as the reservoir and maintains a roughly constant temperature throughout.

3. **Measure of work done**

   The work done in each breaking and reforming of the molecule is given by the integral of the force (measured by the laser deflection) with regard to the position z.

4. **Measure of Gibbs free energy**

   The Gibbs energy is estimated by the experimenters to be (using their notation) $W_{\text{A,rev}}$, which is the work done in taking the molecule from un-stretched to stretched in the reversible slow stretching reaction. Clearly, to be truly reversible the system would have to change infinitely slowly, so the use of $W_{\text{A,rev}}$ is a best estimate rather than an exact measure of ΔG.

5. **Testing the equality**

   The test of the equality is measuring how well it agrees with empirical evidence. As is made explicit in equation (1) of the paper, a theoretical value of ΔG (which we will call W_{JE})

can be obtained by taking the mean of $exp\left[-\beta w_i(z,r)\right]$ over many experiments, such that $W_{JE} = -\frac{1}{\beta}\log(\langle\exp\left[-\beta w_i(z,r)\right]\rangle)$. The disagreement between theory and experiment is encoded in $W_{A,rev} - W_{JE}$, and this difference is then plotted against z. "Agreement" means the difference lies within the error bounds of the experiment, stated in the paper (without justification) to be $k_BT/2$.

6. **Difference between curves**

   Figure 2(a) depicts the force-extension curves of the RNA molecules at two distinct rates of force exertion. The blue curve is produced from a slower rate of force exertion, so the RNA molecule is able to extend in a reversible manner and close to zero energy is lost in the extension and retraction. This can be deduced from the extension and retraction curves being approximately equal, so their integrals with respect to extension (work done) are approximately equal and of opposite sign if one considers the direction of time. So the energy of the system is the same before and after the process. By contrast, the red curve is produced from a faster rate of force exertion, so the RNA molecule extends in an irreversible manner. Energy is lost from the system, as can be deduced by the extension curve having a greater integral than the retraction curve, so more work is done in extending the RNA molecule than is recovered from it retracting.

7. **Predictions versus readings for the reversible reaction**

   Fig.3(a) and (b) shows the difference between theory and experiment for the measure of free energy of the molecule for varying values of z. (a) explores the case of the near-reversible reaction (slow extension). The measure of the difference increases in magnitude in a decay shape. The fact that the rate of divergence doesn't do anything special around the critical range suggests that this divergence isn't related to inefficiencies of the model, but rather something to do with the experimental error. This is also supported by the fact that two separate models for the difference in free energy agree almost precisely with one another. The experimenters note that the time taken for the molecule to fully extend being longer means there is more time for experimental noise to accumulate, which increases the energy of the system by some unknown amount. This random variable increases the variance in values for the work, meaning the actual error is probably somewhat higher than that already stated. From the form of Jarzynski's equality it is clear that the greatest contributions to W_{JE} come from low values for the work (since $\exp(-x)$ takes on larger values for smaller x). Since the measured work for slower extension is more consistent, the instrument noise has a much larger effect on the variance. These two facts combined make the measured value have a much higher error than formally stated.

8. **Predictions versus readings for the irreversible reaction**

   Fig.3(b) explores the case of the irreversible reaction, when the system is not given time to remain close to equilibrium. It plots the difference between prediction and experiment against z for all three models of the Gibbs Energy (the most important being W_{JE}). The solid line predicts no energy dissipation, so obviously it will overestimate the free energy difference in the scenario when energy is lost in the extension-retraction process. This manifests itself in the solid lines lying above the axis. The dotted lines show somewhat better experimental agreement than solid lines, which makes sense since this model is highly effective for dealing with energy dissipation when close to the equilibrium. However this model breaks down once the hysteresis gives way around $z = 15$nm. The prediction from Jarzynski's Equality, however, remains very close to the experimenters approximation for ΔG across almost the entire range of z, only falling outside of error for z close to 30nm (and as we have already discussed the error bounds are fairly conservative).

9. **Probabilities of values of dissipated work**

The switching rate for the blue data set is slower than the switching rate of the red data set, because the rate of force exertion is slower for the blue data set, and so it takes longer for the molecule to fully extend and retract. We expect the average dissipated work in the blue data set to be 0 because as illustrated in figure 2(a), the work done in extending is approximately equal to the work recovered in retracting, so we would expect the system to conserve energy. This is why we define 0 dissipated work to be the mean dissipated work of the blue data set.

In a system with a positive mean dissipated work $\langle W_{dis} \rangle > 0$, we might expect that $\langle \mathrm{e}^{-\beta W_{dis}} \rangle < 1$: a violation of Jarzynski's equality. However, measured values of W_{dis} which are below the mean will increase the value of $\mathrm{e}^{-\beta W_{dis}}$ by more than corresponding values above the mean will decrease it. This is true due to the convexity of the $\mathrm{e}^{-}x$. Applying Jensen's inequality tells us that

$$\langle \mathrm{e}^{-\beta W_{dis}} \rangle \geq \mathrm{e}^{-\beta \langle W_{dis} \rangle}.$$

So even if $W_{dis} > 0$ and the right hand side of the above is less than 1, it is possible that the left hand side is in fact equal to 1, satisfying Jarzynski's equality.

In other words, lower values of W_{dis} are "weighted more" in computing the average.

4.2 Dissipation in computational systems

4.2.4 Operation of the AND gate from the perspective of statistical mechanics

Before initialising, $\langle E \rangle = \epsilon$ since all states have energy E. There are 4 possible states, each with equal probability, and so $\sigma_{\mathrm{Sys}} = \log 4$. Hence:

$$\mathcal{F}_{\text{before init}} = \epsilon - \tau \log 4 \tag{49}$$

After initialisation, one state, say ω_1, has energy ϵ whilst the other three (ω_2, ω_3 and ω_4) have energies which we take to be infinite. Using $p(\omega_1) = \frac{1}{\mathcal{Z}} e^{-\beta E_i}$ we can write:

$$\begin{aligned}
\langle E \rangle &= \sum E_i p(\omega_i) \\
&= \epsilon p(\omega_1) + 3 \lim_{x \to \infty} E \frac{\mathrm{e}^{-\beta E}}{\mathcal{Z}} \\
&= \epsilon p(\omega_1) 3\beta \lim_{\beta E \to \infty} \frac{e^{-\beta E}}{\beta E}
\end{aligned}$$

But using $\lim_{x \to \infty} x\mathrm{e}^{-x}$with $x = \beta E$ gives:

$$\begin{aligned}
\lim_{\beta E \to \infty} \frac{\mathrm{e}^{-\beta E}}{\beta E} &= 0 \\
\text{Thus} \langle E \rangle &= \epsilon p(\omega_1)
\end{aligned}$$

As E_2, E_3 and E_4 go to infinity, the probabilities $p(\omega_2, \omega_3$ and $\omega_4)$ all go to 0, so $p(\omega_1)$ goes to 1. Thus:

$$\langle E \rangle \geq \epsilon \tag{50}$$

Since there is only one possible state after initialisation, $\sigma_{\mathrm{Sys}} = \log 1 = 0$.
Hence: $\mathcal{F}_{\text{before comp}} = \epsilon - 0 = \epsilon$.

So the work done by an external agent in initialising the system is:

$$\begin{aligned}\Delta\mathcal{F} &= \mathcal{F}_{\text{before comp}} - \mathcal{F}_{\text{before init}} \\ &= \epsilon - (\epsilon - \tau \log 4) \\ &= \tau \log 4 \end{aligned} \tag{51}$$

By the same reasoning that we used before computation, the expected energy after computation is ϵ: only the states with energy ϵ make any contribution to the expected value.
The number of possible states is now 2, and both have equal probability. Hence $\sigma_{\text{Sys}} = \log 2$.
Thus $\mathcal{F}_{\text{after comp}} = \epsilon - \tau \log 2$
So during computation, the work done on the system is:

$$\begin{aligned}\mathcal{F}_{\text{after comp}} - \mathcal{F}_{\text{before comp}} &= (\epsilon - \tau \log 2) - \epsilon \\ &= -\tau \log 2. \end{aligned} \tag{52}$$

The negative sign indicates that the computer is in fact extracting energy out of the system. The total work done in initialising the system and running the computation is:

$$\begin{aligned}\mathcal{F}_{\text{after comp}} - \mathcal{F}_{\text{before init}} &= (\epsilon - \tau \log 2) - (\epsilon - \tau \log 4) \\ &= \tau \log(\frac{4}{2}) \\ &= \tau \log 2 \end{aligned} \tag{53}$$

So overall, energy has been put into the system in order to perform the computation.

4.2.5 Where is the energy dissipated?

Although the computation is said to be "irreversible", of course no energy is lost from the computer (under the assumption that it is isolated from the rest of the universe). The work done by the agent in performing the computational cycle is dissipated as heat, which could conceivably be gathered and used to power future computations. However, assuming the heat is is lost and cannot be recovered, the term "irreversible" is suitable.

4.2.6 Generalizing the computational model

Now, instead of 4 states (or the suggested number of 40), let us consider a system with n equally probably states for each combination of part A and part B, thus making a total of 4n possible states in total. During initialisation and computation, the expected energy $\langle E \rangle$ remains constant. Before initialisation there are 4n equally probable states, before computation there are only n, and after computation there there are 2n states. Thus the value of σ_{Sys} changes from $\log 4n$ to $\log n$, and finally to $\log 2n$.
So the total work done on the system during these steps, i.e the energy cost of one sequence of

computation, is:

$$\begin{aligned}\mathcal{F}_{\text{after comp}} - \mathcal{F}_{\text{before init}} &= (\langle E\rangle - \tau\log 2n) - (\langle E\rangle - \tau\log 4n)\\ &= \tau\log(\frac{4n}{2n})\\ &= \tau\log 2\end{aligned} \tag{54}$$

Hence the energy cost remains the same as the system with only 4 states.

4.2.7 Computing without the cost

A reversible gate does not erase any bits, and so the number of possible states after computation is equal to the number of states before computation. In other words, there is a bijective mapping between system states before and after computation. Since we assume each distinct state to be equally probable, the entropy of the system is $\sigma_{\text{Sys}} = \log N$ where N is the number of states: a quantity which we know remains constant. So σ_{Sys} is constant, and since $\langle E\rangle$ is also constant we can conclude that $\mathcal{F} = \langle E\rangle - \tau\sigma_{\text{Sys}}$ is constant, so long as the temperature is unchanged. Hence (theoretically) no work must be done on the system in order to perform a computation with a reversible gate. The computer extracts no energy.

In general, modern computers are not at all close to reaching the thermodynamic limits of computation, and are far more bound by physical implementation than by theoretical limitation.

In terms of RAM, most modern computers do not reach the thermodynamic limits that we have discussed. For example, let us compare the performance of current laptops with their thermodynamic limits; let us assume that the "ideal" laptop has mass 1 kg and volume 1 litre. It follows that $\tau = k_B T = 8.10\cdot 10^{-15}$ joules, or $T = 634$ K.The entropy is $S = 2.04\cdot 108 joule/K$. This corresponds to a memory space with $I = S/k_B\log 2 = 2.13\cdot 10^{31}$ bits available. On the other hand, modern laptops generally operate with about 8GB RAM, which is approximately 10^{10} bits - well below the thermodynamic limits. Though current quantum computers are able to improve storage density by storing bits on individual atoms, this would allow up to around $6\cdot 10^{24}$ bits in a kg machine made of silicon. Furthermore, quantum computers still require huge development in order for them to be able to be used reliably and effectively, many require systems to cool them to extremely low temperatures in order to be able to function.

Laser and Plasma Physics

2 Theory

2.1 Plasma oscillation

Suppose the yz plane has some arbitrary large surface area, A. Then the volume of the region without electrons is given by $2Ad$. Prior to the electrons being displaced, this region had no net charge density. The electrons contributed a total of $n_e e$ to the charge density, and so the net charge density of the region following the electrons being displaced is:

$$0 - n_\mathrm{e}\mathrm{e} = -n_\mathrm{e}\mathrm{e} \tag{55}$$

So the net charge, Q, inside the region is given by:

$$Q = -2A\mathrm{d}n_\mathrm{e}\mathrm{e} \tag{56}$$

We apply Gauss' Law to a closed surface S that just encloses the region without electrons.

$$\Phi_E = \frac{Q}{\epsilon_0} = \oiint \vec{E} \cdot \mathrm{d}\vec{A} \tag{57}$$

As our plane is arbitrarily large we may assume the y and z components of our electric field average to 0, as the ion number density is sufficiently large that a symmetry argument may be applied. So:

$$\oiint \vec{E} \cdot \mathrm{d}\vec{A} = 2A\vec{E} \tag{58}$$

i.e, the electric field multiplied by the total surface area of our surface S, (Which the flux lines are always normal to).
Substituting in $-2A\mathrm{d}n_\mathrm{e}\mathrm{e}$ for Q and $2A\vec{E}$ for $\oiint \vec{E} \cdot \mathrm{d}\vec{A}$ we have;

$$\frac{-2A\mathrm{d}n_\mathrm{e}\mathrm{e}}{\epsilon_0} = 2A\vec{E} \tag{59}$$

at the border of the slab. Dividing by 2A:

$$\frac{-n_\mathrm{e}\mathrm{e}}{\epsilon_0} d = \vec{E} \tag{60}$$

As $\vec{E}q = \vec{F}$ and in the case of an electron, $q = e$:

$$\frac{-n_\mathrm{e}\mathrm{e}^2}{E_0}\mathrm{d} = \vec{E}\mathrm{e} = \vec{F} \tag{61}$$

Without loss of generality with electrons either side of the yz axis, let us refer to the electrons with $x > 0$ at the border of the slab from now on. Let their position, x, be a function of time,

with $x(0) = d$
We see that the above argument applies to any slab width, so for any x(t) we have:

$$-\frac{n_e e^2}{\epsilon_0} x = F \tag{62}$$

Note: We are no longer using vector notation as we only need to consider motion in the x direction. Applying Newton's Second Law to an electron at the border of the slab we have:

$$-\frac{n_e e^2}{\epsilon_0} x = F = m_e \ddot{x}$$
$$-\frac{n_e e^2}{\epsilon_0 m_e} x = \ddot{x} \tag{63}$$

This is the differential equation of simple harmonic motion, with a solution in the form:

$$x(t) = x_0 \cos(\omega t) + \frac{v_0}{\omega} \sin(\omega t) \tag{64}$$

where:

$$\omega^2 = \frac{n_e e^2}{m_e e_0}. \tag{65}$$

Assuming the electron(s) have no initial velocity, this reduces to: $x(t) = d\cos(\omega_{pl} t)$ for electrons with $x > 0$ initially, and $x(t) = -d\cos(\omega_{pl} t)$ for electrons with $x < 0$ initially, with an oscillation frequency of:

$$\omega_{pl} = \sqrt{\frac{n_e e^2}{m_e \epsilon_0}}. \tag{66}$$

2.2 The Langmuir wave: a warm model for the plasma oscillation

We will now simplify and then relate the equations of charge conservation and motion based on certain approximations and assumptions.
Starting with charge conservation:

$$e\frac{\partial n}{\partial t} + \frac{\partial nev}{\partial x} = 0 \tag{67}$$

as e is constant, we may write this as

$$e\frac{\partial n}{\partial t} + e\frac{\partial nv}{\partial x} = 0 \tag{68}$$

which we can divide by e:

$$\frac{\partial n}{\partial t} + \frac{\partial nv}{\partial x} = 0 \tag{69}$$

To this we apply the product rule:

$$\frac{\partial n}{\partial t} + n\frac{\partial v}{\partial x} + v\frac{\partial n}{\partial x} = 0 \tag{70}$$

Now writing n as $n_0 + n'$ and v as $v_0 + v'$:

$$\frac{\partial (n_0 + n')}{\partial t} + (n_0 + n')\frac{\partial (v_0 + v')}{\partial x} + (v_0 + v')\frac{\partial (n_0 + n')}{\partial x} = 0 \tag{71}$$

which we can expand to give

$$\frac{\partial n_0}{\partial t} + \frac{\partial n'}{\partial t} + v_0\frac{\partial n_0}{\partial x} + v_0\frac{\partial n'}{\partial x} + v'\frac{\partial n_0}{\partial x} + v'\frac{\partial n'}{\partial x} + n_0\frac{\partial v_0}{\partial x} + n_0\frac{\partial v'}{\partial x} + n'\frac{\partial v_0}{\partial x} + n'\frac{\partial v'}{\partial x} = 0 \quad (72)$$

Since n_0 and v_0 are constants, their derivatives are precisely zero. We also use the approximation that $n'\frac{\partial v'}{\partial x} \approx v'\frac{\partial n'}{\partial x} \approx 0$, since these are second order terms. This simplifies our equation significantly:

$$\frac{\partial n'}{\partial t} + v_0\frac{\partial n'}{\partial x} + n_0\frac{\partial v'}{\partial x} = 0$$

Since $v_0 = 0$, this becomes

$$\frac{\partial n'}{\partial t} + n_0\frac{\partial v'}{\partial x} = 0$$

Differentiating with respect to time:

$$\frac{\partial^2 n'}{\partial t^2} + n_0\frac{\partial^2 v'}{\partial t\partial x} + \frac{\partial n_0}{\partial t}\cdot\frac{\partial v'}{\partial x} = 0$$

Since $\frac{\partial n_0}{\partial t} = 0$, we get, upon multiplying by the mass of the particle (in this case the electron mass m_e)

$$m\frac{\partial^2 n'}{\partial t^2} + n_0 m\frac{\partial^2 v'}{\partial t\partial x} = 0 \quad (73)$$

Now considering the equation of motion:

$$nm\frac{\partial v}{\partial t} = -\frac{\partial n}{\partial x}E_{kin} + f$$

Using the same $n = n_0 + n'$ and $v = v_0 + v'$ substitutions we find that

$$(n_0 + n')m\frac{\partial(v_0 + v')}{\partial t} = -\frac{\partial(n_0 + n')}{\partial x}E_{kin} + f$$

$$m(n_0\frac{\partial v_0}{\partial t} + n_0\frac{\partial v'}{\partial t} + n'\frac{\partial v_0}{\partial t} + n'\frac{\partial v'}{\partial t}) = -(\frac{\partial n_0}{\partial x} + \frac{\partial n'}{\partial x})E_{kin} + f$$

We again note that $\frac{\partial v_0}{\partial t} = \frac{\partial n_0}{\partial x} = 0$, and that $n'\frac{\partial v'}{\partial t}$ is second order, so can be neglected:

$$n_0 m\frac{\partial v'}{\partial t} = -\frac{\partial n'}{\partial x}E_{kin} + f$$

Differentiating with respect to x this time:

$$n_0 m\frac{\partial^2 v'}{\partial x\partial t} = -\frac{\partial^2 n'}{\partial x^2}E_{kin} - \frac{\partial n'}{\partial x}\cdot\frac{\partial E_{kin}}{\partial x} + \frac{\partial f}{\partial x}$$

Young's Theorem tells us that $\frac{\partial^2 v'}{\partial t\partial x} \equiv \frac{\partial^2 v'}{\partial x\partial t}$ In addition, we may assume that the thermal velocity of the electron is much greater than the velocity of any oscillations, so E_{kin} is roughly constant, that is $\frac{\partial E_{kin}}{\partial x} \approx 0$. This leads us to the following relationship:

$$n_0 m\frac{\partial^2 v'}{\partial t\partial x} = -\frac{\partial^2 n'}{\partial x^2}E_{kin} + \frac{\partial f}{\partial x} \quad (74)$$

We now substitute (73) into (74) by eliminating $n_0 m\frac{\partial^2 v'}{\partial t\partial x}$:

$$-m\frac{\partial^2 n'}{\partial t^2} = -\frac{\partial^2 n'}{\partial x^2}E_{kin} + \frac{\partial f}{\partial x} \quad (75)$$

As $\frac{f}{n_0}$ is the external force on one particle of the fluid, particularly from an electric field in the case of an electron, this is given by eE, where E is the electric field, and e is the electron charge. Thus $f = n_0 e\mathrm{E}$. Since n_0 and e are constants, we can say

$$\frac{\partial f}{\partial x} = n_0 e \frac{\partial \mathrm{E}}{\partial x}$$

Which is equivalent, by Gauss' Law, to the statement

$$\frac{\partial f}{\partial x} = n_0 e \frac{\rho_e}{\epsilon_0}$$

For an electron in the plasma, $n_e = n_0$ indicates a net charge density of 0, so in this case $\rho_e = 0$. Our net charge density ρ_e is hence a measure (proportional to the charge of the electron, e)of the difference between n_e and n_0, namely n_e'. Hence $\rho_e = en_e'$. It then follows that

$$\frac{\partial f}{\partial x} = \frac{n_0 e^2}{\epsilon_0} n_e'$$

which we substitute into (75) to give

$$-m\frac{\partial^2 n_e'}{\partial t^2} = -\frac{\partial^2 n_e'}{\partial x^2} E_{kin} + \frac{n_0 e^2}{\epsilon_0} n_e' \tag{76}$$

Letting $n_e' = Ae^{i(\omega t - kx)}$ we have

$$\frac{\partial^2 n_e'}{\partial t^2} = -\omega^2 n_e'$$

$$\frac{\partial^2 n_e'}{\partial x^2} = -k^2 n_e'$$

which we can substitute back into (76), giving

$$m\omega^2 n_e' = k^2 E_{kin} n_e' + \frac{n_0 e^2}{\epsilon_0} n_e$$

Dividing by mn_e'

$$\omega^2 = \frac{k^2 E_{kin}}{m} + \frac{n_0 e^2}{m\epsilon_0} \tag{77}$$

If we recall from section 2.1 that $\omega_{pl}^2 = \frac{n_0 e^2}{m\epsilon_0}$, (as in our derivation of ω_{pl}^2 we assumed $n_0 = n_e$), we can make a substitution back into (77):

$$\omega^2 = \frac{k^2 E_{kin}}{m} + \omega_{pl}^2$$

Since we assumed also that $v_{therm} \gg v_{oscillations}$, $E_{kin} \approx \frac{1}{2} m v_{therm}^2$. It follows that

$$\omega^2 = \omega_{pl}^2 + \frac{k^2 v_{therm}^2}{2}$$

2.3 Raman scattering and magnetic potential

2.3.1 Raman scattering

The form of the electric field created by the incoming light produced with the polorisability gives the dipole moment

$$\mathbf{P} = \alpha \mathbf{E}_0 \cos(\omega_0 t)$$

Since α is a function of atomic positions and separations, but the bond length is large in comparison to Q_0, we can say

$$\alpha = \alpha_0 + \alpha_1 \, dQ + O(dQ^2)$$

$$\alpha_1 = \frac{d\alpha}{d(dQ)}\Big|_{dQ=0}$$

where α_0 is the contribution to α from the bond length which effectively stays constant (since this is the equilibrium position) , and thus the higher order terms account only for the changes in α resulting from small disturbances dQ. To our first-order approximation, terms of order dQ^2 or higher can be ignored. Therefore, we get

$$\begin{aligned}
\mathbf{P} &= (\alpha_0 + \alpha_1 \, dQ)\mathbf{E}_0 \cos(\omega_0 t) \\
&= (\alpha_0 + \alpha_1 \, Q_0 \cos(\omega_{vib} t))\mathbf{E}_0 \cos(\omega_0 t) \\
&= \alpha_0 \mathbf{E}_0 \cos(\omega_0 t) + \alpha_1 \, \mathbf{E}_0 Q_0 \cos(\omega_{vib} t) \cos(\omega_0 t) \\
&= \alpha_0 \mathbf{E}_0 \cos(\omega_0 t) + \frac{1}{2}\alpha_1 \, \mathbf{E}_0 Q_0 \left(\cos((\omega_0 + \omega_{vib})t) + \cos((\omega_0 - \omega_{vib})t)\right)
\end{aligned}$$

As can be seen, there are 3 modes of vibration, with angular frequencies ω_0, $\omega_0 - \omega_{vib}$, and $\omega_0 + \omega_{vib}$ for the Rayleigh, Stokes, and Anti-Stokes frequencies respectively.
We expect the Rayleigh frequency to dominate the observed scattered light, since the oscillations of the entire molecule dwarf the small vibrations of the atoms. Mathematically we can relate this to the Rayleigh frequency having a scalar factor of α_0, as opposed to $\frac{\alpha_1 Q_0}{2} = \frac{\frac{d\alpha}{d(dQ)}|_{dQ=0} Q_0}{2}$ for the Stokes frequencies.

2.3.2 Magnetic potential

First, we verify that the vector field $\mathbf{A}_0$ satisfies $\nabla \cdot \mathbf{A}_0 = 0$:

$$\nabla \cdot \mathbf{A}_0 = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

Since $A_x = A_z = 0$, this simplifies to

$$\nabla \cdot \mathbf{A}_0 = \frac{\partial A_y}{\partial y} = 0$$

since A_y is solely a function of x and t.
The flux density $\mathbf{B}$ is given by

$$\mathbf{B} = \nabla \times \mathbf{A}$$

But A has only y-components, and is only a function of x, so the only non-zero term is

$$\mathbf{B} = \hat{\mathbf{z}} \frac{\partial A_y}{\partial x} \mathbf{B} = \hat{\mathbf{z}} \frac{m_e c}{e} \, \mathbf{E}^{i(k_0 x - \omega_0 t)} \, (i k_o a_o + \frac{\partial a_0}{\partial x})$$

But we know that terms like $\frac{\partial a_0}{\partial x}$ can be ignored, leaving

$$\mathbf{B} = \hat{\mathbf{z}} i k_o a_o \frac{m_e c}{e} \mathbf{E}^{i(k_0 x - \omega_0 t)}$$

which we clearly only want the real part of:

$$\mathbf{B} = -\hat{\mathbf{z}} k_o a_o \frac{m_e c}{e} \sin(k_0 x - \omega_0 t)$$

It is a well known result that the intensity of an electromagnetic wave is given by the time average of the Poynting vector S [1]:

$$I_0 = \langle S \rangle = \frac{c B^2_{without\ sine}}{2\mu_0}$$

where we have used the fact $\langle \sin^2(k_0 x - \omega_0 t) \rangle = \frac{1}{2}$ when we average over time. Substituting in our expression for B and rearranging, we find that

$$I_0 = \frac{c}{2\mu_0}(-k_o a_o \frac{m_e c}{e})^2$$

$$\begin{aligned} a_0 &= \pm\lambda_0 \sqrt{I_0} \sqrt{\frac{2e^2\mu_0}{4\pi^2 m_e^2 c^3}} \\ &\approx \pm 8.55 \times 10^{-6}\, \lambda_0 \sqrt{I_0} \end{aligned}$$

3 Simulation: laser amplification in plasma using Raman backscattering

The simulation was created with the aim of simulating the interaction between a short, low-power seed pulse and a constant, high-power pump pulse in the presence of a Langmuir wave, resulting in the amplification of the seed and the depletion of the pump.

To start with, we calculated the initial intensity of the pump and seed pulses, using their total energy, time duration and cross-sectional area. We then used this (and their wavelengths) to find the initial values of a_1 and a_2 using the formula

$$a1_0(x,t) = 0.855 \times 10^{-5} \cdot \lambda_0 \sqrt{I_0(x,t)} \tag{78}$$

given on page 9 of the question paper.

We then found the cold plasma oscillation frequency ω_{pl} using the formula we derived in section 2.1, and with this found the Langmuir wave oscillation frequency ω_3 using the Bohm-Gross dispersion relation

$$\omega_3^2 = \omega_{pl}^2 + 3k^2 v_{thm}^2 \tag{79}$$

where k is the wavenumber of the Langmuir wave, equal to the sum of the positive wavenumbers of the seed and pump pulses.

Then having defined our observation period and number of discretising steps, we found the times after which the seed and pump pulses would end.

Next, applying the Euler method for discrete approximation to solutions of differential equations, we found from equations (10), (11), and (12) in the question paper the formulae

$$a_1(x, t+\delta t) = a_1(x,t) + \delta t \left(K a_2(x,t) a_3(x,t) - c \frac{a_1(x,t) - a_1(x-\delta x, t)}{\delta x} \right), \tag{80}$$

$$a_2(x, t+\delta t) = a_2(x,t) + \delta t \left(-K a_1(x,t) a_3(x,t) + c \frac{a_2(x+\delta x, t) - a_2(x,t)}{\delta x} \right), \tag{81}$$

$$a_3(x, t+\delta t) = a_3(x,t) + \delta t[-K a_1(x,t) a_2(x,t)]. \tag{82}$$

in which $K = \frac{\sqrt{\omega_1 \omega_3}}{2}$ and δx and δt are the small amounts that x and t respectively are increased by in each computation step. We use ω_3 instead of ω_{pl} as the frequency for warm plasma oscillations provides a more accurate model than that of cold plasma oscillations. Note that this difference leads to a small change in the value of K, a constant, and hence has no effect on the qualitative aspect of our simulation. Assuming that $\delta x = c \cdot \delta t$, these simplify down quite nicely to

$$a_1(x, t+\delta t) = a_1(x - \delta x, t) + \delta t \cdot K a_2(x,t) a_3(x,t), \tag{83}$$

$$a_2(x, t+\delta t) = a_2(x + \delta x, t) - \delta t \cdot K a_1(x,t) a_3(x,t), \tag{84}$$

$$a_3(x, t+\delta t) = a_3(x,t) - \delta t \cdot K a_1(x,t) a_2(x,t). \tag{85}$$

We set the equations slightly differently at the boundaries of the observation window for a_1 and a_2, calculating $(\frac{\partial a}{\partial x})$ using the two values we actually have, and then set the Langmuir wave a_3 to 0 for all x-values in the section where there is no plasma.

Finally, after initial plots, for the purpose of a more interesting output we adjusted the observation time t_{exp} from 16.5×10^{-12} s to 33×10^{-12} s and increased the length of the plasma section l_{pl} from 0.004 m to 0.009 m. Along with these changes, we increased the time duration of the pump pulse t_p from 15×10^{-12} s to 56×10^{-12} s and so increased its total energy E_p proportionally from 0.5 J to 1.86 J to preserve its power. We also increased the total number of computation steps from 1500 to 6000 to increase accuracy of calculation.

Figure 1 shows output of our simulation over time, and figure 2 shows an enlarged version of the final stage of amplification.

Additionally, we found that the maximum value of a_2 obtained during this time period using these parameters was 0.0385, and so because (78) tells us that

$$I_0(x,t) = \left(\frac{a_0(x,t)}{0.855 \times 10^{-}5 \cdot \lambda_0} \right)^2, \tag{86}$$

it can be seen that seed pulse was amplified to a maximum intensity of

$$\left(\frac{0.0385}{0.855 \times 10^{-5} \cdot 4.000 \times 10^{-7}} \right)^2 = 1.2673 \times 10^{20} \text{ W ms}^{-2}.$$

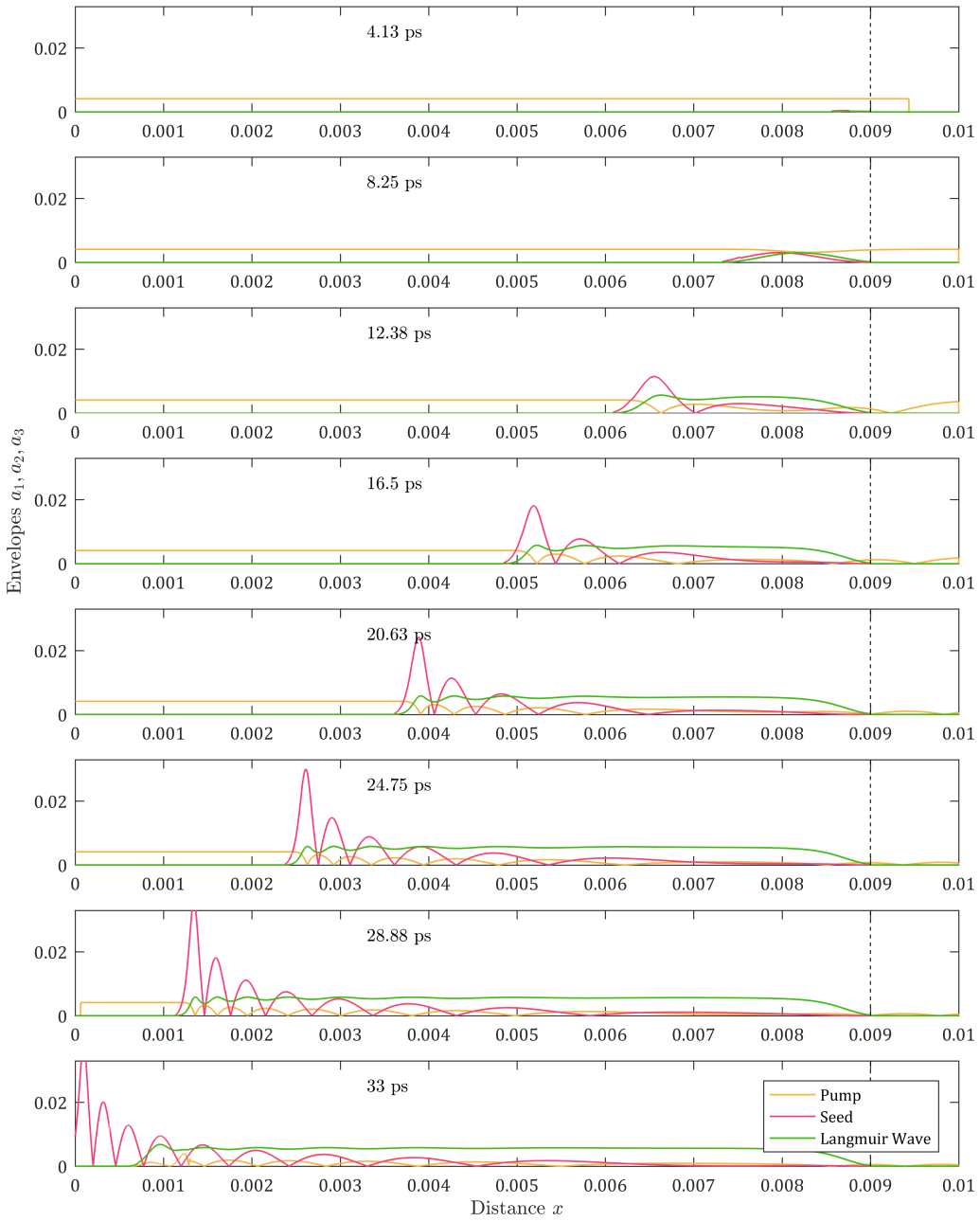


Figure 1: Snapshots of the absolute amplitudes of the envelopes a_1, a_2 and a_3 at various time intervals up to 33 picoseconds.

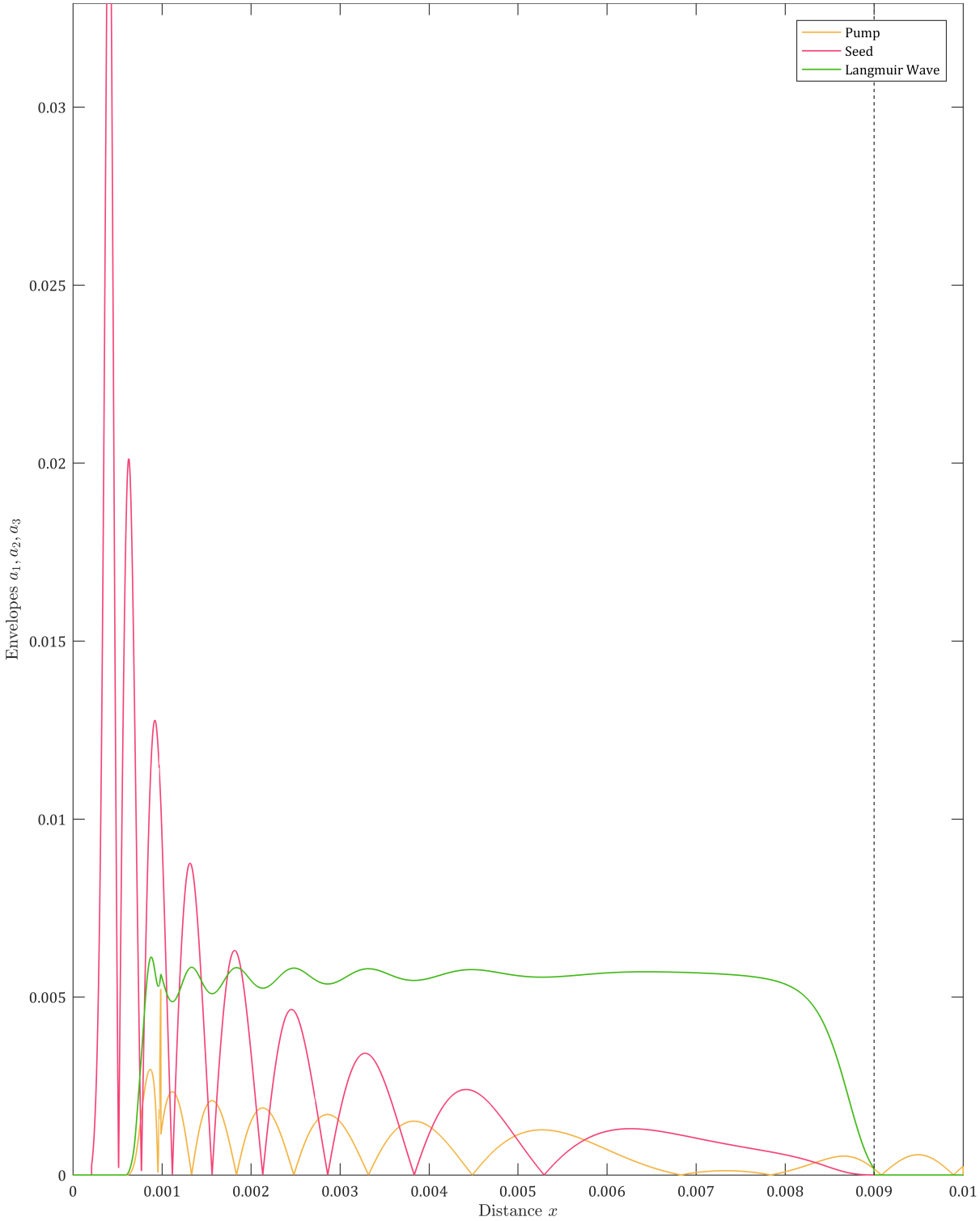


Figure 2: Absolute amplitudes of envelopes of the three waves at $t = 31.97$ ps.

3.8 Explanation of amplification

Other than for small t, we observe local maxima of the Langmuir wave at x positions for which the pump wave has no amplitude. The seed has local maxima (or minima) at x positions slightly to the left of this. As the langmuir wave is roughly a travelling wave, we expect the partial

derivative with respect to position to be 0 when partial derivatives with respect to time are 0. Considering our derived iterative formula for a_3:

$$a_3(x, t + \delta t) = a_3(x, t) + \delta t[-Ka_1(x, t)a_2(x, t)]. \tag{87}$$

We see that for $a_1 = 0$ or $a_2 = 0$, this simplifies to:

$$a_3(x, t + \delta t) = a_3(x, t) \tag{88}$$

I.e, the approximate partial derivative with respect to time is 0. Hence we expect:

$$a_3(x + \delta x, t) \approx a_3(x, t) \tag{89}$$

As a_3 is roughly a travelling wave. For $a_1 = 0$ (pump), we observe local maxima of the langmuir wave, and for $a_2 = 0$ (seed), local minima.

We see that to start with, the electrons in the plasma are 'at rest', i.e, the Langmuir oscillations have no amplitude. The pump then transmits a constant source of photons that act as a source of energy. Initially these don't affect the plasma, since eddies are nullified by the alternating fields created by the light. When they begin to superpose with the seed wave the fields begin to interfere and create large eddies which propagate through the plasma; the Langmuir wave. The oscillating electrons of the plasma then transfer energy into the small packet of photons (the seed), acting as resonance that continually increases the amplitude of their oscillations as they travel in the opposite direction to the pump through the tube.

In the plasma, the Langmuir wave has a varying amplitude, as it describes oscillations of the electrons in the plasma. So the seed wave is amplified by the Langmuir wave as expected. In our model, the pump wave is unaffected by the medium in which it travels.

In the vacuum, the Langmuir wave has amplitude $a_3 = 0$ everywhere, as the Langmuir wave describes oscillations of electrons in the plasma. Hence the partial differential equations (10) and (11) on the question paper simplify to give:

$$(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x})a_1(x, t) = 0 \tag{90}$$

$$(\frac{\partial}{\partial t} - c\frac{\partial}{\partial x})a_2(x, t) = 0 \tag{91}$$

These equations have general solutions $a_1(x, t) = f(x-ct), a_2(x, t) = g(x+ct)$, for any functions f and g of a single variable. Thus in the vacuum, a_1 and a_2 are respectively left and right travelling waveforms, both with speed c. In addition, before the pump and seed waveforms interact there is no Langmuir wave present, so once again we see the pump waveform move rightward and the seed move leftward at a speed c: just as we would expect. Thus no separate code was required to model the system before the waves interacted or to force the waves to move towards each other: this is already encoded by the guiding PDEs.

3.8.1 Creative extensions of the simulation

We start by displaying a graph of smaller dimensions (still at $t = 31.97ps$). This will be continually referenced in comparing our results when certain parameters are changed.

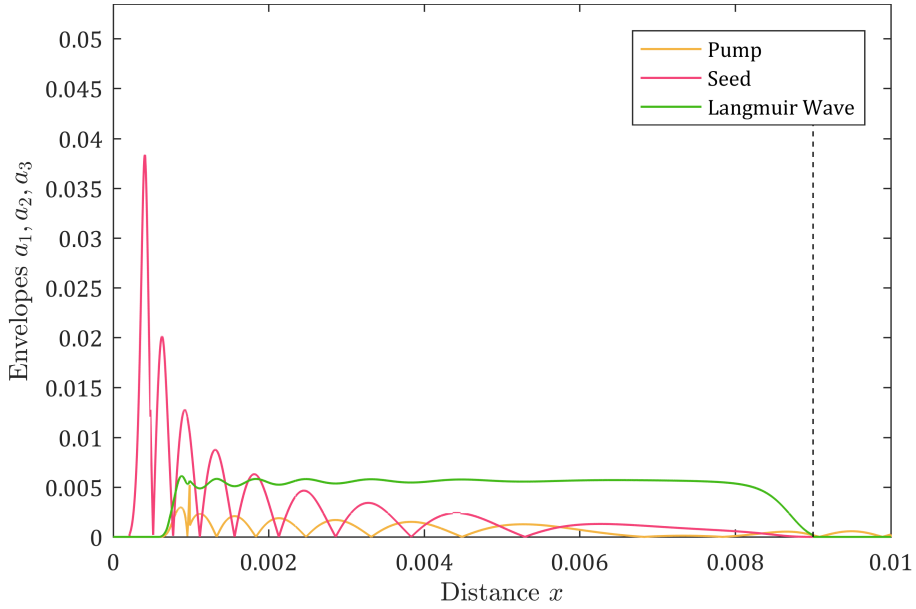
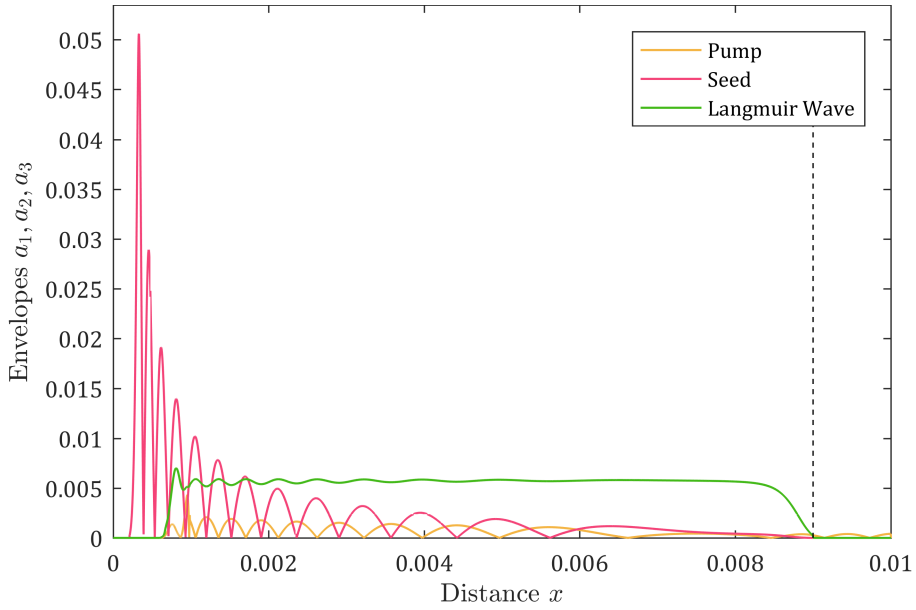


Figure 3: Amplitudes at $t = 31.97$ ps.

Throughout this section the various parameters are altered by seemingly arbitrary amounts to achieve a noticeable effect. Every change is made from the initial conditions shown on the graph above.

We begin our changes by increasing the electron density from 1.1×10^{25} to 4.4×10^{25} electrons per cubic metre; a factor of 4.

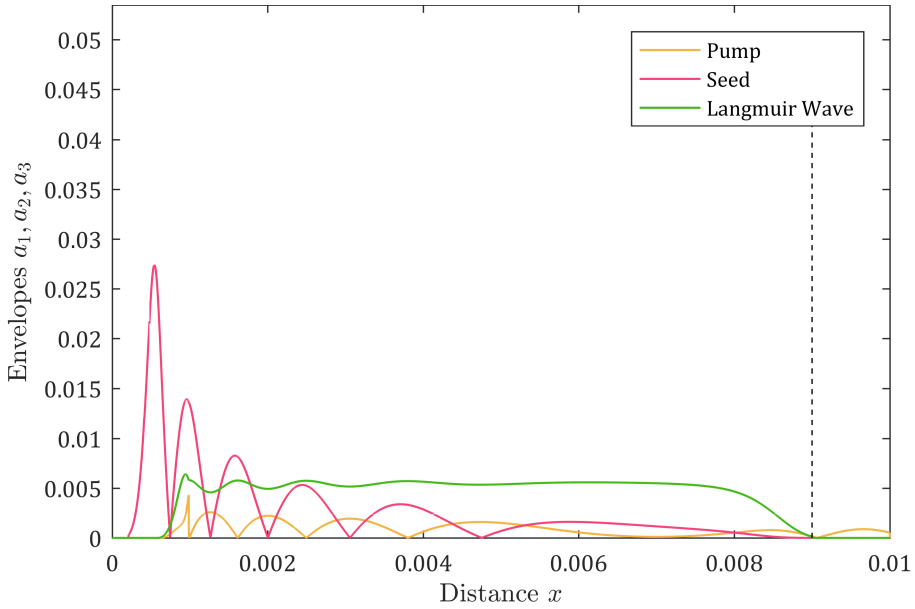


We observe that increasing the electron density increases the frequency of oscillations of the pump, seed and Langmuir wave, as well as the amplitude of the seed wave as it approaches the end of the tube.

Upon increasing electron density, our derived formula for ω_{pl} increases, and hence ω_3 increases. So the electrons in the plasma oscillate with greater frequency (the Langmuir wave has a greater

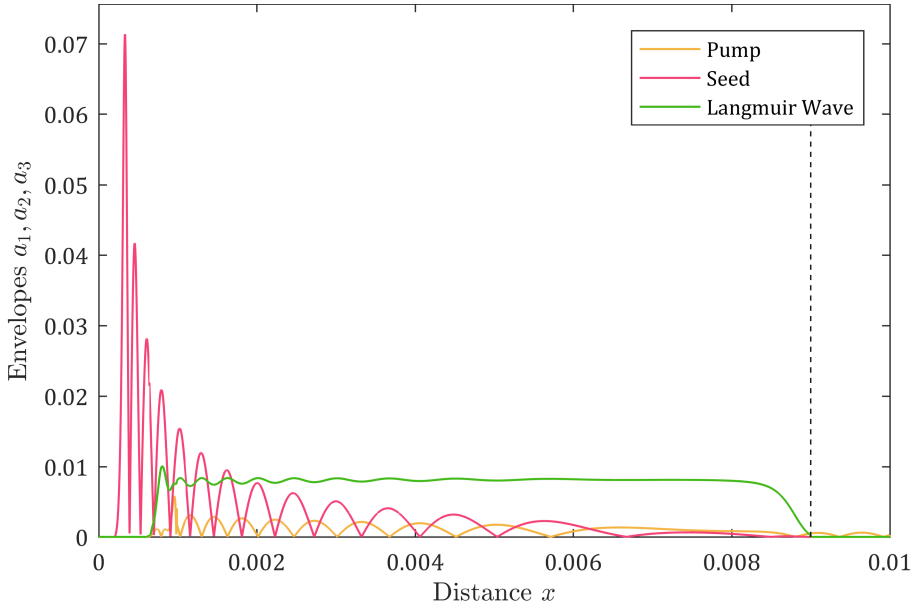
frequency), and as a result of this, the oscillations of our seed wave and pump wave also increase in frequency. Energy from the Langmuir wave can be transferred faster to the seed wave of photons as a result of this increase in electron oscillation frequency, so the seed wave gains a greater amplitude in a given amount of time, as observed.

By contrast, we can change the electron density this time by a factor of $\frac{1}{10}$ from the original density to 0.11×10^{25} electrons per cubic metre:

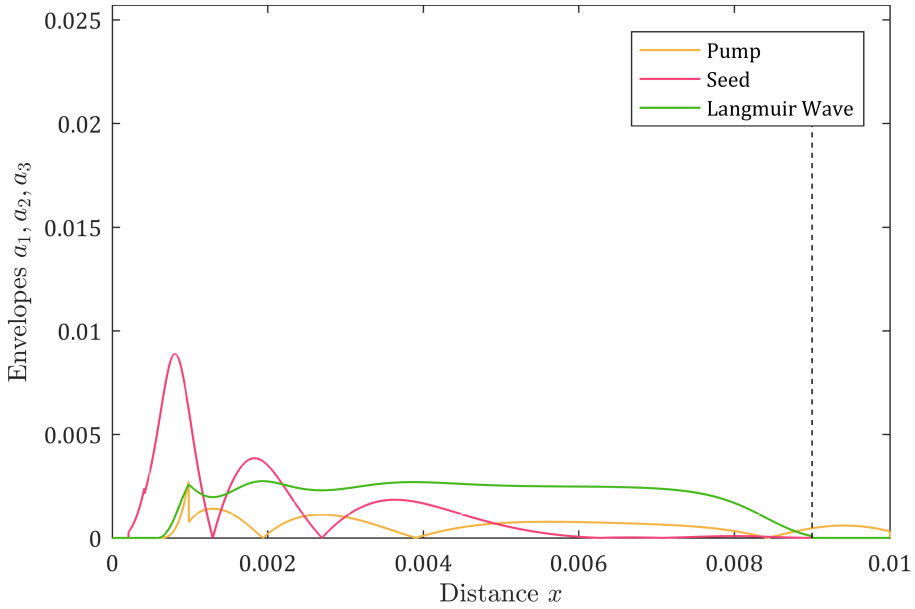


We observe that decreasing the electron density decreases the frequency of oscillations of the pump, seed and Langmuir wave, as well as the amplitude of the seed wave as it approaches the end of the tube. A similar plausibility argument to the one above may be applied.s

Now we move to varying the total energies of the laser pulses. We start by doubling the energy of the pump pulse from 1.86 J to 3.72 J:

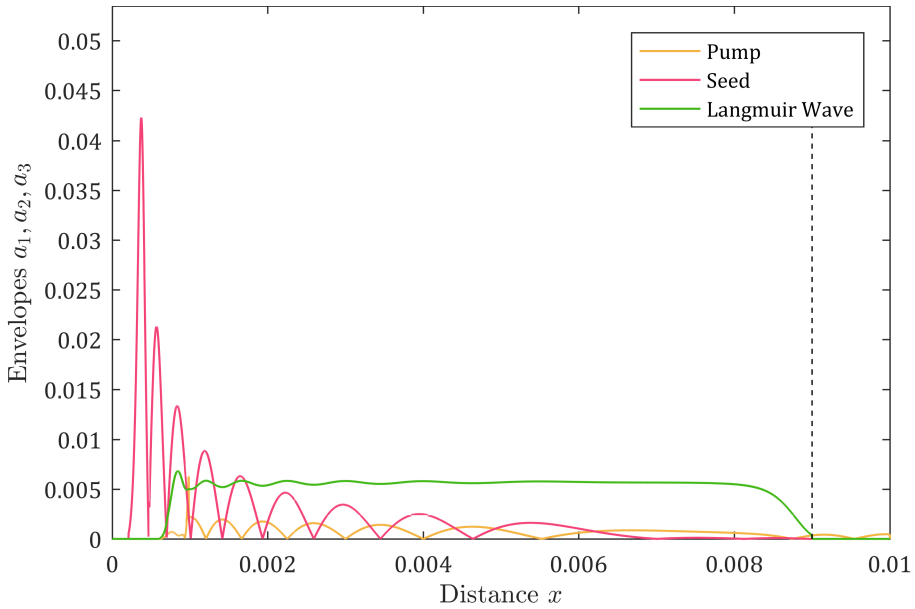


And now halving it to 0.93 J:



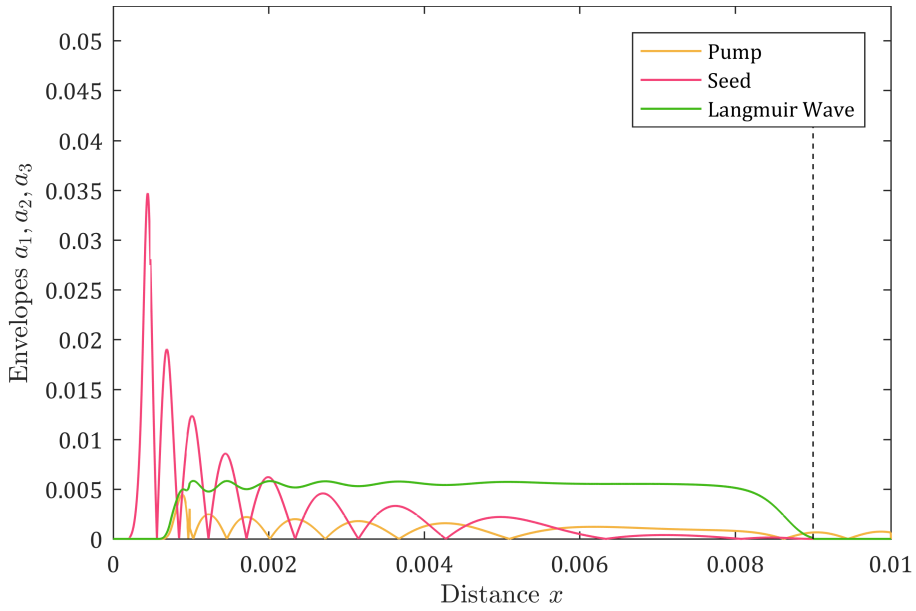
We observe that increasing the energy of the pump pulse increases the amplitudes of the seed wave, and decreasing the energy decreases the amplitude. This is the result we expect, as increasing the energy of the pump pulse increases the energy that is transferred to the seed wave, and the amplitude of the seed wave is proportional to the square root of the energy supplied to the seed wave. In addition, we observe that increasing the energy of the pump pulse dramatically increases the frequency of oscillations of the pump, seed and Langmuir waves.

Now we proceed to do the same thing to the seed pulse, though with different factors. Multiplying the seed energy by 4 from 70×10^{-6} J to 280×10^{-6} produces the following graph:



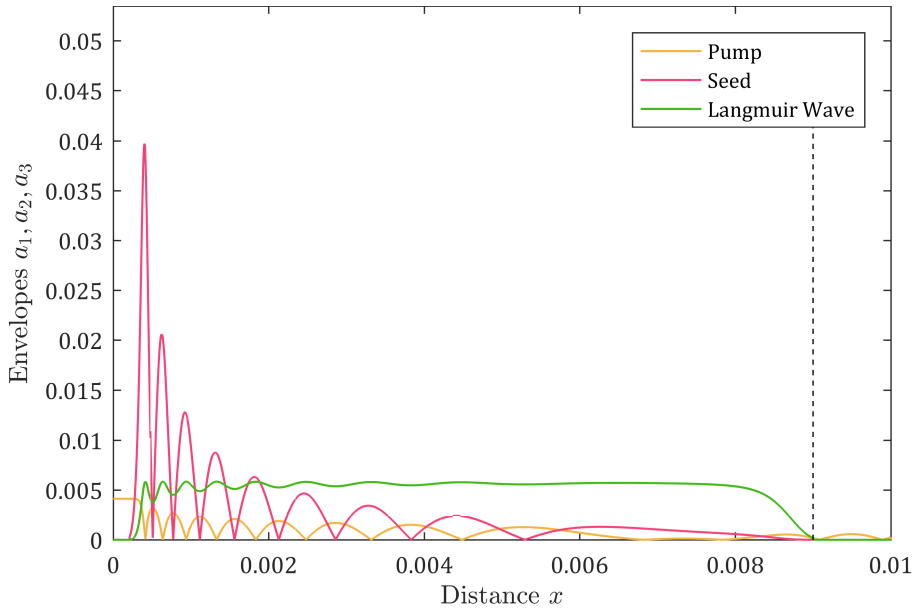
We observe that increasing the energy of the seed pulse has little effect on the system as a whole, as the energy transferred to it from the Langmuir wave is far greater than its own initial energy.

And now dividing it by 4 to 17.5×10^{-6}:



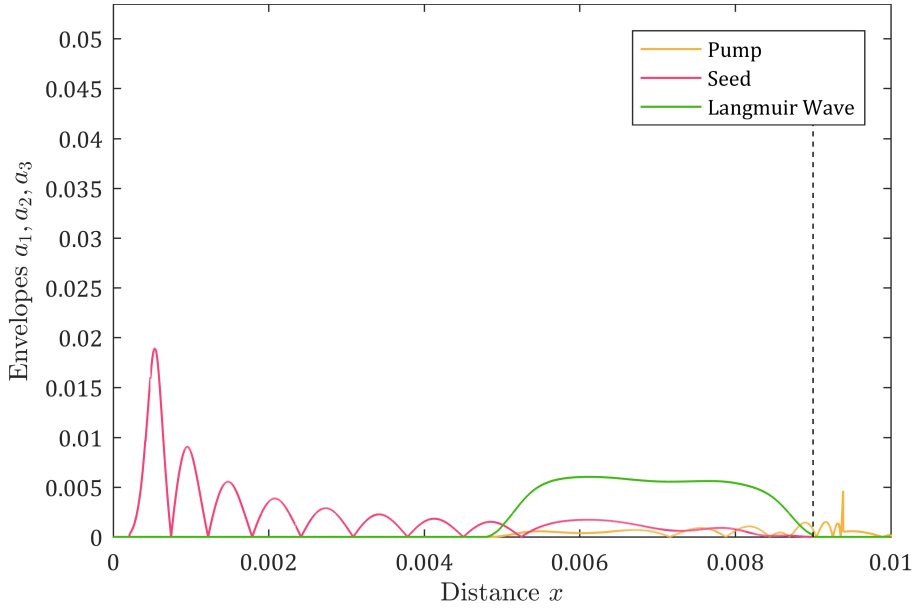
Similarly, decreasing the energy of the seed pulse has little effect on the system as a whole.

Next, we alter the time durations of the two pulses. Starting with the pump pulse, we double the duration of the pump pulse (from 56×10^{-12} s to 112×10^{-12} s so that it does not end within our time window. To do this we also increase its total energy proportionally, from 1.86 J to 3.72 J, so as to maintain a constant power.



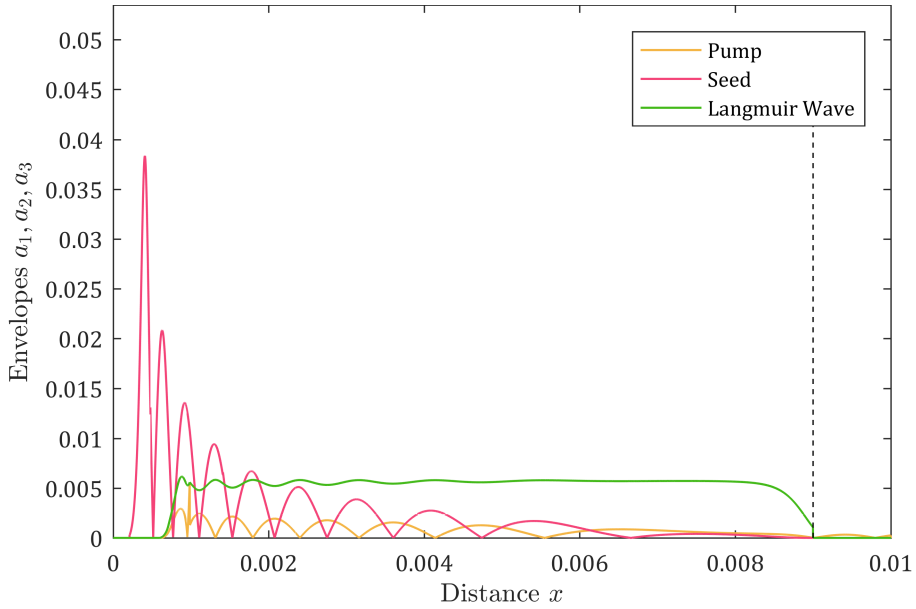
We observe that doubling the duration of the pump pulse has little effect on the system as a whole as our original duration of the pump pulse was sufficient in providing energy to the system for the required amount of time.

Now we halve its duration to 28×10^{-12} s, also halving the energy to 0.93 J:



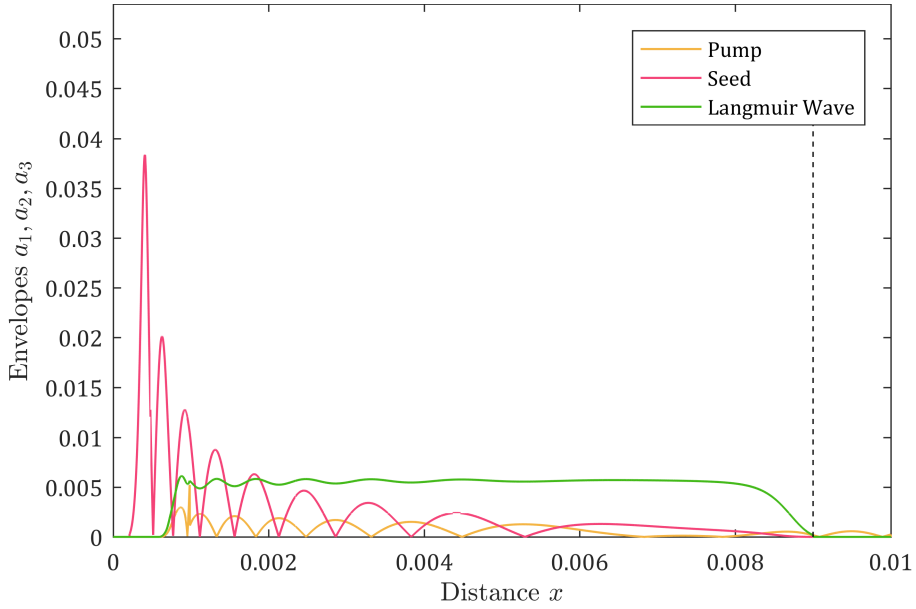
Conversely, halving the duration of the pump pulse has a dramatic effect on our simulation. Not enough energy is supplied to the seed and Langmuir waves and the seed is amplified to a fraction of what it would if supplied with sufficient energy from the pump.

Finally, we alter the duration of the seed pulse. A multiplication by 8 from 500×10^{-15} s to 4000×10^{-15} s along with a corresponding alteration of its energy from 70×10^{-6} J to 560×10^{-6} J produces this graph:



We see that this has little effect on our system as a whole. The seed pulse is there to provide a small initial amount of energy to the system, interacting with the pump. The interaction begins the Langmuir wave, which is what amplifies the seed. The Langmuir wave amplifies the seed, but this amplification is related to the energy of the pump wave. Thus we wouldn't expect the energy of the seed to have much effect on the output.

And now we divide its duration by 8 to 62.5×10^{-15} (and changing the energy to 8.75×10^{-6}):



We observe that decreasing the duration of the seed pulse has little effect on our system as a whole, as a result of a similar line of reasoning to the above argument.

End of Submission.

Bibliography

[1] MIT. *Maxwell's Equations and Electromagnetic Waves: Chapter 13*. URL: http://web.mit.edu/8.02t/www/materials/StudyGuide/guide13.pdf.

Chapter 22

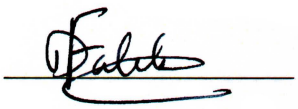
Physics Unlimited Explorer Competition 2017: Part 1

A year later in November 2017, six of us in Year 13 did this competition (a rebranding of the Princeton competition) which was again *huge* fun. It took 2 and a bit weeks and this tidal heating part was a collaboration between many of us. I was very happy with the paper we produced in the end.

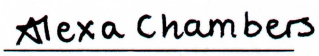
Physics Unlimited Explorer Competition 2017

Tidal Heating Section
Submission of Answers

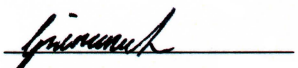
Team: *One Diraction*

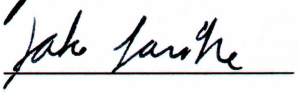
Damon Falck 

Thalia Seale 

Alexa Chambers 

Leon Galli 

Gianmarco Luppi 

Jake Saville 

HIGHGATE

Highgate School

London, United Kingdom

November 2017

This page is intentionally left blank.

Modelling tidal dissipation in Io

Gianmarco Luppi, Damon Falck, Leon Galli, Thalia Seale, Alexa Chambers, and Jake Saville
Highgate School, London
(Dated: November 29, 2017)

Abstract. Jupiter's moon Io is of particular interest due to its unusual heat source. By far the most geologically active body in the solar system, it has a surface heat flux hundreds of times higher than expected from radiogenic heating. Orbital resonance with the other Galilean satellites causes a forced eccentricity in Io's orbit, and as a result of the continuously varying distance from Jupiter the extent of Io's tidal deformation changes throughout each orbital period. The internal friction caused by this warping generates a large amount of heat. In this paper we first investigate the exact nature and time-dependence of Io's tidal deformation by assuming a fluid model, and then using principles of harmonic oscillation we calculate the approximate heat produced as a function of tidal phase angle, subsequently making an estimate for the surface temperature of Io. Our model will also include a qualitative discussion of Io's interior and of implications for the future of the Jovian system.

I. INTRODUCTION

Io is the innermost of the four largest Jovian moons that are known as the Galilean satellites. It is the most geologically active body in the solar system, with hundreds of volcanoes on its surface and a vast magma ocean beneath its thin crust. Indeed, Io has an observed average surface heat flux of between 1 and $2\,\mathrm{J\,m^{-2}}$ [1, 2], compared to Earth's $0.06\,\mathrm{J\,m^{-2}}$. Unlike most natural satellites in the solar system, whose internal heating comes mostly from radiogenic decay, Io's primary source of power comes from the changing tidal forces that act on the body during its elliptical orbit around Jupiter.

Although Io's orbit has a very low free eccentricity of 0.00001 [3], the orbital resonance of Io, Europa and Ganymede (cf. section VI) causes a forced eccentricity of $e = 0.0041$. As a result, Io's distance from Jupiter varies from 4.200×10^8 m at perijove to 4.234×10^8 m at apojove, and so the magnitude of the tidal force exerted on Io by Jupiter oscillates, varying Io's tidal deformation.

The viscoelastic interior of Io generates a resistive force to the tidal movement, and this frictional force dissipates energy through heat as the tides cycle, a process which would circularise its orbit were it not for its orbital resonance with Europa and Ganymede. It is this heat which is responsible for the enormous energy flux observed.

We will start by deriving exactly how Io deforms in Jupiter's gravitational field. After making several approximations, this will allow us to reach an estimate for the general order of magnitude of the heat produced.

Io's tidal heating has been studied in various ways in existing literature; perhaps most notable is the work done by Segatz *et al.* [4] to model Io's interior and the distribution of tidal dissipation rate across the surface, focusing on multilayer Maxwell rheology models. Moore [5] did extensive work on convection currents within Io responsible for allowing the heat produced to flow to the surface. Yoder [6] also did research into the effects of tidal heating on the formation of resonance locks with the other Galilean satellites. These former two papers focus closely on the specifics of Io's interior structure and its effects on the exact nature and distribution of tidal heating. The latter paper focuses on context in the Jovian system. While we will touch on both of these topics, we are primarily interested in a phenomenological model that can be used to predict the approximate nature of Io's behaviour given only common empirical values. This will generate a model that can more easily be transferred to other planetary systems, and as such our approach is unique.

II. DEFORMATION OF IO IN THE GRAVITATIONAL FIELD OF JUPITER

A. Finding the deformed shape of Io

We want to start by considering all of the forces acting on any point on the surface of Io. Let L be the distance between the centres of Io and Jupiter, and let l be the distance between the centre of Io and the barycentre of orbit. Clearly

$$\begin{aligned} l &= \frac{M_J}{M_I + M_J} L \\ &= 0.999953\, L \end{aligned}$$

where M_J is the mass of Jupiter and M_I is the mass of Io, and so we can make the approximation $l \approx L$. That is, we can take Io's orbital barycentre to be the centre of Jupiter.

The spherical polar coordinate system that we'll use is shown in figs. 1 and 2; taking Io's centre as the origin, r is the distance to an arbitrary point on Io's surface, θ is the polar angle (latitude) and ϕ is the azimuth angle (longitude).

Consider the non-inertial frame of reference in which we are at a point directly above the centre of Jupiter in line with the axis of rotation and rotating with angular speed ω such that the Jupiter-Io system appears stationary (ignoring the bodies' varying separation). Note that because of how Io is tidally locked,

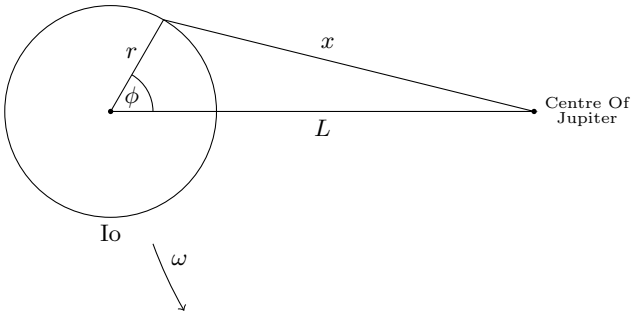


FIG. 1. A cross-sectional view in the plane perpendicular to the axis of rotation and through the centres of Io and Jupiter.

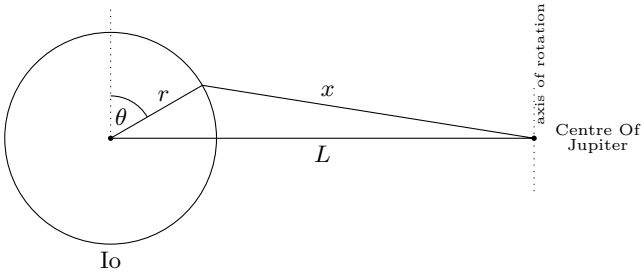


FIG. 2. A cross-sectional view in the plane containing the axis of rotation and through the centres of Io and Jupiter.

in this frame of reference the surface of Io will not appear to move either.

At any point (r, θ, ϕ) on Io's surface, let x be the distance from that point to the centre of Jupiter. The potential energy of a point mass m at coordinates (r, θ, ϕ) consists of three parts:

1. Gravitational potential energy from Io, given by
$$V_I = -G\frac{mM_I}{r}. \tag{1}$$
2. Gravitational potential energy from Jupiter, given by
$$V_J = -G\frac{mM_J}{x}. \tag{2}$$
3. Centrifugal potential energy arising from the fictitious centrifugal force. Since potential energy is given by
$$V'(x) = -F(x),$$
$$\implies V(x) = \int -F(x) \cdot \mathrm{d}x,$$
we can calculate the centrifugal potential energy to be
$$V_C = \int_0^{x_p} -\omega^2 s \, \mathrm{d}s = -\frac{1}{2}m\omega^2 x_p^2$$
where ω is the angular velocity of orbit and x_p is the component of x in the plane of rotation. (Clearly the centrifugal force only depends upon this quantity.) Where h is the perpendicular height above the Io-Jupiter axis,
$$x_p = \sqrt{x^2 - h^2}$$
$$\approx x \qquad \text{as } h \ll x.$$
and so the centrifugal potential energy is given by
$$V_C = -\frac{1}{2}m\omega^2 x^2. \tag{3}$$

Therefore, by combining eqs. (1) to (3), we see that the total potential energy V of a mass m is given by
$$\frac{V}{m} = -\frac{1}{2}\omega^2 x^2 - G\frac{M_J}{x} - G\frac{M_I}{r}. \tag{4}$$

Now let α be the angle between (r, θ, ϕ) and $(L, \frac{\pi}{2}, 0)$ a.k.a. Jupiter. By looking at fig. 3, we see that
$$\cos\alpha = \frac{r\sin\theta\cos\phi}{r}$$
$$= \sin\theta\cos\phi.$$

Now, by the cosine rule, the distance x to Jupiter is

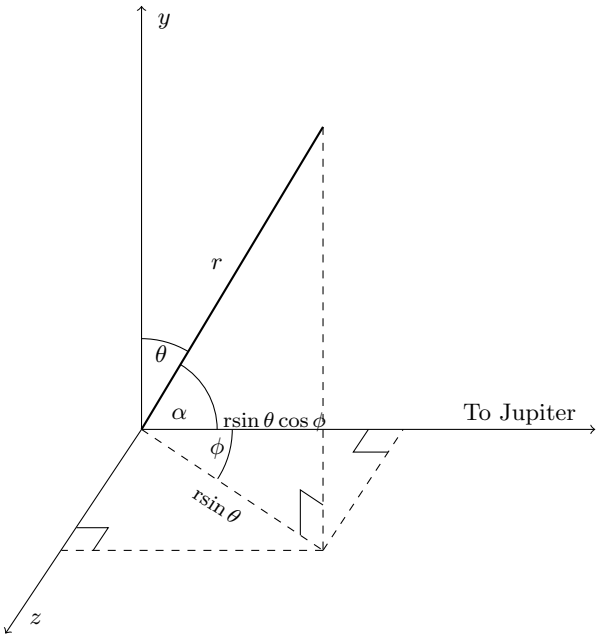


FIG. 3. Our new angle α between (r, θ, ϕ) and Jupiter is shown.

given by
$$x^2 = L^2 + r^2 - 2rL\cos\alpha$$
$$\implies x = \sqrt{L^2 + r^2 - 2rL\cos\alpha}$$
$$= L\sqrt{1 + \left(\frac{r}{L}\right)^2 - 2\frac{r}{L}\cos\alpha}.$$

Using the Taylor expansion

$$\frac{1}{\sqrt{1+a^2-2ab}} = 1+ab+\frac{1}{2}(3b^2-1)a^2+\frac{1}{2}b(5b^2-3)a^3+\cdots$$

where b is kept constant, we come to the approximation valid for $a \ll 1$ of

$$\frac{1}{\sqrt{1+a^2-2ab}} \approx 1+ab+\frac{1}{2}a^2(3b^2-1)$$

and so since $r \ll R$, we can approximate

$$\frac{1}{x} = \frac{1}{L}\left(1+\frac{r}{L}\cos\alpha+\frac{1}{2}\frac{r^2}{L^2}(3\cos^2\alpha-1)\right).$$

Substituting this into eq. (4) gives us

$$\begin{aligned}\frac{V}{m} &= -\frac{1}{2}\omega^2(L^2+r^2-2rL\cos\alpha)-\frac{GM_J}{L}\left(1+\frac{r}{L}\cos\alpha+\frac{1}{2}\frac{r^2}{L^2}(3\cos^2\alpha-1)\right)-G\frac{M_I}{r}\\ &= -\frac{1}{2}\omega^2r^2-G\frac{M_I}{r}-\frac{GM_Jr^2}{2L^3}(3\cos^2\alpha-1)+\omega^2rL\cos\alpha-GM_J\frac{r}{L^2}\cos\alpha-\frac{1}{2}\omega^2L^2-\frac{GM_J}{L}. \qquad (5)\end{aligned}$$

Since by considering circular motion on Io as a body we know

$$\omega^2 = \frac{GM_J}{L^3}, \qquad (6)$$

we have

$$\begin{aligned}\omega^2rL\cos\alpha-GM_J\frac{r}{L^2}\cos\alpha &= r\cos\alpha\left[L\cdot\frac{GM_J}{L^3}-\frac{GM_J}{L^2}\right]\\ &= 0,\end{aligned}$$

and since ω, L, G and M_J are all constants, $-\frac{1}{2}\omega^2L^2-\frac{GM_J}{L}$ is a constant. Hence, eq. (5) becomes

$$\frac{V}{m} = -\frac{1}{2}\omega^2r^2-G\frac{M_I}{r}-G\frac{M_Jr^2}{2L^3}(3\cos^2\alpha-1)+C \quad (7)$$

for some constant C.

Now in order to quantitatively model the shape that Io forms we will first assume it to be a fluid such that the mass will adjust hydrostatically to form an equipotential surface. The effect on the final deformation should be mainly a matter of amplitude as we assume the shape made by both a fluid body and solid body will be of the same topology. If we so wish, we can later multiply the fluid height by a scaling factor in order to reach an approximate value for the tide at any point on the surface of Io.

Where R is the mean radius of Io, the fluid tidal height is therefore $h_f = r - R$. Given that $h_f \ll R$,

$$\begin{aligned}\frac{1}{r} &= \frac{1}{R+h_f}\\ &= \frac{1}{R}\cdot\frac{1}{1+\frac{h_f}{R}}\\ &\approx \frac{1}{R}\left(1-\frac{h_f}{R}\right)\\ &= \frac{1}{R}-\frac{h_f}{R^2} \qquad (8)\end{aligned}$$

and

$$\begin{aligned}r^2 &= R^2+2Rh_f+h_f^2\\ &\approx R^2+2rh_f. \qquad (9)\end{aligned}$$

Combining eqs. (6), (8) and (9) into eq. (7),

$$\begin{aligned}\frac{V}{m} = &-\frac{G(M_I+M_J)R}{L^3}h_f+\frac{GM_I}{R^2}h_f\\ &-\frac{GM_Jr^2}{2L^3}(3\cos^2\alpha-1)+C_2\end{aligned}$$

for some other constant C_2. Since the ratio of the first term to the second terms is

$$\frac{M_I+M_J}{M_I}\frac{R^3}{L^3} \approx 10^{-3},$$

we consider its contribution negligible and so we can now say

$$\frac{V}{m} = \frac{GM_I}{R^2}h_f-\frac{GM_Jr^2}{2L^3}(3\cos^2\alpha-1)+C_2.$$

Because we are modelling the surface as equipotential, V must be a constant at any point a distance r from Io's centre. This means that the first two terms must compensate each other and so

$$\begin{aligned}\frac{GM_I}{R^2}h_f &= \frac{GM_Jr^2}{2L^3}(3\cos^2\alpha-1)\\ \implies h_f &= \frac{M_J}{M_I}\cdot\frac{R^2r^2}{2L^3}(3\cos^2\alpha-1).\end{aligned}$$

Again, given $R \gg h_f$, we can now approximate $r^2 \approx R^2$. Therefore, our fluid tidal height is

$$h_f = \frac{M_J}{M_I}\frac{R^4}{2L^3}(3\cos^2\alpha-1). \qquad (10)$$

This situation is rotationally symmetric around the Jupiter-Io axis as one might expect. A polar plot of Io's tidal deformation is shown in fig. 4.

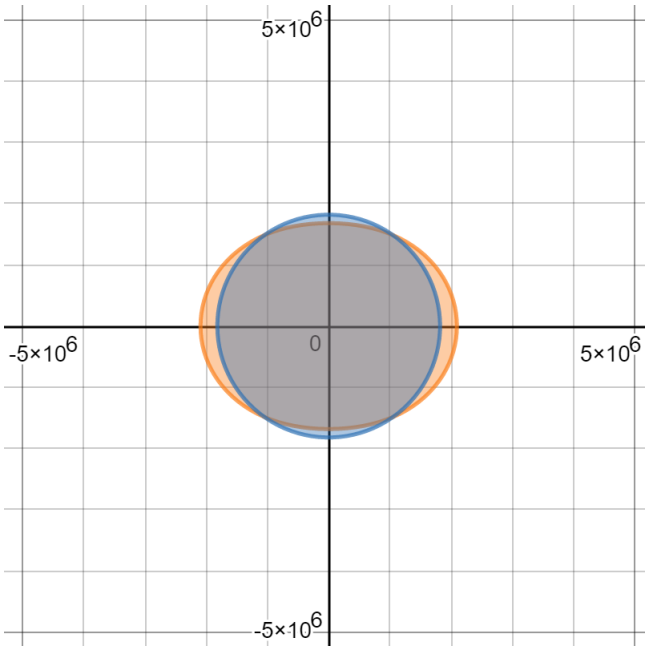


FIG. 4. A polar plot of the deformed Io (given by eq. (10)) overlaid with a spherical Io. The deformation is exaggerated.

B. Estimating Io's maximum tidal amplitude

Now that we know tidal warping as a function of angle (eq. (10)) we will find the maximum tidal amplitude of Io. The maximum tidal amplitude Δh_f is the difference in the height of Io's tidal bulge between perijove and apojove.

We're interested in the tidal bulge i.e. the tidal height at the point on Io closest to Jupiter, so $\alpha = 0 \implies \cos\alpha = 1$. Therefore eq. (10) gives

$$h_f = \frac{M_J R^4}{M_I L^3}.$$

The maximum tidal amplitude (assuming h_f at $\alpha = 0$ is greatest at perijove and least at apojove) is therefore

$$\Delta h_f = \frac{M_J R^4}{M_I L_P^3} - \frac{M_J R^4}{M_I L_A^3} = \frac{M_J R^4}{M_I}\left(\frac{1}{L_P^3} - \frac{1}{L_A^3}\right)$$

where L_P is the Io-Jupiter separation at perijove and L_A is the Io-Jupiter separation at apojove. This leads to a value of

$$\Delta h_f = 75.54\,\mathrm{m}.$$

Note that this value is calculated assuming Io is a fluid, and so is likely an overestimate. The observed height of Io's tidal bulge is about 50 m [4] and so our prediction is quite accurate.

III. QUALITATIVE MODEL OF THE INTERIOR OF IO

There is very little that we know for certain about the interior of Io. The most useful measurements come from the flybys of various spacecraft such as Pioneer 10 and Voyager. They were able to accurately measure Io's mass and size leading to the first estimate of Io's density, now accepted to be $3.53 \times 10^3\,\mathrm{kg\,m^{-3}}$, which is the highest of any moon in the solar system. Later on, during Galileo's 1999 flyby, the onboard magnetometer recorded measurements of the magnetic field along its trajectory giving us valuable insight into what the core of Io consists of.

We begin with a very simple model like that of Peale *et al.* [7] in order to find an approximation for the size of Io's core. We might hypothesise that the large majority of Io's mass comes from an abundance of iron and silicate rock, which are both extremely common in most terrestrial objects such as the rocky planets and the Moon. In this case we can estimate the volume ratio of the two as follows.

Let ρ_I and ρ_S be the densities of iron and silicate rock respectively. If there is a volume V_I of iron and a volume V_S of silicate rock in Io, then

$$\frac{\rho_I V_I + \rho_S V_S}{V_{\mathrm{Io}}} = \rho_{\mathrm{Io}}$$

and

$$V_I + V_S = V_{\mathrm{Io}}$$

where ρ_{Io} and V_{Io} are the density and volume of Io respectively. Therefore,

$$\rho_I \times V_I + \rho_S(V_{\mathrm{Io}} - V_I) = \rho_{\mathrm{Io}} V_{\mathrm{Io}}$$
$$\implies V_I = \frac{\rho_{\mathrm{Io}} - \rho_S}{\rho_I - 1} V_{\mathrm{Io}}.$$

If we assume an almost entirely pure iron core of density $7.9 \times 10^3\,\mathrm{kg\,m^{-3}}$ and take the density of silicate rock as $3.0 \times 10^3\,\mathrm{kg\,m^{-3}}$, then the volume of iron is

$$V_I = 1.5 \times 10^{18}\,\mathrm{m^3}$$

which gives an iron core of radius 710 km. This value is within the accepted range of 600–900 km depending on the concentration of sulphur compounds in the core.

Now in order to improve our model, we look at the plausibility of Io having a subsurface ocean of free iron much like the model by Schubert *et al.* [8]. The first clue that there may lie a molten layer beneath the surface is the active volcanoes that most likely form above molten pockets of rock. However until the investigation done by Khurana *et al.* [9], there was a real lack of scientific evidence to support this. Khurana looked at the warping of Jupiter's rotating magnetic field and in turn noted that Io must contain a global conducting layer. A field decrease of nearly 40 percent of the background Jovian field at closest approach to Io was recorded by the Galileo spacecraft. Kivelson *et al.* [10] showed that plasma sources alone would not warp the field to such an extent but instead an amount of free iron must be present for the induced source. This would act as a conducting layer allowing an induced

field to occur throughout Io and in turn have the affect of weakening Jupiter's field nearby. They went on to estimate the lower bound for the thickness of this layer to be around 50 km. The presence of such a fluid layer would also increase the accuracy of a fluid approximation to Io's deformation under tidal forces.

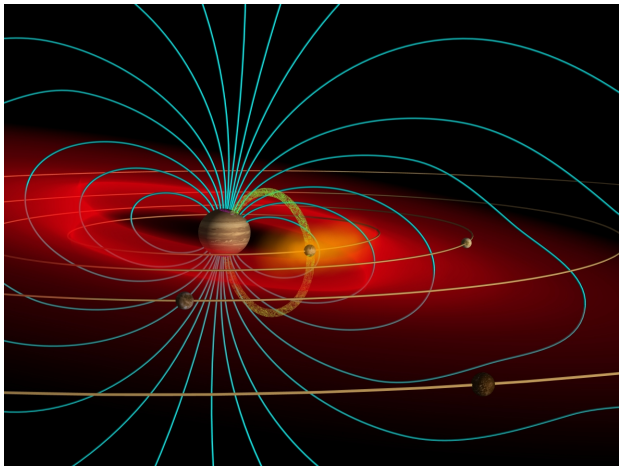


FIG. 5. The magnetosphere of Jupiter. Image from Khurana *et al.* [9].

There is still an element of ambiguity however in that a magnetised core could have the same effect as a global subsurface layer. Evidence both for and against this hypothesis is lacking however and in favour of the more studied model, we will assume the Schubert model.

Io's lithosphere is better understood on the other hand. It consists of a combination of silicate rocks and alkali-rich minerals such as feldspars and nepheline. At the base of this, we begin to see the melting of the rocks to form the magma.

As for the mantle, in the region of 700 - 1750 km from the centre, we know temperatures in the asthenosphere must exceed 1400 K in order to support rock melting whilst the rest of the mantle is solid and silicate rich.

In conclusion, we see that Io is in fact a lot more like terrestrial bodies than other moons in the size of its core and the presence of the molten asthenosphere.

IV. TIDAL DEFORMATION AS A FUNCTION OF TIME

There will be two main effects as Io completes each orbit. Firstly, the varying distance to Jupiter will cause the height of the tidal bulge to change. Secondly, because Io is tidally locked with Jupiter but its orbit is eccentric, the bulge will change position on Io's surface over time, also causing warping. We will ignore the latter effect for the purposes of this paper as its contribution to heating may be considered negligible.

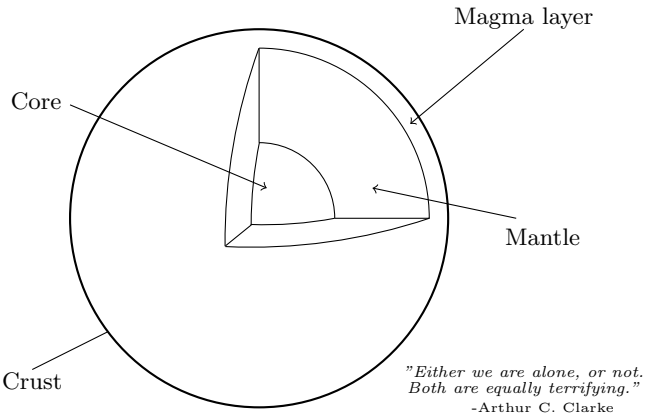


FIG. 6. A diagram showing the inferred structure of Io's interior.

A. Io-Jupiter separation as a function of time

The distance of a body in elliptical orbit to the centre of the body it is orbiting is

$$L(\theta) = a \cdot \frac{1-e^2}{1+e\cos\theta}, \tag{11}$$

where a is the semi-major axis, e the eccentricity of the orbit and θ is the true anomaly. The relation between the true anomaly and the eccentric anomaly E is

$$\cos\theta = \frac{\cos E - e}{1 - e\cos E}.$$

Substituting this into eq. (11), we get

$$\begin{aligned} L(E) &= a \cdot \frac{(1-e^2)}{1+e\dfrac{\cos E - e}{1-e\cos E}} \\ &= a \cdot \frac{(1-e^2)(1-e\cos E)}{(1-e\cos E)+e(\cos E - e)} \\ &= a \cdot \frac{(1-e^2)(1-e\cos E)}{1-e^2} \\ \implies L(E) &= a(1-e\cos E). \end{aligned} \tag{12}$$

Kepler's equation gives

$$M = E - e\sin E$$

where M is the mean anomaly, and so since Io's eccentricity e is small we may approximate $M \approx E$. The mean anomaly is given by

$$M = \omega t$$

if Io is at perijove when $t = 0$. Thus, eq. (12) gives the Io-Jupiter distance as

$$L(t) = a(1 - e\cos\omega t).$$

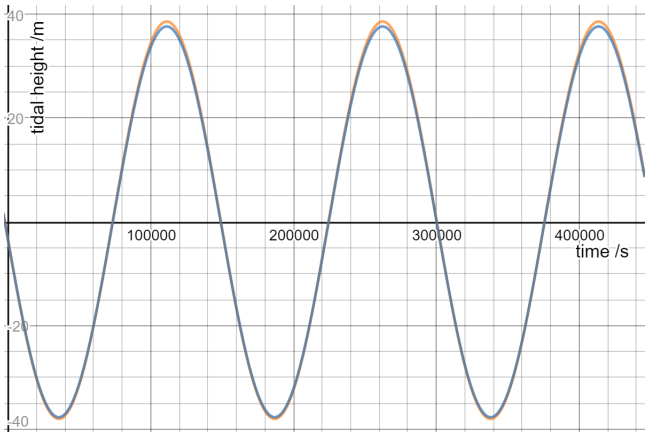


FIG. 7. A plot of fluid tidal height h_f as a function of time at $\alpha = 0$.

B. Tidal height as a function of time

Now combining this expression with eq. (10), the fluid tidal height of a point on Io's surface as a function of time is

$$h_f(t) = \frac{M_J}{M_I}\frac{R^4}{2L^3(t)}(3\cos^2\alpha - 1) = \frac{M_J}{M_I}\frac{R^4(3\cos^2\alpha - 1)}{2a^3(1 - e\cos\omega t)^3}.$$

A plot of this function over time is shown in fig. 7.

C. Making simple harmonic approximations

From fig. 7 it is evident that Io's tidal warping behaves almost exactly as a sinusoidal wave. This shows us that we can approximate Io's behaviour as that of a simple harmonic oscillator, a class of problems very well-studied. We will make this approximation in the next section.

V. HEAT GENERATED BY TIDAL DEFORMATION

We have modelled Io's deformation as a function of time, and have shown that its behaviour is approximately simple harmonic. The power lost to heat as a result of tidal forces at any point in Io is simply the dot product of the velocity of that point and the resistive force at that point.

We will therefore model both the velocity and resistive forces as vector sinusoidal, plus a constant, as a function of both position within Io and of time. Where $\overrightarrow{F_F}$ is the resistive force per unit mass and $\vec{v}$ is the velocity, we will show that the power dissipated at any moment in time by the whole of Io is therefore approximately

$$P_T = \iiint_V \vec{v}\cdot\overrightarrow{F_F}\rho\,\mathrm{d}V$$

where ρ is the density of Io. We will then go on to temporally average this over a single time period, to find the average tidal power dissipated by Io.

A. Finding the vector simple harmonic form of velocity

For the purposes of this section we will define a new angle ε as shown in figs. 8 and 9 and will use the spherical coordinate system indexed by r, α and ε.

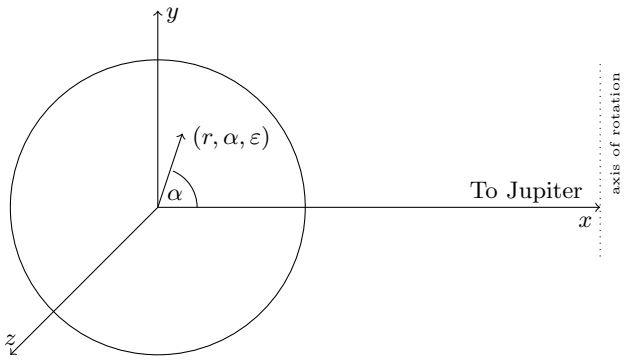


FIG. 8. A side view of our coordinate system.

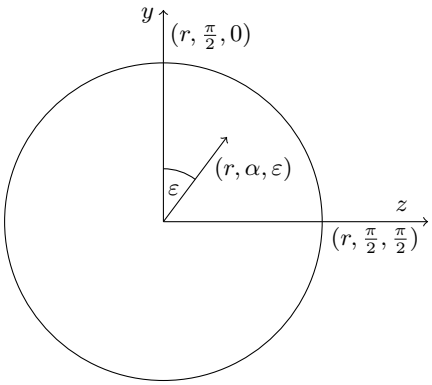


FIG. 9. A head-on view of our coordinate system.

We know from section IV that the tidal height h approximates a sinusoidal wave. The amplitude of this wave is just half the difference between tidal height at perijove and tidal height at apojove. Since eq. (10) gives the tidal height as

$$h = \frac{M_J}{M_I}\cdot\frac{R^4}{2L^3}\left(3\cos^2\alpha - 1\right),$$

the tidal amplitude is therefore

$$\implies h_0(\alpha,\varepsilon) = \frac{M_J}{M_I}\cdot\frac{R^4}{4}\left(3\cos^2\alpha - 1\right)\left(\frac{1}{L_\mathrm{P}^3} - \frac{1}{L_\mathrm{A}^3}\right)$$

where L_P and L_A are the distances between the centres of Io and Jupiter at perijove and apojove respectively. Therefore, defining $t = 0$ to be at apojove, the

vector tidal height is

$$\begin{aligned}\vec{h}(\alpha,\varepsilon,t) &= (-h_0(\alpha,\varepsilon)\cos(\omega t)+\overline{h})\begin{pmatrix}\cos\alpha\\ \sin\alpha\cos\varepsilon\\ \sin\alpha\sin\varepsilon\end{pmatrix}\\ &= (-h_0(\alpha,\varepsilon)\cos(\omega t)+\overline{h})\hat{c} \qquad (13)\end{aligned}$$

where $\hat{c} \coloneqq \begin{pmatrix}\cos\alpha\\ \sin\alpha\cos\varepsilon\\ \sin\alpha\sin\varepsilon\end{pmatrix}$ is the unit vector from any point (r,α,ε) radially outwards from the origin and $\overline{h}$ is the mean tidal height. Here our vectors take Io's centre as the origin, with the x axis through the centre of Jupiter, the y axis parallel to the axis of rotation and the z axis perpendicular to these two.

Now differentiating eq. (13), we find the velocity of a point on the surface of Io to be

$$\dot{\vec{h}}(\alpha,\varepsilon,t) = \omega h_0(\alpha,\varepsilon)\sin(\omega t)\hat{c}.$$

It is clear, however, that velocity will scale linearly with radius — that is, if a point (r,α,ε) has velocity $\dot{\vec{h}}$ then a point $(\frac{r}{2},\alpha,\varepsilon)$ will have velocity $\frac{\dot{\vec{h}}}{2}$ and so on. This principle is shown in fig. 10.

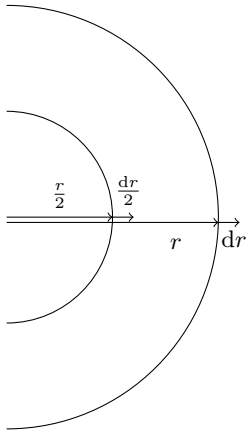


FIG. 10. If when a tide occurs the surface is extended by an amount $\mathrm{d}r$ then a point halfway out from the centre of Io will be extended by an amount $\frac{\mathrm{d}r}{2}$.

Therefore, the velocity at time t of any point inside Io (r,α,ε) is

$$\begin{aligned}\vec{v}(r,\alpha,\varepsilon,t) &= \frac{r}{R}\dot{\vec{h}}(\alpha,\varepsilon,t)\\ &= \frac{\omega r h_0(\alpha,\varepsilon)}{R}\cdot\sin(\omega t)\hat{c}\\ &= v_0\cdot\sin(\omega t)\hat{c} \qquad (14)\end{aligned}$$

where $v_0 = \frac{\omega r h_0(\alpha,\varepsilon)}{R}$ is the magnitude of the velocity amplitude. Indeed, this will be maximised at $(r,\alpha,\varepsilon) = (R,0,\varepsilon)$ and so

$$v_{0_{\max}} \approx 0.001\,56\,\mathrm{m\,s^{-1}}$$

which appears to be a reasonable value.

B. Finding the vector simple harmonic form of force

Using our previous formula in eq. (7) for the potential energy V on the surface of Io and modifying it for any point (r,α,ε) in the interior of Io, we replace the M_I in the potential due to Io's gravitational field with $M_I\frac{r^3}{R^3}$ to account for the lessening of Io's gravitational potential in its interior due to less of its mass having an impact. This is a consequence of the shell theorem [11]. The potential energy becomes

$$\begin{aligned}\frac{V(r,\alpha,\varepsilon,L)}{m} = &-\frac{1}{2}\omega^2 r^2 - \frac{GM_I\left(\frac{r}{R}\right)^3}{r}\\ &- \frac{GM_J\,r^2}{2L^3}\cdot(3\cos^2\alpha-1)+C.\end{aligned}$$

Since each concentric shell to the surface will be approximately equipotential, we know that the tidal force will be acting radially away from (or towards) Io's centre.

Therefore differentiating with respect to r, the total tidal force acting on a point mass m at (r,α,ε) will be

$$\begin{aligned}&\overrightarrow{F_T} = -\vec{\nabla}V\\ \implies &\frac{\overrightarrow{F_T}}{m} = \left(\omega^2 r + \frac{2GM_I\,r}{R^3} + \frac{GM_J\,r}{L^3}(3\cos^2\alpha-1)\right)\hat{c}.\end{aligned}$$

Plotting a graph of tidal force against angle α shows that it can be approximated as simple harmonic in time, but off by a constant.

We will from now take all forces we talk about as being on a unit mass. The force can be expressed in harmonic form as

$$\overrightarrow{F_T} = (F_{T_{\mathrm{amp}}}\cos(\omega t)+\overline{F_T})\hat{c}$$

where the mean tidal force is

$$\overline{F_T} = \frac{GM_J r}{2\overline{L}^3}(3\cos^2\alpha-1).$$

and the amplitude of the tidal force $F_{T_{\mathrm{amp}}}$ is

$$F_{T_{\mathrm{amp}}}(r,\alpha,\varepsilon) = \frac{GM_J r}{2}(3\cos^2\alpha-1)\left(\frac{1}{L_P^3}-\frac{1}{L_A^3}\right). \qquad (15)$$

The displacement of any point in Io is approximately sinusoidal in time. This means that the acceleration at any point will likewise be sinusoidal. Since for any point mass at arbitrary position acceleration is proportional to the force applied by Newton's second law of motion (see Newton [11]), we know that the resultant force must be likewise sinusoidal. Therefore, in some time convention t, the resultant force on a unit mass is

$$\overrightarrow{F_R} = \left(\frac{\omega^2 r h_0(\alpha,\varepsilon)}{R}\cdot\cos(\omega t)\right)\hat{c}.$$

Since the only forces acting on any point (r, α, ε) are the tidal forces and the material frictional tension forces, this allows us to calculate the frictional force, as the motion of the object and the tidal force are both sinusoidal with some phase difference ζ. The frictional force is therefore

$$\begin{aligned}\overrightarrow{F_F} &= \overrightarrow{F_R} - \overrightarrow{F_T} \\ \implies \overrightarrow{F_F} m &= \left(\frac{\omega^2 r h_0(\alpha,\varepsilon)}{R} \cdot \cos(\omega t)\right)\hat{c} \\ &\quad - \left(F_{T_{amp}} \cos(\omega t + \zeta) + \overline{F_T}\right)\hat{c} \\ &= \left(F_0 \cos(\omega t + \delta) - \overline{F_T}\right)\hat{c}, \qquad (16)\end{aligned}$$

where

$$F_0 = \sqrt{\omega^2 v_0^2 + F_{T_{\text{amp}}}^2} \qquad (17)$$

is the amplitude of oscillation of the frictional force, and $\delta = \frac{\pi}{2} + \zeta$ is the phase difference between the velocity and force oscillations.

C. Calculating power dissipated as a function of phase difference

We now have harmonic expressions for force (on a mass m) and velocity in eqs. (14) and (16). To calculate the average power dissipated by tidal forces, we have to integrate over all points (r, α, ε) in Io and then take the temporal average over one time period T:

$$\langle P_T \rangle = \frac{1}{T}\int_0^T \iiint_V \vec{v} \cdot \overrightarrow{F_F}\rho \,\mathrm{d}V \,\mathrm{d}t.$$

Here $\mathrm{d}V$ is a small volume of Io, so that where ρ is the density of Io, $\rho\,\mathrm{d}V$ is the small mass on which the force acts. The volume element in our spherical polar coordinate system is $\mathrm{d}V = r^2 \sin\alpha\,\mathrm{d}r\,\mathrm{d}\varepsilon\,\mathrm{d}\alpha$ and so this integral becomes

$$\langle P_T \rangle = \frac{1}{T}\int_0^T \int_0^\pi \int_0^{2\pi} \int_0^R \vec{v}\cdot\overrightarrow{F_F}\rho r^2 \sin\alpha\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha\ \mathrm{d}t.$$

We now substitute in our expressions for frictional force per unit mass $\overrightarrow{F_F}$ and velocity $\vec{v}$ from eqs. (14) and (16) and rearrange the integral:

$$\begin{aligned}\langle P_T \rangle &= \frac{1}{T}\int_0^T \int_0^\pi \int_0^{2\pi} \int_0^R v_0 \cos(\omega t)\left[F_0\cos(\omega t+\delta) - \overline{F_T}\right]\rho r^2 \sin\alpha\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha\ \mathrm{d}t \\ &= \frac{1}{T}\int_0^T \cos(\omega t)\cos(\omega t+\delta)\int_0^\pi\int_0^{2\pi}\int_0^R \rho r^2 \sin\alpha v_0 F_0\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha\ \mathrm{d}t \\ &\qquad - \frac{1}{T}\int_0^T \cos(\omega t)\int_0^\pi\int_0^{2\pi}\int_0^R v_0\overline{F_T}\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha\ \mathrm{d}t \\ &= \frac{1}{T}\int_0^T \cos(\omega t)\cos(\omega t+\delta)\,\mathrm{d}t\int_0^\pi\int_0^{2\pi}\int_0^R \rho r^2 \sin\alpha v_0 F_0\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha\,. \qquad (18)\end{aligned}$$

The second half of the penultimate line above is discarded because the temporal average of $\cos(\omega t)$ over one time period is zero. Looking at eq. (18), we now define the temporal part as

$$I_t \coloneqq \frac{1}{T}\int_0^T \cos(\omega t)\cos(\omega t+\delta)\,\mathrm{d}t$$

and the spatial part as

$$I_s \coloneqq \int_0^\pi\int_0^{2\pi}\int_0^R \rho r^2 \sin\alpha v_0 F_0\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha$$

so that the average power is

$$\langle P_T \rangle = I_t I_s.$$

We start by evaluating I_t:

$$\begin{aligned}I_t &= \frac{1}{T}\int_0^T \cos(\omega t)\cos(\omega t+\delta)\,\mathrm{d}t \\ &= \frac{1}{T}\int_0^T \cos(\omega t)(\cos(\omega t)\cos\delta - \sin(\omega t)\sin\delta)\,\mathrm{d}t \\ &= \frac{1}{T}\int_0^T \cos^2(\omega t)\cos\delta - \frac{1}{2}\sin(2\omega t)\sin\delta\,\mathrm{d}t \\ &= \frac{1}{T}\left[\frac{\cos\delta}{2\omega}(\omega t + \sin(\omega t)\cos(\omega t)) + \frac{\sin\delta}{4\omega}\cos(2\omega t)\right]_0^T \\ &= \frac{1}{T}\left(\frac{T\cos\delta}{2} + \frac{\sin\delta}{4\omega} - \frac{\sin\delta}{4\omega}\right) \\ &= \frac{\cos\delta}{2}.\end{aligned}$$

Now we will evaluate the spatial part I_s. We previously derived that the mean velocity is

$$v_0 = \frac{\omega r h_0}{R} = \frac{\omega r}{R}\frac{M_J}{M_I}\cdot\frac{R^4}{4}\left(3\cos^2\alpha - 1\right)\left(\frac{1}{L_{\mathrm{P}}^3} - \frac{1}{L_{\mathrm{A}}^3}\right)$$

and so we define

$$k_v \coloneqq \frac{\omega}{R}\frac{M_J}{M_I}\cdot\frac{R^4}{4}\left(\frac{1}{L_{\mathrm{P}}^3} - \frac{1}{L_{\mathrm{A}}^3}\right)$$

such that

$$v_0 = k_v r(3\cos\alpha - 1)^2.$$

Similarly, the force amplitude is given by eq. (15) as

$$F_{T_{\mathrm{amp}}} = \frac{GM_J r}{2}(3\cos^2\alpha - 1)\left(\frac{1}{L_P^3} - \frac{1}{L_A^3}\right)$$

so we define

$$k_F \coloneqq \frac{GM_J}{2}\left(\frac{1}{L_P^3} - \frac{1}{L_A^3}\right)$$

such that

$$\implies F_{T_{\mathrm{amp}}} = k_F r(3\cos^2\alpha - 1).$$

Therefore, our frictional force amplitude F_0 is given by eq. (17) as

$$\begin{aligned} F_0 &= \sqrt{\omega^2 v_0^2 + F_{T_{\mathrm{amp}}}^2} \\ &= \sqrt{\omega^2 k_v^2 r^2(3\cos^2\alpha - 1)^2 + k_F^2 r^2(3\cos^2\alpha - 1)^2} \\ &= r(3\cos^2\alpha - 1)\sqrt{\omega^2 k_v^2 + k_F^2}. \end{aligned}$$

This means that our spatial integral takes the form

$$\begin{aligned} I_s &= \int_0^{\pi}\int_0^{2\pi}\int_0^{R} \rho r^2 \sin\alpha v_0 F_0 \,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha \\ &= \int_0^{\pi}\int_0^{2\pi}\int_0^{R} \rho r^2 \sin\alpha k_v r(3\cos\alpha - 1)^2 r(3\cos^2\alpha - 1)\sqrt{\omega^2 k_v^2 + k_F^2}\,\mathrm{d}r\ \mathrm{d}\varepsilon\ \mathrm{d}\alpha \\ &= 2\pi\int_0^{\pi}\int_0^{R} r^4 \sin\alpha(3\cos^2\alpha - 1)^2\left(\rho k_v\sqrt{\omega^2 k_v^2 + k_F^2}\right)\mathrm{d}r\ \mathrm{d}\alpha. \end{aligned}$$

Defining

$$k_I \coloneqq \rho k_v\sqrt{\omega^2 k_v^2 + k_F^2},$$

this integral becomes

$$\begin{aligned} I_s &= 2\pi\int_0^{\pi}\int_0^{R} r^4 \sin\alpha(3\cos^2\alpha - 1)^2 k_I \,\mathrm{d}r\ \mathrm{d}\alpha \\ &= 2\pi k_I\frac{R^5}{5}\int_0^{\pi}\sin\alpha(3\cos^2\alpha - 1)^2\,\mathrm{d}\alpha \\ &= 2\pi k_I\frac{R^5}{5}\left[-\frac{9\cos^5\alpha}{5} + 2\cos^3\alpha - \cos\alpha\right]_0^{\pi} \\ &= 2\pi k_I\frac{8R^5}{25}. \end{aligned}$$

This leaves us the following formula for the power loss in terms of the phase difference:

$$\begin{aligned}
\langle P_T\rangle &= \pi k_I \frac{8R^5}{25}\cos\delta \\
&= \pi\frac{8R^5}{25}\cos\delta\cdot\rho\frac{M_J}{M_I}\cdot\frac{\omega R^3}{4}\left(\frac{1}{L_{\mathrm{P}}^3}-\frac{1}{L_{\mathrm{A}}^3}\right)\sqrt{w^2\left(\frac{M_J}{M_I}\cdot\frac{\omega R^3}{4}\left(\frac{1}{L_{\mathrm{P}}^3}-\frac{1}{L_{\mathrm{A}}^3}\right)\right)^2+\left(\frac{GM_J}{2}\left(\frac{1}{L_P^3}-\frac{1}{L_A^3}\right)\right)^2} \\
&= \rho\frac{\pi\omega R^8}{25}\frac{M_J^2}{M_I}\left(\frac{1}{L_P^3}-\frac{1}{L_A^3}\right)^2\sqrt{\frac{\omega^4R^6}{4M_I^2}+G^2}\cdot\cos\delta \\
&= \rho\frac{\pi\omega R^8}{25}\frac{M_J^2}{M_I^2}\left(\frac{1}{L_P^3}-\frac{1}{L_A^3}\right)^2\sqrt{\frac{\omega^4R^6}{4}+(GM_I)^2}\cdot\cos\delta \\
&= \rho\frac{\pi\omega R^9}{25}\frac{M_J^2}{M_I^2}\left(\frac{1}{L_P^3}-\frac{1}{L_A^3}\right)^2\sqrt{\left(\frac{\omega^2R^2}{2}\right)^2+\left(\frac{GM_I}{R}\right)^2}\cdot\cos\delta. \qquad (19)
\end{aligned}$$

Therefore, the maximum value that could possibly be obtained for the tidal thermal power would be

$$\begin{aligned}
\langle P_T\rangle_{max} &= \pi k_I\frac{8R^5}{25} \\
&\approx 6.1876\cdot10^{14}\mathrm{W}.
\end{aligned}$$

If Io were a perfectly elastic body, the resistive force would be exactly $\frac{\pi}{2}$ out of phase since the acceleration and tidal forces would be exactly in phase. This would give us, by our formula, a power loss of zero, which fits with our conventional understanding of perfectly elastic materials.

D. Estimating average power dissipated

The phase angle δ between the tidal force and the velocity is dependent on the viscoelasticity of the interior of Io. For perfectly elastic oscillators, the tidal force and the resulting deformation will in phase (so $\delta=\frac{\pi}{2}$), and for perfectly viscous oscillators, the tidal force and the resulting deformation will be $\frac{\pi}{2}$ out of phase (so $\delta=\pi$).

We can realistically expect the actual phase difference to be somewhere in between. Phase difference between force and response is inversely proportional to the tidal quality factor Q [12] which is a notoriously difficult value to calculate. The tidal quality factor is generally smaller for smaller bodies, and so we will make an approximate order-of-magnitude estimate of $Q\approx10$. This leads to a phase lag of $\delta=\frac{\pi}{2}-\frac{1}{10}=84.2°$. The resulting power dissipated is given by eq. (19) as

$$\begin{aligned}
\langle P_T\rangle &= \pi k_I\frac{8R^5}{25}\cos(84.2°) \\
&= 6.3\times10^{13}\,\mathrm{W}. \qquad (20)
\end{aligned}$$

Dividing this by the surface area of Io leads to a predicted surface heat flux of $1.5\,\mathrm{W\,m^{-2}}$ which is in good agreement with the observed value of 1 to $2\,\mathrm{W\,m^{-2}}$ [1].

E. Surface temperature of Io

The power absorbed by Io from the Sun is

$$P_{\mathrm{Sun}}=\frac{L_{\mathrm{Sun}}R^2(1-A)}{4D^2}=1.94\times10^{14}\,\mathrm{W}$$

where $L_{\mathrm{Sun}}=3.83\times10^{26}\,\mathrm{W}$ is the luminosity of the Sun and $A=0.63$ is Io's albedo. Combining this with tidal heating in eq. (20), the total power absorbed by Io is

$$P_{\mathrm{in}}=2.57\times10^{14}\,\mathrm{W}.$$

It is worth noting that this total power input is if anything an underestimate as we have neither taken into account the tidal heating caused by the movement of Io's tidal bulge across its surface (cf. section IV) nor the effects of other heat sources such as radiogenic heating.

Io radiates as a blackbody and so the Stefan-Boltzmann law gives the power radiated as

$$P_{\mathrm{out}}=4\pi R^2\sigma T^4$$

where σ is the Stefan-Boltzmann constant and T is the effective temperature of Io. Setting $P_{\mathrm{in}}=P_{\mathrm{out}}$ yields an effective surface temperature of

$$T=\sqrt[4]{\frac{2.57\times10^{14}\,\mathrm{W}}{4\pi R^2\sigma}}=103\,\mathrm{K}.$$

This is in good agreement with the observed mean surface temperature of 110 K. An estimate not taking tidal heating into account yields a value of 96.1 K.

Therefore tidal heating does not actually cause a significant rise in the effective temperature of Io, but because of the large amount of internally generated heat there is extreme volcanic activity on the moon.

VI. ORBITAL RESONANCE OF THE GALILEAN SATELLITES

We know that Io loses about $6.3 \times 10^{13}\,\mathrm{W} \cdot (365.25 \cdot 24 \cdot 3600\,\mathrm{s}) = 2.0 \times 10^{21}\,\mathrm{J}$ to heat through the tides every Earth year. While tidal heating would usually result in the decay of the moon's angular momentum and orbital decay as energy from a moon's elliptic orbit and spin is converted into tidal heating, Io's orbit is both already synchronous and continuously eccentric.

By the principle of conservation of energy, we therefore know that this much energy must be put into the Ionian system every year by an external agent. Given that Io is in a synchronous orbit, the energy input must come from the eccentricity of Io's orbit, and therefore the forces that keep Io's orbit elliptical are the sources of the tidal heating of Io. We also know that Io's orbital period is exactly twice that of Europa and exactly four times that of Ganymede. Io, as a result, experiences an oscillating force from these two moons, which keeps Io's orbit eccentric.

It is therefore likely that the orbits of Io, Europa and Ganymede will circularise over time as the energy from their elliptic orbit is transferred to Io through gravitation and then dissipated through tidal heating.

VII. CONCLUSION

In this paper, we have used classical methods to find the deformation of Io under Jupiter. Using only simple experimental measurements, we found an estimate of the surface temperature of Io due to tidal heating.

This is particularly significant, as these are measurements that we can take for any moon system with a stable elliptic orbit. Therefore, this method can be extended for any such system to find theoretical surface temperature if tidal heating were the main heating effect (apart from the system's central star). If the recorded mean temperature were significantly lower than that theorised by tidal heating, we could conclude that the material of the planet was significantly stiffer or softer than that which would optimise heat generated. On the other hand, if the recorded mean temperature were significantly higher, we would know that other factors such as radiogenic heating are dissipating heat energy, and so we could know to study the body in more detail.

Therefore, the modelling techniques detailed in this paper are not only accurate in the context of Ionian tidal heating, but could be extended in studying moons in extrasolar systems to determine the material properties of those moons. The natural continuation of this technique would be to find some explicit formula for the all-important phase difference δ in terms of better known properties of materials and structures, such as Young's modulus and shear modulus. Given such a link, we would be able to analyse with very few measurements the materials with which any moon was made.

This paper is clearly a step towards understanding not only the qualitative reasoning behind tidal analysis, but also creating a solid, quantitative approach that can allow others to more accurately study the properties of planetary bodies.

[1] G. J. Veeder, D. L. Matson, T. V. Johnson, D. L. Blaney, and J. D. Goguen, Journal of Geophysical Research: Planets **99**, 17095 (1994).

[2] J. R. Spencer, J. A. Rathbun, L. D. Travis, L. K. Tamppari, L. Barnard, T. Z. Martin, and A. S. McEwen, Science **288**, 1198 (2000).

[3] W. de Sitter, Annalen van de Sterrewacht te Leiden **16**, B1 (1928).

[4] M. Segatz, T. Spohn, M. Ross, and G. Schubert, Icarus **75**, 187 (1988).

[5] W. Moore, Journal of Geophysical Research: Planets **108** (2003).

[6] C. F. Yoder, Nature **279**, 767 (1979).

[7] S. J. Peale, P. Cassen, and R. T. Reynolds, Science **203**, 892 (1979).

[8] G. Schubert, J. Anderson, T. Spohn, and W. McKinnon, Jupiter: The planet, satellites and magnetosphere **1**, 281 (2004).

[9] K. K. Khurana, X. Jia, M. G. Kivelson, F. Nimmo, G. Schubert, and C. T. Russell, Science **332**, 1186 (2011).

[10] M. Kivelson, K. Khurana, R. Walker, C. Russell, J. Linker, D. Southwood, and C. Polanskey, Science **273** (1996).

[11] I. Newton, Principia (1687).

[12] M. Efroimsky and V. Lainey, Journal of Geophysical Research: Planets **112** (2007).

Chapter 23

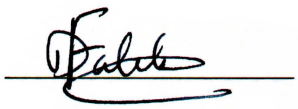
Physics Unlimited Explorer Competition 2017: Part 2

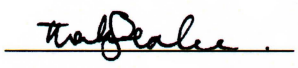
The second part was apparently standard relativistic electrodynamics, though of course it didn't feel standard to us at the time. This was the result of a week's work from me and a friend, and the typesetting was done in large part by some very helpful Year 12s.

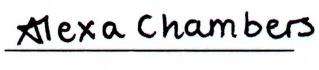
Physics Unlimited Explorer Competition 2017

Relativistic Electrodynamics Section
Submission of Answers

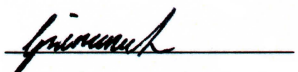
Team: *One Diraction*

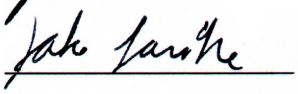
Damon Falck 

Thalia Seale 

Alexa Chambers 

Leon Galli 

Gianmarco Luppi 

Jake Saville 

HIGHGATE

Highgate School

London, United Kingdom

November 2017

This page is intentionally left blank.

PUEC 2017 Relativistic Electrodynamics

Damon Falck & Thalia Seale

November 2017

"Einstein, my upset stomach hates your theory — it almost hates you yourself! How am I to provide for my students? What am I to answer to the philosophers?"

— Paul Ehrenfest, November 1919

Contents of Submission

2.1 Basics of Special Relativity

2.1.1 Conceptual Basics of Special Relativity

Problem: The Barn Paradox

We start by considering the classic relativistic 'paradox' of a runner carrying a pole slightly longer than a barn trying to fit the pole inside the barn while a farmer closes the barn doors instantaneously.

From the reference frame of the farmer, the runner is travelling at a high speed. The phenomenon of length contraction means that the farmer observes the length of the pole as being shorter than the length of the barn, so it is possible for him to close both barn doors while the entire pole fits inside the barn. Let's assume the farmer does this.

From the reference frame of the runner, however, it is the barn that is moving at a high speed, and so experiences length contraction. Therefore, from this frame it seems impossible that the farmer was able to close both doors simultaneously around the pole, as the pole is most definitely longer than the barn. Thus, the 'paradox' arises.

However, the two doors close (and open) simultaneously in the reference frame of the farmer and thus the two doors close and open in succession in the reference frame of the runner; in special relativity, simultaneity is relative and two events that are simultaneous in one reference frame are not in another. Therefore from the point of view of the runner the far door closes and opens first, followed by the near door, and so the paradox is resolved.

2.1.4 The Spacetime Interval

For reference, when performing a Lorentz boost with velocity v in the x-direction, the transformation is given as follows:

$$\begin{pmatrix} c\,\mathrm{d}t' \\ \mathrm{d}x' \\ \mathrm{d}y' \\ \mathrm{d}z' \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c\,\mathrm{d}t \\ \mathrm{d}x \\ \mathrm{d}y \\ \mathrm{d}z \end{pmatrix}$$

$$\implies c\,\mathrm{d}t' = \gamma(c\,\mathrm{d}t - \beta\,\mathrm{d}x), \tag{1}$$

$$\mathrm{d}x' = \gamma(\mathrm{d}x - \beta c\,\mathrm{d}t), \tag{2}$$

$$\mathrm{d}y' = \mathrm{d}y, \tag{3}$$

$$\mathrm{d}z' = \mathrm{d}z \tag{4}$$

where $\gamma = \frac{1}{\sqrt{1-\beta^2}}$ and $\beta = \frac{v}{c}$.

Problem: Invariance of the Spacetime Interval

We are told that writing $\mathrm{d}x_\mu\,\mathrm{d}x^\mu$ implies a summation from $\mu = 0$ to 3; this can be defined as a dot product.

Expanding the given product according to the rules of the Einstein summation convention gives

the spacetime interval as

$$\begin{aligned} \mathrm{d}s^2 &= \mathrm{d}x_\mu \, \mathrm{d}x^\mu \\ &= \sum_{\mu=0}^{3} \mathrm{d}x_\mu \, \mathrm{d}x^\mu \\ &= \mathrm{d}x_0 \, \mathrm{d}x^0 + \mathrm{d}x_1 \, \mathrm{d}x^1 + \mathrm{d}x_2 \, \mathrm{d}x^2 + \mathrm{d}x_3 \, \mathrm{d}x^3 \\ &= (-c \, \mathrm{d}t)(c \, \mathrm{d}t) + (\mathrm{d}x)(\mathrm{d}x) + (\mathrm{d}y)(\mathrm{d}y) + (\mathrm{d}z)(\mathrm{d}z) \\ &= -c^2 \, \mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \end{aligned} \tag{5}$$

If we apply a Lorentz boost in the x-direction with speed v, then our new spacetime coordinates are given by eqs. (1) to (4) as

$$\begin{aligned} c \, \mathrm{d}t' &= \gamma(c \, \mathrm{d}t - \beta \, \mathrm{d}x), \\ \mathrm{d}x' &= \gamma(\mathrm{d}x - \beta c \, \mathrm{d}t), \\ \mathrm{d}y' &= \mathrm{d}y, \\ \mathrm{d}z' &= \mathrm{d}z \end{aligned}$$

and so our new spacetime interval is

$$\begin{aligned} \mathrm{d}s'^2 &= \mathrm{d}x'_\mu \, \mathrm{d}x'^\mu \\ &= -(c \, \mathrm{d}t')^2 + \mathrm{d}x'^2 + \mathrm{d}y'^2 + \mathrm{d}z'^2 \\ &= -\gamma^2(c \, \mathrm{d}t - \beta \, \mathrm{d}x)^2 + \gamma^2(\mathrm{d}x - \beta c \, \mathrm{d}t)^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \\ &= \gamma^2(-(c \, \mathrm{d}t) - \beta^2 \, \mathrm{d}x^2 + \mathrm{d}x^2 + \beta^2 c^2 \, \mathrm{d}t^2) \\ &= \gamma^2(1 - \beta^2) \, \mathrm{d}x^2 - \gamma^2(1 - \beta^2) c^2 \, \mathrm{d}t^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \end{aligned}$$

However, we defined $\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$ and $\beta = \frac{v}{c}$, so

$$\gamma^2(1 - \beta^2) = \frac{1 - \frac{v^2}{c^2}}{1 - \frac{v^2}{c^2}} = 1$$

and so our transformed spacetime interval is simply

$$\begin{aligned} \mathrm{d}s'^2 &= \mathrm{d}x^2 - c^2 \, \mathrm{d}t^2 + \mathrm{d}y^2 + \mathrm{d}z^2 \\ &= -c \, \mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2. \end{aligned}$$

Therefore, by comparison with eq. (5) we see that our Lorentz transformation had no effect on the value of our spacetime interval.

Indeed, the situation is symmetric in the three spatial coordinates (we could have chosen our x-direction as anything) and so the spacetime interval must be invariant under *all* Lorentz boosts.

Problem: Time Dilation

Let us consider two consecutive instantaneous events, the first occurring at $t = 0$ in both frames the second occurring at $t = \mathrm{d}t$ in frame S or $t = \mathrm{d}t'$ in frame S'. These two events, for instance, could be two ticks of a clock. Since we are interested only in time dilation, we shall say the two events occur at the same point in space in the unprimed frame S, so that $\mathrm{d}x = 0$. (This is equivalent to saying that in frame S the clock isn't moving.)

Hence, due to the invariance of the interval,

$$\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2 - c^2\,\mathrm{d}t^2 = \mathrm{d}x'^2 + \mathrm{d}y'^2 + \mathrm{d}z'^2 - c^2\,\mathrm{d}t'^2$$

but the reference frames are moving relative to each other only in the x-direction, and Lorentz transformations do not alter perpendicular distances, so performing a Lorentz boost in the x-direction gives $\mathrm{d}y = \mathrm{d}y'$ and $\mathrm{d}z = \mathrm{d}z'$. Hence,

$$\mathrm{d}x^2 - c^2\,\mathrm{d}t^2 = \mathrm{d}x'^2 - c^2\,\mathrm{d}t'^2. \tag{6}$$

However, we have set $\mathrm{d}x = 0$, and so

$$c^2\,\mathrm{d}t^2 = c^2\,\mathrm{d}t'^2 - \mathrm{d}x'^2. \tag{7}$$

The second Lorentz transformation equation, eq. (2), gives us

$$\mathrm{d}x' = \gamma(\mathrm{d}x - \beta c\,\mathrm{d}t)$$

and therefore we know, substituting this into eq. (7), that

$$c^2\,\mathrm{d}t^2 = c^2\,\mathrm{d}t'^2 - \gamma^2(\mathrm{d}x - \beta c\,\mathrm{d}t)^2$$

but since $\mathrm{d}x = 0$,

$$\begin{aligned} c^2\,\mathrm{d}t^2 &= c^2\,\mathrm{d}t'^2 - \gamma^2\beta^2 c^2\,\mathrm{d}t^2 \\ \implies \mathrm{d}t^2 &= \mathrm{d}t'^2 - \gamma^2\beta^2\,\mathrm{d}t^2 \\ \implies \mathrm{d}t'^2 &= (1 + \gamma^2\beta^2)\,\mathrm{d}t^2 \\ \implies \mathrm{d}t' &= \sqrt{1 + \gamma^2\beta^2}\,\mathrm{d}t. \end{aligned}$$

Using our definitions $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$ and $\beta = \frac{v}{c}$, we see that

$$\begin{aligned} \sqrt{1 + \beta^2\gamma^2} &= \sqrt{1 + \frac{\frac{v^2}{c^2}}{1 - \frac{v^2}{c^2}}} \\ &= \sqrt{\frac{1 - \frac{v^2}{c^2} + \frac{v^2}{c^2}}{1 - \frac{v^2}{c^2}}} \\ &= \sqrt{\frac{1}{1 - \frac{v^2}{c^2}}} = \gamma \end{aligned}$$

and so we come finally to our formula for time dilation,

$$\mathrm{d}t' = \gamma\,\mathrm{d}t. \tag{8}$$

Since the Lorentz factor γ varies between 1 (at $v = 0$) and ∞ (at $v = c$), as shown in fig. 1, the time measured by the primed frame will increase without bound as the relative velocity of the two frames approaches the speed of light.

This, rather unintuitively, implies that the faster an object is moving relative to you, the slower time will appear to pass for that object. For instance, if you look at two clocks, one stationary relative to you and one moving very quickly away from or towards you, the moving clock's ticks will be much further apart than those of the stationary clock.

A real life example of this effect can be observed in satellites in Earth's orbit, especially GPS satellites (although some of this time dilation is due to gravity). Objects higher in Earth's orbit have relatively higher speeds, and hence time runs more slowly on the satellites' clocks (relative to clocks on the surface of the Earth). This results in onboard clocks requiring adjustment in order to match clocks on the Earth's surface.

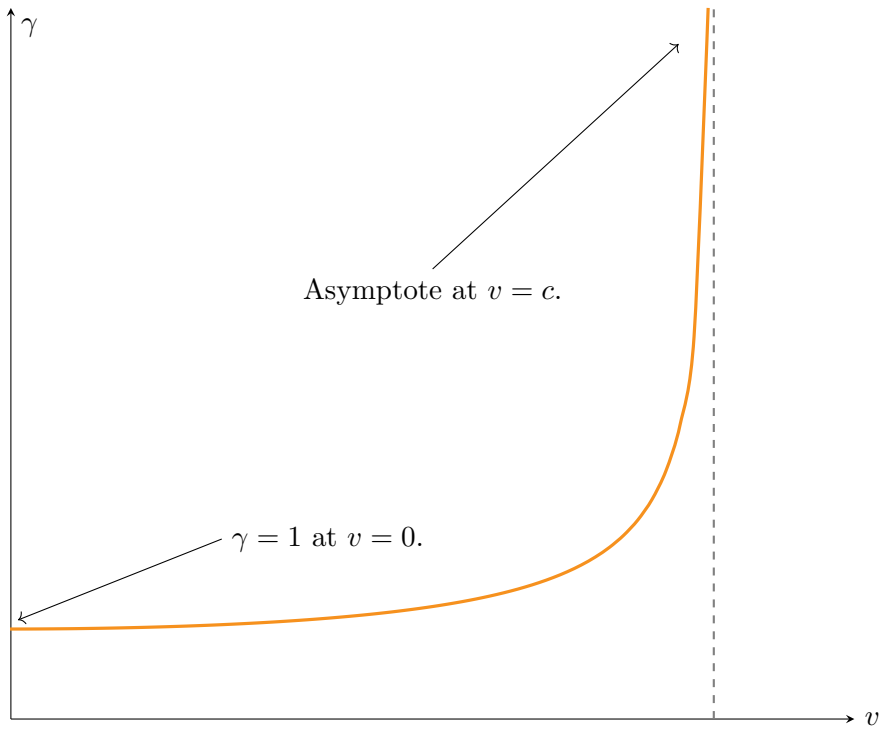


Figure 1: A graph of the Lorentz factor $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$ plotted against v.

Problem: Length Contraction

We are now interested in length contraction, so let's consider a beam of wood with one end at $x = 0$ in both frames and the other end at $x = \mathrm{d}x$ in frame S or $x = \mathrm{d}x'$ in frame S'. Suppose there are two events, one at either end of the beam of wood, that are simultaneous in the moving frame S', so that $\mathrm{d}t' = 0$, so that the length of the wood as measured from the moving frame S' is $\mathrm{d}x'$ but the *proper* length of the wood, as measured from its rest frame S, is $\mathrm{d}x$.

We know from eq. (6) that

$$\mathrm{d}x^2 - c^2\,\mathrm{d}t^2 = \mathrm{d}x'^2 - c^2\,\mathrm{d}t'^2$$

and so since $\mathrm{d}t' = 0$,

$$\mathrm{d}x'^2 = \mathrm{d}x^2 - c^2\,\mathrm{d}t^2. \tag{9}$$

Then by the first Lorentz transformation (eq. (2)),

$$\begin{aligned} c\,\mathrm{d}t' &= \gamma(c\,\mathrm{d}t - \beta\,\mathrm{d}x) \\ \implies c\,\mathrm{d}t &= \frac{c\,\mathrm{d}t'}{\gamma} + \beta\,\mathrm{d}x \end{aligned}$$

and so substituting this into eq. (9),

$$\mathrm{d}x'^2 = \mathrm{d}x^2 - \left(\frac{c\,\mathrm{d}t'}{\gamma} + \beta\,\mathrm{d}x\right)^2.$$

We have set $\mathrm{d}t' = 0$ however, and thus

$$\begin{aligned} \mathrm{d}x'^2 &= \mathrm{d}x^2 - \beta^2\,\mathrm{d}x^2 \\ \implies \mathrm{d}x'^2 &= (1-\beta^2)\,\mathrm{d}x^2 \\ \implies \mathrm{d}x' &= \sqrt{1-\beta^2}\,\mathrm{d}x. \end{aligned}$$

However, $\gamma = \frac{1}{\sqrt{1-\beta^2}}$ by definition and so $\sqrt{1-\beta^2} = \frac{1}{\gamma}$, meaning

$$\mathrm{d}x' = \frac{\mathrm{d}x}{\gamma}.$$

This equation for length contraction is beautifully symmetric with that for time dilation given in eq. (8), and just as time for some object will appear to pass increasingly slowly to an observer as an object's relative speed nears the speed of light, so the length of that object will decrease until it takes up apparently no space when travelling at the speed of light.

As an example, consider a coach moving at some velocity with lights at the front and back. When either of the lights flash, that light drops a marker.

An observer sees the coach pass. When the midpoint of the coach is in line with the observer, the lights flash. The light has to travel equal distances and since the speed of light is constant, the flashes reach the observer simultaneously.

However, an observer inside the coach placed midway between the two lights would appear to see first the back light and then the front light because they are travelling at a velocity, and the simultaneity of events is not conserved because of time dilation. The lights drop markers as they flash, but because they now occur in succession the distance between the markers seems reduced. Since from the passenger's frame of reference it is the outside of the train which is moving quickly, there is a length contraction at higher relative velocity.

Problem: Relativity and Rotations

For this question, we will be using the Minkowskian space of metric signature $(+,-,-,-)$. For the remainder of the submission we will return to the signature $(-,+,+,+)$.

Under this metric, the spacetime interval is given by:

$$\mathrm{d}s^2 = c^2\,\mathrm{d}t^2 - \mathrm{d}x^2 = c^2\,\mathrm{d}t'^2 - \mathrm{d}x'^2 \tag{10}$$

since the spacetime interval is preserved under Lorentz transformations.

By analogy with Euclidean space (consider how the trigonometric functions relate the components of a vector to its magnitude), we would like $\mathrm{d}x$ and $c\,\mathrm{d}t$ to be parameters of $\mathrm{d}s$ for some functions a, b:

$$\begin{aligned} \mathrm{d}x &= a\,\mathrm{d}s, \\ c\,\mathrm{d}t &= b\,\mathrm{d}s. \end{aligned}$$

Substituting this into eq. (10),

$$\begin{aligned} \mathrm{d}s^2 &= (a\,\mathrm{d}s)^2 - (b\,\mathrm{d}s)^2 \\ &= a^2\,\mathrm{d}s^2 - b^2\,\mathrm{d}s^2 \\ \implies 1 &= a^2 - b^2. \end{aligned}$$

Since $\cosh^2\phi - \sinh^2\phi = 1$, we see that the hyperbolic functions satisfy the relation given, and so we say $a = \cosh\phi$ and $b = \sinh\phi$:

$$\therefore \qquad c\,\mathrm{d}t = \mathrm{d}s\cosh\phi, \tag{11}$$

$$\mathrm{d}x = \mathrm{d}s\sinh\phi. \tag{12}$$

By eqs. (11) and (12), we see that

$$\begin{aligned} \frac{\sinh\phi}{\cosh\phi} &= \frac{1}{c}\frac{\mathrm{d}x}{\mathrm{d}t} \\ \implies \tanh\phi &= \frac{v}{c} = \beta \end{aligned}$$

as $\tanh\phi = \frac{\sinh\phi}{\cosh\phi}$, so

$$\begin{aligned} \gamma &= \frac{1}{\sqrt{1-\beta^2}} \\ &= \frac{1}{\sqrt{1-\tanh^2\phi}} \\ &= \cosh\phi \end{aligned}$$

by the identity $\cosh\theta \equiv \frac{1}{\sqrt{1-\tanh^2\theta}}$.

Hence the Lorentz boost for a time dimension and and one parallel spatial dimension is given by:

$$\begin{pmatrix} \gamma & -\gamma\beta \\ -\gamma\beta & -\gamma \end{pmatrix} = \begin{pmatrix} \cosh\phi & -\sinh\phi \\ -\sinh\phi & \cosh\phi \end{pmatrix}$$

seeing as $\sinh\phi = \cosh\phi\tanh\phi$. Indeed, this is the hyperbolic rotation matrix for two dimensions, and so we have successfully created a way to think of Lorentz transformations purely as rotations.

If we add back in the other two spatial dimensions, our full transformation matrix becomes

$$\begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cosh\phi & -\sinh\phi & 0 & 0 \\ -\sinh\phi & \cosh\phi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

which generates the equations

$$\begin{aligned} c\,\mathrm{d}t' &= c\,\mathrm{d}t\cosh\phi - \mathrm{d}x\sinh\phi, \\ \mathrm{d}x' &= \mathrm{d}x\cosh\phi - c\,\mathrm{d}t\sinh\phi, \\ \mathrm{d}y' &= \mathrm{d}y, \\ \mathrm{d}z' &= \mathrm{d}z. \end{aligned}$$

This is a fairly nice way to compute Lorentz transformations. The parameter ϕ is usually known as the *rapidity* and, unlike the actual speed, rapidities add linearly under boosts.

2.1.5 Mechanics in the Language of Four-Vectors

Problem: Four-Velocity

(a) As demonstrated, time dilation means that an object's velocity relative to a frame S determines how quickly time passes for that object, as observed by S. Hence, to fix the passage of time at a constant rate we want to use the *proper time* — that is, the time as measured by a frame that is always stationary relative to the object. Indeed, if we were to

differentiate with respect to the time measured from S instead then the zeroth component of the four-velocity vector would be

$$u^0 = \frac{c\,\mathrm{d}t}{\mathrm{d}t} = c$$

and so would be constant, which makes no sense. We must differentiate with respect to proper time because it is the only *invariant* quantity of time.

A rather nice interpretation of the four-velocity (differentiating with respect to proper time τ, of course), is that it's simply the unit tangent vector to the world line of the object — that is, its path through all of spacetime.

(b) Consider an object moving at velocity u in the x-direction relative to a reference frame S. We wish to find its velocity relative to a frame S' travelling at a velocity v in the x-direction relative to S.

If in frame S the object has four-velocity $u^\mu = (u^0, u^1, u^2, u^3)$ and in frame S' has four-velocity $u'^\mu = (u'^0, u'^1, u'^2, u'^3)$, then a Lorentz boost between the frames gives

$$\begin{pmatrix} u'^0 \\ u'^1 \\ u'^2 \\ u'^3 \end{pmatrix} = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u^0 \\ u^1 \\ u^2 \\ u^3 \end{pmatrix}$$

or, in expanded form,

$$u'^0 = \gamma(u^0 - \beta u^1), \tag{13}$$
$$u'^1 = \gamma(u^1 - \beta u^0), \tag{14}$$
$$u'^2 = u^2,$$
$$u'^3 = u^3.$$

By the definition of four-velocity, $u^\mu = \dfrac{\mathrm{d}x^\mu}{\mathrm{d}\tau}$ where τ is the proper time of the object, so

$$u^0 = c\,\frac{\mathrm{d}t}{\mathrm{d}\tau} \tag{15}$$

and

$$u^1 = \frac{\mathrm{d}x}{\mathrm{d}\tau}\,. \tag{16}$$

Now, we have derived that where λ is the Lorentz factor between S and the object's proper frame, time dilation means

$$\mathrm{d}t = \lambda\,\mathrm{d}\tau$$

and so we can simplify eq. (15) and eq. (16), giving

$$u^0 = c\,\frac{\mathrm{d}t}{\mathrm{d}\tau} = c\lambda\,\frac{\mathrm{d}\tau}{\mathrm{d}\tau} = c\lambda$$

and similarly,

$$u^1 = \frac{\mathrm{d}x}{\mathrm{d}\tau} = \lambda\,\frac{\mathrm{d}x}{\mathrm{d}t} = \lambda u$$

since u is the x-velocity of the object in frame S. So, the Lorentz transformation in eq. (13) and eq. (14) gives

$$u'^0 = \gamma(c\lambda - \beta\gamma u),$$
$$u'^1 = \gamma(\lambda u - \beta c\lambda),$$

and hence the new velocity of the object in the x-direction as measured from S' is simply

$$\begin{aligned} u' = \frac{\mathrm{d}x'}{\mathrm{d}t'} &= \frac{c\dfrac{\mathrm{d}x'}{\mathrm{d}\tau}}{c\dfrac{\mathrm{d}t'}{\mathrm{d}\tau}} = \frac{cu'^1}{u'^0} \\ &= \frac{c\gamma\lambda(u-\beta c)}{\gamma\lambda(c-\beta u)} \\ &= \frac{cu - vc}{c - \frac{vu}{c}} \\ &= \frac{u-v}{1-\frac{uv}{c^2}}. \end{aligned}$$

Hence we have derived one of the Einstein addition laws. For the other laws we must use an inverse Lorentz transformation on the four-velocity to boost from frame S' to frame S:

$$\begin{aligned} \begin{pmatrix} u^0 \\ u^1 \\ u^2 \\ u^3 \end{pmatrix} &= \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} u'^0 \\ u'^1 \\ u'^2 \\ u'^3 \end{pmatrix} \\ &= \begin{pmatrix} \gamma & \gamma\beta & 0 & 0 \\ \gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u'^0 \\ u'^1 \\ u'^2 \\ u'^3 \end{pmatrix}. \end{aligned}$$

This matrix inversion results in the equations

$$u^0 = \gamma(u'^0 + \beta u'^1), \tag{17}$$
$$u^1 = \gamma(u'^1 + \beta u'^0), \tag{18}$$
$$u^2 = u'^2,$$
$$u^3 = u'^3.$$

By the same argument as before, where now μ is the Lorentz factor between S' and our moving object's proper frame, we see

$$u'^0 = \mu c$$

and

$$u'^1 = \mu u'$$

since u' is the x-velocity of the object in the frame S. So, eq. (17) and eq. (18) give

$$u^0 = \gamma(\mu c + \beta\mu u')$$
$$u^1 = \gamma(\mu u' + \beta\mu c)$$

and thus our transformed velocity in frame S is:

$$\begin{aligned} u = \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{c\dfrac{\mathrm{d}x}{\mathrm{d}\tau}}{c\dfrac{\mathrm{d}t}{\mathrm{d}\tau}} \\ &= \frac{cu'}{u^0} \\ &= \frac{c\gamma\mu(u'+\beta c)}{\gamma\mu(c+\beta u')} \\ &= \frac{cu'+cv}{c+\frac{vu}{c}} \\ &= \frac{u'+v}{1+\frac{uv}{c^2}} \end{aligned}$$

which is the other Einstein velocity addition law as desired.

Problem: Invariance of Energy and Momentum

We are given that the definition of four-momentum is

$$p^\mu = m_0 u^\mu \tag{19}$$

where m_0 is the rest mass, and also that equivalently

$$p^\mu = \begin{pmatrix} \frac{E}{c} \\ p_x \\ p_y \\ p_z \end{pmatrix}. \tag{20}$$

So, we can calculate the (square) length of the four-momentum in two different ways. Indeed, the length of a spacetime four-vector is a Lorentz invariant as shown previously.
First, by eq. (19), the squared length is

$$p_\mu p^\mu = m_0^2 u_\mu u^\mu. \tag{21}$$

However, we know from the previous problem that the four velocity is

$$u^\mu = \begin{pmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{pmatrix}$$

and so the squared length of the four-velocity is just

$$\begin{aligned} u_\mu u^\mu &= \begin{pmatrix} -\gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{pmatrix} \cdot \begin{pmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{pmatrix} \\ &= -\gamma^2 c^2 + \gamma^2 u_x^2 + \gamma^2 u_y^2 + \gamma^2 u_z^2 \\ &= -\gamma^2 c^2 + \gamma^2 u^2 \end{aligned}$$

where u is the magnitude of the three-velocity. However, simplifying further, this becomes

$$\begin{aligned} u_\mu u^\mu &= \gamma^2(u^2 - c^2) \\ &= \gamma^2 c^2 \left(\frac{u^2}{c^2} - 1\right) \\ &= \gamma^2 c^2 \left(-\frac{1}{\gamma^2}\right) \\ &= -c^2 \end{aligned}$$

and so the magnitude of the four-momentum as given by eq. (21) is

$$\begin{aligned} p_\mu p^\mu &= m_0^2(-c^2) \\ &= -m_0^2 c^2. \end{aligned} \tag{22}$$

Now, we shall work out the same quantity using the definition in eq. (20). Taking the squared magnitude of this four-vector gives

$$\begin{aligned} p_\mu p^\mu &= \begin{pmatrix} -\frac{E}{c} \\ p_x \\ p_y \\ p_z \end{pmatrix} \cdot \begin{pmatrix} \frac{E}{c} \\ p_x \\ p_y \\ p_z \end{pmatrix} \\ &= -\frac{E^2}{c^2} + p_x^2 + p_y^2 + p_z^2 \\ &= -\frac{E^2}{c^2} + p^2 \end{aligned} \tag{23}$$

where p is the magnitude of the three-momentum. So, equating eq. (22) and eq. (23),

$$\begin{aligned} -m_0^2 c^2 &= -\frac{E^2}{c^2} + p^2 \\ \implies E^2 &= p^2 c^2 + m_0^2 c^4, \end{aligned}$$

which is the well-known relativistic energy-momentum relation that we were looking for.

Problem: Four-Acceleration

Four-acceleration is simply the derivative of four-velocity with respect to proper time. We have shown that the magnitude of the four-velocity is always $-c^2$, that is

$$u_\mu u^\mu = -c^2$$

and so, differentiating this using the chain rule, we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}(u_\mu u^\mu) &= \frac{\mathrm{d}}{\mathrm{d}\tau}(-c^2) \\ \implies 2u_\mu \frac{\mathrm{d}u^\mu}{\mathrm{d}\tau} &= 0 \\ \implies u_\mu a^\mu &= 0 \end{aligned}$$

since four-acceleration is the derivative of four-velocity with respect to proper time.

In other words, the dot product $u_\mu a^\mu$ of four-acceleration and four-velocity is always identically zero.

2.2 Relativistic Electrodynamics and Tensors

2.2.2 Four-Current

Problem: The Continuity Equation

(a) Consider some volume V bounded by surface S in three-dimensional space. Now, the total current out of V at any point in time is

$$-\frac{\partial q}{\partial t} = \oiint_S \vec{j} \cdot \hat{n}\, \mathrm{d}S.$$

where q is the total charge contained by V, $\vec{j}$ is the flux (current density) and $\hat{n}$ is the unit normal vector to the surface S. By the divergence theorem, this becomes

$$-\frac{\partial q}{\partial t} = \iiint_V \operatorname{div}(\vec{j})\, \mathrm{d}V \tag{24}$$

which makes sense intuitively; the total charge flow out of the shape is going to be the same as the sum of all the net charge flows out of every point in the shape (the sum of divergences at every point in V).

Now we use the fact that the total charge is

$$q = \iiint_V \rho\, \mathrm{d}V$$

given charge density ρ, so that by eq. (24),

$$-\frac{\partial}{\partial t} \iiint_V \rho\, \mathrm{d}V = \iiint_V \operatorname{div} \vec{j}\, \mathrm{d}V$$
$$\implies \iiint_V \left[\frac{\partial \rho}{\partial t} + \operatorname{div} \vec{j}\right] \mathrm{d}V = 0.$$

Since this must hold true for any value, V, it is clear that the integrand itself must be identically zero; that is,

$$\frac{\partial \rho}{\partial t} + \operatorname{div} \vec{j} = 0$$
$$\implies \vec{\nabla} \cdot \vec{j} = -\frac{\partial \rho}{\partial t}$$

which is what was to be shown.

Intuitively, this was obvious: all this law says is that the only way the charge density at some point will increase is if there is a net flow of charge into that point (a negative divergence).

(b) If we are now making relativistic considerations, our charge density becomes

$$\rho = \gamma \rho_0$$

due to length contraction, where ρ_0 is the rest charge density. Hence, the continuity equation we just derived expands out to become

$$\frac{\partial}{\partial x}(\gamma\rho_0 u_x) + \frac{\partial}{\partial y}(\gamma\rho_0 u_y) + \frac{\partial}{\partial z}(\gamma\rho_0 u_z) = -\frac{\partial}{\partial t}(\gamma\rho_0)$$
$$\implies \frac{\partial}{\partial t}(\gamma\rho_0) + \frac{\partial}{\partial x}(\gamma\rho_0 u_x) + \frac{\partial}{\partial y}(\gamma\rho_0 u_y) + \frac{\partial}{\partial z}(\gamma\rho_0 u_z) = 0.$$

using the definition of current density.

Now, using the definition of four-velocity as

$$u^\mu = \begin{pmatrix} \gamma c \\ \gamma u_x \\ \gamma u_y \\ \gamma u_z \end{pmatrix}$$

we re-write this:

$$\frac{1}{c}\frac{\partial}{\partial t}(\rho_0 u^0) + \frac{\partial}{\partial x}(\rho_0 u^1) + \frac{\partial}{\partial y}(\rho_0 u^2) + \frac{\partial}{\partial z}(\rho_0 u^3) = 0$$

and since our standard four-vector components are

$$x^0 = ct,$$
$$x^1 = x,$$
$$x^2 = y,$$
$$x^3 = z,$$

we may re-write this again as

$$\frac{\partial}{\partial x^0}(\rho_0 u^0) + \frac{\partial}{\partial x^1}(\rho_0 u^1) + \frac{\partial}{\partial x^2}(\rho_0 u^2) + \frac{\partial}{\partial x^3}(\rho_0 u^3) = 0$$

or equivalently,

$$\partial_\mu(\rho_o u^\mu) = 0 \iff \partial_\mu j^\mu = 0.$$

This tells us that the four-dimensional divergence of the current density four-vector is identically zero, a rather beautiful way of explaining conservation of charge: at any point in spacetime there is no net flow of charge into or out of that point. In other words, there are no sources or sinks of charge in the universe; charge cannot be created or destroyed.

2.2.3 Four-Potential

For reference, Maxwell's four equations of electromagnetism are as follows:

$$\vec{\nabla}\cdot\vec{E} = \frac{\rho}{\epsilon_0}, \tag{25}$$
$$\vec{\nabla}\cdot\vec{B} = 0, \tag{26}$$
$$\vec{\nabla}\times\vec{E} = -\frac{\partial\vec{B}}{\partial t}, \tag{27}$$
$$c^2\vec{\nabla}\times\vec{B} = \frac{\vec{j}}{\epsilon_0} + \frac{\partial\vec{E}}{\partial t}. \tag{28}$$

Problem: Maxwell's Equations in Terms of the Potentials

(a) Gauss' law for magnetism (eq. (25)) specifies that the divergence of a magnetic field is identically zero — that is, it is a solenoidal vector field. This implies that there exists a vector potential $\vec{A}$ such that

$$\vec{B} = \vec{\nabla} \times \vec{A} \tag{29}$$

as Helmholtz's theorem [1] implies that the divergence of the curl of a vector field is identically zero.

Now substituting this into Faraday's law (eq. (27)), we get

$$\begin{aligned} \vec{\nabla} \times \vec{E} &= -\frac{\partial}{\partial t}\left(\vec{\nabla} \times \vec{A}\right) \\ \implies \vec{\nabla} \times \vec{E} &= -\vec{\nabla} \times \frac{\partial \vec{A}}{\partial t} \\ \implies \vec{\nabla} \times \left(\vec{E} + \frac{\partial \vec{A}}{\partial t}\right) &= 0. \end{aligned}$$

This is true as both the cross product and derivative functions are distributive over addition.

This implies that $\vec{E} + \frac{\partial \vec{A}}{\partial t}$ is a conservative vector field and so, also because of Helmholtz's theorem, may be written as the gradient of some scalar field ϕ that we will call the scalar potential.[1] So,

$$\begin{aligned} \vec{E} + \frac{\partial \vec{A}}{\partial t} &= -\vec{\nabla}\phi \\ \implies \vec{E} &= -\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t} \end{aligned} \tag{30}$$

as desired.

(b) To achieve this reformulation of Maxwell's equations in terms of potential, we start by substituting our newly derived eq. (30) into Gauss' law (the first Maxwell equation), giving

$$\begin{aligned} \vec{\nabla} \cdot \left(-\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t}\right) &= \frac{\rho}{\epsilon_0} \\ \implies -\frac{\partial}{\partial t}\left(\vec{\nabla} \cdot \vec{A}\right) - \nabla^2\phi &= \frac{\rho}{\epsilon_0} \end{aligned} \tag{31}$$

and similarly, we now substitute eq. (29) into Ampere's law (the fourth Maxwell equation):

$$c^2 \vec{\nabla} \times \left(\vec{\nabla} \times \vec{A}\right) = \frac{\vec{j}}{\epsilon_0} + \frac{\partial \vec{E}}{\partial t}.$$

Dividing by c^2 and substituting in eq. (30) gives

$$\vec{\nabla} \times \left(\vec{\nabla} \times \vec{A}\right) = \frac{\vec{j}}{c^2 \epsilon_0} + \frac{1}{c^2}\frac{\partial}{\partial t}\left(-\vec{\nabla}\phi - \frac{\partial \vec{A}}{\partial t}\right)$$

[1] We made ϕ negative to preserve its physical meaning.

and using the identity $\vec{\nabla} \times (\vec{\nabla} \times \vec{C}) \equiv \vec{\nabla}(\vec{\nabla} \cdot \vec{C}) - \nabla^2 \vec{C}$, this becomes

$$\vec{\nabla}(\vec{\nabla} \cdot \vec{A}) - \nabla^2 \vec{A} = \frac{\vec{j}}{c^2 \epsilon_0} - \frac{1}{c^2} \vec{\nabla} \frac{\partial \phi}{\partial t} - \frac{1}{c^2} \frac{\partial^2 \vec{A}}{\partial t^2}$$

$$\implies \vec{\nabla} \left(\vec{\nabla} \cdot \vec{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t} \right) + \frac{1}{c^2} \frac{\partial^2 \vec{A}}{\partial t^2} - \nabla^2 \vec{A} = \frac{\vec{j}}{c^2 \epsilon_0}. \tag{32}$$

The values of A and ϕ we are using here are not unique[2] and so we have some freedom of choice which we will choose to exercise by setting the value of $\vec{\nabla} \cdot \vec{A}$ — that is, we will fix the divergence of the vector potential $\vec{A}$. This choice us what is known as gauge freedom. Looking at eqs. (31) and (32), and especially at the leftmost term of eq. (32), we see that everything will simplify down very nicely if that term is zero - that is, if

$$\vec{\nabla} \cdot A = -\frac{1}{c^2} \frac{\partial \phi}{\partial t}.$$

In fact, this choice is known as the Lorenz gauge. In this case, eq. (32) becomes

$$0 + \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} - \nabla^2 \vec{A} = \frac{\vec{j}}{c^2 \epsilon_0} \tag{33}$$

and eq. (31) becomes

$$\frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} - \nabla^2 \phi = \frac{\rho}{\epsilon_0}. \tag{34}$$

Indeed, using the definition of the d'Alembertian and the fact that $c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}$, eqs. (33) and (34) simplify to

$$\Box^2 \vec{A} = \vec{j} \mu_0 \tag{35}$$

and

$$\Box^2 \phi = c^2 \mu_0 \rho$$

$$\implies \Box^2 \left(\frac{\phi}{c} \right) = c \mu_0 \rho. \tag{36}$$

These are the equations we were to derive (the constants as printed in the question paper are incorrect).

(c) i. Suppose we apply the d'Alembertian operator to a Lorentz-boosted four-vector x^μ. To say that $\Box^2$ is Lorentz invariant means this gives the same result as applying the d'Alembertian operator before the Lorentz transformation.

[2] What I mean by this is that there are many potentials which will generate the same field and so we may choose one such potential.

We shall prove this; where $\Lambda^{\nu}{}_{\mu}$ is the Lorentz transformative matrix,

$$\begin{aligned}
\Box^2\Lambda^{\nu}{}_{\mu}x^{\mu} &= \Box^2\begin{pmatrix}\gamma & -\gamma\beta & 0 & 0\\ -\gamma\beta & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}x^0\\x^1\\x^2\\x^3\end{pmatrix}\\
&= \Box^2\begin{pmatrix}\gamma(x^0-\beta x^1)\\ \gamma(x^1-\beta x^0)\\ x^2\\ x^3\end{pmatrix}\\
&= \begin{pmatrix}\Box^2\gamma(x^0-\beta x^1)\\ \Box^2\gamma(x^1-\beta x^0)\\ \Box^2x^2\\ \Box^2x^3\end{pmatrix}\\
&= \begin{pmatrix}\Box^2x^0-\beta\Box^2x^1)\\ \gamma(\Box^2x^1-\beta\Box^2x^0)\\ \Box^2x^2\\ \Box^2x^3\end{pmatrix}\\
&= \begin{pmatrix}\gamma & -\gamma\beta & 0 & 0\\ -\gamma\beta & \gamma & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\Box^2x^0\\ \Box^2x^1\\ \Box^2x^2\\ \Box^2x^3\end{pmatrix}\\
&= \Lambda^{\nu}{}_{\mu}\Box^2x^{\mu}.
\end{aligned}$$

Thus, we have used the fact that $\Box^2$ is distributive over addition (since it is made up of sums of second partial derivatives, all of which are themselves distributive over addition) to show that the d'Alembertian operator is indeed Lorentz invariant.

Now, using the given definition of the four-potential A^{μ},

$$\Box^2A^{\mu} = \begin{pmatrix}\Box^2\left(\frac{\phi}{c}\right)\\ \Box^2\vec{A}\end{pmatrix}$$

which according to eqs. (35) and (36) gives

$$\Box^2A^{\mu} = \begin{pmatrix}\mu_0\rho c\\ \mu_0\vec{j}\end{pmatrix} = \begin{pmatrix}\mu_0\rho c\\ \mu_0\rho\vec{u}\end{pmatrix}.$$

The charge density ρ in the moving frame is related to the rest charge density ρ_0 due to the length contraction by

$$\rho = \gamma\rho_0$$

and therefore

$$\Box^2A^{\mu} = \begin{pmatrix}\mu_0\rho_0\gamma c\\ \mu_0\rho_0\gamma\vec{u}\end{pmatrix} = \mu_0\rho_0u^{\mu}$$

where u^{μ} is the four-velocity. Since u^{μ} is a valid four-vector, $\Box^2A^{\mu}$ is a four-vector, and since $\Box^2$ is Lorentz invariant, A^{μ} is also a four-vector.

ii. The quantity $\frac{\mathrm{d}V}{r}$ is in fact *not* Lorentz invariant; consider the following counterexample.

Let $\mathrm{d}V$ be some small volume $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ and r be the distance from the small volume to some point in the y direction. If we perform a Lorentz boost with velocity v in the x-direction then due to length contraction the volume $\mathrm{d}V$ will contract. However, the length r is perpendicular to the boost direction and so its value will be unaffected. Thus the quantity $\frac{\mathrm{d}V}{r}$ will be decreased. It is therefore clear that $\frac{\mathrm{d}V}{r}$ cannot be Lorentz invariant.

We assume that we are instead asked to show that $\frac{\mathrm{d}V}{r}$ is Lorentz covariant, meaning that it transforms according to the rules of the Lorentz transformations.

Scalar potential is given by

$$\phi = \int_V \frac{\rho\,\mathrm{d}V}{4\pi\epsilon_0 r}$$

and vector potential is given by

$$\vec{A} = \int_V \frac{\mu_0\,\vec{j}\,\mathrm{d}V}{4\pi r},$$

so contravariant four-potential is

$$A^\mu = \begin{pmatrix} \int \frac{\rho\,\mathrm{d}V}{4\pi c\epsilon_0 r} \\ \int \frac{\mu_0\,\vec{j}\,\mathrm{d}V}{4\pi r} \end{pmatrix} = \frac{\mu_0}{4\pi}\int \begin{pmatrix} \rho c \\ \vec{j} \end{pmatrix} \frac{\mathrm{d}V}{r} = \frac{\mu_0}{4\pi}\int j^\mu \frac{\mathrm{d}V}{r}. \tag{37}$$

Performing a Lorentz boost in the x-direction gives the new four-potential therefore as

$$\begin{aligned} A'^\mu &= \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \int \frac{\rho\,\mathrm{d}V}{4\pi c\epsilon_0 r} \\ \int \frac{\mu_0 j_x\,\mathrm{d}V}{4\pi r} \\ \int \frac{\mu_0 j_y\,\mathrm{d}V}{4\pi r} \\ \int \frac{\mu_0 j_z\,\mathrm{d}V}{4\pi r} \end{pmatrix} \\ &= \begin{pmatrix} \gamma\left(\int \frac{\rho\,\mathrm{d}V}{4\pi c\epsilon_0 r} - \beta\int \frac{\mu_0 j_x\,\mathrm{d}V}{4\pi r}\right) \\ \gamma\left(\int \frac{\mu_0 j_x\,\mathrm{d}V}{4\pi r} - \beta\int \frac{\rho\,\mathrm{d}V}{4\pi c\epsilon_0 r}\right) \\ \int \frac{\mu_0 j_y\,\mathrm{d}V}{4\pi r} \\ \int \frac{\mu_0 j_z\,\mathrm{d}V}{4\pi r} \end{pmatrix}. \end{aligned}$$

Combining integrals and using the fact that $c = \frac{1}{\sqrt{\mu_0\epsilon_0}}$, this becomes

$$A'^\mu = \begin{pmatrix} \int \gamma\left(\frac{\mu_0\rho c}{4\pi} - \beta\frac{\mu_0 j_x}{4\pi}\right)\frac{\mathrm{d}V}{r} \\ \int \gamma\left(\frac{\mu_0 j_x}{4\pi} - \beta\frac{\mu_0\rho c}{4\pi}\right)\frac{\mathrm{d}V}{r} \\ \int \frac{\mu_0 j_y}{4\pi}\frac{\mathrm{d}V}{r} \\ \int \frac{\mu_0 j_z}{4\pi}\frac{\mathrm{d}V}{r} \end{pmatrix}.$$

Taking the integral outside of the vector, this becomes

$$A'^{\mu} = \int \begin{pmatrix} \gamma(\rho c - \beta j_x) \\ \gamma(j_x - \beta \rho c) \\ j_y \\ j_z \end{pmatrix} \frac{\mu_0 \, \mathrm{d}V}{4\pi r}$$
$$= \frac{\mu_0}{4\pi} \int j'^{\mu} \frac{\mathrm{d}V}{r}.$$

We have shown that boosting the four-current and then integrating it is equivalent to integrating it and then boosting it, and so the quantity $\frac{\mathrm{d}V}{r}$ must be Lorentz covariant, since it transforms according to the Lorentz transformations. Indeed, this means every part of the right hand side of eq. (37) transforms according to the Lorentz transformations, and therefore so must A^{μ} itself. Thus, A^{μ} is a four-vector.

(d) By combining eqs. (35) and (36) in vector form, we get

$$\begin{pmatrix} \Box^2 \frac{\phi}{c} \\ \Box^2 \vec{A} \end{pmatrix} = \begin{pmatrix} c\rho\mu_0 \\ \vec{j}\,\mu_0 \end{pmatrix},$$

which we can write as

$$\Box^2 \begin{pmatrix} \frac{\phi}{c} \\ \vec{A} \end{pmatrix} = \begin{pmatrix} c\rho\mu_0 \\ \vec{j}\,\mu_0 \end{pmatrix}$$
$$\implies \Box^2 A^{\mu} = \begin{pmatrix} c\rho\mu_0 \\ \vec{j}\,\mu_0 \end{pmatrix} \tag{38}$$

as this is the definition of A^{μ}. Now, finding an expanded form for the current density four-vector,

$$j^{\mu} = \rho_0 u^{\mu} = \begin{pmatrix} \rho_0 \gamma c \\ \rho_0 \gamma \vec{u} \end{pmatrix}$$

where $\vec{u}$ is the three-velocity. Due to length contraction, the charge density in the moving frame is given by

$$\rho = \gamma \rho_0$$

and so

$$j^{\mu} = \begin{pmatrix} \rho c \\ \rho \vec{u} \end{pmatrix} = \begin{pmatrix} \rho c \\ \vec{j} \end{pmatrix}.$$

Therefore by eq. (38),

$$\Box^2 A^{\mu} = \mu_0 j^{\mu}$$

as desired. This is a beautiful single equation representing all of electrodynamics.

Problem: Forces in Different Frames

We shall first consider how a general momentum four-vector behaves under a Lorentz transformation. Let p^{μ} be the four-momentum in the unprimed frame and let p'^{ν} be the four-momentum

in the primed (proper) frame. Then,

$$\begin{aligned} p'^{\nu} &= \Lambda^{\nu}{}_{\mu} p^{\mu} \\ \implies \begin{pmatrix} \frac{E'}{c} \\ p'_x \\ p'_y \\ p'_z \end{pmatrix} &= \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{E}{c} \\ p_x \\ p_y \\ p_z \end{pmatrix} \\ \implies \begin{pmatrix} \frac{E'}{c} \\ p'_x \\ p'_y \\ p'_z \end{pmatrix} &= \begin{pmatrix} \gamma\left(\frac{E}{c} - \beta p_x\right) \\ \gamma\left(p_x - \beta\frac{E}{c}\right) \\ p_y \\ p_z \end{pmatrix} \end{aligned}$$

and so the boosted three-momentum is

$$\vec{p'} = \begin{pmatrix} p'_x \\ p'_y \\ p'_z \end{pmatrix} = \begin{pmatrix} \gamma(p_x - \frac{\beta}{c}E) \\ p_y \\ p_z \end{pmatrix}.$$

Note that we have assumed the primed frame is travelling with velocity v relative to the unprimed frame in the x-direction only: this is without loss of generality as we can just choose the direction of our x-axis.

So if $\vec{F_e}$ is the force in the unprimed frame and $\vec{F'}$ is the force in the primed frame, then by Newton II

$$\vec{F'} = \frac{\mathrm{d}\vec{p'}}{\mathrm{d}t'} = \frac{\left(\frac{\mathrm{d}\vec{p'}}{\mathrm{d}t}\right)}{\left(\frac{\mathrm{d}t'}{\mathrm{d}t}\right)}$$

but from the Lorentz transformation for four-displacement we know

$$\begin{aligned} c\,\mathrm{d}t' &= \gamma(c\,\mathrm{d}t - \beta\,\mathrm{d}x) \\ \implies \mathrm{d}t' &= \gamma\,\mathrm{d}t - \gamma\frac{v}{c^2}\,\mathrm{d}x \\ \implies \frac{\mathrm{d}t'}{\mathrm{d}t} &= \gamma - \gamma\frac{v}{c^2}\frac{\mathrm{d}x}{\mathrm{d}t} \\ \implies \frac{\mathrm{d}t'}{\mathrm{d}t} &= \gamma\left(1 - \frac{v^2}{c^2}\right) = \frac{\gamma}{\gamma^2} = \frac{1}{\gamma} \end{aligned}$$

and so

$$\begin{aligned} \vec{F'} &= \gamma\frac{\mathrm{d}\vec{p'}}{\mathrm{d}t} \\ &= \gamma\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \gamma\left(p_x - \frac{\beta}{c}E\right) \\ p_y \\ p_z \end{pmatrix} \\ &= \gamma\begin{pmatrix} \gamma\left(\frac{\mathrm{d}p_x}{\mathrm{d}t} - \frac{\beta}{c}\frac{\mathrm{d}E}{\mathrm{d}t}\right) \\ \frac{\mathrm{d}p_y}{\mathrm{d}t} \\ \frac{\mathrm{d}p_z}{\mathrm{d}t} \end{pmatrix}. \end{aligned}$$

However if $\overrightarrow{F_e} = \begin{pmatrix} F_x \\ F_y \\ F_z \end{pmatrix}$ then by Newton II the momentum time derivatives become forces and so

$$\begin{aligned} \overrightarrow{F'} &= \gamma \begin{pmatrix} \gamma \left(F_x - \frac{\beta}{c} \frac{\mathrm{d}E}{\mathrm{d}t} \right) \\ F_y \\ F_z \end{pmatrix} \\ &= \begin{pmatrix} \gamma^2 \left(F_x - \frac{v}{c^2} \frac{\mathrm{d}E}{\mathrm{d}t} \right) \\ \gamma F_y \\ \gamma F_z \end{pmatrix}. \end{aligned}$$

This is almost very nice; now we may use the fact that the change in energy over time of the particle (since this change is only due to the force $\overrightarrow{F_e}$ acting on it) is just the power,

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \overrightarrow{F_e} \cdot \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix} = F_x v$$

and so our primed force becomes

$$\begin{aligned} \overrightarrow{F'} &= \begin{pmatrix} \gamma^2 \left(F_x - \frac{v}{c^2} (F_x v) \right) \\ \gamma F_y \\ \gamma F_z \end{pmatrix} \\ &= \begin{pmatrix} \gamma^2 F_x \left(1 - \frac{v^2}{c^2} \right) \\ \gamma F_y \\ \gamma F_z \end{pmatrix} \\ &= \begin{pmatrix} F_x \\ \gamma F_y \\ \gamma F_z \end{pmatrix}. \end{aligned} \tag{39}$$

Thus, when a Lorentz boost is applied to a three-dimensional force vector, the component of that force in the direction of the boost is unchanged and the other two components are increased by a factor of γ.

Problem: Particles in a Wire

Note: this question is not particularly clear. We assume that firstly $\lambda+$ and $\lambda-$ are the same charge density but with opposite signs, and secondly that we are to take both electric and magnetic forces into account in both frames.

(a) We consider here a wire containing stationary positive particles and negative particles moving at a velocity u. In the lab frame the positive particles have charge density λ and the negative particles have charge density $-\lambda$, and there is a test particle of charge $+q$ a distance r from the wire moving at a velocity v parallel to the wire.

Clearly in the lab frame, there is no net charge in the wire and so the test charge will experience no electric force. However, the test charge *is* moving and so will experience a magnetic force of

$$\overrightarrow{F_B} = q\,\vec{v} \times \vec{B}$$

where $\vec{B}$ is the magnetic field. A simple application of Ampère's law or Biot-Savart gives this field as

$$B = \frac{\mu_0 I}{2\pi r}$$

radially outwards, where I is the conventional current. So, as the flow of positive charge is at speed $-u$,

$$\begin{aligned} I &= (-\lambda)(-u) \\ &= \lambda u \end{aligned}$$

and hence

$$B = \frac{\mu_0 \lambda u}{2\pi r}.$$

So, the magnetic field is of magnitude

$$F_B = qvB = \frac{qv\mu_0\lambda u}{2\pi r}.$$

(b) Now let us consider the rest frame of the test charge. The current density four-vector in the lab frame due to the positive charges is

$$j_+^\mu = \begin{pmatrix} \lambda c \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and due to the negative charges is

$$j_-^\mu = \begin{pmatrix} -\lambda c \\ -\lambda(-u) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -\lambda c \\ \lambda u \\ 0 \\ 0 \end{pmatrix},$$

assuming the motion is in the x-direction. Performing a Lorentz boost (with velocity v) to the rest frame of the test charge gives the new current density four-vector for the positive

charges as

$$\begin{aligned} j'^{\mu}_{+} &= \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -\lambda c \\ \lambda u \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \gamma(\lambda c - \beta \cdot 0) \\ \gamma(0 - \beta \cdot \lambda c) \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \gamma\lambda c \\ -\beta\lambda\gamma c \\ 0 \\ 0 \end{pmatrix} \end{aligned}$$

and the current density four-vector for the negative charges is

$$\begin{aligned} j'^{\mu}_{-} &= \begin{pmatrix} \gamma(-\lambda c - \beta\lambda u) \\ \gamma(\lambda u - \beta(-\lambda c)) \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} -\gamma\lambda(\beta u + c) \\ \gamma\lambda(u + \beta c) \\ 0 \\ 0 \end{pmatrix}. \end{aligned}$$

Hence, the new charge density of the positive charges is given by the first component of j'^{μ}_{+}:

$$\begin{aligned} c\lambda'_{+} &= \gamma\lambda c \\ \implies \lambda'_{+} &= \gamma\lambda \end{aligned}$$

and similarly for the negative particles,

$$\begin{aligned} c\lambda'_{-} &= -\gamma\lambda(\beta u + c) \\ \implies \lambda'_{-} &= -\gamma\lambda\left(\frac{vu}{c^2} + 1\right). \end{aligned}$$

The net charge density in the wire in the test charge's rest frame is therefore

$$\begin{aligned} \lambda' &= \lambda'_{+} + \lambda'_{-} \\ &= \gamma\lambda - \gamma\lambda\left(\frac{vu}{c^2} + 1\right) \\ &= \gamma\lambda\left(1 - \frac{vu}{c^2} - 1\right) \\ &= \gamma\lambda\frac{vu}{c^2}. \end{aligned}$$

Now, in this frame the test charge is stationary so cannot experience a magnetic force. The electric field near a line of charge is given by Coulomb's law as

$$E' = \frac{\lambda'}{2\pi\epsilon_0 r}$$

and so the test charge experiences an electric force of

$$F'_E = \frac{q\lambda'}{2\pi\epsilon_0 r} = \frac{q\gamma\lambda vu}{2\pi\epsilon_0 c^2 r}$$

but $\frac{1}{\epsilon_0 c^2} = \mu_0$ and thus

$$F'_E = \frac{q\gamma\lambda\mu_0 vu}{2\pi r} = \gamma F_B.$$

Comparison with our force transformation law in eq. (39) shows us that this means the magnetic force in the unprimed frame is equivalent to the electric force in the primed frame; the electric force is just the result of transforming the magnetic force, and vice versa.

This very nicely shows that electric and magnetic fields are really one and the same.

2.2.4 The Electromagnetic Field Tensor

a) We will use eqs. (29) and (30) to find expressions for each component of the electric and magnetic fields in terms of the vector and scalar potential.

Let $\vec{B} = \begin{pmatrix} B_x \\ B_y \\ B_z \end{pmatrix}$, $\vec{E} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix}$ and $\vec{A} = \begin{pmatrix} A_x \\ A_y \\ A_z \end{pmatrix}$. Evaluating the cross product in eq. (29) gives

$$B_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}, \tag{40}$$

$$B_y = \frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}, \tag{41}$$

$$B_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}. \tag{42}$$

Now similarly evaluating each component of eq. (30) leads to three more relationships:

$$E_x = -\frac{\partial \phi}{\partial x} - \frac{\partial A_x}{\partial t}, \tag{43}$$

$$E_y = -\frac{\partial \phi}{\partial y} - \frac{\partial A_y}{\partial t}, \tag{44}$$

$$E_z = -\frac{\partial \phi}{\partial z} - \frac{\partial A_z}{\partial t}. \tag{45}$$

Now, we defined contravariant four-potential as

$$A^\mu = \begin{pmatrix} \frac{\phi}{c} \\ A_x \\ A_y \\ A_z \end{pmatrix} = \begin{pmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}$$

and so since we are using the Minowski metric signature $(-,+,+,+)$ the covariant four-potential is

$$A_\mu = \begin{pmatrix} -\frac{\phi}{c} \\ A_x \\ A_y \\ A_z \end{pmatrix} = \begin{pmatrix} -A_0 \\ A_1 \\ A_2 \\ A_3 \end{pmatrix}.$$

Using this and the definition of four-gradient, we can re-write eqs. (40) to (45) using only four-gradient and four-potential:

$$B_x = \partial_2 A_3 - \partial_3 A_2, \tag{46}$$
$$B_y = \partial_3 A_1 - \partial_1 A_3, \tag{47}$$
$$B_z = \partial_1 A_2 - \partial_2 A_1, \tag{48}$$
$$E_x = c\partial_1 A_0 - c\partial_0 A_1, \tag{49}$$
$$E_y = c\partial_2 A_0 - c\partial_0 A_2, \tag{50}$$
$$E_z = c\partial_3 A_0 - c\partial_0 A_3. \tag{51}$$

b) Therefore, we can write all of the field components in one vector:

$$\begin{pmatrix} B_x \\ B_y \\ B_z \\ \frac{E_x}{c} \\ \frac{E_y}{c} \\ \frac{E_z}{c} \end{pmatrix} = \begin{pmatrix} \partial_2 A_3 - \partial_3 A_2 \\ \partial_3 A_1 - \partial_1 A_3 \\ \partial_1 A_2 - \partial_2 A_1 \\ \partial_1 A_0 - \partial_0 A_1 \\ \partial_2 A_0 - \partial_0 A_2 \\ \partial_3 A_0 - \partial_0 A_3. \end{pmatrix} \tag{52}$$

The symmetry in the right hand side inspires us to define a tensor to represent all of this; a useful quantity that we call a field strength tensor:

$$T_{\mu\nu} \coloneqq \partial_\mu A_\nu - \partial_\nu A_\mu.$$

c) If we switch the positions of the indices μ and ν, then we have

$$\begin{aligned} T_{\nu\mu} &= \partial_\nu A_\mu - \partial_\mu A_\nu \\ &= -(\partial_\mu A_\nu - \partial_\nu A_\mu) \\ &= -T_{\mu\nu} \end{aligned}$$

That is, flipping the indices negates the tensor. (The tensor is 'anti-symmetric'.) If the indices are the same ($\nu = \mu$) then (not using the summation convention)

$$\begin{aligned} T_{\nu\mu} = T_{\mu\mu} &= \partial_\mu A_\mu - \partial_\mu A_\mu \\ &= 0 \end{aligned}$$

and so the diagonals of the tensor are zero.

The field strength tensor is indexing through $\mu = 0, 1, 2, 3$ and $\nu = 0, 1, 2, 3$ and so has 16 entries altogether. Of these, flipping the indices accounts for 8 and setting the indices equal accounts for a further 2 (four in total) leaving 6 distinct entries. This makes sense as the vector in eq. (52) has 6 components!

d) We will now write the field strength tensor $T_{\mu\nu}$ out as a matrix, indexing the rows with μ

and the columns with ν:

$$
\begin{aligned}
T_{\mu\nu} &= \begin{pmatrix} T_{00} & T_{01} & T_{02} & T_{03} \\ T_{10} & T_{11} & T_{12} & T_{13} \\ T_{20} & T_{21} & T_{22} & T_{23} \\ T_{30} & T_{31} & T_{32} & T_{33} \end{pmatrix} \\
&= \begin{pmatrix} (\partial_0 A_0 - \partial_0 A_0) & (\partial_0 A_1 - \partial_1 A_0) & (\partial_0 A_2 - \partial_2 A_0) & (\partial_0 A_3 - \partial_3 A_0) \\ (\partial_1 A_0 - \partial_0 A_1) & (\partial_1 A_1 - \partial_1 A_1) & (\partial_1 A_2 - \partial_2 A_1) & (\partial_1 A_3 - \partial_3 A_1) \\ (\partial_2 A_0 - \partial_0 A_2) & (\partial_2 A_1 - \partial_1 A_2) & (\partial_2 A_2 - \partial_2 A_2) & (\partial_2 A_3 - \partial_3 A_2) \\ (\partial_3 A_0 - \partial_0 A_3) & (\partial_3 A_1 - \partial_1 A_3) & (\partial_3 A_2 - \partial_2 A_3) & (\partial_3 A_3 - \partial_3 A_3) \end{pmatrix} \\
&= \begin{pmatrix} 0 & -\frac{E_x}{c} & -\frac{E_y}{c} & -\frac{E_z}{c} \\ \frac{E_x}{c} & 0 & B_z & -B_y \\ \frac{E_y}{c} & -B_z & 0 & B_x \\ \frac{E_z}{c} & B_y & -B_x & 0 \end{pmatrix}.
\end{aligned}
$$

This tensor behaves very nicely under Lorentz transformations and can be used to represent all of the electromagnetic fields at a point in space.

2.2.5 The Transformations of Fields

a) We are given that to Lorentz boost the electromagnetic field matrix we apply the transformation, in tensor notation,

$$T'_{\mu\nu} = \Lambda_\mu{}^\lambda T_{\lambda\sigma} \Lambda^\sigma{}_\nu,$$

or in matrix notation

$$T' = \Lambda T \Lambda^\intercal$$

where $\Lambda^\intercal$ is the matrix transpose of Λ. Therefore, for a boost in the x-direction,

$$
\begin{aligned}
T' &= \Lambda \begin{pmatrix} 0 & -\frac{Ex}{c} & -\frac{Ey}{c} & -\frac{Ez}{c} \\ \frac{Ex}{c} & 0 & B_z & -B_y \\ \frac{Ey}{c} & -B_z & 0 & B_x \\ \frac{Ez}{c} & B_y & -B_x & 0 \end{pmatrix} \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\gamma\beta}{c}E_x & \frac{-\gamma}{c}E_x & \frac{-E_y}{c} & \frac{-E_z}{c} \\ \frac{\gamma}{c}E_x & \frac{-\beta\gamma}{c}E_x & B_z & -B_y \\ \frac{\gamma}{c}E_y + \beta\gamma B_z & -\gamma B_z - \frac{\beta\gamma}{c}E_y & 0 & B_x \\ \frac{\gamma}{c}E_z - \beta\gamma B_y & \gamma B_y - \frac{\beta\gamma}{c}E_z & -B_x & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \frac{\gamma^2\beta^2}{c}E_x - \frac{\gamma^2}{c}E_x & -\beta\gamma B_z - \frac{\gamma}{c}E_y & \gamma\beta B_y - \frac{\gamma}{c}E_z \\ \frac{\gamma^2}{c}E_x - \frac{\gamma^2\beta^2}{c}E_x & 0 & \gamma B_z - \frac{\gamma\beta}{c}E_y & \frac{\gamma\beta}{c}E_z - \gamma B_y \\ \frac{\gamma}{c}E_y + \gamma\beta B_z & -\gamma B_z - \frac{\gamma\beta}{c}E_y & 0 & B_x \\ \frac{\gamma}{c}E_z\gamma\beta B_y & \gamma B_y - \frac{\gamma\beta}{c}E_z & -B_x & 0 \end{pmatrix}.
\end{aligned}
$$

This is a rather large matrix, but by comparing its entries with the definition of the electromagnetic field tensor, we can identify the effects of the transformation:

$$E'_x = \gamma^2(1-\beta^2)E_x = E_x, \tag{53}$$
$$E'_y = \gamma(E_y + vB_z), \tag{54}$$
$$E'_z = \gamma(E_z - vB_y), \tag{55}$$
$$B'_x = B_x, \tag{56}$$
$$B'_y = \gamma\left(B_y - \frac{\beta}{c}E_z\right), \tag{57}$$
$$B'_z = \gamma\left(B_z + \frac{\beta}{c}E_y\right). \tag{58}$$

The rightmost terms of these six equations can be thought of as elements of the cross product $\vec{v} \times \vec{E}$ or $\vec{v} \times \vec{B}$ where $\vec{v} = \begin{pmatrix} v \\ 0 \\ 0 \end{pmatrix}$, and formulating the equations in this way leads nicely to the generalised field transformations given in the question paper.

b) We will now use the general form of the field transformation equations as given in the question paper. Consider a particle travelling at some velocity $\vec{v}$. In the particle's rest frame, the magnetic field is zero and the electric field given by Coulomb's law is

$$\vec{E} = \frac{q}{4\pi\epsilon_0 r^3}\vec{r}$$

where $\vec{r}$ is the position vector relative to the particle. Applying[3] a Lorentz transformation with velocity $-\vec{v}$ from frame S (where the particle has zero velocity) to frame S' (where

[3]We boost with velocity $-\vec{v}$ as we are going from the particle's rest frame 'back' to the frame in which it is travelling at velocity $\vec{v}$. This is essentially an *inverse* Lorentz transformation.

the particle has velocity $\vec{v}$), we know from the field transformation equations that in the direction of the boost, the new magnetic field is

$$B'_{\parallel} = B_{\parallel} = 0$$

and in the two directions perpendicular to the boost the new magnetic field is the two-dimensional vector

$$\begin{aligned}\vec{B}'_{\perp} &= \gamma\left(\vec{B}_{\perp} - \frac{1}{c^2}(-\vec{v}\times\vec{E})_{\perp}\right)\\ &= \gamma\left(0 + \frac{1}{c^2}\left(\vec{v}\times\frac{q}{4\pi\epsilon_0 r^3}\vec{r}\right)_{\perp}\right)\\ &= \frac{\gamma}{c^2}\left(\frac{q}{4\pi\epsilon_0 r^3}\right)(\vec{v}\times\vec{r})_{\perp}\\ &= \frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r})_{\perp}\end{aligned}$$

and so since the component of $\vec{v}\times\vec{r}$ parallel to the boost direction is zero by the definition of the cross product, we may write

$$\begin{aligned}\vec{B}' = \begin{pmatrix} B_{\parallel} \\ \vec{B}_{\perp}\end{pmatrix} &= \begin{pmatrix} 0 \\ \frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r})_{\perp}\end{pmatrix}\\ &= \frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r}).\end{aligned}$$

This is the Biot-Savart law for point charges. If we assume the direction parallel to the boost is the x-direction, then

$$\begin{aligned}\vec{B}' &= \frac{\gamma\mu_0 q}{4\pi r^3}\begin{pmatrix} v\\0\\0\end{pmatrix}\times\begin{pmatrix} r_x\\r_y\\r_z\end{pmatrix}\\ &= \frac{\gamma\mu_0 q}{4\pi r^3}\begin{pmatrix} 0\\-vr_z\\vr_y\end{pmatrix}\\ &= \frac{\gamma\mu_0 qv}{4\pi r^3}\begin{pmatrix} 0\\-r_z\\r_y\end{pmatrix}\end{aligned}$$

and this expanded form shows that the magnetic field is essentially circular in shape, with its centre of rotation as the axis of motion of the point charge. In other words, an electric current moving in a straight line will produce a circular magnetic field around it.

c) The above proof is really 'if and only if' anyway, meaning that we have already proven Coulomb's law given Biot-Savart's law for point charges. We'll do it explicitly for clarification though. If in our primed frame S' the magnetic field is given by Biot-Savart as

$$\vec{B}' = \frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r}),$$

then the fourth field transformation equation gives

$$\frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r})_{\perp} = \gamma\left(\vec{B}_{\perp} - \frac{1}{c^2}(-\vec{v}\times\vec{E})_{\perp}\right).$$

Since we know the magnetic field in frame S is zero,

$$\begin{aligned} \frac{\gamma\mu_0 q}{4\pi r^3}(\vec{v}\times\vec{r})_\perp &= -\gamma\frac{1}{c^2}(\vec{v}\times\vec{E})_\perp \\ \implies \frac{q}{4\pi\epsilon_0 r^3}(\vec{v}\times\vec{r})_\perp &= (\vec{v}\times\vec{E})_\perp \\ \implies \vec{E}_\perp &= \frac{q}{4\pi\epsilon_0 r^3}\vec{r}_\perp \end{aligned}$$

which is Coulomb's law as desired. Of course, this is only in the two dimensions perpendicular to the boost, because the forces are unchanged in the parallel direction and so there would be no way to derive Coulomb's law in this direction from Biot-Savart.[4]

The interchangeability of Coulomb's law and Biot-Savart's law in different reference frames shows, as did the question about particles in a wire, that electric and magnetic forces are truly the same phenomenon, with the magnetic force becoming an electric force if you boost to the right reference frame and vice versa.

Indeed, although neither the electric nor magnetic forces are covariant[5] under Lorentz transformations, the total electromagnetic force *is*.

2.2.6 Field Transformation Problems

Problem: Moving Solenoid

Consider a stationary densely-wound solenoid of infinite length with N turns *per unit length* through which is flowing a current I. At every point there is a uniform charge density, and so in the centre of the solenoid there is no net electric field (due to the principle of superposition).

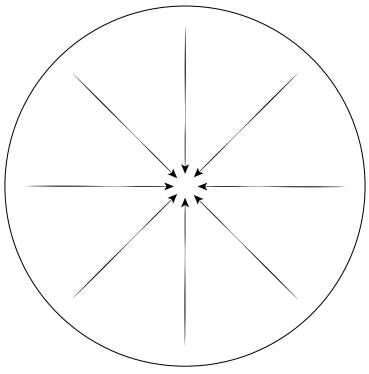


Figure 2: The electric field is radially inwards at all points and so cancels completely at the centre.

However, there is a circular flow of current and so there is *indeed* a magnetic field. As we have seen, a circular magnetic field is generated perpendicular to a moving point charge (and therefore a current). Consider one 'ring' of the solenoid (we ignore the fact that it doesn't actually join up).

[4] Alternatively, we could have first found the electric field in S' and used the full fields in S' to do a reverse transformation back to S, showing that Coulomb's law holds. However, such an argument would be inescapably circular since the fields in frame S' depend on Coulomb's law in S in the first place.

[5] Here covariant is defined to mean 'transforms according to the Lorentz transformations'. In this case specifically it means for the force to transform according to the force law derived in eq. (39).

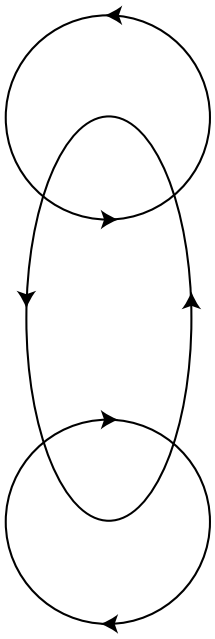


Figure 3: The middle loop shows the direction of current flow, and the top and bottom loops show the generated magnetic field lines.

This circular magnetic field generated from every segment of current on that ring will be pointing in the same direction at the centre of the ring, and so by the principle of superposition there will be a very strong homogeneous magnetic field pointing along the central axis of the solenoid. No such reinforcement occurs outside the solenoid, however, and so the magnetic field will be quite weak.

Knowing the direction of the internal magnetic field, we may apply Ampère's law (the fourth Maxwell equation, given in eq. (28)) to find its strength.

This gives

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t}$$

but the electric field is always zero, so

$$\vec{\nabla} \times B = \mu_0 \vec{j}\,.$$

Using the divergence theorem gives the integral form of this equation as

$$\oint_C \vec{B} \cdot d\vec{l} = \mu_0 I_C$$

where C is some closed loop, $\vec{l}$ is an infinitesimal tangent element of C and I_C is the current enclosed passing through the surface enclosed by C.

Let's pick our closed loop to be a rectangle through the centre of the solenoid with two sides of length h parallel to the solenoid.

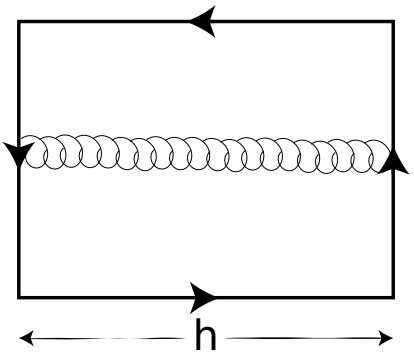


Figure 4: We choose a rectangle of width h as our closed loop for integration.

Clearly the magnetic field has no component parallel to the two short sides of the rectangle, and the magnetic field outside the solenoid is negligible, so the total line integral is just

$$\oint_C \vec{B} \cdot d\vec{l} = Bh$$

where B is the magnitude of the magnetic field in the middle of the solenoid. So, Ampère's law gives

$$Bh = \mu_0 I_C.$$

This length h of rectangle contains Nh turns of the solenoid and so the enclosed current is

$$I_C = NhI$$

so that

$$\begin{aligned} Bh &= \mu_0 NhI \\ \implies B &= \mu_0 NI. \end{aligned}$$

Hence, in the solenoid's rest frame there is in the middle of the solenoid a magnetic field parallel to the x-axis of magnitude $\mu_0 NI$, and no other electromagnetic fields.

Let's now consider the frame in which the solenoid is moving along the x-axis with velocity v. Performing a Lorentz boost to this frame, the equations for field transformations show that the perpendicular components remain unchanged, so there will still be a magnetic field of magnitude $\mu_0 NI$ along the x-axis, and similarly after the boost there will still be no other field components.

Correction for Maxwell's Equations

Ampère's law (the fourth Maxwell equation, eq. (28)) states that

$$c^2 \vec{\nabla} \times \vec{B} = \frac{\vec{j}}{\epsilon_0} + \frac{\partial \vec{E}}{\partial t}$$

and dividing through by c^2 gives

$$\vec{\nabla} \times \vec{B} = \mu_0 \vec{j} + \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t}.$$

Now taking the divergence of both sides,

$$\begin{aligned}\vec{\nabla}\cdot(\vec{\nabla}\times\vec{B}) &= \vec{\nabla}\cdot(\mu_0\vec{j}+\frac{1}{c^2}\frac{\partial\vec{E}}{\partial t})\\ &= \mu_0\vec{\nabla}\cdot\vec{j}+\frac{1}{c^2}\vec{\nabla}\cdot\frac{\partial\vec{E}}{\partial t}\\ &= \mu_0\vec{\nabla}\cdot\vec{j}+\frac{1}{c^2}\frac{\partial}{\partial t}\vec{\nabla}\cdot\vec{E}\end{aligned}$$

and substituting in Gauss' law (the first Maxwell equation),

$$\begin{aligned}\vec{\nabla}\cdot(\vec{\nabla}\times\vec{B}) &= \mu_0\vec{\nabla}\cdot\vec{j}+\frac{1}{c^2}\frac{\partial}{\partial t}\frac{\rho}{\epsilon_0}\\ &= \mu_0\vec{\nabla}\cdot\vec{j}+\mu_0\frac{\partial\rho}{\partial t}\,.\end{aligned}$$

However, the divergence of the curl of a vector field is always zero[6]. Indeed, this makes sense as $\vec{P}\cdot(\vec{P}\times\vec{Q})\equiv 0$ for any vectors $\vec{P}$ and $\vec{Q}$, because the cross product $\vec{P}\times\vec{Q}$ is perpendicular to both $\vec{P}$ and $\vec{Q}$ and so makes a zero dot product with $\vec{P}$. Using $\vec{\nabla}\cdot(\vec{\nabla}\times\vec{B})\equiv 0$ reduces the equation to

$$\begin{aligned}0 &= \mu_0\vec{\nabla}\cdot\vec{j}+\mu_0\frac{\partial\rho}{\partial t}\\ \iff \vec{\nabla}\cdot\vec{j} &= -\frac{\partial\rho}{\partial t}\end{aligned}$$

and this is simply the continuity equation we derived earlier; that is, it ensures conservation of charge.

Since this must always be true, we *do* require the $\frac{\partial\vec{E}}{\partial t}$ term otherwise we would come to

$$\vec{\nabla}\cdot\vec{j}=0$$

which is not always true. In fact, this second equation holds true only when

$$\frac{\partial\rho}{\partial t}=0$$

or in other words, when the charge density is static (does not change over time). So the corrected form of Ampère's law must be used whenever we have time-dependent charge density.

An example of such a problem would be a circuit with a capacitor. The charge density on the plates of the capacitor is certainly not constant and so we must use the converted version of Ampère's law to find the magnetic field around a capacitor, no matter what the reference frame.

End of Submission.

References

[1] Yong Feng Gui and Wen-Bin Dou. "A rigorous and completed statement on helmholtz theorem". In: *Progress in Electromagnetics Research* 69 (2007), pp. 287–304.

[6] This is a consequence of Helmholtz's theorem, otherwise known as the fundamental theorem of vector calculus.

Part IV

Solutions to problems

Chapter 24

PROMYS Europe 2017 — application problem set

Miss Brownlee recommended this summer course to me in January 2017, and it looked amazing so I decided to apply. I worked through the application problems for about a month and wrote them up over half term. Sadly I didn't get in but I still had a lot of fun doing the problems (though they were a bit too number-theoretic for my tastes).

PROMYS Europe 2017 — Application Problem Set

Damon Falck
Highgate School, London

March 2017

In presenting my solutions to this problem set, I have attempted to combine 'formal proof' with describing my thought process. For this reason, some parts may go into perhaps irrelevant detail as to the order of my reasoning.

Throughout these solutions I use $\mathbb{Z}^+$ to refer to the set of *positive* integers and $\mathbb{Z}^*$ to refer to that of *nonnegative* integers.

Question 1

Calculate each of the following:

$$1^3 + 5^3 + 3^3 = ??$$
$$16^3 + 50^3 + 33^3 = ??$$
$$166^3 + 500^3 + 333^3 = ??$$
$$1666^3 + 5000^3 + 3333^3 = ??$$

What do you see? Can you state and prove a generalization of your observations?

To start off with, we'll calculate the equations given:

$$1^3 + 5^3 + 3^3 = 153$$
$$16^3 + 50^3 + 33^3 = 165033$$
$$166^3 + 500^3 + 333^3 = 166500333$$
$$1666^3 + 5000^3 + 3333^3 = 166650003333.$$

It can be seen that from left to right, the digits in each term of the left hand side are the same as the digits in the right hand side. We can rewrite the above equations as follows:

$$(1)^3 + (5)^3 + (3)^3 = (1) \cdot 10^2 + (5) \cdot 10^1 + (3)$$
$$(10+6)^3 + (50)^3 + (30+3)^3 = (10+6) \cdot 10^4 + (50) \cdot 10^2 + (10+6)$$

et cetera, and so as we continue downwards, we propose that

$$\begin{aligned}(10^n + 6 \cdot 10^{n-1} + 6 \cdot 10^{n-2} + \cdots + 6 \cdot 10^0)^3 + (5 \cdot 10^n)^3 + (3 \cdot 10^n + 3 \cdot 10^{n-1} + \cdots + 3 \cdot 10^0)^3 \\ = (10^n + 6 \cdot 10^{n-1} + 6 \cdot 10^{n-2} + \cdots + 6 \cdot 10^0) \cdot 10^{2(n+1)} \\ + (5 \cdot 10^n) \cdot 10^{n+1} + (3 \cdot 10^n + 3 \cdot 10^{n-1} + \cdots + 3 \cdot 10^0)\end{aligned}$$

for any $n \in \mathbb{Z}^+$. Writing this formally with summation notation, we get

$$\left(10^n + \sum_{i=0}^{n-1} 6 \cdot 10^i\right)^3 + (5 \cdot 10^n)^3 + \left(\sum_{i=0}^{n} 3 \cdot 10^i\right)^3 = \left(10^n + \sum_{i=0}^{n-1} 6 \cdot 10^i\right) \cdot 10^{2(n+1)} + (5 \cdot 10^n) \cdot 10^{n+1} + \sum_{i=0}^{n} 3 \cdot 10^i.$$

Now we want to make this look a bit nicer. Manipulating it a little to replace all the sums with $\sum_{i=0}^{n} 10^i$, we now come to the following.

Proposition 1.1. *For any $n \in \mathbb{Z}^+$,*

$$\left(10^n - 6 \cdot 10^n + 6\sum_{i=0}^{n} 10^i\right)^3 + (5 \cdot 10^n)^3 + \left(3\sum_{i=0}^{n} 10^i\right)^3 = \left(10^n - 6 \cdot 10^n + 6\sum_{i=0}^{n} 10^i\right) \cdot 10^{2(n+1)} + (5 \cdot 10^n) \cdot 10^{n+1} + 3\sum_{i=0}^{n} 10^i.$$

Let $\sigma = \sum_{i=0}^{n} 10^i$ and let $\mu = 10^n$. Now we can rearrange and simplify to a much nicer equation:

$$\begin{aligned}
&(10^n - 6 \cdot 10^n + 6\sigma)^3 + (5 \cdot 10^n)^3 + (3\sigma)^3 = (10^n - 6 \cdot 10^n + 6\sigma) \cdot 10^{2n+2} + (5 \cdot 10^n) \cdot 10^{n+1} + 3\sigma \\
&\implies (-5\mu + 6\sigma)^3 + (5\mu)^3 + (3\sigma)^3 = (-5\mu + 6\sigma)(10^2\mu^2) + (5\mu)(10\mu) + 3\sigma \\
&\implies (6\sigma - 5\mu)^3 + 125\mu^3 + 27\sigma^3 = 100\mu^2(6\sigma - 5\mu) + 50\mu^2 + 3\sigma.
\end{aligned}$$

Expanding fully,

$$\begin{aligned}
&(6\sigma)^3 + 3(6\sigma)^2(-5\mu) + 3(6\sigma)(-5\mu)^2 + (-5\mu)^3 + 125\mu^3 + 27\sigma^3 = 600\mu^2\sigma - 500\mu^3 + 50\mu^2 + 3\sigma \\
&\implies 216\sigma^3 - 540\mu\sigma^2 + 450\mu^2\sigma - 125\mu^3 + 125\mu^3 + 27\mu^3 - 600\mu^2\sigma + 500\mu^3 - 50\mu^2 - 3\sigma = 0.
\end{aligned}$$

Collecting terms of σ,

$$\begin{aligned}
&243\sigma^3 - 540\mu\sigma^2 - (150\mu^2 + 3)\sigma + 50\mu^2(10\mu - 1) = 0 \\
&\implies \sigma^3 - \frac{20}{9}\sigma^2 - \frac{50\mu^2 + 1}{3}\sigma + \frac{50\mu^2(10\mu - 1)}{9} = 0 \qquad (1)
\end{aligned}$$

and hence we arrive at a nice monic cubic in σ. Remembering that $\sigma = \sum_{i=0}^{n} 10^i$, perhaps solving for σ will let us find a nice expression for the summation.

Let $f(\sigma) = \sigma^3 - \frac{20}{9}\sigma^2 - \frac{50\mu^2+1}{3}\sigma + \frac{50\mu^2(10\mu-1)}{9}$, the cubic from eq. (1). After trying a few factors of the constant term, we find that $f\left(\frac{10\mu-1}{9}\right) = 0$ and so by the factor theorem, $\left(\sigma - \frac{10\mu-1}{9}\right)$ must be a factor. Hence, either

$$\sigma = \frac{10\mu - 1}{9} \qquad (2)$$

or

$$\frac{9f(\sigma)}{10\mu - 1} = 0.$$

We'll leave the second solution (a quadratic) for now, as the first looks like it will more likely lead us somewhere. From eq. (2), by the definitions of μ and σ we come to

$$\sum_{i=0}^{n} 10^i = \frac{10 \cdot 10^n - 1}{9}.$$

We now propose the following theorem based on our observations of the initial equations, which we will then prove by induction.

Theorem 1.2. *Let n be any positive integer. Then,*

$$\sum_{i=0}^{n} 10^i = \frac{10^{n+1}-1}{9}.$$

Proof. We will prove by induction the above theorem . For our base case, let $n = 1$. So,

$$\sum_{i=0}^{n} 10^i = 10^0 + 10^1 = 11$$

and

$$\frac{10^{n+1}-1}{9} = \frac{100-1}{9} = 11,$$

therefore the equality is valid in the case $n = 1$. Now assume the equality holds for $n = k$ (our induction hypothesis):

$$\sum_{i=0}^{k} 10^i = \frac{10^{k+1}-1}{9}.$$

We can now show it also to be true for $n = k + 1$ (our inductive step):

$$\begin{aligned}\sum_{i=0}^{k+1} 10^i &= 10^{k+1} + \sum_{i=0}^{k} 10^i \\ &= 10^{k+1} + \frac{10^{k+1}-1}{9} \\ &= \frac{9 \cdot 10^{k+1} + 10^{k+1} - 1}{9} \\ &= \frac{10 \cdot 10^{k+1} - 1}{9} \\ &= \frac{10^{(k+1)+1} - 1}{9}.\end{aligned}$$

Hence by mathematical induction,

$$\sum_{i=0}^{n} 10^i = \frac{10^{n+1}-1}{9}$$

for all n. □

Question 2

> The *repeat* of a positive integer is obtained by writing it twice in a row (so, for example, the repeat of 2017 is 20172017). Is there a positive integer whose repeat is a perfect square? If so, how many such positive integers can you find?

Let us start by constructing an n-digit positive integer s. Such an integer can be expressed as

$$s = \sum_{i=0}^{n-1} 10^i s_i$$

where $\{s_i\}_{i=0}^{n-1}$ are the digits of s from right to left. Now the repeat $R(s)$ of integer s is obtained by summing s with $s \cdot 10^n$ (as $s \cdot 10^n$ is s 'shifted' by n digits to the left). So,

$$\begin{aligned} R(s) &= s + s \cdot 10^n \\ &= s(10^n + 1). \end{aligned} \tag{3}$$

We now come to the following result.

Lemma 2.1. *Let s be an n-digit positive integer. If $10^n - 1$ is square-free, then the repeat of s cannot be a perfect square.*

Proof. If the repeat of s is a perfect square, then by eq. (3) we can say

$$s(10^n + 1) = k^2$$

for some $k \in \mathbb{Z}^+$. We know s has n digits by definition, and $10^n + 1$ must have $n + 1$ digits, and so

$$s < 10^n + 1. \tag{4}$$

If the prime factorisation of k is

$$k = \prod_{i=1}^{m} p_i^{\alpha_i}$$

where $p_1, p_2, \ldots, p_m$ are primes and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are nonnegative integers (and m represents how many distinct primes make up the prime factorisation of k) then we know

$$s(10^n + 1) = \left(\prod_{i=1}^{m} p_i^{\alpha_i} \right)^2 = \prod_{i=1}^{m} p_i^{2\alpha_i}.$$

Due to eq. (4), at least one of the prime factors $p_i^{\alpha_i}$ must become $p_i^{\alpha_i + \beta}$ where $\beta \in \mathbb{Z}^+$ (that is, must occur at least twice) in the prime factorisation of $10^n + 1$; otherwise, we'd require $s \geqslant 10^n + 1$, which is a contradiction. So, the repeat $R(s) = s(10^n + 1)$ can only be a perfect square if there is a squared number (prime or otherwise) that divides $10^n + 1$. □

Our next result follows:

Lemma 2.2. *For every n such that $10^n + 1$ has a square factor, there exists at least one n-digit integer s whose repeat is a perfect square.*

Proof. From lemma 2.1, let

$$10^n + 1 = a^2 b$$

where $a, b \in \mathbb{Z}^+$ and b is square-free. (Note that, assuming we know the prime factorisation of $10^n + 1$, we have 'absorbed' all square factors into a^2, so that a is not necessarily prime.)

Suppose that $R(s)$ *is* a perfect square; we will show that this is possible. By eq. (3), $a^2 b \cdot s$ must be a perfect square, so bs must too. We know b is square-free, so $b^2 \mid bs$ which implies $b \mid s$. So let $s = br$ with $r \in \mathbb{Z}^+$. Therefore $b \cdot bc = b^2 r$ is a perfect square, and thus so is r. So, letting $c^2 = r$ with $c \in \mathbb{Z}^+$,

$$R(s) = a^2 b^2 c^2 = (abc)^2$$

where we know a and b from the prime factorisation of $10^n + 1$. We therefore want to choose $c \in \mathbb{Z}^+$ such that s is precisely n digits long. If we can find c that satisfies this, then we know s and its repeat $R(s)$, and we know that $R(s)$ is a perfect square.

We require $s = bc^2$ to have n digits and so we need

$$10^{n-1} \leqslant bc^2 < 10^n.$$

We also know that $10^n + 1 = a^2 b$ is $n+1$ digits long, so dividing bc^2 by $a^2 b$, we get

$$\frac{10^{n-1}}{10^n + 1} \leqslant \frac{bc^2}{a^2 b} < \frac{10^n}{10^n + 1}.$$

Because we're dealing with integers, we can simplify this to

$$\frac{1}{10} < \frac{c^2}{a^2} < 1.$$

Therefore, we must choose $\sqrt{r}$ such that

$$\frac{a}{\sqrt{10}} < c < a.$$

As long as $a > 1$ this is possible, and this must always be true for otherwise a^2 would not count as a square number. Hence, it is always possible to choose c such that s has n digits and $R(s)$ is a perfect square. □

The problem is therefore reduced to how many integers n there are such that $10^n + 1$ has a square factor.

By trial and error, we find that $n = 11$ produces the result desired, and so there is at least one n for which this is true. After some experimenting, we notice that all $10^1 1w + 1$ where w is odd appear to have a square factor. We can prove this as follows.

Lemma 2.3. *If we find some n for which $10^n + 1$ is not square-free, then we may generate infinitely many such numbers.*

Proof. Let $f(x) = x^w + 1$. By the factor theorem, $(x + 1)$ is a factor of $x^w + 1$ if and only if $f(-1) = 0$. We see that $f(-1) = (-1)^w + 1$ and so $f(-1) = 0$ if and only if w is odd.

Therefore, if $x = 10^n$ then we know $10^n + 1 \mid 10^w n + 1$. Consequently, if $10^n + 1$ has a square factor then so will $10^w n + 1$ for all odd w. Since there is an infinity of odd numbers, we can produce infinitely many numbers of the form $10^w n + 1$ that are not square-free. □

We have shown all that we need to; our main theorem now follows.

Theorem 2.4. *There are infinitely many positive integers s whose repeats are perfect squares.*

Such a repeated number can always be found by the formula

$$R(s) = (10^{11w} + 1)bc^2$$

where w is odd and positive, b is the square-free part of $10^{11w} + 1$ and c is any positive integer such that

$$\sqrt{\frac{10^{11w} + 1}{10b}} < c < \sqrt{\frac{10^{11w} + 1}{b}}.$$

Proof. By lemma 2.3 there are infinitely many integers n such that $10^n + 1$ is a perfect square. By lemma 2.2, for every such number there is at least one integer s whose repeat is a perfect square. Therefore there are infinitely many such integers s. The formula given follows from lemma 2.3, the proof of lemma 2.2 and the fact that $11^2 \mid 10^{11} + 1$. □

We can now generate the first few square repeat numbers. The prime factorisation of $10^1 1 + 1$ is $11^2 \cdot 23 \cdot 4093 \cdot 8779$, and so $a = 11$ and $b = 23 \cdot 4093 \cdot 8779 = 826,446,281$. Hence we require

$$\frac{11}{\sqrt{10}} < c < 11$$

which is the same as

$$4 \leqslant c \leqslant 10.$$

Hence the following four examples come from $c = 4, 5, 6, 7$ respectively:

$$13223140496\ 13223140496 = 36,363,636,364^2$$
$$20661157025\ 20661157025 = 45,454,545,455^2$$
$$29752066116\ 29752066116 = 54,545,454,546^2$$
$$40495867769\ 40495867769 = 63,636,363,637^2.$$

Question 3

> A lattice point is a point (x, y) in the plane, both of whose coordinates are integers. It is easy to see that every lattice point can be surrounded by a small circle which excludes all other lattice points from its interior. It is not much harder to see that it is possible to draw a circle that has exactly two lattice points in its interior, or exactly 3, or exactly 4.
>
> (figure)
>
> Do you think that for every positive integer n there is a circle in the plane containing exactly n lattice points in its interior? Justify your answer.

After much thought, we note the possibility of choosing a point $(x, y) \in \mathbb{R}^2$ such that no circle with centre (x, y) and radius $r \in \mathbb{R}$ has more than one lattice point on its circumference. In other words, the distance between every lattice point and the point (x, y) is unique. If such a point exists, then by starting with $r = 0$ and gradually increasing r to infinity, one by one a new lattice point will be reached by the circumference of the circle, and so one by one the number of lattice points contained within the circle will increase.

Lemma 3.1. *It is possible to choose $(x, y) \in \mathbb{R}^2$ such that no circle with centre (x, y) can have more than one lattice point on its circumference.*

Proof. Let us assume that two lattice points (a, b) and (c, d) such that $a, b, c, d \in \mathbb{Z}$ lie on the circumference of a circle with centre (x, y). Thus, we can say

$$\left\| \begin{pmatrix} a \\ b \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} \right\| = \left\| \begin{pmatrix} c \\ d \end{pmatrix} - \begin{pmatrix} x \\ y \end{pmatrix} \right\|$$

and so

$$\left\| \begin{pmatrix} a-x \\ b-y \end{pmatrix} \right\| = \left\| \begin{pmatrix} c-x \\ d-y \end{pmatrix} \right\|$$
$$\implies \sqrt{(a-x)^2 + (b-y)^2} = \sqrt{(c-x)^2 + (d-y)^2}$$
$$\implies (a-x)^2 + (b-y)^2 = (c-x)^2 + (d-y)^2.$$

We are searching for some way to choose x and y so as to lead to a contradiction here. Expanding and rearranging, we come to

$$a^2 - 2ax + x^2 + b^2 - 2by + y^2 = c^2 - 2cx + x^2 + d^2 - 2dy + y^2$$
$$\implies a^2 + b^2 - c^2 - d^2 = 2ax + 2by - 2cx - 2dy.$$

Now by the definition of a, b, c, d the left hand side must be integral, and so

$$2ax + 2by - 2cx - 2dy = m$$

where $m \in \mathbb{Z}$. This is the same as

$$x(2a - 2c) + y(2b - 2d) = m$$

and so as $2a - 2c \in \mathbb{Z}$ and $2b - 2d \in \mathbb{Z}$ for the same reason, we can write

$$px + qy = m$$

with $p, q \in \mathbb{Z}$. So we want to choose x and y so that $px+qy$ cannot possibly be an integer. Choosing two irrational numbers will often satisfy this property, and so let us for instance choose $\sqrt{2}$ and $\sqrt{3}$.

Then $px + qy = \sqrt{3}p + \sqrt{2}q$ which cannot be integral. However, m is an integer by definition so we reach a contradiction. Therefore if $x = \sqrt{2}$ and $y = \sqrt{3}$ then it is impossible to have two lattice points simultaneously on the circumference of any circle with centre (x, y). □

Therefore we come to our main theorem:

Theorem 3.2. *For any $n \in \mathbb{Z}^+$ it is possible to draw a circle which contains precisely n lattice points within its circumference.*

Proof. Every lattice point is a finite distance from the centre of a circle, so by gradually increasing the radius of a circle from zero, the circumference at some point pass through any given lattice point.

It follows from lemma 3.1 therefore that for a circle with centre $(\sqrt{2}, \sqrt{3})$ (or indeed any similar combination of irrational numbers), by increasing the radius gradually from zero any arbitrary number of lattice points may be contained within the circle's circumference. □

Question 4

According to the Journal of Irreproducible Results, any obtuse angle is a right angle!

(figure)

Here is their argument. Given the obtuse angle x, we make a quadrilateral $ABCD$ with $\angle DAB = x$, and $\angle ABC = 90°$, and $AD = BC$. Say the perpendicular bisector to DC meets the perpendicular bisector to AB at P. Then $PA = PB$ and $PC = PD$. So the triangles PAD and PBC have equal sides and are congruent. Thus $\angle PAD = \angle PBC$. But PAB is isosceles, hence $\angle PAB = \angle PBA$. Subtracting, gives $x = \angle PAD - \angle PAB = \angle PBC - \angle PBA = 90°$. This is a preposterous conclusion — just where is the mistake in the "proof" and why does the argument break down there?

The critical error in this "proof" is the assumption that $x = \angle PAD - \angle PBA$. (All statements up until this point are correct.) In fact, in the setup described, $x = 360° - \angle PAD - \angle PAB$. This leads to the alternative, correct, conclusion that

$$x = 360° - \angle PBC - \angle PBA = 360° - (\angle PBA + 90°) - \angle PBA = 270° - 2 \cdot \angle PBA.$$

We come to this finding by noticing that the construction of the diagram given in the question is incorrect; in reality we must have $\angle ADC < \angle PDC$ whereas in the diagram $\angle PDC < \angle ADC$ (ray DP is impossibly placed). The proof of this is the very contradiction reached in the question.

Theorem 4.1. $x = 360° - \angle PAD - \angle PAB$.

Proof. By the (correct) argument given in the question, $\triangle PAD \cong \triangle PBC$. Therefore either

$$x = \angle PAD - \angle PBA \tag{5}$$

or

$$x = 360^\circ - \angle PAD - \angle PAB \tag{6}$$

depending on which side of ray DP point A lies at. Equation 5 leads to the absurd conclusion in the question, and so by contradiction we must have eq. (6). □

Question 5

A unit fraction is a fraction of the form $\frac{1}{n}$ where n is a positive integer. Note that the unit fraction $\frac{1}{11}$ can be written as the sum of two unit fractions in the following three ways:

$$\frac{1}{11} = \frac{1}{12} + \frac{1}{132} = \frac{1}{22} + \frac{1}{22} = \frac{1}{132} + \frac{1}{12}.$$

Are there any other ways of decomposing $\frac{1}{11}$ into the sum of two unit fractions? In how many ways can we write $\frac{1}{60}$ as the sum of two unit fractions? More generally, in how many ways can the unit fraction $\frac{1}{n}$ be written as the sum of two unit fractions? In other words, how many ordered pairs (a, b) of positive integers a, b are there for which

$$\frac{1}{n} = \frac{1}{a} + \frac{1}{b} \ ?$$

The first thing we notice is the following.

Lemma 5.1. *Where N is the number of distinct positive integer factors of n^2, the number of ordered pairs (a, b) such that*

$$\frac{1}{n} = \frac{1}{a} + \frac{1}{b}$$

is N.

Proof. Let us do some algebraic experimenting. If

$$\frac{1}{n} = \frac{1}{a} + \frac{1}{b} \tag{7}$$

with $a, b, n \in \mathbb{Z}^+$, then we note that

$$\frac{1}{a} < \frac{1}{n} \implies a > n$$

and

$$\frac{1}{b} < \frac{1}{n} \implies b > n.$$

Therefore, we can rewrite a and b as

$$a = n + p, \tag{8}$$
$$b = n + q \tag{9}$$

where $p, q \in \mathbb{Z}^+$, and so

$$\frac{1}{n} = \frac{1}{n+p} + \frac{1}{n+q}. \tag{10}$$

This leads us to a rather nice result; combining the fractions gives

$$\frac{1}{n} = \frac{(n+p)+(n+q)}{(n+p)(n+q)}$$

and so

$$\begin{aligned} n &= \frac{(n+p)(n+q)}{2n+p+q} \\ \implies 2n^2 + np + nq &= n^2 + np + nq + pq \\ \implies n^2 &= pq. \end{aligned} \tag{11}$$

Now because of eqs. (8) and (9), we know that for any particular n the number of ordered pairs (p, q) satisfying eq. (11) is equal to the number of ordered pairs (a, b) satisfying eq. (7).

The number of distinct ordered pairs of positive factors of any natural number is equal to the number of distinct positive factors of that number. Therefore if N is the number of distinct positive factors of N then the number of ordered pairs (p, q) is N and so is the number of ordered pairs (a, b). □

So, in the case $n = 11$, we can list the factor pairs (p, q) of $11^2 = 121$:

$$(p, q) = \begin{cases} (121, 1), \\ (11, 11), \\ (1, 121). \end{cases}$$

Thus, by eq. (10) the possible ways to split up $\frac{1}{11}$ are:

$$\frac{1}{11} = \frac{1}{11+p} + \frac{1}{11+q} = \begin{cases} \frac{1}{11+1} + \frac{1}{11+121} = \frac{1}{12} + \frac{1}{132}, \\ \frac{1}{11+11} + \frac{1}{11+11} = \frac{1}{22} + \frac{1}{22}, \\ \frac{1}{11+121} + \frac{1}{11+1} = \frac{1}{132} + \frac{1}{12}. \end{cases}$$

(In this case, $N = 3$.) Therefore we can confirm that there are no ways other than those in the question to decompose $\frac{1}{11}$.

Next, in the case $n = 60$, our factor pairs of $60^2 = 3600$ are, somewhat more laboriously,

$$(p, q) = \begin{cases} (1, 3600),\ (3600, 1), \\ (2, 1800),\ (1800, 2), \\ (3, 1200),\ (1200, 3), \\ (4, 900),\ (900, 4), \\ (5, 720),\ (720, 5), \\ (6, 600),\ (600, 6), \\ (8, 450),\ (450, 8), \\ (9, 400),\ (400, 9), \\ (10, 360),\ (360, 10), \\ (12, 300),\ (300, 12), \\ (15, 240),\ (240, 15), \\ (16, 225),\ (255, 16), \\ (18, 200),\ (200, 18), \\ (20, 180),\ (180, 20), \\ (24, 150),\ (150, 24), \\ (25, 144),\ (144, 25), \\ (30, 120),\ (120, 30), \\ (36, 100),\ (100, 36), \\ (40, 90),\ (90, 40), \\ (45, 80),\ (80, 45), \\ (48, 75),\ (75, 48), \\ (50, 72),\ (72, 50), \\ (60, 60). \end{cases}$$

Note: these are only listed here for demonstration purposes — we will confirm this result anyway by generalising.

It can be seen there are a whopping 45 such pairs, and so there are precisely $N = 45$ ways to write $\frac{1}{60}$ as the sum of two unit fractions. I shan't list them here.

Now looking at any n, we can ask the question of how many unique positive factors n^2 has. Our next result is just this:

Lemma 5.2. *If the prime factorisation of $n \in \mathbb{Z}^+$ is*

$$n = \prod_{i=1}^{m} s_i^{\alpha_i}$$

then the number N of distinct positive factors of n^2 is

$$N = \prod_{i=1}^{m} (2\alpha_i + 1).$$

Proof. Let us assume that we know the prime factorisation of n. Thus n can always be written as

$$n = \prod_{i=1}^{m} s_i^{\alpha_i}$$

where $s_1, s_2, \ldots, s_m$ are primes and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are integers (and m reflects how many *distinct* primes make up the prime factorisation of n). It follows, therefore, that the prime factorisation of n^2 is

$$n^2 = \left(\prod_{i=1}^{m} s_i^{\alpha_i} \right)^2 = \prod_{i=1}^{m} s_i^{2\alpha_i}$$

Now, any positive factor of n^2 will therefore have the form

$$\prod_{i=1}^{m} s_i^{\beta_i}$$

where $\beta_i \in \mathbb{Z}^*, \beta_i \leqslant 2\alpha_i$. So, there are $2\alpha_i + 1$ possible values for every β_i; namely, $0, 1, 2, \ldots, 2\alpha_i$. Therefore, the number of distinct positive factors of n^2 must be

$$(2\alpha_1 + 1)(2\alpha_2 + 1) \cdots (2\alpha_m + 1) = \prod_{i=1}^{m} (2\alpha_i + 1) = N,$$

equal to the number of distinct ordered positive factor pairs. □

Therefore by combination of lemmas 5.1 and 5.2 we come to the conclusion that:

Theorem 5.3. *The number N of ordered pairs (a, b) with $a, b \in \mathbb{Z}^+$ satisfying the Diophantine equation*

$$\frac{1}{n} = \frac{1}{a} + \frac{1}{b}$$

is

$$N = \prod_{i=1}^{m} (2\alpha_i + 1)$$

where the prime factorisation of n is

$$n = \prod_{i=1}^{m} s_i^{\alpha_i}.$$

Proof. Can be proved by combination of the proofs of lemmas 5.1 and 5.2. □

Reconsidering the case $n = 60$, we know that the prime factorisation of 60 is

$$60 = 2^2 \cdot 3^1 \cdot 5^1.$$

Therefore, in this case

$$N = (2 \cdot 2 + 1)(2 \cdot 1 + 1)(2 \cdot 1 + 1) = 45,$$

confirming our previous finding.

Question 6

> Let's agree to say that a positive integer is *prime-like* if it is not divisible by 2, 3, or 5. How many prime-like positive integers are there less than 100? less than 1000? A positive integer is *very prime-like* if it is not divisible by any prime less than 15. How many very prime-like positive integers are there less than 90000? Without giving an exact answer, can you say *approximately* how many very prime-like positive integers are less than 10^{10}? less than 10^{100}? Explain your reasoning as carefully as you can.

We start by noting that the number of multiples $M(n, k)$ of $n \in \mathbb{Z}^+$ less than $k \in \mathbb{Z}^+$ is

$$M(n, k) = \left\lfloor \frac{k-1}{n} \right\rfloor.$$

We subtract 1 to adjust for the case that $n \mid k$, in which case our count would be one too high (since we're only considering multiples *less* than k, not equal to). Similarly, the number of multiples of both n and $m \in \mathbb{Z}^+$ less than k is the same as the number of multiples of nm less than k, which is

$$M(nm, k) = \left\lfloor \frac{k-1}{nm} \right\rfloor.$$

Say now that we want to find the number of multiples of n, m, $p \in \mathbb{Z}^+$, or any combination thereof. We can draw a Venn diagram (fig. 1) of such to bring us to the first lemma.

Lemma 6.1. *The number of multiples of n, m or p less than k is*

$$M(n, k) + M(m, k) + M(p, k) - M(nm, k) - M(np, k) - M(mp, k) + M(nmp, k).$$

Proof. The Venn diagram in fig. 1 shows us that, taking into account any 'overlaps' betwen the three sets, the number of multiples of n, m or p less than k is

$$\begin{aligned} &M(nmp, k) + [M(nm, k) - M(nmp, k)] + [M(np, k) - M(nmp, k)] + [M(mp, k) - M(nmp, k)] \\ &\quad + \big[M(n, k) - [M(np, k) - M(nmp, k)] - [M(nm, k) - M(nmp, k)] - M(nmp, k)\big] \\ &\quad + \big[M(m, k) - [M(mp, k) - M(nmp, k)] - [M(nm, k) - M(nmp, k)] - M(nmp, k)\big] \\ &\qquad + \big[M(p, k) - [M(np, k) - M(nmp, k)] - [M(mp, k) - M(nmp, k)] - M(nmp, k)\big]. \end{aligned}$$

This simplifies very readily down to

$$M(n, k) + M(m, k) + M(p, k) - M(nm, k) - M(np, k) - M(mp, k) + M(nmp, k). \tag{12}$$

This makes sense intuitively too: in the sum of the first three terms (in eq. (12)), the numbers divisible by any two of n, m and p are counted twice, so we must subtract each of these set intersections once. Initially the the numbers divisible by all three of n, m and p were counted thrice, but we've now eliminated them three times so we must add this intersection back once. □

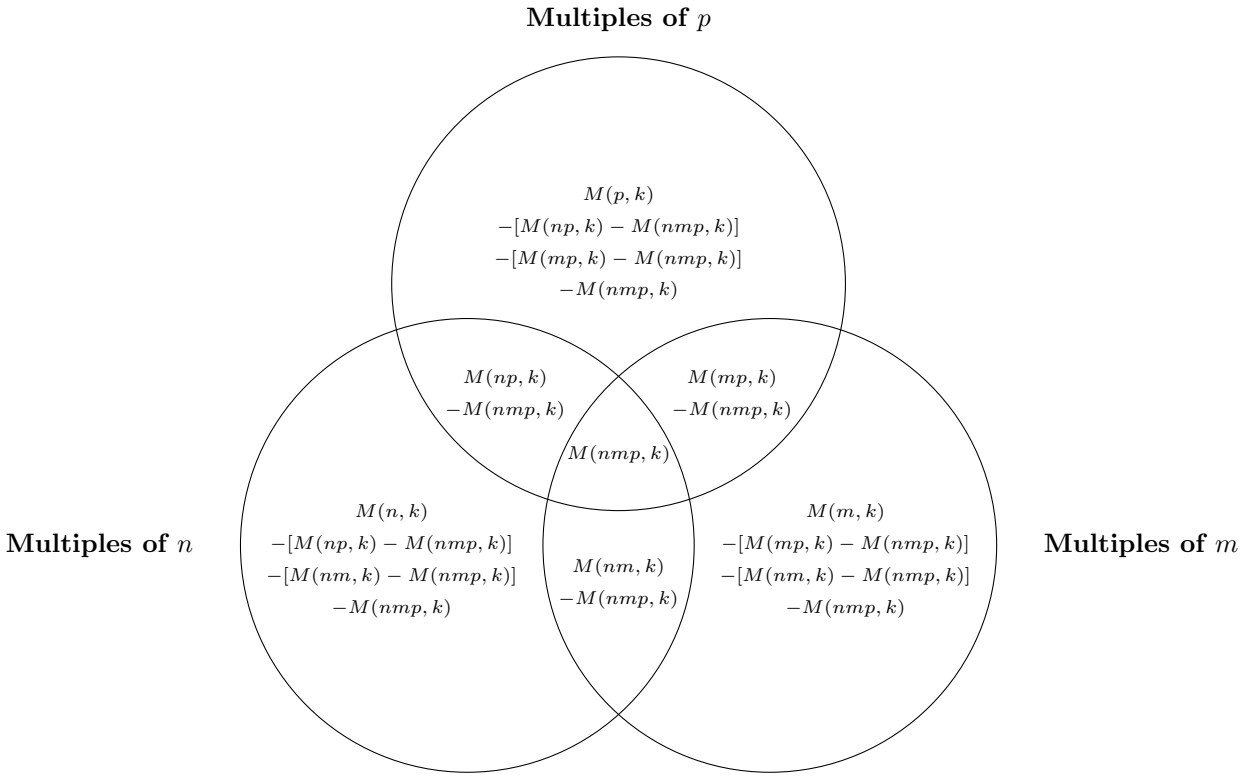


Figure 1: Multiples of n, m and p.

Hence, the number of integers less than k *not* divisible by n, m or p is

$$(k-1) - M(n,k) - M(m,k) - M(p,k) + M(nm,k) + M(np,k) + M(mp,k) - M(nmp,k).$$

Now with $k = 100$, $n = 2$, $m = 3$ and $p = 5$, we can find that the number of 'prime-like' integers less than 100 is

$$\begin{aligned} &99 - \left\lfloor \frac{99}{2} \right\rfloor - \left\lfloor \frac{99}{3} \right\rfloor - \left\lfloor \frac{99}{5} \right\rfloor + \left\lfloor \frac{99}{2\cdot 3} \right\rfloor + \left\lfloor \frac{99}{2\cdot 5} \right\rfloor + \left\lfloor \frac{99}{3\cdot 5} \right\rfloor - \left\lfloor \frac{99}{2\cdot 3\cdot 5} \right\rfloor \\ &= 99 - 49 - 33 - 19 + 16 + 9 + 6 - 3 \\ &= 99 - 73 \\ &= 26. \end{aligned}$$

Thus there are 26 prime-like integers less than 100. Now for $k = 1000$, our count is

$$\begin{aligned} &999 - \left\lfloor \frac{999}{2} \right\rfloor - \left\lfloor \frac{999}{3} \right\rfloor - \left\lfloor \frac{999}{5} \right\rfloor + \left\lfloor \frac{999}{2\cdot 3} \right\rfloor + \left\lfloor \frac{999}{2\cdot 5} \right\rfloor + \left\lfloor \frac{999}{3\cdot 5} \right\rfloor - \left\lfloor \frac{999}{2\cdot 3\cdot 5} \right\rfloor \\ &= 999 - 499 - 333 - 199 + 166 + 99 + 66 - 33 \\ &= 999 - 865 \\ &= 134, \end{aligned}$$

and so there are 134 prime-like integers less than 1000.

We'll now expand our purview to include the 'very prime-like' numbers. The primes less than 15 are

$$2, 3, 5, 7, 11, 13$$

and so we are dealing with 6 divisors now. This ruins any hopes of a Venn diagram but we can still think about the situation conceptually — it is essentially a combinatorics problem.

For the sake of clarity, let us briefly call these primes a, b, c, d, e, f. So our total number of multiples of any of these numbers is:

Lemma 6.2. *The number of multiples of a, b, c, d, e or f less than k is*

$$M(a,k)+M(b,k)+M(c,k)+M(d,k)+M(e,k)+M(f,k) - \left[\sum \textit{2-set intersections}\right] + \left[\sum \textit{3-set intersections}\right] - \left[\sum \textit{4-set intersections}\right] + \left[\sum \textit{5-set intersections}\right] - [\textit{6-set intersection}] \quad (13)$$

where $\left[\sum i\text{-set intersections}\right]$ refers to the total number of multiples of precisely i primes in the set $\{a,b,c,d,e,f\}$.

Proof. This follows the principle that the intersection of any 2 sets is going to be counted twice initially, so we must subtract one 'lot' of these 2-set intersections. In doing so, however, we have eliminated all of the 3-set intersections, so we must add them back. Hence we proceed in an alternating fashion. □

We should perhaps work out each summation separately, as there are going to be rather a lot of terms ($\binom{6}{3}=20$ so there are 20 ways of multiplying 3 of the 6 primes together, meaning that there will be 20 terms in the 3-set intersection sum alone).

So, with $k=90,000$, let's start working these out. I include my mechanical calculation here so as to make my answer more easily verifiable.

$$\begin{aligned}\sum \text{2-set intersections} &= M(ab,k)+M(ac,k)+M(ad,k)+M(ae,k)+M(af,k)\\ &\quad+M(bc,k)+M(bd,k)+M(be,k)+M(bf,k)+M(cd,k)\\ &\quad+M(ce,k)+M(cf,k)+M(de,k)+M(df,k)+M(ef,k)\\ &=\left\lfloor\frac{89,999}{2\cdot3}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot5}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot7}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot13}\right\rfloor\\ &\quad+\left\lfloor\frac{89,999}{3\cdot5}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot7}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{5\cdot7}\right\rfloor\\ &\quad+\left\lfloor\frac{89,999}{5\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{5\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{7\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{7\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{11\cdot13}\right\rfloor\\ &=61,682.\end{aligned}$$

$$\begin{aligned}\sum \text{3-set intersections} &= M(abc,k)+M(abd,k)+M(abe,k)+M(abf,k)+M(acd,k)\\ &\quad+M(ace,k)+M(acf,k)+M(ade,k)+M(adf,k)+M(aef,k)\\ &\quad+M(bcd,k)+M(bce,k)+M(bcf,k)+M(bde,k)+M(bdf,k)\\ &\quad+M(bef,k)+M(cde,k)+M(cdf,k)+M(cef,k)+M(def,k)\\ &=\left\lfloor\frac{89,999}{2\cdot3\cdot5}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot3\cdot7}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot3\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot3\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot5\cdot7}\right\rfloor\\ &\quad+\left\lfloor\frac{89,999}{2\cdot5\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot5\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot7\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot7\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{2\cdot11\cdot13}\right\rfloor\\ &\quad+\left\lfloor\frac{89,999}{3\cdot5\cdot7}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot5\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot5\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot7\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{3\cdot7\cdot13}\right\rfloor\\ &\quad+\left\lfloor\frac{89,999}{3\cdot11\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{5\cdot7\cdot11}\right\rfloor+\left\lfloor\frac{89,999}{5\cdot7\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{5\cdot11\cdot13}\right\rfloor+\left\lfloor\frac{89,999}{7\cdot11\cdot13}\right\rfloor\\ &=15,278.\end{aligned}$$

$$\begin{aligned}\sum \text{4-set intersections} &= M(abcd,k) + M(abce,k) + M(abcf,k) + M(abde,k) + M(abdf,k) \\ &\quad + M(abef,k) + M(acde,k) + M(acdf,k) + M(acef,k) + M(adef,k) \\ &\quad + M(bcde,k) + M(bcdf,k) + M(bcef,k) + M(bdef,k) + M(cdef,k) \\ &= \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 7} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 11} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 7\cdot 11} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 7\cdot 13} \right\rfloor \\ &\quad + \left\lfloor \frac{89,999}{2\cdot 3\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 5\cdot 7\cdot 11} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 5\cdot 7\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 5\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 7\cdot 11\cdot 13} \right\rfloor \\ &\quad + \left\lfloor \frac{89,999}{3\cdot 5\cdot 7\cdot 11} \right\rfloor + \left\lfloor \frac{89,999}{3\cdot 5\cdot 7\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{3\cdot 5\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{3\cdot 7\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{5\cdot 7\cdot 11\cdot 13} \right\rfloor \\ &= 1,941.\end{aligned}$$

$$\begin{aligned}\sum \text{5-set intersections} &= M(abcde,k) + M(abcdf,k) + M(abcef,k) \\ &\quad + M(abdef,k) + M(acdef,k) + M(bcdef,k) \\ &= \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 7\cdot 11} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 7\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 11\cdot 13} \right\rfloor \\ &\quad + \left\lfloor \frac{89,999}{2\cdot 3\cdot 7\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{2\cdot 5\cdot 7\cdot 11\cdot 13} \right\rfloor + \left\lfloor \frac{89,999}{3\cdot 5\cdot 7\cdot 11\cdot 13} \right\rfloor \\ &= 117.\end{aligned}$$

Finally,

$$\text{6-set intersection} = M(abcdef,k) = \left\lfloor \frac{89,999}{2\cdot 3\cdot 5\cdot 7\cdot 11\cdot 13} \right\rfloor = 2.$$

So, looking back at eq. (13), the number of multiples below $90,000$ of these primes must be

$$\begin{aligned}&\left\lfloor \frac{89,999}{2} \right\rfloor + \left\lfloor \frac{89,999}{3} \right\rfloor + \left\lfloor \frac{89,999}{5} \right\rfloor + \left\lfloor \frac{89,999}{7} \right\rfloor + \left\lfloor \frac{89,999}{11} \right\rfloor + \left\lfloor \frac{89,999}{13} \right\rfloor \\ &\quad - 61,682 + 15,278 - 1,941 + 117 - 2 \\ &= 120,958 - 48,230 = 72,728.\end{aligned}$$

Therefore, the number of very prime-like integers below $90,000$ is

$$90,000 - 72,728 = 17,272.$$

Let us now consider how we might approximate this process. The floor function merely rounds numbers down to the nearest integer, so for large values of k not much precision will be lost by removing the floor functions entirely. In doing so we can bring out a common denominator.

For ease of notation, let $P(k)$ be the number of very prime-like integers less than k. So,

$$\begin{aligned}P(k) = (k-1) &- \left\lfloor \frac{k-1}{2} \right\rfloor - \left\lfloor \frac{k-1}{3} \right\rfloor - \left\lfloor \frac{k-1}{5} \right\rfloor - \left\lfloor \frac{k-1}{7} \right\rfloor - \left\lfloor \frac{k-1}{11} \right\rfloor - \left\lfloor \frac{k-1}{13} \right\rfloor \\ &+ \left[\sum \text{2-set intersections}\right] - \left[\sum \text{3-set intersections}\right] + \left[\sum \text{4-set intersections}\right] \\ &- \left[\sum \text{5-set intersections}\right] + \left[\text{6-set intersection}\right].\end{aligned}$$

Let's get rid of the floors and bring out a common denominator of $2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 \cdot 13 = 30,030$:

$$\begin{aligned}
P(k) \approx{}& \frac{30,030(k-1) - 15,015(k-1) - 10,010(k-1)}{30,030} \\
&- \frac{6,006(k-1) - 4290(k-1) - 2730(k-1) - 2310(k-1)}{30,030} \\
&+ \left[\sum \text{2-set intersections}\right] - \left[\sum \text{3-set intersections}\right] + \left[\sum \text{4-set intersections}\right] \\
&- \left[\sum \text{5-set intersections}\right] + [\text{6-set intersection}] \\
\approx{}& -\frac{10,331}{30,030}(k-1) + \left[\sum \text{2-set intersections}\right] - \left[\sum \text{3-set intersections}\right] \\
&+ \left[\sum \text{4-set intersections}\right] - \left[\sum \text{5-set intersections}\right] + [\text{6-set intersection}]\,.
\end{aligned}$$

Now we'll do the same to each summation separately.

$$\begin{aligned}
\sum \text{2-set intersections} = {}& \left\lfloor \frac{k-1}{2 \cdot 3} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 5} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 7} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 13} \right\rfloor \\
&+ \left\lfloor \frac{k-1}{3 \cdot 5} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 7} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{5 \cdot 7} \right\rfloor \\
&+ \left\lfloor \frac{k-1}{5 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{5 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{7 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{7 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{11 \cdot 13} \right\rfloor \\
\approx{}& \frac{5,005(k-1) + 3,003(k-1) + 2,145(k-1) + 1,365(k-1) + 1,155(k-1)}{30,030} \\
&+ \frac{2,002(k-1) + 1,430(k-1) + 910(k-1) + 770(k-1) + 858(k-1)}{30,030} \\
&+ \frac{546(k-1) + 462(k-1) + 390(k-1) + 330(k-1) + 210(k-1)}{30,030} \\
\approx{}& \frac{20,581}{30,030}(k-1).
\end{aligned}$$

$$\begin{aligned}
\sum \text{3-set intersections} = {}& \left\lfloor \frac{k-1}{2 \cdot 3 \cdot 5} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 3 \cdot 7} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 3 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 3 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 5 \cdot 7} \right\rfloor \\
&+ \left\lfloor \frac{k-1}{2 \cdot 5 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 5 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 7 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 7 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{2 \cdot 11 \cdot 13} \right\rfloor \\
&+ \left\lfloor \frac{k-1}{3 \cdot 5 \cdot 7} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 5 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 5 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 7 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{3 \cdot 7 \cdot 13} \right\rfloor \\
&+ \left\lfloor \frac{k-1}{3 \cdot 11 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{5 \cdot 7 \cdot 11} \right\rfloor + \left\lfloor \frac{k-1}{5 \cdot 7 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{5 \cdot 11 \cdot 13} \right\rfloor + \left\lfloor \frac{k-1}{7 \cdot 11 \cdot 13} \right\rfloor \\
\approx{}& \frac{1,001(k-1) + 715(k-1) + 455(k-1) + 385(k-1) + 429(k-1)}{30,030} \\
&+ \frac{273(k-1) + 231(k-1) + 195(k-1) + 165(k-1) + 105(k-1)}{30,030} \\
&+ \frac{286(k-1) + 182(k-1) + 154(k-1) + 130(k-1) + 110(k-1)}{30,030} \\
&+ \frac{70(k-1) + 78(k-1) + 66(k-1) + 42(k-1) + 30(k-1)}{30,030} \\
\approx{}& \frac{5,102}{30,010}(k-1).
\end{aligned}$$

$$\begin{aligned}\sum \text{4-set intersections} &= \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 7}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 11}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 7\cdot 11}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 7\cdot 13}\right\rfloor \\ &\quad + \left\lfloor \frac{k-1}{2\cdot 3\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 5\cdot 7\cdot 11}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 5\cdot 7\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 5\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 7\cdot 11\cdot 13}\right\rfloor \\ &\quad + \left\lfloor \frac{k-1}{3\cdot 5\cdot 7\cdot 11}\right\rfloor + \left\lfloor \frac{k-1}{3\cdot 5\cdot 7\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{3\cdot 5\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{3\cdot 7\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{5\cdot 7\cdot 11\cdot 13}\right\rfloor \\ &\approx \frac{143(k-1)+91(k-1)+77(k-1)+65(k-1)+55(k-1)}{30,030} \\ &\quad + \frac{35(k-1)+39(k-1)+33(k-1)+21(k-1)+15(k-1)}{30,030} \\ &\quad + \frac{26(k-1)+22(k-1)+14(k-1)+10(k-1)+6(k-1)}{30,030} \\ &\approx \frac{652}{30,030}(k-1).\end{aligned}$$

$$\begin{aligned}\sum \text{5-set intersections} &= \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 7\cdot 11}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 7\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 11\cdot 13}\right\rfloor \\ &\quad + \left\lfloor \frac{k-1}{2\cdot 3\cdot 7\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{2\cdot 5\cdot 7\cdot 11\cdot 13}\right\rfloor + \left\lfloor \frac{k-1}{3\cdot 5\cdot 7\cdot 11\cdot 13}\right\rfloor \\ &\approx \frac{13(k-1)+11(k-1)+7(k-1)+5(k-1)+3(k-1)+2(k-1)}{30,030} \\ &\approx \frac{41}{30,030}(k-1).\end{aligned}$$

$$\text{6-set intersection} = \left\lfloor \frac{k-1}{2\cdot 3\cdot 5\cdot 7\cdot 11\cdot 13}\right\rfloor \approx \frac{1}{30,030}(k-1).$$

So, we can now put these all together and simplify:

$$\begin{aligned}P(k) &\approx -\frac{10,331}{30,030}(k-1) + \frac{20,581}{30,030}(k-1) - \frac{5,102}{30,010}(k-1) \\ &\quad + \frac{652}{30,030}(k-1) - \frac{41}{30,030}(k-1) + \frac{1}{30,030}(k-1) \\ &\approx \frac{5760}{30,030}(k-1) \\ &\approx \frac{192}{1,001}(k-1).\end{aligned}$$

Hence we come to the following theorem.

Theorem 6.3. *The approximate number of very prime-like integers less than k is*

$$P(k) \approx \frac{192}{1,001}(k-1).$$

Proof. This approximation is come to by ignoring the floor functions and bringing out a common denominator, as detailed above. □

For $k = 10^{10}$ this gives us

$$P(10^{10}) \approx \frac{192}{1,001}\cdot 10^{10} \approx 1.9181\cdot 10^{10}$$

and for $k = 10^{100}$,

$$P(10^{100} \approx \frac{192}{1,001}\cdot 10^{100} \approx 1.9181\cdot 10^{100}.$$

We can also check our previous answer (remembering that this is now only an approximation). With $k = 90,000$,

$$P(90,000) \approx \frac{192}{1,001} \cdot 90,000 \approx 17,262.7$$

and so we see that our previous answer of $17,272$ was in the right range.

Question 7

The triangular numbers are the numbers $1, 3, 6, 10, 15, \ldots$. The square numbers are the numbers $1, 4, 9, 16, 25, \ldots$. The pentagonal numbers are $1, 5, 12, 22, 35, \ldots$. The geometrical language is justified by the following diagrams:

(figure)

a. What are the first five hexagonal numbers? What are the first five septagonal numbers? What are the first five r-gonal numbers? Give a formula for the nth triangular number. Give a formula for the nth square number. Give a formula for the nth pentagonal number. In general, give a formula for the nth r-gonal number.

b. How many numbers can you find that are simultaneously triangular and square? How many numbers can you find that are simultaneously square and pentagonal?

We will start by considering the diagrams given representing the first five pentagonal numbers. We'll approach this from a purely geometric point of view.

Let

$$s(r,n) : \left[\mathbb{Z}^+ \setminus \{1,2\}\right] \cdot \mathbb{Z}^+ \to \mathbb{Z}^+$$

be the number of points on the *outer edge* of the nth r-gonal number, and let also

$$A(r,n) : \left[\mathbb{Z}^+ \setminus \{1,2\}\right] \cdot \mathbb{Z}^+ \to \mathbb{Z}^+$$

be the nth r-gonal number. Generalising from the start, the first thing we'll do is find a recurrence relation for the nth r-gonal number.

Lemma 7.1. *Given the $(n-1)$th r-gonal number, the nth r-gonal number is*

$$A(r,n) = A(r,n-1) + nr - 2n - r + 3. \tag{14}$$

where the 1st r-gonal number is always

$$A(r,1) = 1. \tag{15}$$

Proof. It can be seen from the pentagonal numbers that the number of edge points is $5(n-1)$, and so

$$s(5,n) = 5(n-1)$$

which we can immediately generalise to

$$s(r,n) = r(n-1).$$

(By simple geometry this pattern must always be true.) Now looking at the diagrams again we see that every pentagonal number is the sum of its edge points and the previous pentagonal number, subtract the number of edge points shared by the previous (inner) pentagon. This "overlap" is

$$2(n-2) + 1$$

which by drawing similar diagrams for higher polygonal numbers we see is the same for any r.

Therefore, we come to the equation

$$A(r,n) = s(r,n) + A(r,n-1) - 2(n-2) - 1.$$

Simplifying, we come to

$$A(r,n) = r(n-1) + A(r,n-1) - 2(n-2) - 1.$$

The first r-gonal number must always be 1 by the geometric justification given in the question. □

Now that we have a recurrence relation for the nth r-gonal number, we want to solve it to find $A(r,n)$ explicitly in terms of r and n only.

We'll try to guess a solution and prove it by induction. Iterating the recurrence relation backwards may give us some insight:

$$\begin{aligned}
A(r,n) &= A(r,n-1) + nr - 2n - r + 3 \\
&= \big[A(r,n-2) + (n-1)r - 2(n-1) - r + 3\big] + nr - 2n - r + 3 \\
&= A(r,n-2) + 2nr - 4n - 3r + 8 && (16)\\
&= \big[A(r,n-3) + (n-2)r - 2(n-2) - r + 3\big] + 2nr - 4n - 3r + 8 \\
&= A(r,n-3) + 3nr - 6n - 6r + 15 && (17)\\
&= \big[A(r,n-4) + (n-3)r - 2(n-3) - r + 3\big] + 3nr - 6n - 6r + 15 \\
&= A(r,n-4) + 4nr - 8n - 10r + 24 && (18)\\
&= \big[A(r,n-5) + (n-4)r - 2(n-4) - r + 3\big] + 4nr - 8n - 10r + 24 \\
&= A(r,n-5) + 5nr - 10n - 15r + 35. && (19)
\end{aligned}$$

If we keep going we'll eventually get to an expression

$$A(r,n) = A(r,n-k) + a_k nr + b_k n + c_k r + d_k \tag{20}$$

where $k, a_k, b_k, c_k, d_k \in \mathbb{Z}$. We want to find a_k, b_k, c_k, d_k in terms of k. To do this, we can construct a table of values from eqs. (14) and (16) to (19) as below:

k	a_k	b_k	c_k	d_k
1	1	-2	-1	3
2	2	-4	-3	8
3	3	-6	-6	15
4	4	-8	-10	24
5	5	-10	-15	35

It can be seen immediately that

$$a_k = k \tag{21}$$

and

$$b_k = -2k \tag{22}$$

for all k.

However, sequences $\{c_k\}$ and $\{d_k\}$ appear to be quadratic. The second difference of $\{c_k\}$ is

$$\Delta^2(c_k) = -1$$

and so the leading term must be $-\frac{1}{2}k^2$ (because in quadratic sequences the leading coefficient is half the second difference). Subtracting this from the values of c_k in the table leaves us with a linear sequence:

k	1	2	3	4	5
$c_k - (-\frac{1}{2}k^2)$	$-\frac{1}{2}$	-1	$-\frac{3}{2}$	-2	$-\frac{5}{2}$

We observe that

$$c_k + \frac{1}{2}k^2 = -\frac{1}{2}k$$

and so we now know that

$$c_k = -\frac{1}{2}k^2 - \frac{1}{2}k. \tag{23}$$

Similarly, the sequence $\{d_k\}$ has second difference

$$\Delta^2(d_k) = 2$$

and so the leading term must be k^2. Subtracting this from d_k, we have a linear sequence:

k	1	2	3	4	5
$d_k - k^2$	2	4	6	8	10

This linear sequence has equation

$$d_k - k^2 = 2k$$

and so we come to

$$d_k = k^2 + 2k. \tag{24}$$

Substituting eqs. (21) to (24) into eq. (20), we get

$$A(r,n) = A(r,n-k) + knr - 2kn - (\frac{1}{2}k^2 + \frac{1}{2}k)r + (k^2 + 2k).$$

Hence if $k = n-1$,

$$A(r,n) = A(r,1) + (n-1)nr - 2(n-1)n - \left[\frac{1}{2}(n-1)^2 + \frac{1}{2}(n-1)\right] r + (n-1)^2 + 2(n-1)$$

which we can simplify:

$$\begin{aligned} A(r,n) &= A(r,1) + n^2r - nr - 2n^2 + 2n - \frac{1}{2}n^2r + nr - \frac{1}{2}r - \frac{1}{2}nr + \frac{1}{2}r + n^2 - 2n + 1 + 2n - 2 \\ &= A(r,1) + n^2\left(r - 2 - \frac{1}{2}r + 1\right) + n\left(-r + 2 + r - \frac{1}{2}r - 2 + 2\right) + \left(-\frac{1}{2}r + \frac{1}{2}r + 1 - 2\right) \\ &= A(r,1) + n^2\left(\frac{1}{2}r - 1\right) + n\left(\frac{1}{2}r + 2\right) - 1. \end{aligned}$$

Therefore as by eq. (15) we know $A(r,1) = 1$, we finally come to our main theorem, which we can prove by induction:

Theorem 7.2. *The nth r-gonal number is*

$$A(r,n) = \left(\frac{1}{2}r - 1\right)n^2 - \left(\frac{1}{2}r - 2\right)n.$$

Proof. Assume lemma 7.1. We want to show that

$$A(r,n-1) + nr - 2n - r + 3 = \left(\frac{1}{2}r - 1\right)n^2 - \left(\frac{1}{2}r - 2\right)n.$$

As our base case, with $n = 2$,

$$\text{LHS} = A(r,1) + 2r - 2 \cdot 2 - r + 3 = r$$

and

$$\text{RHS} = 4\left(\frac{1}{2}r - 1\right) - 2\left(\frac{1}{2}r - 2\right) = r$$

thereby confirming that the equality holds for $n = 2$. Now as our induction hypothesis, let us assume that

$$A(r, j) = \left(\frac{1}{2}r - 1\right) j^2 - \left(\frac{1}{2}r - 2\right) j.$$

Evaluating now for $j = 1$ (our inductive step):

$$\begin{aligned}
A(r, j+1) &= A(r, j) + (j+1)r - 2(j+1) - r + 3 \\
&= A(r, j) + jr - 2j + 1 \\
&= \left(\frac{1}{2}r - 1\right) j^2 - \left(\frac{1}{2}r - 2\right) j + jr - 2j + 1 \\
&= \left(\frac{1}{2}r - 1\right) j^2 - \left(\frac{1}{2}r - 2\right) j + jr - 2j + 1 \\
&\quad + \left(\frac{1}{2}r - 1\right)(2j+1) - \left(\frac{1}{2}r - 1\right)(2j+1) \\
&= \left(\frac{1}{2}r - 1\right)(j+1)^2 - \left(\frac{1}{2}r - 2\right) j + jr - 2j + 1 \\
&\quad - \left(\frac{1}{2}r - 1\right)(2j+1) \\
&= \left(\frac{1}{2}r - 1\right)(j+1)^2 - \left(\frac{1}{2}r - 2\right) j + jr - 2j + 1 \\
&\quad - \left(\frac{1}{2}r - 1\right)(2j+1) + \left(\frac{1}{2}r - 2\right) - \left(\frac{1}{2}r - 2\right) \\
&= \left(\frac{1}{2}r - 1\right)(j+1)^2 - \left(\frac{1}{2}r - 2\right)(j+1) \\
&\quad + jr - 2j + 1 - \left(\frac{1}{2}r - 1\right)(2j+1) + \left(\frac{1}{2}r - 2\right) \\
&= \left(\frac{1}{2}r - 1\right)(j+1)^2 - \left(\frac{1}{2}r - 2\right)(j+1) \\
&\quad + jr - 2j + 1 - jr - \frac{1}{2}r + 2j + 1 + \frac{1}{2}r - 2 \\
&= \left(\frac{1}{2}r - 1\right)(j+1)^2 - \left(\frac{1}{2}r - 2\right)(j+1).
\end{aligned}$$

Hence, by mathematical induction the theorem must be correct. □

We can therefore now answer the questions in part (a). The nth triangular number is

$$A(3, n) = \frac{1}{2}n^2 + \frac{1}{2}n. \tag{25}$$

The nth square number is, predictably,

$$A(4, n) = n^2.$$

Lastly, the nth pentagonal number is

$$A(5, n) = \frac{3}{2}n^2 - \frac{1}{2}n. \tag{26}$$

Let us now address the question of numbers that are simultaneously triangular and square. After some more experimentation we come to the following lemma.

Lemma 7.3. *For any integer m, m^2 is triangular (and square) if and only if $8m^2 + 1$ is a perfect square.*

Proof. If a perfect square m^2 with $m \in \mathbb{Z}^+$ is also triangular, then by eq. (25),

$$m^2 = \frac{1}{2}n^2 + \frac{1}{2}n$$

for some $n \in \mathbb{Z}^+$. Since this is a quadratic, we suspect that by making n the subject we might gain some information. Completing the square,

$$\begin{aligned} & n^2 + n - 2m^2 = 0 \\ \implies & \left(n + \frac{1}{2}\right)^2 - \frac{1}{4} - 2m^2 = 0 \\ & \implies n = \pm\sqrt{\frac{1}{4} + 2m^2} - \frac{1}{2}. \end{aligned}$$

As $m^2 \geqslant 1$ by definition, for n to be positive (which it must be) the surd must be positive, so we can disregard the negative sign. Thus, the equation becomes

$$n = \sqrt{\frac{1}{4} + 2m^2} - \frac{1}{2}.$$

Since n is an integer, $\sqrt{\frac{1}{4} + 2m^2}$ must be a half-integer and so

$$2\sqrt{\frac{1}{4} + 2m^2} = \sqrt{8m^2 + 1}$$

must be an integer. Every step in this proof is reversible, and so there is a two-way implication. □

This leads directly to the following conclusion:

Theorem 7.4. *There are infinitely many numbers that are simultaneously square and triangular.*

Proof. By lemma 7.3, we can say that

$$8m^2 + 1 = \ell^2$$

or

$$\ell^2 - 8m^2 = 1$$

for some $\ell \in \mathbb{Z}^+$. This is an instance of Pell's equation, which we can show has infinitely many solutions.[1] □

Similarly if a square number p^2 with $p \in \mathbb{Z}^+$ is also pentagonal, then by eq. (26),

$$p^2 = \frac{3}{2}n^2 - \frac{1}{2}n$$

for some (different) $n \in \mathbb{Z}^+$. Therefore,

$$3n^2 - n - 2p^2 = 0$$

and by the quadratic formula,

$$n = \frac{1 \pm \sqrt{1 + 24p^2}}{6}$$

as before, we can neglect the negative sign, leading us to

$$n = \frac{1 + \sqrt{1 + 24p^2}}{6}$$

which implies that $\sqrt{1 + 24p^2}$ is an integer and so

$$1 + 24p^2 = q^2 \implies q^2 - 24p^2 = 1$$

for some $q \in \mathbb{Z}^+$. This is another instance of Pell's equation, and so by the same reasoning there is an infinity of numbers simultaneously square and pentagonal.

[1] https://en.wikipedia.org/wiki/Pell's_equation#Solutions

Question 8

10 people are to be divided into 3 committees, in such a way that every committee must have at least one member, and no person can serve on all three committees. (Note that we do not require everybody to serve on at least one committee.) In how many ways can this be done?

Let there be n people $p_1, p_2, \ldots, p_n$ to be distributed among 3 committees A, B, C. (We avoid generalising the number of committees so as to aid visualisation later.)

In this question we come quickly to the following theorem.

Theorem 8.1. *The number of ways of dividing n people into 3 committees such that no committee may be empty and no person may serve on all three committees is*

$$7^n + 3 \cdot 2^n - 3 \cdot 4^n - 1.$$

Proof. Considering a particular individual p_i where $i \in \mathbb{Z}, 1 \leqslant i \leqslant n$, we know that p_i can be a member of any combination of committees (including none) *with the exception of* being a member of all 3 committees.

Thus, the number of possible combinations of memberships for p_i is

$$\binom{3}{0} + \binom{3}{1} + \binom{3}{2} = 7.$$

where $\binom{a}{b}$ is the binomial coefficient. In fact, we can list all 7 cases:

$$p_i \in \begin{cases} \varnothing \\ A \\ B \\ C \\ A, B \\ A, C \\ B, C. \end{cases}$$

Therefore, as there are n people, the total number of possible membership combinations for everyone is

$$\text{perms}_{\text{total}} = 7^n.$$

However, we have not yet taken into account the condition that no committee may be empty. For this we must consider the number of permutations that include one or more empty committee.

For all three committees to be empty, we require

$$p_i \in \varnothing$$

for every i, and so there is only one such combination. For any two committees to be empty, we require

$$p_i \in \begin{cases} \varnothing \\ A \end{cases} \quad \textbf{or} \quad p_i \in \begin{cases} \varnothing \\ B \end{cases} \quad \textbf{or} \quad p_i \in \begin{cases} \varnothing \\ C \end{cases}$$

for every i, and so there are $2^n \cdot 3$ such combinations overall. Finally, for exactly one committee to be empty we need

$$p_i \in \begin{cases} \varnothing \\ B \\ C \\ B, C \end{cases} \quad \textbf{or} \quad p_i \in \begin{cases} \varnothing \\ A \\ C \\ A, C \end{cases} \quad \textbf{or} \quad p_i \in \begin{cases} \varnothing \\ A \\ B \\ A, B \end{cases}$$

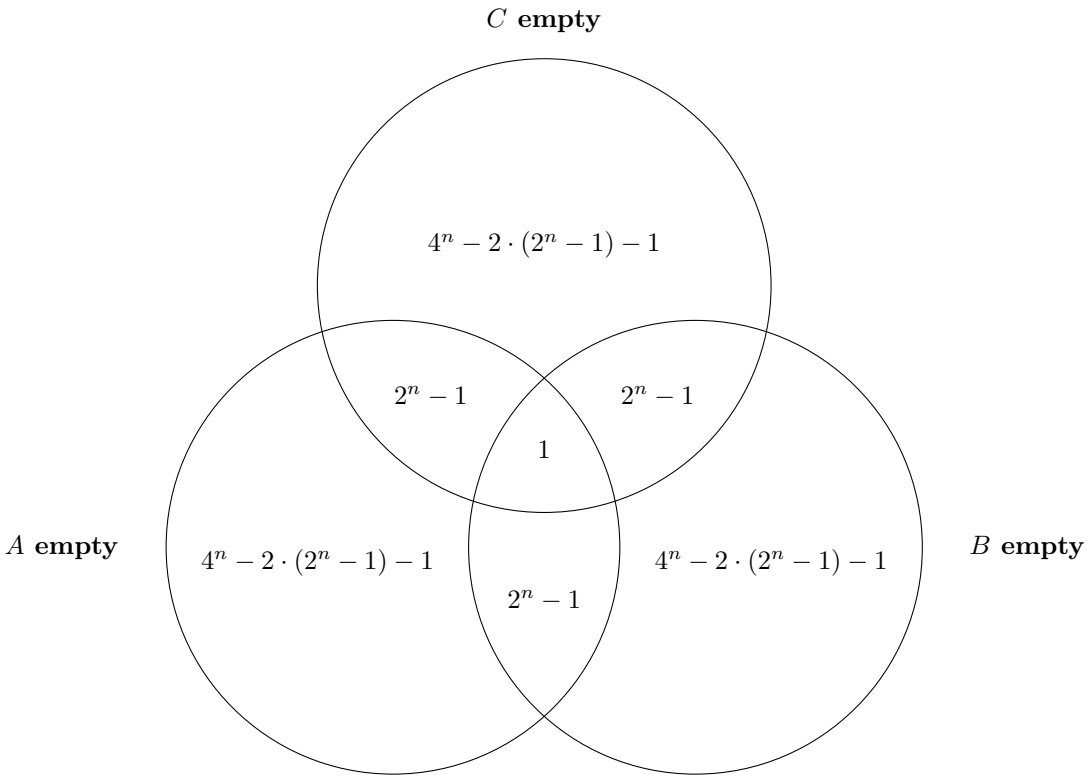


Figure 2: The number of membership combinations resulting in empty committees.

for every i, and so there are $4^n \cdot 3$ combinations overall.

These cases, however, are not mutually exclusive, and so we use a Venn diagram to clear up overlaps (in fig. 2). Thus we see that the total number of 'illegal' permutations — permutations in which one or more committee is empty — is

$$\text{perms}_{\text{illegal}} = 1 + 3 \cdot (2^n - 1) + 3 \cdot [4^n - 2 \cdot (2^n - 1) - 1].$$

Therefore we know that the total number of legal permutations is

$$\begin{aligned}
\text{perms}_{\text{legal}} &= \text{perms}_{\text{total}} - \text{perms}_{\text{illegal}} \\
&= 7^n - \big(1 + 3 \cdot (2^n - 1) + 3 \cdot [4^n - 2 \cdot (2^n - 1) - 1]\big) \\
&= 7^n - 1 - 3 \cdot 2^n + 3 - 3 \cdot 4^n + 6 \cdot 2^n - 6 + 3 \\
&= 7^n + 3 \cdot 2^n - 3 \cdot 4^n - 1.
\end{aligned}$$

□

Hence with $n = 10$ as in the question,

$$\begin{aligned}
\text{perms}_{\text{legal}} &= 7^{10} + 3 \cdot 2^{10} - 3 \cdot 4^{10} - 1 \\
&= 279,332,592.
\end{aligned}$$

Question 9

Let S be a set of positive real numbers. If S contains at least four distinct elements show that there are elements $x, y \in S$ such that

$$0 < \frac{x - y}{1 + xy} < \sqrt{3}/3.$$

What can you say if S has at least 7 elements? What if S has at least n elements, where $n > 2$?

Upon seeing this inequality we immediately note its similarity to the tangent compound angle identity. That is, that

$$\tan(\alpha - \beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta}.$$

We suspect that using this fact may simplify the problem significantly.

Lemma 9.1. *A set $S \subset \mathbb{R}^+$ with four distinct elements will always contain two elements $x, y \in S$ such that*

$$0 < \frac{x - y}{1 + xy} < \sqrt{3}/3.$$

Proof. Let's consider a set $S \subset \mathbb{R}^+$ with 4 distinct elements x_1, x_2, x_3, x_4 such that

$$x_1 < x_2 < x_3 < x_4. \tag{27}$$

So, we want to show that there exists $x_i, x_j \in S$ given $i, j \in \{1, 2, 3, 4\}$ and $i \neq j$ such that

$$\frac{|x_i - x_j|}{1 + x_i x_j} < \frac{\sqrt{3}}{3}.$$

(If $x_i - x_j$ is negative we can simply swap i and j to make it positive, and the left hand side will never equal zero because $x_i \neq x_j$ by definition.) Now let $x_i = \tan\psi_i$ for every i, with $0 < \psi_i < \frac{\pi}{2}$ so that $\psi_i = \arctan x_i$. Then by the tangent compound angle identity, we are now looking for

$$\tan(|\psi_i - \psi_j|) < \frac{\sqrt{3}}{3}$$

which is the same as

$$|\psi_i - \psi_j| < \frac{\pi}{6}.$$

By the nature of the tangent function between 0 and $\frac{\pi}{2}$, we know from eq. (27) that

$$0 < \psi_1 < \psi_2 < \psi_3 < \psi_4 < \frac{\pi}{2}.$$

Now we split the interval $\left(0, \frac{\pi}{2}\right)$ into three equally wide intervals $\left(0, \frac{\pi}{6}\right]$, $\left(\frac{\pi}{6}, \frac{\pi}{3}\right]$ and $\left(\frac{\pi}{3}, \frac{\pi}{2}\right)$. By the pigeonhole principle, since we have 4 elements and 3 intervals, at least one of these intervals must contain two or more elements. The width of each interval is $\frac{\pi}{6}$ and so these two elements must have an absolute difference less than $\frac{\pi}{6}$.

Consequently, there are always two elements x_i, x_j in S such that

$$\begin{aligned} 0 < \psi_i - \psi_j &< \frac{\pi}{6} \\ \implies 0 < \tan(\psi_i - \psi_j) &< \frac{\sqrt{3}}{3} \\ \implies 0 < \frac{\tan\psi_i - \tan\psi_j}{1 + \tan\psi_i\tan\psi_j} &< \frac{\sqrt{3}}{3} \\ \implies 0 < \frac{x_i - x_j}{1 + x_i x_j} &< \frac{\sqrt{3}}{3}. \end{aligned}$$

□

We will now generalise the above finding to a set S with n distinct elements.

Theorem 9.2. *A set $S \subset \mathbb{R}^+$ with $n \geqslant 2$ distinct elements will always contain two elements $x, y \in S$ such that*

$$0 < \frac{x - y}{1 + xy} < \tan\left(\frac{\pi}{2n - 2}\right).$$

Proof. Consider a set $S \subset \mathbb{R}^+$ now with $n \geqslant 2$ distinct elements $y_1, y_2, \ldots, y_n$ such that

$$i < j \iff y_i < y_j$$

for any $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$. Let $y_i = \tan \omega_i$ for every i, where $0 < \omega_i < \frac{\pi}{2}$ such that $\omega_i = \arctan y_i$. Splitting the interval $\left(0, \frac{\pi}{2}\right)$ into $n-1$ equally wide intervals, each of width $\frac{\pi/2}{n-1}$, by the pigeonhole principle since there are n elements and $n-1$ intervals, there must be at least one interval containing two or more elements. These two elements must therefore have an absolute difference less than $\frac{\pi/2}{n-1}$.

Therefore there exist always two elements $y_i, y_j \in S$ such that

$$\begin{aligned} 0 < \omega_i - \omega_j &< \frac{\pi/2}{n-1} \\ \implies 0 < \tan(\omega_i - \omega_j) &< \tan\left(\frac{\pi/2}{n-1}\right) \\ \implies 0 < \frac{\tan\omega_i - \tan\omega_j}{1 + \tan\omega_i \tan\omega_j} &< \tan\left(\frac{\pi}{2n-2}\right) \\ \implies 0 < \frac{y_i - y_j}{1 + y_i y_j} &< \tan\left(\frac{\pi}{2n-2}\right). \end{aligned}$$

□

Question 10

> The tail of a giant kangaroo is tied to a pole in the ground by an infinitely stretchy elastic cord. A flea sits on the pole watching the kangaroo (hungrily). The kangaroo sees the flea, leaps into the air and lands one kilometre from the pole (with its tail still attached to the pole by the elastic cord). The flea gives chase and leaps into the air landing on the stretched elastic cord one centimetre from the pole. The kangaroo, seeing this, again leaps into the air and lands another kilometre away from the pole (i.e., a total of two kilometres from the pole). Undaunted, the flea bravely leaps into the air again, landing on the elastic cord one centimetre further along. Once again the kangaroo jumps another kilometre and the flea jumps another centimetre along the cord. If this continues indefinitely, will the flea ever catch up to the kangaroo? (Assume the earth is flat and extends infinitely far in all directions.)

We'll first define a few variables. Let α be the distance the kangaroo leaps every time, and let β be the distance the flea leaps.

Let also ℓ_n and x_n be the distance of the kangaroo and the flea respectively from the pole after $n \in \mathbb{Z}^*$ leaps.

Lemma 10.1. *After n leaps of the kangaroo, given the flea's displacement from the pole after $n-1$ leaps, the flea's new displacement from the pole is*

$$\frac{n}{n-1} x_{n-1} + \beta$$

where $x_1 = \beta$.

Proof. From the information in the question, we can say that

$$\begin{aligned} \ell_0 &= 0, \\ \ell_1 &= \alpha \end{aligned}$$

and more generally, as the kangaroo moves α further away every leap,

$$\ell_n = n\alpha.$$

Similarly, we know that

$$x_0 = 0$$

and

$$x_1 = \beta.$$

Now after every leap of the kangaroo, the cord is stretched by a factor of $\frac{\ell_n}{\ell_{n-1}}$ and hence the flea's position increases by the same factor; and then the flea leaps a further β along the string. Hence after n leaps, the flea's displacement is

$$\begin{aligned} x_n &= \frac{\ell_n}{\ell_{n-1}} x_{n-1} + \beta \\ &= \frac{n\alpha}{(n-1)\alpha} x_{n-1} + \beta \\ &= \frac{n}{n-1} x_{n-1} + \beta. \end{aligned}$$

□

By iterating, we notice a pattern in the expansion of this recurrence relation:

$$\begin{aligned} x_n &= \frac{n}{n-1}\left[\frac{n-1}{n-2}x_{n-2} + \beta\right] + \beta \\ &= \frac{n(n-1)}{(n-1)(n-2)}x_{n-2} + \frac{n}{n-1}\beta + \beta \\ &= \frac{n(n-1)}{(n-1)(n-2)}\left[\frac{n-2}{n-3}x_{n-3} + \beta\right] + \frac{n}{n-1}\beta + \beta \\ &= \frac{n(n-1)(n-2)}{(n-1)(n-2)(n-3)}x_{n-3} + \frac{n(n-1)}{(n-1)(n-2)}\beta + \frac{n}{n-1}\beta + \beta \\ &\vdots \\ &= \frac{n(n-1)\cdots[n-(k-1)]}{(n-1)(n-2)\cdots(n-k)}x_{n-k} + \frac{n(n-1)\cdots[n-(k-2)]}{(n-1)(n-2)\cdots[n-(k-1)]}\beta \\ &\quad + \cdots + \frac{n(n-1)}{(n-1)(n-2)}\beta + \frac{n}{n-1}\beta + \beta \\ &= \frac{n}{n-k}x_{n-k} + \frac{n}{n-(k-1)}\beta + \cdots + \frac{n}{n-2}\beta + \frac{n}{n-1}\beta + \beta. \end{aligned}$$

If the placeholder $k \in \mathbb{Z}^+$ is equal to $n-1$, then

$$x_n = \frac{n}{n-(n-1)}x_1 + \frac{n}{n-[(n-1)-1]}\beta + \cdots + \frac{n}{n-2} + \frac{n}{n-1} + \beta$$

and since $x_1 = \beta$,

$$\begin{aligned} x_n &= \frac{n}{1}\beta + \frac{n}{2}\beta + \cdots + \frac{n}{n-2}\beta + \frac{n}{n-1}\beta + \frac{n}{n}\beta \\ &= \sum_{i=1}^{n} \frac{n}{i}\beta. \end{aligned}$$

Hence, we can state and prove the next result:

Lemma 10.2. *The flea's displacement from the pole after n leaps is*

$$x_n = n\beta \sum_{i=1}^{n} i^{-1}. \tag{28}$$

Proof. We can prove this is analagous to lemma 10.1 by induction. In the base case $n = 2$, our recurrence relation gives us

$$x_2 = \frac{2}{2-1}\beta + \beta = 3\beta$$

and eq. (28) gives us

$$x_2 = 2\beta \sum_{i=1}^{2} i^{-1} = 2\beta \left(1 + \frac{1}{2}\right) = 3\beta.$$

So, the equality holds for $n = 2$. Now as our induction hypothesis, let us assume that

$$x_r = \frac{r}{r-1} x_{r-1} + \beta = r\beta \sum_{i=1}^{r} i^{-1}.$$

We'll now show that the equality holds for $r + 1$ (our inductive step):

$$\begin{aligned} x_{r+1} = \frac{r+1}{r} x_r + \beta &= \beta + \frac{r+1}{r} r\beta \sum_{i=1}^{r} i^{-1} \\ &= \beta + (r+1)\beta \sum_{i=1}^{r} i^{-1} \\ &= \frac{1}{r+1}(r+1)\beta + (r+1)\beta \sum_{i=1}^{r} i^{-1} \\ &= (r+1)\beta \left[(r+1)^{-1} + \sum_{i=1}^{r} i^{-1} \right] \\ &= (r+1)\beta \sum_{i=1}^{r+1} i^{-1}. \end{aligned}$$

Hence by mathematical induction, the lemma is shown to be true. □

We can now immediately answer the question.

Theorem 10.3. *The flea catches up with the kangaroo.*

Proof. If the flea is ever to catch up with the kangaroo, then at some point,

$$\begin{aligned} & x_n \geqslant \ell_n \\ \implies & n\beta \sum_{i=1}^{n} i^{-1} \geqslant n\alpha \\ \implies & \sum_{i=1}^{n} i^{-1} \geqslant \frac{\alpha}{\beta}. \end{aligned}$$

Since the left hand side is just the harmonic series, the sequence never converges and so as the right hand side is constant, the flea *will* indeed catch up with the kangaroo — no matter what the leaping distances are. □

However, this may take a while in practice. Given $\alpha = 1000\,\text{m}$ and $\beta = 0.01\,\text{m}$, the ratio of the leaping distances is

$$\frac{\alpha}{\beta} = 100,000.$$

The partial sum of the harmonic series can be approximated by

$$\sum_{i=1}^{n} i^{-1} \approx \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2}$$

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant. (This stems from an approximation of the area under $\frac{1}{x}$, which for large values of x is similar to the partial sum of the harmonic series.) So,

$$\ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} \approx 100,000.$$

As n will clearly be very large, we can reduce this approximation to

$$\begin{aligned} \ln n &\approx 100,000 \\ \implies n &\approx e^{100,000} \approx 10^{43,429} \\ \implies \ell_n &\approx 10^{43,432} \text{ m}. \end{aligned}$$

Considering that the diameter of the observable universe is about 10^{27} m, it is clear that the flea will only catch up with the kangaroo after a truly unfathomable distance.

Chapter 25

STEP I 2007 solutions

Miss Brownlee and Dr Brown made us, quite sadistically, do all the questions of a STEP 1 paper over the course of one weekend in preparation for the exam at the end of Year 12. These were probably a bit rushed but I managed to do them in time.

STEP I 2007 — Solutions

Damon Falck

May 8, 2017

Section A: Pure Mathematics

1. **(i)** Using the digits 1 to 4, there are eight possible sums of last two (and so first two) digits of such a balanced number. There are a number of ways of achieving each sum, and we have to be careful to make sure we do not allow a zero at the start of the number. We will list the possibilities:

- There are 2 ways for the last two digits to sum to 1 (01,10), and 1 way for the first two digits to sum to 1 (10).
- There are 3 ways for the last two digits to sum to 2 (02,11,20), and 2 ways for the first two digits to sum to 2 (11,20).
- There are 4 ways for the last two digits to sum to 3 (03,12,21,30), and 3 ways for the first two digits to sum to 3 (12,21,30).
- There are 5 ways for the last two digits to sum to 4 (04,13,22,31,40), and 4 ways for the first two digits to sum to 4 (13,22,31,40).
- There are 4 ways for the last two digits or the first two digits to sum to 5 (14,23,32,41).
- There are 3 ways for the last two digits or the first two digits to sum to 6 (24,33,42).
- There are 2 ways for the last two digits or the first two digits to sum to 7 (34,43).
- There is 1 way for the last two digits or the first two digits to sum to 8 (44).

The number of ways N_4 of achieving a 4-digit balanced number with each sum is equal to the product of the number of choices for the first two digits and the number of choices for the last two digits. So,

$$\begin{aligned} N_4 &= 2 \cdot 1 + 3 \cdot 2 + 4 \cdot 3 + 5 \cdot 4 + 4 \cdot 4 + 3 \cdot 3 + 2 \cdot 2 + 1 \cdot 1 \\ &= 2 + 6 + 12 + 20 + 16 + 9 + 4 + 1 \\ &= 70 \end{aligned}$$

as we wanted to show.

(ii) Noticing a pattern in the above list, we can generalise our result. For two-digit sums of $r = 1$ to k, the number of balanced numbers is $r(r+1)$, and for sums of $r = k+1$ to $2k$ (the maximum), the number of balanced numbers is $[(2k+1) - r]^2$. So, the total number of possible 4-digit balanced numbers N_k using k digits is

$$N_k = \sum_{r=1}^{k} r(r+1) + \sum_{r=k+1}^{2k} (2k+1-r)^2.$$

We see from the list above that the second sum must just be the sum of the square numbers up to k^2. Trying out a few values in hope of simplifying this sum indeed

does show that $[2k+1-(k+1)]^2 = k^2$ all the way down to $[2k+1-(2k)]^2 = 1$, and so the second sum simplifies very nicely. Hence,

$$\begin{aligned} N_k &= \sum_{r=1}^{k} r(r+1) + \sum_{r=1}^{k} r^2 \\ &= 2\sum_{r=1}^{k} r^2 + \sum_{r=1}^{k} r. \end{aligned}$$

Using the arithmetic series summation formula and the identity given in the question,

$$\begin{aligned} N_k &= 2 \cdot \frac{1}{6}k(k+1)(2k+1) + \frac{1}{2}k(k+1) \\ &= k(k+1)\left[\frac{1}{3}(2k+1) + \frac{1}{2}\right] \\ &= \frac{1}{6}k(k+1)[2(2k+1)+3] \\ &= \frac{1}{6}k(k+1)(4k+5) \end{aligned}$$

as desired.

2. **(i)** If $A = \arctan\frac{1}{2}$ and $B = \arctan\frac{1}{3}$ (and A and B are acute), then $\tan A = \frac{1}{2}$ and $\tan B = \frac{1}{3}$. Now by the tangent compound angle identity,

$$\begin{aligned} \tan(A+B) &= \frac{\tan A + \tan B}{1 - \tan A \tan B} \\ &= \frac{\frac{1}{2} + \frac{1}{3}}{1 - \frac{1}{2} \cdot \frac{1}{3}} \\ &= 1 \\ \implies A + B &= \arctan 1 \\ &= \frac{\pi}{4} \end{aligned}$$

as desired.

Now, if $\arctan\frac{1}{p} + \arctan\frac{1}{q} = \frac{\pi}{4}$, then applying the tangent function to both sides and using the same identity,

$$\begin{aligned} \frac{\frac{1}{p} + \frac{1}{q}}{1 - \frac{1}{p} \cdot \frac{1}{q}} &= \arctan\frac{\pi}{4} \\ \implies \frac{\frac{p+q}{pq}}{pq-1} pq &= 1 \\ \implies \frac{p+q}{pq-1} &= 1 \\ \implies pq - p - q - 1 &= 0 \\ \implies (p-1)(q-1) - 2 &= 0 \\ \implies (p-1)(q-1) &= 2 \end{aligned}$$

as required.

The only two factors of 2 are 2 and 1, so that $(p, q) = (2, 3)$ or $(3, 2)$.

(ii) Similarly to before, the given equation leads us to

$$\begin{aligned}
&\frac{\frac{1}{r}+\frac{s}{s+t}}{1-\frac{1}{r}\cdot\frac{s}{s+t}} = 1 \\
\implies &\frac{rs+s+t}{rs+rt} = \frac{rs+rt-s}{rs+rt} \\
\implies &s+t = rt-s \\
\implies &r = \frac{2s}{t}+1.
\end{aligned}$$

For r to be an integer we require $\frac{2s}{t}$ to be an integer. However, we know s and t are coprime, so either $t=1$ or $t=2$.

If $t=1$ then $r=2s+1$ and if $t=2$ then $r=s+1$.

3. We first want to prove that $\cos^4\theta - \sin^4\theta \equiv \cos 2\theta$. Taking the difference of two squares,

$$\text{LHS} \equiv (\cos^2\theta+\sin^2\theta)(\cos^2\theta-\sin^2\theta)$$

and since $\sin^2\theta+\cos^2\theta\equiv 1$,

$$\text{LHS} \equiv \cos^2\theta-\sin^2\theta \equiv \cos 2\theta \equiv \text{RHS}$$

using the inverse cosine double angle identity, and we're done. We can prove that $\cos^4\theta+\sin^4\theta \equiv 1-\frac{1}{2}\sin^2 2\theta$ similarly: factorising,

$$\text{LHS} \equiv (\cos^2\theta+\sin^2\theta)^2 - 2\cos^2\theta\sin^2\theta = 1-2\cos^2\theta\sin^2\theta$$

and so using the inverse sine double angle identity,

$$\text{LHS} \equiv 1-\frac{1}{2}\sin^2 2\theta \equiv \text{RHS}.$$

Now we move to evaluating the given integrals.

Let $A=\displaystyle\int_0^{\frac{\pi}{2}}\cos^4\theta\,\mathrm{d}\theta$ and let $B=\displaystyle\int_0^{\frac{\pi}{2}}\sin^4\theta\,\mathrm{d}\theta$. So,

$$\begin{aligned}
A-B &= \int_0^{\frac{\pi}{2}}(\cos^4\theta-\sin^4\theta)\,\mathrm{d}\theta \\
&= \int_0^{\frac{\pi}{2}}\cos 2\theta\,\mathrm{d}\theta \\
&= \left[\frac{1}{2}\sin 2\theta\right]_0^{\frac{\pi}{2}} \\
&= \frac{1}{2}\sin\pi \\
&= 0
\end{aligned}$$

and so $A=B$. Similarly,

$$\begin{aligned}
A+B &= \int_0^{\frac{\pi}{2}}(\cos^4\theta+\sin^4\theta)\,\mathrm{d}\theta \\
&= \int_0^{\frac{\pi}{2}}(1-\frac{1}{2}\sin^2 2\theta)\,\mathrm{d}\theta.
\end{aligned}$$

Using the identity $\sin^2 x \equiv \frac{1}{2} - \frac{1}{2}\cos 2x$,

$$\begin{aligned} A + B &= \int_0^{\frac{\pi}{2}} \left(1 - \frac{1}{2}\left(\frac{1}{2} - \frac{1}{2}\cos 4\theta\right)\right) \mathrm{d}\theta \\ &= \frac{1}{4}\int_0^{\frac{\pi}{2}} (\cos 4\theta + 3)\,\mathrm{d}\theta \\ &= \frac{1}{4}\left[\frac{1}{4}\sin 4\theta + 3\theta\right]_0^{\frac{\pi}{2}} \\ &= \frac{1}{4}\left(\frac{1}{4}\sin 2\pi + \frac{3\pi}{2}\right) \\ &= \frac{3\pi}{8}. \end{aligned}$$

Therefore, $A = B = \frac{3\pi}{16}$.

We now use a similar method for the next two integrals. We start by trying to simplify $\cos^6\theta - \sin^6\theta$ into something easily integrable. Factorising, we have

$$\begin{aligned} \cos^6\theta - \sin^6\theta &\equiv (\cos^2\theta - \sin^2\theta)^3 + 3\cos^4\theta\sin^2\theta - 3\cos^2\theta\sin^4\theta \\ &\equiv (\cos^2\theta - \sin^2\theta)^3 + 3\cos^2\theta\sin^2\theta(\cos^2\theta - \sin^2\theta) \\ &\equiv \cos^3 2\theta + 3\cos^2\sin^2\cos 2\theta \\ &\equiv \cos^3 2\theta + \frac{3}{4}\sin^2 2\theta\cos 2\theta \\ &\equiv (1 - \sin^2 2\theta)\cos 2\theta + \frac{3}{4}\sin^2 2\theta\cos 2\theta \\ &\equiv \cos 2\theta - \frac{1}{4}\sin^2 2\theta\cos 2\theta. \end{aligned}$$

Similarly,

$$\begin{aligned} \cos^6\theta + \sin^6\theta &\equiv (\cos^2\theta + \sin^2\theta)^3 - 3\cos^2\theta\sin^4\theta - 3\cos^4\theta\sin^2\theta \\ &\equiv (\cos^2\theta + \sin^2\theta)^3 - 3\cos^2\theta\sin^2\theta(\sin^2\theta + \cos^2\theta) \\ &\equiv 1 - 3\cos^2\theta\sin^2\theta \\ &\equiv 1 - \frac{3}{4}\sin^2 2\theta \\ &\equiv 1 - \frac{3}{4}\left(\frac{1}{2} - \frac{1}{2}\cos 4\theta\right) \\ &\equiv \frac{5}{8} + \frac{3}{8}\cos 4\theta. \end{aligned}$$

Now let $C = \int_0^{\frac{\pi}{2}} \cos^6\theta\,\mathrm{d}\theta$ and let $D = \int_0^{\frac{\pi}{2}} \sin^6\theta\,\mathrm{d}\theta$. Hence,

$$\begin{aligned} C - D &= \int_0^{\frac{\pi}{2}} \left(\cos^6\theta - \sin^6\theta\right)\mathrm{d}\theta \\ &= \int_0^{\frac{\pi}{2}} \left(\cos 2\theta - \frac{1}{4}\sin^2 2\theta\cos 2\theta\right)\mathrm{d}\theta \\ &= \left[\frac{1}{2}\sin 2\theta - \frac{1}{24}\sin^3 2\theta\right]_0^{\frac{\pi}{2}} \\ &= \frac{1}{2}\sin\pi - \frac{1}{24}\sin^3\pi \\ &= 0 \end{aligned}$$

and so $C = D$. Similarly,

$$\begin{aligned} C + D &= \int_0^{\frac{\pi}{2}} \left(\cos^6\theta + \sin^6\theta\right) \mathrm{d}\theta \\ &= \int_0^{\frac{\pi}{2}} \left(\frac{5}{8} + \frac{3}{8}\cos 4\theta\right) \mathrm{d}\theta \\ &= \left[\frac{5}{8}\theta + \frac{3}{32}\sin 4\theta\right]_0^{\frac{\pi}{2}} \\ &= \frac{5}{8}\left(\frac{\pi}{2}\right) + \frac{3}{32}\sin 2\pi \\ &= \frac{5\pi}{16}. \end{aligned}$$

Therefore $C = D = \frac{5\pi}{32}$.

4. We first factorise $x + b + c$ out of the expression $x^3 - 3xbc + b^3 + c^3$ using a multiplication grid.

	x^2	$-bx - cx$	$-bc + b^2 + c^2$
x	x^3	$-bx^2 - cx^2$	$-bcx + b^2x + c^2x$
b	bx^2	$-b^2x - bcx$	$-b^2c + b^3 + bc^2$
c	cx^2	$-bcx - c^2x$	$-bc^2 + b^2c + c^3$

Hence we see that

$$x^3 - 3xbc + b^3 + c^3 = (x + b + c)(x^2 - bx - cx - bc + b^2 + c^2$$

and so our quadratic expression is

$$Q(x) = x^2 - bx - cx - bc + b^2 + c^2.$$

Therefore,

$$\begin{aligned} 2Q(x) &= 2x^2 - 2bx - 2cx - 2bc + 2b^2 + 2c^2 \\ &= (x^2 - 2bx + b^2) + (x^2 - 2cx + c^2) + (b^2 - 2bc + c^2) \\ &= (x - b)^2 + (x - c)^2 + (b - c)^2, \end{aligned}$$

an expression with three square terms as desired. Now, if the equations $ay^2 + by + c = 0$ and $by^2 + cy + a = 0$ have a common root k, then they are simultaneously true when $y = k$. So, we have

$$ak^2 + bk + c = 0 \tag{1}$$

and

$$bk^2 + ck + a = 0. \tag{2}$$

If $k = 0$ then $c = a = 0$, but we're given $a \neq 0$ so k cannot be zero and therefore also $c \neq 0$. Setting k^2 equal in eqs. (1) and (2), we come to

$$\begin{aligned} \frac{-bk - c}{a} &= \frac{-ck - a}{b} \\ \implies b^2k + bc &= ack + a^2 \\ \implies (ac - b^2)k &= bc - a^2 \end{aligned} \tag{3}$$

as we wanted. (Here we use the fact that $a, b \neq 0$.) Similarly, setting k equal we come to

$$\begin{aligned} \frac{-ak^2 - c}{b} &= \frac{-bk^2 - a}{c} \\ \implies ack^2 + c^2 &= b^2k^2 + ab \\ \implies (ac - b^2)k^2 &= ab - c^2, \end{aligned} \tag{4}$$

a similar expression involving k^2 (using the fact that $b, c \neq 0$). Squaring eq. (3) gives

$$(ac - b^2)^2k^2 = (bc - a^2)^2 \tag{5}$$

and now dividing eq. (5) by eq. (4) (which we can do as we're given $ac \neq b^2$ and we know $k \neq 0$, so also $ab - c^2 \neq 0$), we come to

$$\begin{aligned} \frac{(ac - b^2)^2k^2}{(ac - b^2)k^2} &= \frac{(bc - a^2)^2}{ab - c^2} \\ \implies (ac - b^2)(ab - c^2) &= (bc - a^2)^2, \end{aligned}$$

as required. This implies that

$$(ac - b^2)(ab - c^2) - (bc - a^2)^2 = 0.$$

Expanding fully, we get

$$a^2bc - ac^3 - bc^3 + b^2c^2 - b^2c^2 + a^2bc + a^2bc - a^4 = 0$$

which, collecting and cancelling terms, becomes

$$3a^2bc - ab^3 - ac^3 - a^4 = 0.$$

Finally, dividing by $-a$, we have

$$a^3 - 3abc + b^3 + c^3 = 0$$

as required. Hence, as shown earlier, this can be written as

$$(a + b + c)Q(a) = 0$$

which is the same as

$$(a + b + c)\left(\frac{(a - b)^2 + (b - c)^2 + (a - c)^2}{2}\right) = 0.$$

Therefore either

$$\left(\frac{(a - b)^2 + (b - c)^2 + (a - c)^2}{2}\right) = 0$$

which implies $a = b = c$ (and so the two equations are identical as all of the coefficients are the same), or

$$a + b + c = 0.$$

If $a + b + c = 0$, then we must have $k^2 = k = 1$ and so $k = 1$ as this is the only value of k that will satisfy both eqs. (1) and (2). Hence, either $k = 1$ or the equations are identical.

5. **(i)** Let x be the length of one edge of the octahedron. As shown in the fig. 1, we start by constructing medians from two opposite vertices to the midpoint of the edge in between them.

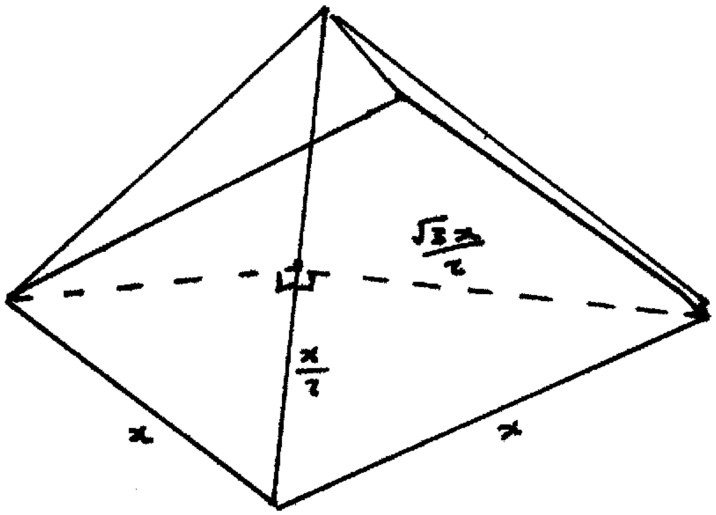


Figure 1

This forms two right triangles with hypoteneuse x and common side length $\frac{x}{2}$. Pythagoras gives that the third side of each triangle (the length of the median we constructed) is $\sqrt{x^2 - \frac{x^2}{4}} = \frac{\sqrt{3}x}{2}$. Now in fig. 2 we join the two opposite corners directly:

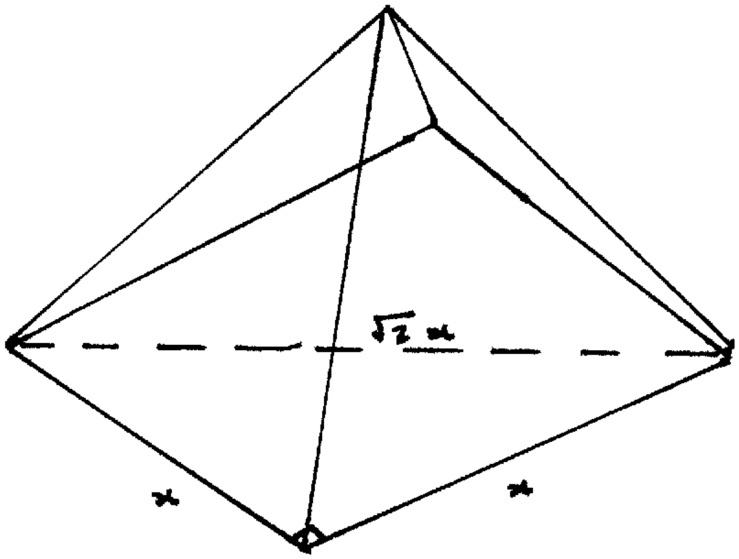


Figure 2

This forms a right triangle with the two edges, and so by Pythagoras the length of this connecting line is $\sqrt{x^2 + x^2} = \sqrt{2}x$.

Therefore if θ is the angle between any two faces on the octahedron, we have constructed an isosceles triangle as shown in fig. 3:

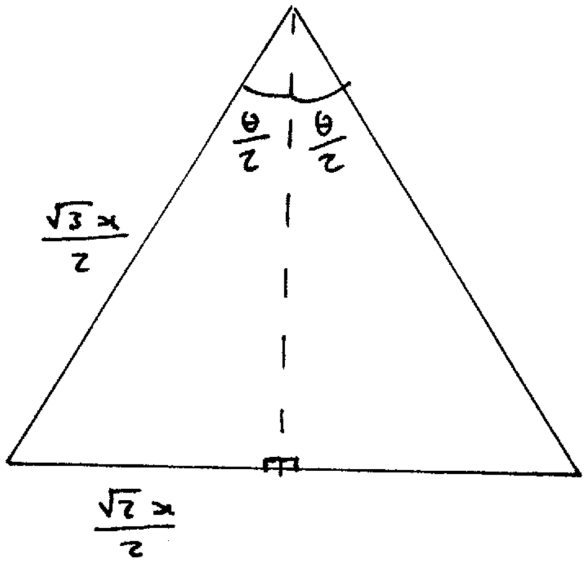


Figure 3

Splitting the triangle into two right triangles, by Pythagoras the median is $\sqrt{\frac{3x^2}{4} - \frac{2x^2}{4}} = \frac{x}{2}$. So,

$$\begin{aligned} \cos\left(\frac{\theta}{2}\right) &= \frac{\frac{x}{2}}{\frac{\sqrt{3}x}{2}} \\ &= \frac{\sqrt{3}}{3} \\ \implies \cos^2\left(\frac{\theta}{2}\right) &= \frac{1}{3}. \end{aligned}$$

Using the identity $\cos 2x \equiv \cos^2 x - \sin^2 x$,

$$\begin{aligned} \cos\theta &= \cos^2\left(\frac{\theta}{2}\right) - \sin^2\left(\frac{\theta}{2}\right) \\ &= \cos^2\left(\frac{\theta}{2}\right) - \left(1 - \cos^2\left(\frac{\theta}{2}\right)\right) \\ &= 2\cos^2\left(\frac{\theta}{2}\right) - 1 \\ &= 2\left(\frac{\sqrt{3}}{3}\right)^2 - 1 \\ &= \frac{2}{3} - 1 \\ &= -\frac{1}{3} \end{aligned}$$

and so

$$\theta = \arccos\left(-\frac{1}{3}\right)$$

as desired.

(ii) The volume of an octahedron, which is made up of two square-based pyramids, is given by $V_O = 2 \cdot \frac{x^2 h}{3}$ where h is the distance from the center of the octahedron to one of its vertices. As found earlier, the distance between two vertices is $\sqrt{2}x$ and so we must have

$$\begin{aligned} h &= \frac{\sqrt{2}x}{2} \\ \implies V_O &= 2 \cdot \frac{x^2 \frac{\sqrt{2}x}{2}}{3} \\ &= \frac{\sqrt{2}x^3}{3}. \end{aligned}$$

Now, the centre of an equilateral triangle is $\frac{2}{3}$ of the way along any median from a vertex. We constructed a triangle earlier (in fig. 3) with two such medians as two of its sides, and the third side length — the distance between two opposite vertices of the octahedron — was $\sqrt{2}x$. So, using similar triangles, the distance y between two points $\frac{2}{3}$ of the way down the median, that is the distance between the centres of two adjacent faces, is

$$y = \frac{1}{3}\sqrt{2}x = \frac{\sqrt{2}x}{3}.$$

This is the side length of a cube with its corners on the centre of each face of the octahedron, and so the volume of such a cube is

$$V_C = \left(\frac{\sqrt{2}}{3}\right)^3 = \frac{2\sqrt{2}x^3}{27}.$$

So, the ratio of the volumes is

$$\frac{V_O}{V_C} = \frac{\left(\frac{\sqrt{2}x^3}{3}\right)}{\left(\frac{2\sqrt{2}x^3}{27}\right)} = \frac{27}{2 \cdot 3} = \frac{9}{2}.$$

6. (i) We're given the two equations

$$x^2 - y^2 = (x - y)^3 \tag{6}$$

and

$$x - y = d \tag{7}$$

where $d \neq 0$. Taking the difference of two squares, eq. (6) simplifies to

$$\begin{aligned} (x - y)(x + y) &= (x - y)^3 \\ \implies x + y &= (x - y)^2 \end{aligned} \tag{8}$$

and substituting eq. (7) into eq. (8) gives

$$x + y = d^2. \tag{9}$$

So, adding eqs. (7) and (9),

$$\begin{aligned} 2x &= d^2 + d \\ \implies x &= \frac{d(d+1)}{2} \end{aligned}$$

and subtracting them,

$$\begin{aligned} 2y &= d^2 - d \\ \implies y &= \frac{d(d-1)}{2}. \end{aligned}$$

Now let $x = \sqrt{m}$ and $y = \sqrt{n}$. We want x^2 and y^2 to both be integers larger than 100. We have formulas for x and y in terms of their difference d, so let us pick such a difference that is large enough to ensure x^2 and y^2 are sufficiently large; let $d = 10$. Then,

$$x = \frac{10(10+1)}{2} = 55$$

and

$$y = \frac{10(10-1)}{2} = 45.$$

So, one possible pair of solutions is $m = x^2 = 55^2 = 3025$ and $n = y^2 = 45^2 = 2025$. We can check this:

$$3025 - 2025 = (55-45)^3 = 10^3 = 1000.$$

(ii) We're given

$$x^3 - y^3 = (x-y)^4 \tag{10}$$

and

$$x - y = d \tag{11}$$

where $d \neq 0$. Taking the difference of two cubes now on eq. (10), we come to

$$\begin{aligned} (x-y)(x^2+xy+y^2) &= (x-y)^4 \\ \implies x^2+xy+y^2 &= (x-y)^3. \end{aligned}$$

The left hand side of this factorises further to

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

and so substituting in eq. (11),

$$\begin{aligned} d^2 + 3xy &= d^3 \\ \implies 3xy &= d^3 - d^2 \end{aligned}$$

as required.

Next, substituting in $y = x - d$ leads us to

$$\begin{aligned} 3x(x-d) &= d^3 - d^2 \\ \implies 3x^2 - 3dx + (d^2 - d^3) &= 0 \\ \implies x &= \frac{3d \pm \sqrt{9d^2 - 4(3)(d^2-d^3)}}{2(3)} \\ \implies 2x &= d \pm \frac{\sqrt{9d^2 - 12(d^2-d^3)}}{3} \\ &= d \pm \frac{\sqrt{12d^3 - 2d^2}}{3} \\ &= d \pm \frac{d\sqrt{3}\sqrt{4d-1}}{3} \\ &= d \pm d\sqrt{\frac{4d-1}{3}} \end{aligned} \tag{12}$$

as desired.
From this we also know that

$$2y = 2x - 2d = -d \pm d\sqrt{\frac{4d-1}{3}}. \tag{13}$$

Now let $m = x$ and $n = y$, with $m, n \in \mathbb{Z}^+$. With reference to eqs. (12) and (13), we see that we must choose an integer difference d such that $\frac{4d-1}{3}$ is a perfect square. So,

$$\frac{4d-1}{3} = k^2$$

where $k \in \mathbb{Z}^+$, and so

$$d = \frac{3k^2+1}{4}.$$

After $k = 0$ and $k = 1$ (neither of which work because we require $m, n \neq 0$), the first value of k which satisfies this equation is $k = 3$ which gives $d = 7$.
So, using eq. (12),

$$\begin{aligned} 2m &= 7 + 7\sqrt{\frac{4(7)-1}{3}} \\ &= 7 + 7\sqrt{9} \\ \implies m &= \frac{7+7(3)}{2} = 14. \end{aligned}$$

Similarly, using eq. (13),

$$\begin{aligned} 2n &= -7 + 7\sqrt{\frac{4(7)-1}{3}} \\ &= -7 + 7\sqrt{9} \\ \implies n &= \frac{7(3)-7}{2} = 7. \end{aligned}$$

We can check that these values of m and n work:

$$14^3 - 7^3 = 2744 - 343 = 2401$$

and

$$(14-7)^4 = 7^4 = 2401.$$

7. **(i)** The distance D between a point on L_1 and a point on L_2 is given by the modulus of

the vector difference of these two points:

$$\begin{aligned}
D &= \left\| \left[\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix} \right] - \left[\begin{pmatrix} 4 \\ -2 \\ 9 \end{pmatrix} + \mu \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} \right] \right\| \\
&= \left\| \begin{pmatrix} 1 + 2\lambda - 4 - \mu \\ 0 + 2\lambda + 2 - 2\mu \\ 2 - 3\lambda - 9 + 2\mu \end{pmatrix} \right\| \\
\implies D^2 &= (2\lambda - \mu - 3)^2 + (2\lambda - 2\mu + 2)^2 + (-3\lambda + 2\mu - 7)^2 \\
&= 4\lambda^2 - 2\lambda\mu - 6\lambda - 2\lambda\mu + \mu^2 + 3\mu - 6\lambda + 3\mu + 9 \\
&\quad + 4\lambda^2 - 4\lambda\mu + 4\lambda - 4\lambda\mu + 4\mu^2 - 4\mu + 4\lambda - 4\mu + 4 \\
&\quad + 9\lambda^2 - 6\lambda\mu + 21\lambda - 6\lambda\mu + 4\mu^2 - 14\mu + 21\lambda - 14\mu + 49 \\
&= 17\lambda^2 + 9\mu^2 - 24\lambda\mu + 38\lambda - 30\mu + 62 \\
&= 16\lambda^2 + 9\mu^2 - 24\lambda\mu + 36\lambda - 30\mu + 25 + \lambda^2 - 2\lambda + 1 + 36 \\
&= 16\lambda^2 + 9\mu^2 - 24\lambda\mu + 36\lambda - 30\mu + 25 + (\lambda - 1)^2 + 36 \\
&= (3\mu - 4\lambda - 5)^2 + (\lambda - 1)^2 + 36,
\end{aligned}$$

as required. The minimum distance between these two lines, therefore, is when the first two terms on the right hand side are zero, and so $D^2 = 36 \implies D = 6$. For this to occur, we require

$$3\mu - 4\lambda - 5 = 0$$

and

$$\lambda - 1 = 0. \tag{14}$$

Equation (14) gives $\lambda = 1$, and substituting this value into eq. (14) gives us

$$\begin{aligned}
&3\mu - 4 \cdot 1 - 5 = 0 \\
&\implies \mu = \frac{4 + 5}{3} = \frac{9}{3} = 3.
\end{aligned}$$

Hence, using the original vector equations from the question, when $(\lambda, \mu) = (1, 3)$, the point on line L_1 has coordinates

$$\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + 1 \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix}$$

and the point on line L_2 has coordinates

$$\begin{pmatrix} 4 \\ -2 \\ 9 \end{pmatrix} + 3 \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 7 \\ 4 \\ 3 \end{pmatrix}.$$

(ii) We will follow a similar method to above. Let the distance between a point on L_3

and a point on L_4 be S. So, as before,

$$\begin{aligned} S &= \left\| \left[\begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} + \alpha \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right] - \left[\begin{pmatrix} 3 \\ 3 \\ -2 \end{pmatrix} + \beta \begin{pmatrix} 4k \\ 1-k \\ -3k \end{pmatrix} \right] \right\| \\ &= \left\| \begin{pmatrix} 2+0-3-4\beta k \\ 3+\alpha-3-\beta(1-k) \\ 5+0+2+3\beta k \end{pmatrix} \right\| \\ \implies S^2 &= (-1-4\beta k)^2 + (3+\alpha-3-\beta+\beta k)^2 + (7+3\beta k)^2 \\ &= (4\beta k+1)^2 + (\beta k+\alpha-\beta)^2 + (3\beta k+7)^2 \\ &= 16\beta^2k^2 + 4\beta k + 4\beta k + 1 \\ &\quad + \beta^2k^2 + \alpha\beta k - \beta^2 k + \alpha\beta k + \alpha^2 - \alpha\beta - \beta^2 k - \alpha\beta + \beta^2 \\ &\quad + 9\beta^2k^2 + 21\beta k + 21\beta k + 49 \\ &= 26\beta^2k^2 - 2\beta^2 k + 50\beta k + 2\alpha\beta k - 2\alpha\beta + \alpha^2 + \beta^2 + 50 \\ &= (5\beta k+5)^2 + \beta^2k^2 - 2\beta^2 k + 2\alpha\beta k - 2\alpha\beta + \alpha^2 + \beta^2 + 25 \\ &= (5\beta k+5)^2 + (\beta k - \beta + \alpha)^2 + 25. \end{aligned}$$

This is an expression quite similar to the one derived in the first part. The minimum distance between the two lines if $k \neq 0$ is $S = \sqrt{25} = 5$, and this occurs when both

$$5\beta k + 5 = 0 \tag{15}$$

and

$$\beta k - \beta + \alpha = 0, \tag{16}$$

and so by eq. (15), $\beta = -\frac{1}{k}$ and by eq. (16), $\alpha = \beta - \beta k = 1 - \frac{1}{k}$.
However, if $k = 0$ then eq. (15) cannot be true, and so the minimum distance is different. In the case $k = 0$, the two lines are parallel, as the direction vector of L_4 becomes $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ like that of L_3. So, since they are parallel, we can take any arbitrary point on L_4 and work from there; let's use $\beta = 0$.
In this case, the distance of separation becomes

$$S^2 = (0+5)^2 + (0+0+\alpha)^2 + 25 = 50 + \alpha^2$$

so the minimum distance is when $\alpha^2 = 0$ (and therefore whenever $\alpha = \beta$), and so

$$\begin{aligned} S^2 &= 50 \\ \implies S &= \sqrt{50}. \end{aligned}$$

8. Let $f(x) = ax^3 - 6ax^2 + (12a+12)x - (8a+16)$ and let $g(x) = x^3$. So,

$$f'(x) = 3ax^2 - 12ax + (12a+12)$$

and

$$g'(x) = 3x^2.$$

For the two curves to 'touch' it is necessary that both their values and their derivatives are the same. At $x = 2$,

$$\begin{aligned} f(2) &= a(2)^3 - 6a(2)^2 + (12a + 12)(2) - (8a + 16) \\ &= 8a - 24a + 24a + 24 - 8a - 16 \\ &= 24 - 16 \\ &= 8 \end{aligned}$$

and $g(2) = 2^3 = 8$. Also,

$$\begin{aligned} f'(2) &= 3a(2)^2 - 12a(2) + (12a + 12) \\ &= 12a - 24a + 12a + 12 \\ &= 12 \end{aligned}$$

and $g'(x) = 3(2)^2 = 12$. So, the curves touch at $(2, 8)$. Now, let us set $f(x) = g(x)$ to find their other intersections.

$$\begin{aligned} & ax^3 - 6ax^2 + (12a + 12)x - (8a + 16) = x^3 \\ \implies & (a - 1)x^3 - 6ax^2 + (12a + 12)x - (8a + 16) = 0 \end{aligned}$$

We know $x = 2$ is a solution, so we can factorise out $(x - 2)$:

$$(x - 2)\left[(a - 1)x^2 + (2 + 4a)x + (4a + 8)\right] = 0.$$

In fact, since the curves touch at $x = 2$, we know that $x = 2$ is a double root — that is, we can pull out another factor of $(x - 2)$:

$$(x - 2)(x - 2)\left[(a - 1)x - (2a + 4)\right] = 0.$$

So, our other intersection point is given by

$$\begin{aligned} & (a - 1)x - (2a + 4) = 0 \\ \implies & x = \frac{2(a + 2)}{a - 1}. \end{aligned}$$

At this point, the y-value is

$$y = x^3 = \left(\frac{2(a + 2)}{a - 1}\right)^3.$$

So, the curves intersect again at $\left(\frac{2(a+2)}{a-1}, \left[\frac{2(a+2)}{a-1}\right]^3\right)$.

(i) When $a = 2$, then

$$f(x) = 2x^3 - 12x^2 + 36x - 32$$

and

$$f'(x) = 6x^2 - 24x + 36.$$

Clearly $f(0) = -32$. At any turning points,

$$\begin{aligned} & 6x^2 - 24x + 36 = 0 \\ \implies & x^2 - 4x + 6 = 0 \end{aligned}$$

which has a negative discriminant, implying there are no turning points. When $a = 2$, the second point of intersection is at $(8, 8^3) = (8, 512)$ using the general result found before. The curves also touch at $(2, 8)$ as for any value of a.

Hence, fig. 4 shows a sketch of the two curves:

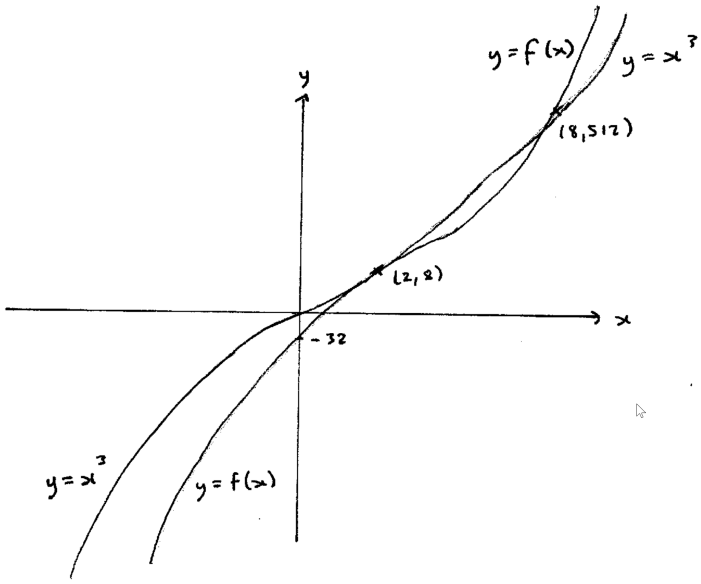


Figure 4

(ii) When $a = 1$, then

$$f(x) = x^3 - 6x^2 + 24x - 24$$

and

$$f'(x) = 3x^2 - 12x + 24.$$

We see that $f(0) = -24$ and wherever the derivative is zero,

$$\begin{aligned} 3x^2 - 12x + 24 &= 0 \\ \implies x^2 - 4x + 8 &= 0 \end{aligned}$$

which again has a negative discriminant, so there are no turning points. When $a = 1$, the coordinates of the second point of intersection found before are undefined (as the denominator is $a - 1$); so, the curves touch at $(2, 8)$ only.

Hence, fig. 5 shows a sketch of the two curves:

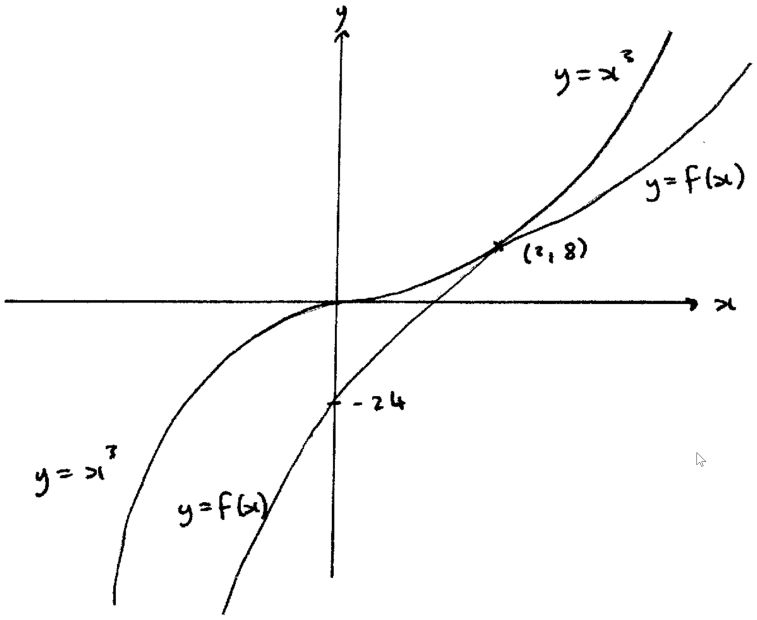


Figure 5

(iii) When $a = -2$, then

$$f(x) = -2x^3 + 12x^2 - 12x$$

and

$$f'(x) = -6x^2 + 24x - 12.$$

Also,

$$f''(x) = -12x + 24.$$

Clearly $f(0) = 0$ so the curve passes through the origin. At any turning points,

$$\begin{aligned} -6x^2 + 24x - 12 &= 0 \\ \implies x^2 - 4x + 2 &= 0 \\ \implies x &= \frac{4 \pm \sqrt{16 - 4(1)(2)}}{2} \\ &= 2 \pm \sqrt{2}. \end{aligned}$$

We know the nature of these turning points from the negative coefficient of x^3. At $a = -2$, the second intersection is $(0, 0^3) = (0, 0)$, and the curves still meet at $(2, 8)$. So, fig. 6 shows a sketch of the two curves:

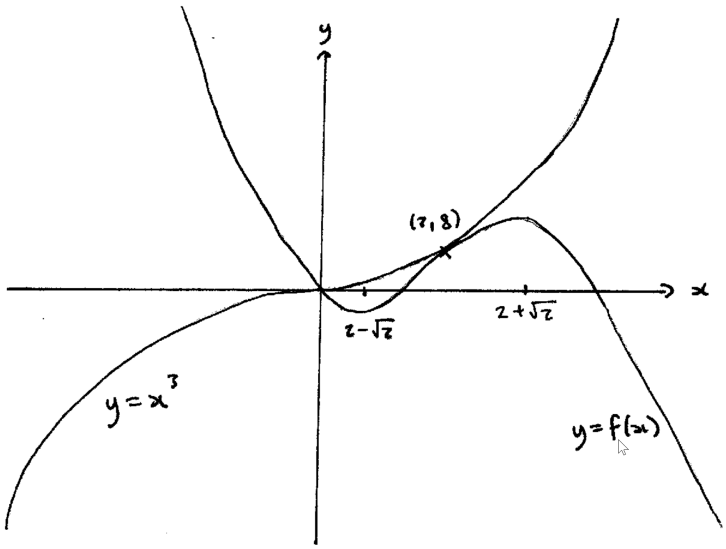


Figure 6

Section B: Mechanics

9. When the particle is about to slide down the plain, fig. 7 models the situation:

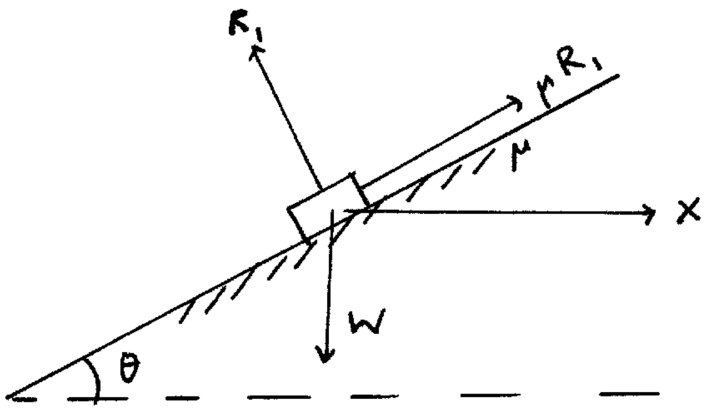


Figure 7

Similarly, fig. 8 shows when the particle is about to slide up the plane:

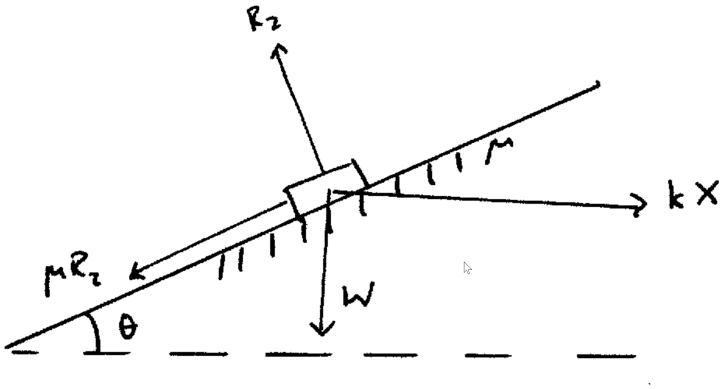


Figure 8

(Both of these diagrams use the face that at the slipping point, the frictional force is equal to the product of the coefficient of friction and the reaction force. The difference is that the frictional forces are in opposite directions.)

Applying Newton II vertically on the particle in the first diagram,

$$R_1 \cos\theta + \mu R_1 \sin\theta = W \tag{17}$$

and horizontally,

$$R_1 \sin\theta = \mu R_1 \cos\theta + X. \tag{18}$$

Similarly, applying Newton II vertically on the particle in the second diagram,

$$R_2 \cos\theta = \mu R_2 \sin\theta + W \tag{19}$$

and horizontally,

$$R_2 \sin\theta + \mu R_2 \cos\theta = kX. \tag{20}$$

Eliminating W from eqs. (17) and (19), we get

$$R_1 \cos\theta + \mu R_1 \sin\theta = R_2 \cos\theta - \mu R_2 \sin\theta \tag{21}$$

and eliminating X from eqs. (18) and (20), we get

$$R_1 \sin\theta - \mu R_1 \cos\theta = \frac{R_2 \sin\theta + \mu R_2 \cos\theta}{k}. \tag{22}$$

Equation (21) implies

$$R_1 = \frac{R_2 \cos\theta - \mu R_2 \sin\theta}{\cos\theta + \mu \sin\theta}$$

and eq. (22) implies

$$R_1 = \frac{R_2 \sin\theta + \mu R_2 \cos\theta}{k(\sin\theta - \mu\cos\theta)},$$

so we come to

$$\frac{R_2 \cos\theta - \mu R_2 \sin\theta}{\cos\theta + \mu \sin\theta} = \frac{R_2 \sin\theta + \mu R_2 \cos\theta}{k(\sin\theta - \mu\cos\theta)}$$

$$\implies k(\sin\theta - \mu\cos\theta)(\cos\theta - \mu\sin\theta) = (\cos\theta + \mu\sin\theta)(\sin\theta + \mu\cos\theta).$$

Expanding, we come to

$$k\sin\theta\cos\theta - k\mu\sin^2\theta - k\mu\cos^2\theta + k\mu^2\sin\theta\cos\theta \\ = \sin\theta\cos\theta + \mu\cos^2\theta + \mu\sin^2\theta + \mu^2\sin\theta\cos\theta$$

and bringing all terms to the left hand side,

$$k\sin\theta\cos\theta - \sin\theta\cos\theta + k\mu^2\sin\theta\cos\theta - \mu^2\sin\theta\cos\theta \\ - k\mu\sin^2\theta - \mu\sin^2\theta - k\mu\cos^2\theta - \mu\cos^2\theta = 0.$$

Now we can factorise:

$$(k-1)\sin\theta\cos\theta + (k-1)\mu^2\sin\theta\cos\theta - (k+1)\mu\sin^2\theta - (k+1)\mu\cos^2\theta = 0$$
$$\implies (k-1)(1+\mu^2)\sin\theta\cos\theta - (k+1)\mu(\sin^2\theta + \cos^2\theta) = 0$$
$$\implies (k-1)(1+\mu^2)\sin\theta\cos\theta = \mu(k+1)$$

as required.

Using the double angle identity, this simplifies to

$$\frac{(k-1)(1+\mu^2)\sin 2\theta}{2} = \mu(k+1)$$

which rearranges to

$$\frac{2\mu(k+1)}{(1+\mu^2)(k-1)} = \sin 2\theta.$$

(The denominator is non-zero as $k = 1$ would be a contradiction as the two situations would be the same, and μ^2 must be positive.)

Now for the situation to be physical we must have $0 \leqslant \theta \leqslant \frac{\pi}{2}$, and so $0 \leqslant 2\theta \leqslant \pi$, which implies

$$0 \leqslant \sin 2\theta \leqslant 1.$$

So,

$$\frac{2\mu(k+1)}{(1+\mu^2)(k-1)} \leqslant 1$$
$$\implies 2k\mu + 2\mu \leqslant k - 1 + k\mu^2 - \mu^2$$
$$\implies 2k\mu - k - k\mu^2 \leqslant -\mu^2 - 2\mu - 1$$
$$\implies k(\mu^2 - 2\mu + 1) \geqslant \mu^2 + 2\mu + 1$$
$$\implies k \geqslant \frac{(1+\mu)^2}{(1-\mu)^2}$$

as required.

10. Figure 9 shows a displacement-time graph of the Norman army and the two horsemen before any of them meet. The displacement s is always given from the Saxon army (which is at rest).

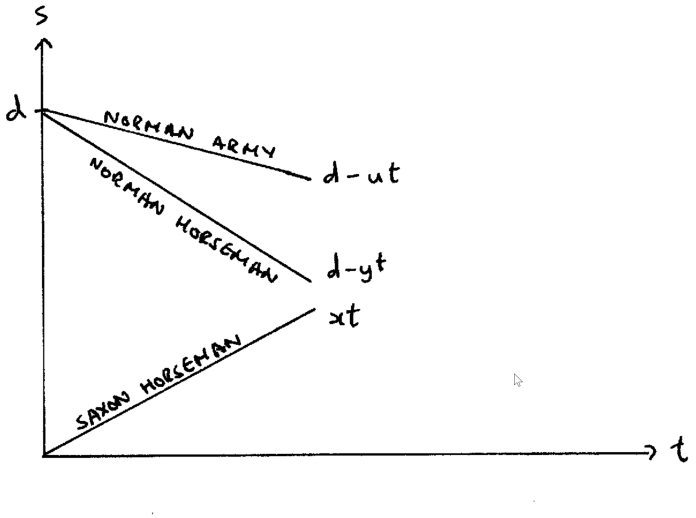


Figure 9

The Saxon horseman meets the Norman army when

$$\begin{aligned} xt &= d - ut \\ \implies t &= \frac{d}{x+u}. \end{aligned} \tag{23}$$

So, at this moment his displacement (from the Saxon army) is

$$s = xt = \frac{xd}{x+u}.$$

At this point he turns back at the same speed x and so his displacement is thereafter given by

$$\begin{aligned} s &= \frac{xd}{x+u} - x\left(t - \frac{d}{x+u}\right) \\ &= \frac{2xd}{x+u} - xt. \end{aligned}$$

Similarly, the Norman horseman meets the Saxon army when

$$\begin{aligned} d - yt &= 0 \\ \implies t &= \frac{d}{y}. \end{aligned} \tag{24}$$

At this point his displacement is zero, and he turns back with the same speed y towards the Norman army. So, after this moment, his displacement is given by

$$\begin{aligned} s &= y\left(t - \frac{d}{y}\right) \\ &= yt - d. \end{aligned}$$

Hence, the two horsemen meet whenever their displacements are equal; that is, when any of

$$xt = d - yt, \tag{25}$$
$$xt = yt - d, \tag{26}$$
$$\frac{2xd}{x+u} - xt = d - yt, \tag{27}$$
$$\frac{2xd}{x+u} - xt = yt - d \tag{28}$$

are true.

Equation (25) gives

$$\begin{aligned} xt + yt &= d \\ \implies t &= \frac{d}{x+y} \\ \implies s &= \frac{xd}{x+y}. \end{aligned}$$

This represents them passing when they are both on their outward journeys. This is their first meeting point no matter what their speeds, as they must pass each other to reach their respective destination armies.

However, when they pass for the second time could take a few different cases.

Equation (28) gives

$$\begin{aligned} \frac{2xd}{x+u} + d &= xt + yt \\ \implies t &= \frac{2xd}{(x+u)(x+y)} + \frac{d}{x+y} \qquad (29) \\ \implies s &= y\left(\frac{2xd}{(x+u)(x+y)} + \frac{d}{x+y}\right) - d \\ &= \frac{2xyd + yd(x+u) - d(x+u)(x+y)}{(x+u)(x+y)} \\ &= \frac{2xyd + xyd + uyd - x^2d - xyd - uxd - uyd}{(u+x)(x+y)} \\ &= \frac{xd(2y - x - u)}{(u+x)(x+y)}. \end{aligned}$$

Here they pass for the second time when they are both on their return journeys. So, the time at which this occurs, given by eq. (29), is greater than the times given by eqs. (23) and (24) when they meet their respective armies. So,

$$\begin{aligned} \frac{2xd}{(x+u)(x+y)} + \frac{d}{x+y} &> \frac{d}{y} \\ \implies \frac{d(3x+u)}{(u+x)(x+y)} &> \frac{d}{y} \\ \implies (3x+u)y &> (u+x)(x+y) \\ \implies 3xy + uy &> ux + uy + x^2 + xy \\ \implies 2xy &> ux + x^2 \\ \implies u &< 2y - x \end{aligned}$$

and also

$$\begin{aligned}\frac{2xd}{(x+u)(x+y)} + \frac{d}{x+y} &> \frac{d}{x+u} \\ \implies \frac{d(3x+u)}{(x+u)(x+y)} &> \frac{d}{x+u} \\ \implies \frac{3x+u}{x+y} &> 1 \\ \implies 2x+u &> y \\ \implies u &> y-2x\end{aligned}$$

and thus we have

$$y - 2x < u < 2y - x,$$

the condition in the question.

(i) However, if $u > 2y - x$ the first inequality is contradicted, and so the horsemen must collide for the second time before the Norman horseman meets the Saxon army (and thus still after the Saxon horseman turns back from the Norman army), and we would use eq. (27) instead. In this case the Saxon horseman is riding especially quickly.

(ii) Similarly, if $u < y - 2x$, the second equality is contradicted and the horsemen must collide for the second time before the Saxon horseman meets the Norman army (and after the Norman horseman turns back from the Saxon army), and we would use eq. (26) instead. In this case the Norman horseman is riding especially quickly.

11. The situation is shown in fig. 10:

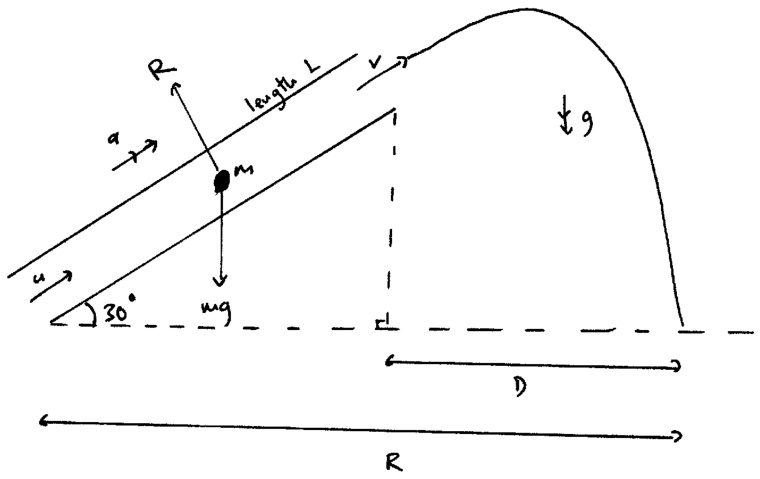


Figure 10

Let the particle have mass m and let the reaction force from the tube (upwards, perpendicular to the slope) be R. Let also the acceleration acting on the particle along the axis of the tube be a.

Applying Newton II to the particle in the tube, in the direction upwards parallel to the slope, we have

$$\begin{aligned} 0 - mg\sin 30^\circ &= ma \\ \implies a &= -g\sin 30^\circ = -\frac{g}{2}. \end{aligned}$$

Now let v be the final speed of the particle when it emerges from the tube. Therefore, using the equations of motion along the length of the tube L, we have

$$\begin{aligned} v^2 &= u^2 + 2aL \\ &= u^2 - gL. \end{aligned} \tag{30}$$

Now we consider the projectile motion of the particle after it emerges from the tube. Let the height of the upper end of the tube be h and let the time of flight of the particle from emerging from the tube to landing be T. So, using the equations of motion vertically,

$$\begin{aligned} h &= v\sin 30^\circ T - \frac{1}{2}gT^2 \\ &= \frac{v}{2}T - \frac{g}{2}T^2 \end{aligned} \tag{31}$$

and horizontally,

$$\begin{aligned} D &= v\cos 30^\circ T \\ &= \frac{\sqrt{3}vT}{2} \\ \implies T &= \frac{2D}{\sqrt{3}v}. \end{aligned} \tag{32}$$

Hence, substituting eq. (32) into eq. (31), we come to

$$\begin{aligned} h &= \frac{v}{2}\left(\frac{2D}{\sqrt{3}v}\right) - \frac{g}{2}\left(\frac{2D}{\sqrt{3}v}\right)^2 \\ &= \frac{\sqrt{3}D}{3} - \frac{4gD^2}{6v^2}. \end{aligned}$$

Now simple trigonometry gives us that $h = \frac{L}{2}$ and also substituting in eq. (30), we now have

$$\begin{aligned} \frac{L}{2} &= \frac{\sqrt{3}D}{3} - \frac{4gD^2}{6(u^2 - gL)} \\ \implies \frac{6L(u^2 - gL)}{2} &= \frac{6\sqrt{3}D(u^2 - gL)}{3} - 4gD^2 \\ \implies 4gD^2 - 2\sqrt{3}(u^2 - gL)D - 3L(u^2 - gL) &= 0 \end{aligned} \tag{33}$$

as desired.

Now differentiating implicitly with respect to L,

$$\begin{aligned} 4g\left[2D\cdot\frac{\mathrm{d}D}{\mathrm{d}L}\right] - 2\sqrt{3}\left[\frac{\mathrm{d}D}{\mathrm{d}L}\cdot(u^2 - gL) + D\cdot(-g)\right] - 3\left[1\cdot(u^2 - gL) + L\cdot(-g)\right] &= 0 \\ \implies 8gD\frac{\mathrm{d}D}{\mathrm{d}L} - 2\sqrt{3}\frac{\mathrm{d}D}{\mathrm{d}L}(u^2 - gL) + 2\sqrt{3}gD - 3(u^2 - gL) + 3gL &= 0. \end{aligned}$$

Collecting terms of $\frac{\mathrm{d}D}{\mathrm{d}L}$,

$$\frac{\mathrm{d}D}{\mathrm{d}L}\left[8gD - 2\sqrt{3}(u^2 - gL)\right] = 3(u^2 - gL) - 3gL - 2\sqrt{3}gD$$

and now rearranging and simplifying,

$$\begin{aligned}\frac{\mathrm{d}D}{\mathrm{d}L} &= \frac{3(u^2 - gL) - 3gL - 2\sqrt{3}gD}{8gD - 2\sqrt{3}(u^2 - gL)} \\ &= \frac{3(u^2 - 2gL) - 2\sqrt{3}gD}{8gD - 2\sqrt{3}(u^2 - gL)} \\ &= -\frac{2\sqrt{3}gD - 3(u^2 - 2gL)}{8gD - 2\sqrt{3}(u^2 - gL)}\end{aligned}$$

like we were hoping for.

By trigonometry, clearly

$$R = L\cos 30^\circ + D = \frac{\sqrt{3}L}{2} + D \tag{34}$$

and so

$$\begin{aligned}\frac{\mathrm{d}R}{\mathrm{d}L} &= \frac{\sqrt{3}}{2} + \frac{\mathrm{d}D}{\mathrm{d}L} \\ &= \frac{\sqrt{3}}{2} - \frac{2\sqrt{3}gD - 3(u^2 - 2gL)}{8gD - 2\sqrt{3}(u^2 - gL)}.\end{aligned}$$

Hence if $2D = L\sqrt{3}$, then $D = \frac{\sqrt{3}L}{2}$ and so

$$\begin{aligned}\frac{\mathrm{d}R}{\mathrm{d}L} &= \frac{\sqrt{3}}{2} + \frac{\mathrm{d}D}{\mathrm{d}L} \\ &= \frac{\sqrt{3}}{2} - \frac{2\sqrt{3}g\left(\frac{\sqrt{3}L}{2}\right) - 3(u^2 - 2gL)}{8g\left(\frac{\sqrt{3}L}{2}\right) - 2\sqrt{3}(u^2 - gL)} \\ &= \frac{\sqrt{3}}{2} - \frac{3gL - 3(u^2 - 2gL)}{4\sqrt{3}gL - 2\sqrt{3}(u^2 - gL)} \\ &= \frac{\sqrt{3}}{2} - \frac{9gL - 3u^2}{\sqrt{3}(6gL - 2u^2)} \\ &= \frac{\sqrt{3}}{2} - \frac{\sqrt{3}(9gL - 3u^2)}{6(3gL - u^2)} \\ &= \frac{\sqrt{3}}{2}\left[1 - \frac{3(3gL - u^2)}{3(3gL - u^1)}\right] \\ &= \frac{\sqrt{3}}{2}(1 - 1) \\ &= 0\end{aligned}$$

as we wanted to show.

Substituting the given value of $D = \frac{\sqrt{3}L}{2}$ into eq. (33), we have

$$\begin{aligned} 4gD^2 - 2\sqrt{3}(u^2 - gL)D - 3L(u^2 - gL) &= 0 \\ \implies 4g\left(\frac{\sqrt{3}L}{2}\right)^2 - 2\sqrt{3}(u^2 - gL)\left(\frac{\sqrt{3}L}{2}\right) - 3L(u^2 - gL) &= 0 \\ \implies 3gL^2 - 6L(u^2 - gL) &= 0 \\ \implies 3gL &= 6(u^2 - gL) \\ \implies 3gL + 6gL &= 6u^2 \\ \implies L &= \frac{6u^2}{9g} = \frac{2u^2}{3g}. \end{aligned}$$

Therefore, by eq. (34),

$$\begin{aligned} R &= \frac{\sqrt{3}L}{2} + D \\ &= \frac{\sqrt{3}L}{2} + \frac{\sqrt{3}L}{2} \\ &= \sqrt{3}L \\ &= \frac{2\sqrt{3}u^2}{3g}. \end{aligned}$$

Section C: Probability and Statistics

12. **(i)** Figure 11 shows a tree diagram of the possible outcomes:

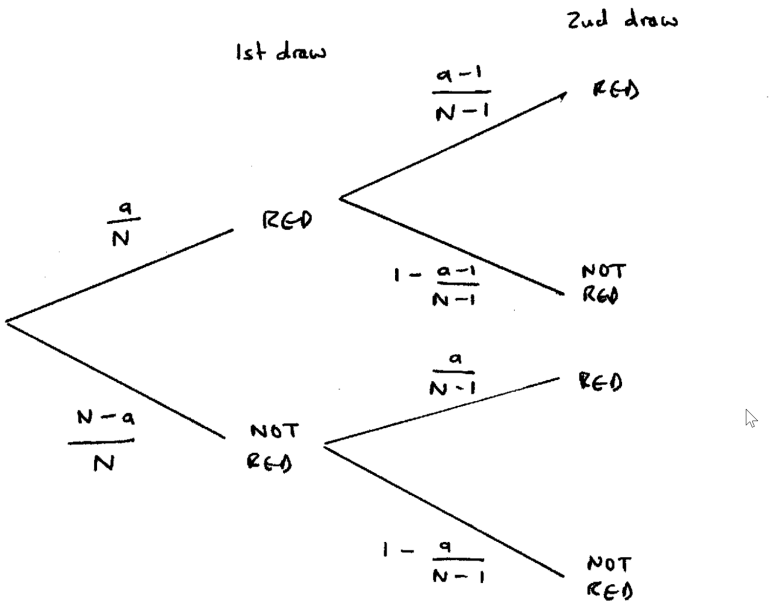


Figure 11

There are N sweets in total and a red sweets initially, so the probability that the

first sweet is red is just

$$\mathrm{P}(\text{1st sweet red}) = \frac{a}{N}.$$

If the first sweet is red, there are $N-1$ sweets left and $a-1$ red sweets left. However, if the first sweet is not red, there are $N-1$ sweets left and still a red sweets. So,

$$\begin{aligned}
\mathrm{P}(\text{2nd sweet red}) &= \left(\frac{a}{N}\right)\left(\frac{a-1}{N-1}\right) + \left(1-\frac{a}{N}\right)\left(\frac{a}{N-1}\right) \\
&= \frac{a(a-1)}{N(N-1)} + \frac{a(N-a)}{N(N-1)} \\
&= \frac{a\left[(a-1)+(N-a)\right]}{N(N-1)} \\
&= \frac{a(N-1)}{N(N-1)} \\
&= \frac{a}{N} \\
&= \mathrm{P}(\text{1st sweet red})
\end{aligned}$$

as we were trying to show.

(ii) The following tree diagram in fig. 12 is of the first coin toss and sweet draw.

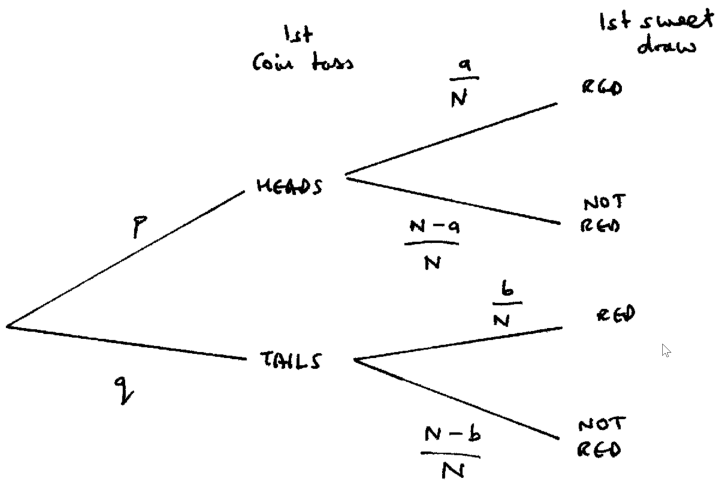


Figure 12

The probability that the first sweet is red is therefore

$$\mathrm{P}(\text{1st sweet red}) = \frac{pa}{N} + \frac{qb}{N} = \frac{pa+qb}{N}.$$

Now, covering all eight cases, the probability that the second sweet is red is

$$\begin{aligned}
\text{P(2nd sweet red)} &= \text{P(heads,red,heads,red)} + \text{P(heads,red,tails,red)} \\
&\quad + \text{P(heads,not red,heads,red)} + \text{P(heads,not red,tails,red)} \\
&\quad + \text{P(tails,red,heads,red)} + \text{P(tails,red,tails,red)} \\
&\quad + \text{P(tails,not red,heads,red)} + \text{P(tails,not red,tails,red)} \\
&= \left(\frac{pa}{N}\right)\left(\frac{p(a-1)}{N-1}\right) + \left(\frac{pa}{N}\right)\left(\frac{q(b+1)}{N+1}\right) \\
&\quad + \left(\frac{p(N-a)}{N}\right)\left(\frac{pa}{N-1}\right) + \left(\frac{p(N-a)}{N}\right)\left(\frac{qb}{N+1}\right) \\
&\quad + \left(\frac{qb}{N}\right)\left(\frac{p(a+1)}{N+1}\right) + \left(\frac{qb}{N}\right)\left(\frac{q(b-1)}{N-1}\right) \\
&\quad + \left(\frac{q(N-b)}{N}\right)\left(\frac{pa}{N+1}\right) + \left(\frac{q(N-b)}{N}\right)\left(\frac{qb}{N-1}\right) \\
&= \frac{p^2a(a-1) + p^2a(N-a) + q^2b(b-1) + q^2b(N-b)}{N(N-1)} \\
&\quad + \frac{pqa(b+1) + pqb(N-a) + pqb(a+1) + pqa(N-b)}{N(N+1)} \\
&= \frac{p^2a(N-1) + q^2b(N-1)}{N(N-1)} + \frac{pqa(N+1) + pqb(N+1)}{N(N+1)} \\
&= \frac{p^2a + q^2b + pqa + pqb}{N} \\
&= \frac{(p+q)(pa+qb)}{N} \\
&= \frac{pa+qb}{N} \\
&= \text{P(1st sweet red)}
\end{aligned}$$

as we hoped for.

13. **(i)** There are $\binom{11}{4}$ ways of choosing four discs overall. Out of these, there are $\binom{5}{4}$ ways of choosing four numbered discs, and so the probability that all four discs are numbered is

$$\text{P(all four discs numbered)} = \frac{\binom{5}{4}}{\binom{11}{4}} = \frac{1}{66}.$$

(ii) After the disc numbered "3" is chosen, there are $\binom{10}{3}$ ways to choose the remaining 3 discs, and there are $\binom{4}{3}$ ways for the remaining 3 discs to be numbered. So,

$$\text{P(all four discs numbered} \mid \text{disc "3" chosen 1st)} = \frac{\binom{4}{3}}{\binom{10}{3}} = \frac{4}{120} = \frac{1}{30}.$$

(iii) If the disc numbered "3" is chosen first, we need precisely one more numbered disc and 2 more non-numbered discs. So,

$$\text{P(exactly two numbered} \mid \text{disc "3" chosen 1st)} = \frac{\binom{4}{1}\cdot\binom{6}{2}}{\binom{10}{3}} = \frac{1}{2}.$$

(iv) There are $\binom{1}{1}\cdot\binom{4}{1}\cdot\binom{6}{2}$ ways to choose exactly two numbered discs given number "3" is chosen. There are $\binom{11}{4} - \binom{10}{4}$ ways to choose four discs including number "3". So,

$$\text{P(exactly two numbered} \mid \text{disc "3" chosen)} = \frac{\binom{1}{1}\cdot\binom{4}{1}\cdot\binom{6}{2}}{\binom{11}{4} - \binom{10}{4}} = \frac{1}{2}.$$

(v) There are $\binom{5}{1}\cdot\binom{4}{1}\cdot\binom{6}{2}$ ways to choose two numbered discs given a numbered disc is chosen first. There are $\binom{5}{1}\cdot\binom{10}{3}$ ways to choose a numbered disc first. So,

$$\mathrm{P}(\text{exactly two numbered}|\text{ numbered chosen 1st}) = \frac{\binom{5}{1}\cdot\binom{4}{1}\cdot\binom{6}{2}}{\binom{5}{1}\cdot\binom{10}{3}} = \frac{1}{2}.$$

(vi) There are $\binom{5}{2}\cdot\binom{6}{2}$ ways to take exactly two numbered discs. There are $\binom{11}{4}-\binom{6}{4}$ ways to take a numbered disc. So,

$$\mathrm{P}(\text{exactly two numbered}|\text{ numbered chosen}) = \frac{\binom{5}{2}\cdot\binom{6}{2}}{\binom{11}{4}-\binom{6}{4}} = \frac{10}{21}.$$

14. **(i)** If $y = (x+1)\mathrm{e}^{-x}$, then

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \mathrm{e}^{-x} - (x+1)\mathrm{e}^{-x} = -x\mathrm{e}^{-x}.$$

Also,

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = -\mathrm{e}^{-x} + x\mathrm{e}^{-x} = (x-1)\mathrm{e}^{-x}.$$

So, when $\frac{\mathrm{d}y}{\mathrm{d}x} = 0$, $x = 0$ and this is the only turning point. At $x = 0$, $\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = -1 < 0$ and $y = 1$ and so this turning point $(0, 1)$ is a local maximum.
At $y = 0$, $x = -1$. As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to -\infty$.
So, we can sketch the graph as in fig. 13:

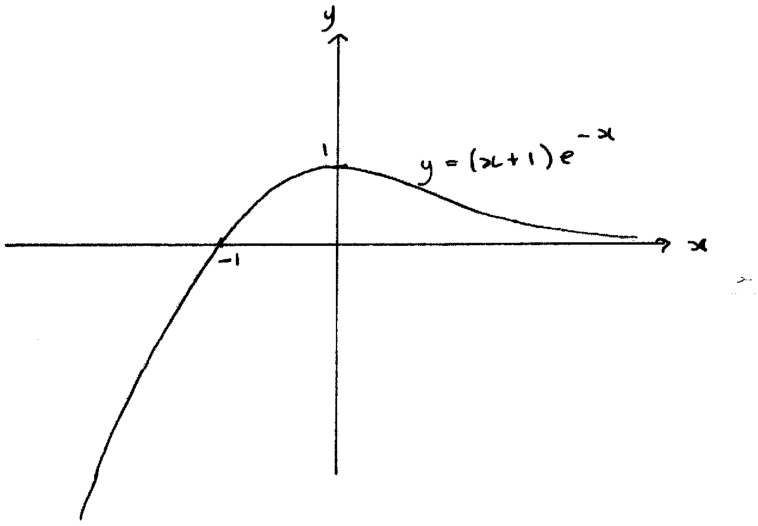


Figure 13

If $\mathrm{P}(X \geqslant 2) = 1 - p$ then

$$\begin{aligned} 1 - \mathrm{P}(X \leqslant 1) &= 1 - p \\ \implies \mathrm{P}(X \leqslant 1) &= p \\ \implies \mathrm{P}(X = 0) + \mathrm{P}(X = 1) &= p \\ \implies \frac{\mathrm{e}^{-\lambda}\lambda^0}{0!} + \frac{\mathrm{e}^{-\lambda}\lambda^1}{1!} &= p \\ \implies p &= (\lambda+1)\mathrm{e}^{-\lambda}. \end{aligned}$$

Now letting $x = \lambda$, we can see from the graph we just sketched that for any p in the range $0 < p < 1$ there is a unique solution for λ (as λ must be positive).

(ii) Now, if $\mathrm{P}(X = 1) = q$, then

$$\begin{aligned} &\frac{\mathrm{e}^{-\lambda}\lambda^1}{1!} = q \\ \implies\ &q = \lambda \mathrm{e}^{-\lambda}. \end{aligned}$$

To find out more about this, we will sketch the graph of $y = x\mathrm{e}^{-x}$. We have

$$\frac{\mathrm{d}y}{\mathrm{d}x} = (1 - x)\mathrm{e}^{-x}$$

(similarly to before before) and

$$\begin{aligned} \frac{\mathrm{d}^2y}{\mathrm{d}x^2} &= -\mathrm{e}^{-x} - (1 - x)\mathrm{e}^{-x} \\ &= (x - 2)\mathrm{e}^{-x}. \end{aligned}$$

So, when $\frac{\mathrm{d}y}{\mathrm{d}x} = 0$, $x = 1$ and at this point, $\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = -\frac{1}{e} < 0$ and $y = \frac{1}{e}$ and so $(1, \frac{1}{e})$ is a local maximum. At $x = 0$, $y = 0$. As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to -\infty$. Hence we can sketch the function as in fig. 14:

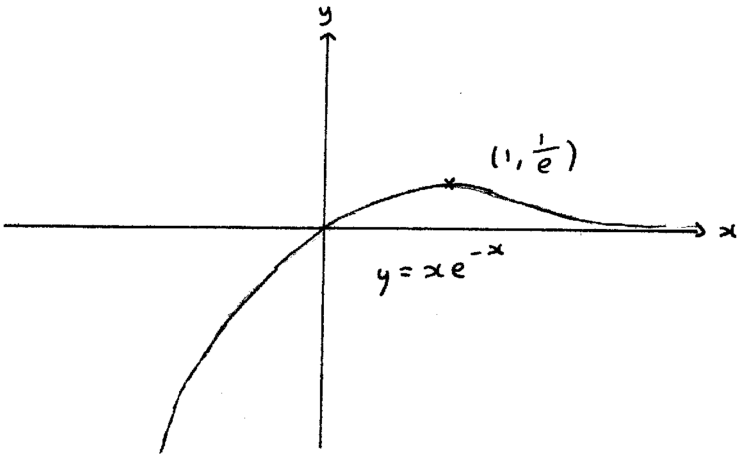


Figure 14

Clearly, the only value of y for which the function is bijective for $x > 0$ is at the turning point $(1, \frac{1}{e})$. So, our uniquely determined values must be $\lambda = 1$ and $q = \frac{1}{e}$.

(iii) We will follow a similar method. If $\mathrm{P}(X = 1 | X \leqslant 2) = r$, then by Bayes' theorem,

$$\begin{aligned} &\frac{\mathrm{P}(X = 1 \cap X \leqslant 2)}{\mathrm{P}(X \leqslant 2)} = r \\ \implies\ &\frac{\mathrm{P}(X = 1)}{\mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2)} = r. \end{aligned}$$

So,

$$\begin{aligned} r &= \frac{\frac{\mathrm{e}^{-\lambda}\lambda^1}{1!}}{\frac{\mathrm{e}^{-\lambda}\lambda^0}{0!} + \frac{\mathrm{e}^{-\lambda}\lambda^1}{1!} + \frac{\mathrm{e}^{-\lambda}\lambda^2}{2!}} \\ &= \frac{\lambda \mathrm{e}^{-\lambda}}{\mathrm{e}^{-\lambda} + \lambda \mathrm{e}^{-\lambda} + \frac{1}{2}\lambda^2 \mathrm{e}^{-\lambda}} \\ &= \frac{2\lambda}{\lambda^2 + 2\lambda + 2}. \end{aligned}$$

As before, we will consider the graph of $y = \frac{2x}{x^2+2x+2}$. At $x = 0$, $y = 0$; as $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^-$. Differentiating,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{2(x^2 + 2x + 2) - 2x(2x + 2)}{(x^2 + 2x + 2)^2}$$

and so where the derivative is zero,

$$\begin{aligned} &\frac{2(x^2 + 2x + 2) - 2x(2x + 2)}{(x^2 + 2x + 2)^2} = 0 \\ \implies &x^2 + 2x + 2 - x(2x + 2) = 0 \\ \implies &-x^2 + 2 = 0 \\ \implies &x = \pm\sqrt{2}. \end{aligned}$$

At $x = \sqrt{2}$, $y = \frac{\sqrt{2}}{\sqrt{2}+2} = \sqrt{2} - 1$ and at $x = -\sqrt{2}$, $y = \frac{\sqrt{2}}{\sqrt{2}-2} = -\sqrt{2} - 1$. Because of the direction from which the curve tends to zero as x tends to $\pm\infty$, the turning point at $(\sqrt{2}, \sqrt{2} - 1)$ must be a local maximum and the turning point at $(-\sqrt{2}, -\sqrt{2} - 1)$ must be a local minimum. Figure 15 shows a sketch of the curve based on this information:

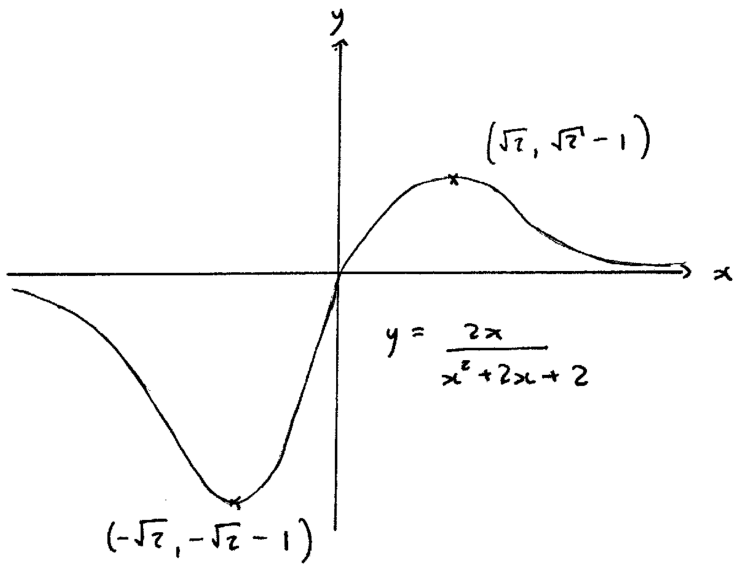


Figure 15

Therefore, the only point for $x > 0$ for which the curve is bijective is the turning point $(\sqrt{2}, \sqrt{2} - 1)$. It follows that our uniquely determined values must be $\lambda = \sqrt{2}$, $r = \sqrt{2} - 1$.

Chapter 26

Project Euler problem solutions

Over the summer of 2017 I did quite a few of the Project Euler problems since they seemed to be great preparation for Maths and Computer Science. Later in the summer I explained my solutions to several of them in writing, and I've included those explanations here.

Project Euler problem 1

August 24, 2017 D Falck Project Euler

Source code available on GitHub.

This first problem gets us off to an easy start. It asks:

> If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23.
>
> Find the sum of all the multiples of 3 or 5 below 1000.

We want to iterate through all the numbers below 1000, so we'll use a `for` loop:

```
1. for i in range(1,1000):
```

This tells Python to start with `i=1` and keep incrementing `i` every loop up until `i=999`. The easiest way to check for divisibility is with the modulus operator, `%`. Writing `x % y` returns the value of $x \mod y$ – that is, the remainder when dividing x by y.

So, using an `if` statement, we write

```
1. if i % 3 == 0 or i % 5 == 0:
```

which will return true when the number is a multiple of (had zero remainder when divided by) either 3 or 5. If this happens, we want to add `i` to some sum so that when the `for` loop terminates, that value is our final value. We'll call it `running_sum`.

Lastly, we want to print the result. Here's the final code:

```
1. running_sum = 0
2. for i in range(1,1000):
3.     if i % 3 == 0 or i % 5 == 0:
4.         running_sum += i
5. print(running_sum)
```

Project Euler problem 2

August 24, 2017 D Falck Project Euler

Source code available on GitHub.

The second problem is not much harder:

> Each new term in the Fibonacci sequence is generated by adding the previous two terms. By starting with 1 and 2, the first 10 terms will be:
>
> $$1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$$
>
> By considering the terms in the Fibonacci sequence whose values do not exceed four million, find the sum of the even-valued terms.

We basically want to run through the Fibonacci sequence until we hit 4 million, all the while adding every even-valued term to a running sum.

So, we start by defining the first two terms of the sequence as 1 and 2 and then making a rule to generate the rest. Making an empty list `fibonacci = [0,0]`, we set `fibonacci[0] = 1` and `fibonacci[1] = 2`. Now, we want to implement a loop that will keep going until we tell it to stop. For this we can use `while True:`, as since `True` is always true, the loop will always keep running.

That means we'll have to do our incrementing manually, unlike a `for` loop. Setting our index as `i = 2` (the *third* term in the sequence, as we've defined the first two), we'll do stuff in the loop and then increment it by one using `i += 1` at the end:

```
1.  fibonacci = [0,0]
2.  fibonacci[0] = 1
3.  fibonacci[1] = 2
4.
5.  i = 2
6.  while True:
7.      stuff
8.      i += 1
```

Next, we want to define the rule to keep the sequence going. Since each term is the sum of the previous two, we know that `fibonacci[i] == fibonacci[i-1] + fibonacci[i-2]` for every `i`. The catch is, the list `fibonacci` doesn't actually *have* an index `i` yet, so we need to append this value to the end of the list instead:

```
1.  fibonacci = [0,0]
2.  fibonacci[0] = 1
3.  fibonacci[1] = 2
4.
```

```
i = 2
while True:
    fibonacci.append(fibonacci[i-1] + fibonacci[i-2])
    i += 1
```

Nice. We know we only want to go up to 4 million, so if we reach or exceed that number we use `break` to terminate the `while` loop:

```
if fibonacci[i] >= 4000000:
        break
```

All that's left is our actual task: sum all the even terms. Taking a similar approach to problem 1, we initialise a variable called `summation` and every loop, if the current term is even, we add it in.

Add `print(summation)` at the end and we're done!

```
fibonacci = [0,0]
fibonacci[0] = 1
fibonacci[1] = 2
summation = 0
i = 2
while True:
    fibonacci.append(fibonacci[i-1] + fibonacci[i-2])
    if fibonacci[i] >= 4000000:
        break
    if fibonacci[i] % 2 == 0:
        summation += fibonacci[i]
    i += 1
print(summation)
```

Project Euler problem 3

August 24, 2017 D Falck Project Euler

Source code available on GitHub.

Here we go:

> The prime factors of 13195 are 5, 7, 13 and 29.
>
> What is the largest prime factor of the number 600851475143 ?

This is the first problem we're going to need to split up into functions. We need a function to tell us the prime factors of any number, and to do that we need a function to tell us whether a particular number is prime or not.

Let's start with the latter. While there are many fancy prime checks out there, we're going to use the very simplest, most obvious one: just try dividing the number by every integer below its square root: if none of them go into it, then the number must be prime. (We only need to check up to the square root because *factors come in pairs*: if the number has a factor larger than its square root, there will be a corresponding one below it.)

So, if `number` is the number we want to check, we want a `for` loop to do what we just described:

```
for i in range(1,math.ceil(math.sqrt(number))+1):
```

In case the square root isn't an integer, we round it up with `math.ceil()` and in case the square root *is* an integer, we add 1 so that the for loop will actually reach it.

Now, within this loop, we want to check whether `i` divides `number`, because if it ever does then `number` cannot be prime. Now is a good time to put it all into a function which will return `True` if the number is prime and `False` if it isn't:

```
def isPrime(number):
    for i in range(1,math.ceil(math.sqrt(number))+1):
        if number % i == 0:
            return False
    return True
```

After testing this I realised that the numbers 1 and 2 confuse things, because 1 is its own square root and 2 is its own square root rounded up. So, we just add an `if`-`else` statement to protect against all that, and we're done:

```
def isPrime(number):
    if number == 1:
        return False
    elif number == 2:
        return True
    else:
```

```
        for i in range(2,math.ceil(math.sqrt(number))+1):
            if number % i == 0:
                return False
    return True
```

Next, we want to find the prime factors of a number `number`. We can do this iteratively: as soon as we find one prime factor, we divide it out of `number` and do the whole process again with this new value. Eventually we'll get to the last prime factor and `number` itself will be prime; this signals we've got them all.

We'll follow a very similar approach to the function `isPrime(number)`. Starting with an empty list of prime factors, `primefactors = []`, we want to put everything inside a `while` loop so that when we change `number` we can do the whole thing again. Then, we want to check each possible factor of `number` just like we did above.

```
def primeFactors(number):
    primeFactors = []
    while True:
        for i in range(1,math.ceil(math.sqrt(number))+1):
```

This value `i` is a prime factor of `number` if and only if it divides evenly into `number` *and* it is prime. If both these conditions are met, we want to append it to our list of prime factors and then divide it out of `number` as mentioned above. Finally, we `break` out of the for loop and the enclosing `while` loop will make the whole process repeat with the new, smaller value of `number`:

```
def primeFactors(number):
    primefactors = []
    while True:
        for i in range(1,math.ceil(math.sqrt(number))+1): # Only check up to the square root: any other factors can be found from the existing ones
            if number % i == 0 and isPrime(i):
                primefactors.append(i)
                number = int(number/i)
                break
```

Note that we use `int(number/i)` instead of just `number/i` to get rid of any floating points that might accidentally be introduced in the division.

At the moment this will carry on forever: we need a termination condition. Once `number` is itself prime, it must be the last prime factor, so we can append it to the list and `break` out of the whole `while` loop. Then, we want our function to return our final list `primefactors`:

```
def primeFactors(number):
    primefactors = []
    while True:
        if isPrime(number):
            primefactors.append(number)
            break
        for i in range(1,math.ceil(math.sqrt(number))+1): # Only check up to the square root: any other factors can be found from the existing ones
            if number % i == 0 and isPrime(i):
                primefactors.append(i)
                number = int(number/i)
                break
    return primefactors
```

We're done! All that's left is to print off the maximum of this list for the specific value given in the problem, which we can do by taking `max(primeFactors(value))`. We also have to do `import math` at the beginning to allow use of the square root and ceiling functions.

Here's the final code:

```
import math

def isPrime(number):
    if number == 1:
        return False
    elif number == 2:
        return True
    else:
        for i in range(2,math.ceil(math.sqrt(number))+1):
            if number % i == 0:
                return False
    return True

def primeFactors(number):
    primefactors = []
    while True:
        if isPrime(number):
            primefactors.append(number)
            break
        for i in range(1,math.ceil(math.sqrt(number))+1): # Only check up to the square root: any other factors can be found from the existing ones
            if number % i == 0 and isPrime(i):
                primefactors.append(i)
                number = int(number/i)
                break
    return primefactors

print(max(primeFactors(600851475143)))
```

Project Euler problem 4

August 24, 2017 D Falck Project Euler

Source code available on GitHub.

This problem reads as follows:

> A palindromic number reads the same both ways. The largest palindrome made from the product of two 2-digit numbers is $9009 = 91 \times 99$.
>
> Find the largest palindrome made from the product of two 3-digit numbers.

The last bit says 'made from the product of two 3-digit numbers'. There aren't that many 3-digit numbers around, so it's perfectly fine for us to just have two `for` loops inside each other:

```
for i in range(100,1000):
    for j in range(i,1000):
        prod = i*j
```

Now in order to determine whether `prod` is a palindrome or not, we're going to have to separate it out into its digits. There is a very simple way of doing this: convert `prod` to a string, and convert that string to a list (or a tuple, as it's not going to change). You just do `tuple(str(prod))` and you're done, you have a tuple containing all of its digits. If you want to convert each of those digits back from a string into an integer, you instead do `tuple(int(i) for i in str(prod))`. You can see how that gets the job done.

Anyway, I didn't think of that when I did this problem, so I'll give my long-winded mathematical way here that I actually used.

The product of two 3-digit numbers is going to be either 6 digits long or 5 digits long. So, we initialise a list as `digits = [0,0,0,0,0,0]`. Now, the last digit of some integer n with digits $a_0, a_1, a_2, a_3, a_4, a_5$ is going to be $a_5 \equiv n \mod 10$. If you don't know modular arithmetic, that just meant it's the remainder when we divide n by 10. If our number is 54, for example, then dividing by 10 gives you a remainder of 4, which is indeed the last digit.

It's not hard to see that the last *two* digits together are $n \mod 100$ in a similar way – for instance, divide 154 by 100 and you get a remainder of 54. To get rid of the 4 at the end, we subtract the last digit (which we already know) and divide by 10: $(54 - 4)/10 = 5$.

The pattern continues all the way down. Our code implementation of this is something like follows:

```
prod = i*j
digits = [0,0,0,0,0,0] # Separating the product out into its digits
digits[5] = int(prod % 10)
digits[4] = int((prod % 100 - digits[5])/10)
digits[3] = int((prod % 1000 - 10*digits[4])/100)
digits[2] = int((prod % 10000 - 100*digits[3])/1000)
digits[1] = int((prod % 100000 - 1000*digits[2])/10000)
digits[0] = int((prod % 1000000 - 10000*digits[1])/100000)
```

I hope it makes sense how that works! Once you see how the first two happen, you can just keep adding zeroes every line.

Anyway, now we know the digits. There are still two cases: if the first of those 6 digits is a zero (so `digits[0] == 0`) then `prod` is really a five-digit number; otherwise it's clearly six digits long.

Fine. Given either of these cases, it's not hard to check whether the number is a palindrome:

```
if digits[0] == 0:
            if digits[1] == digits[5] and digits[2] == digits[4]:
                # PALINDROME TRUE
        else:
            if digits[0] == digits[5] and digits[1] == digits[4] and digits[2] == digits[3]:
                # PALINDROME TRUE
```

Then, using a new variable – let's call it `largest_pallen` – if the new palendrome is the largest yet, we need to do `largest_pallen = prod`. At the end of the whole thing we'll be left with `largest_pallen` as our answer!

Full code below:

```
largest_pallen = 0
for i in range(100,1000):
    for j in range(i,1000):
        prod = i*j
        digits = [0,0,0,0,0,0] # Separating the product out into its digits
        digits[5] = int(prod % 10)
        digits[4] = int((prod % 100 - digits[5])/10)
        digits[3] = int((prod % 1000 - 10*digits[4])/100)
        digits[2] = int((prod % 10000 - 100*digits[3])/1000)
        digits[1] = int((prod % 100000 - 1000*digits[2])/10000)
        digits[0] = int((prod % 1000000 - 10000*digits[1])/100000)
        if digits[0] == 0:
            if digits[1] == digits[5] and digits[2] == digits[4]:
                if prod > largest_pallen:
                    largest_pallen = prod
        else:
            if digits[0] == digits[5] and digits[1] == digits[4] and digits[2] == digits[3]:
                if prod > largest_pallen:
                    largest_pallen = prod
print(largest_pallen)
```

Project Euler problem 5

August 24, 2017 D Falck Project Euler

Source code available on GitHub.

This problem seems simpler than the last few to me.

> 2520 is the smallest number that can be divided by each of the numbers from 1 to 10 without any remainder.
>
> What is the smallest positive number that is evenly divisible by all of the numbers from 1 to 20?

It's fairly clear how to proceed. We can just check every multiple of 20, starting at 20 and adding 20 every time. If the multiple is divisible by all of the numbers 1 through 19, we stop and print the output.

To check every multiple of 20 indefinitely, we use a `while` loop:

```
i = 20
while True:
    # Check stuff here
    i += 20
```

To check whether `i` is divisible by all the numbers 19, 18, 17 and so on, we just do a `for` loop:

```
for j in range(19,0,-1):
    if i % j != 0:
        # Failed test
# Passed test
```

Note that in the `range` function we're starting at 20 and going down to 1 rather than the other way round (the `-1` tells `range()` to add `-1` each time rather than `1`): this is more efficient as it's much less likely for a number to be divisible by, say, 19 than it is for it to be divisible by, say, 3.

The number `i` has only passed the test if we get through the whole for loop without the `if` statement ever being triggered. We only want to do the increment and try again if `i` fails the check, so we put `i += 20` *inside* the `if` statement:

```
i = 20
while True:
    for j in range(19,0,-1):
        if i % j != 0:
            i += 20
            break
```

The `break` statement forces us out of the `for` loop and consequently restarts the `while` loop with the new value of `i`.

Now, if we do indeed get to the end of the `for` loop with every `j` being a factor of `i`, the current value of `j` will be `1` (as the `for` loop iterates downwards). Hence, if this is the case, `i` is our answer. We'll hold it in some other variable `smallest_multiple` and break out of the `for loop:`

```
i = 20
while True:
    for j in range(19,0,-1):
        if i % j != 0:
            i += 20
            break
        if j == 1: # If checked all of them and still here
            smallest_multiple = i
            break
```

The last thing to do is check whether `smallest_multiple` actually holds a value, and if so break out of the whole `while` loop and print it off. We'll set `smallest_multiple = 0` initially to help us.

We're done! Code below:

```
smallest_multiple = 0
i = 20
while True:
    for j in range(19,0,-1):
        if i % j != 0:
            i += 20
            break
        if j == 1: # If checked all of them and still here
            smallest_multiple = i
            break
    if smallest_multiple != 0:
        break
print(smallest_multiple)
```

Project Euler problem 6

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

Here's the problem:

> The sum of the squares of the first ten natural numbers is,
>
> $$1^2 + 2^2 + \cdots + 10^2 = 385$$
>
> The square of the sum of the first ten natural numbers is,
>
> $$(1 + 2 + \cdots + 10)^2 = 55^2 = 3025$$
>
> Hence the difference between the sum of the squares of the first ten natural numbers and the square of the sum is $3025 - 385 = 2640$. Find the difference between the sum of the squares of the first one hundred natural numbers and the square of the sum.

Well, the question looks long but I think this is actually the simplest problem we've had yet. The sum of all integers from 1 to n is

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}.$$

It's not hard to see why. In fact, there's a brilliant story about how the young Gauss found this trick when an annoyed teacher gave him this problem to keep him occupied, expecting it to take him a long time to manually sum all the numbers.

Let this sum be S, so that

$$S = 1 + 2 + 3 + \cdots + (n-1) + n.$$

We write the sum backwards as

$$S = n + (n-1) + (n-2) + \cdots + 2 + 1$$

and add the two equations term-by-term, getting

$$\begin{aligned} 2S &= (n+1) + (2+n-1) + (3+n-2) + \cdots + (n-1+2) + (n+1) \\ &= (n+1) + (n+1) + (n+1) + \cdots + (n+1) + (n+1). \end{aligned} \tag{1}$$

It's easy to see that this is n lots of $n+1$ and so $2S = n(n+1)$. Divide by two, and we have $S = \frac{n(n+1)}{2}$ just as we were hoping for.

A quick implementation of this as a function is:

```
1.  def sumofintegers(n):
2.      return n*(n+1)/2
```

A very similar result for the sum of squares takes a just a little more effort to show, but I won't prove it here:

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

A similar Python function would be:

```
def sumofsquares(n):
    return n*(n+1)*(2*n+1)/6
```

So, all we need to do is take the difference of the sum of squares and the *square* of the regular sum, up to `n`. We're done!

```
def sumofsquares(n):
    return n*(n+1)*(2*n+1)/6

def sumofintegers(n):
    return n*(n+1)/2

def specialdifference(n): # Finds the difference between the sum of squares and the square sum up to n
    return abs(sumofsquares(n) - (sumofintegers(n))**2)

print(specialdifference(100))
```

Project Euler problem 7

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

A very succinct question:

> By listing the first six prime numbers: 2, 3, 5, 7, 11, and 13, we can see that the 6th prime is 13.
>
> What is the 10 001st prime number?

Well, in problem 3 we already made a function `isPrime(number)` to check if a number is prime, so let's port that right in:

```
import math

def isPrime(number):
    if number == 1:
        return False
    else:
        if number == 2:
            return True
        else:
            for i in range(2,math.ceil(math.sqrt(number))+1):
                if number % i == 0:
                    return False
            return True
```

Now we just want to keep going through all the positive integers, generating primes until we have 10,001 of them. This `while` loop will make an infinite list of primes if you leave it to go on for long enough:

```
primes = []

i = 1
while True:
    if isPrime(i):
        primes.append(i)
    i += 1
```

We just add a termination if `len(primes)` is above 10,001 and print off our 10,001st prime! This is the whole program

```
import math

def isPrime(number):
    if number == 1:
        return False
    else:
        if number == 2:
```

```
            return True
        else:
            for i in range(2,math.ceil(math.sqrt(number))+1):
                if number % i == 0:
                    return False
            return True

primes = []

i = 1
while True:
    if isPrime(i):
        primes.append(i)
    if len(primes) > 10001:
        break
    i += 1

print(primes[10000])
```

Project Euler problem 8

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

Ah, an interesting-looking one!

> The four adjacent digits in the 1000-digit number that have the greatest product are $9 \times 9 \times 8 \times 9 = 5832$.
>
> ```
> 7316717653133062491922511967442657474235534919493496983520312774506326239578318016984801869478851843
> 8586156078911294949545950173795833195285320880551125406987471585238630507156932909632952274430435576689
> 6489504452445232161731856403098711121722383113622298934233803081353362766142828064444866452387493035890
> 7296290491560440772390713810515859307960866701724271218839987979087922749219016997208880937765726733300
> 1053367881220235421809751254540594752243525849077116705560136048395864467063244157221553975369781797784
> 6174064955149290862569321978468622482839722413756570560574902614079729686524145351004748216637048440319
> 9890008895243450658541227588666881164271714799244429282308634656748139191231628245861786458359124566529
> 4765456828489128831426076900422421902267105562632111110937054421750694165896040807198403850962455444362
> 9812309878799272442849091887858013660979191338754992005240636899125607176060588611646710940507754100225
> 6983155200055935729725716362695618826704282524836008232575304207529634450
> ```
>
> Find the thirteen adjacent digits in the 1000-digit number that have the greatest product. What is the value of this product?

So, immediately we want to get this big number and open it in a text file. In Python, while you could just open the file, use it and then close it again, there's a handy construct called `with – as` which is perfect for dealing with text files. We copy and paste that big number into a file called `input.txt` and then do:

```
1. with open('input.txt') as f:
```

We do what we want to the text `f` and as soon as we leave the `with` statement, Python closes the file down and cleans up after us.

We want to read the file with `f.read()` and get rid of all the new line characters (`\n`). This does it:

```
1. with open('input.txt') as f:
2.     data = f.read().replace('\n','')
```

Now `data` contains a big string of our number and nothing else. We happen to know the number is 1000 digits long and we want 13 adjacent digits at a time, but these numbers could be anything so we'll be general. The following `for` loop will iterate through all the possible 13 adjacent digits, find their product, and store it if it's the biggest yet.

```
for i in range(len(data)-digitsLength):
    digits = [int(j) for j in data[i:i+digitsLength]]
    product = prod(digits)
    if product > maxProduct:
        maxProduct = product
```

You may notice that `prod()` isn't a built-in function: we'll define that. It just finds the products of all the elements in an iterable variable like a list, tuple or set. It's a simple definition:

```
def prod(iterable):
    product = 1
    for i in iterable:
        product *= i
    return product
```

Fine, that's all done. Our program is as follows:

```
def prod(iterable):
    product = 1
    for i in iterable:
        product *= i
    return product

with open('input.txt') as f:
    data = f.read().replace('\n','')

digitsLength = 13 # The number of adjacent digits we want to find the product of
maxProduct = 0

for i in range(len(data)-digitsLength):
    digits = [int(j) for j in data[i:i+digitsLength]]
    product = prod(digits)
    if product > maxProduct:
        maxProduct = product

print(maxProduct)
```

Project Euler problem 9

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

This problem is as follows:

> A Pythagorean triplet is a set of three natural numbers, $a < b < c$, for which,
>
> $$a^2 + b^2 = c^2$$
>
> For example, $32 + 42 = 9 + 16 = 25 = 52$.
>
> There exists exactly one Pythagorean triplet for which $a + b + c = 1000$.
>
> Find the product abc.

I initially looked at this problem and tried to overcomplicate it quite a lot. I know of ways of generating Pythagorean triplets and tried to implement something much more fancy than necessary. I kept not getting the answer, possibly because in its base form the Euclid method does not generate *all* triplets.

Anyway, I got rid of that solution and went back to the start; a really quite simple program is more than sufficient for this program.

I like breaking things up into functions, partly because it looks nice but also because it means I can much more easily use bits from my previous programs in later ones. So, I started by making a function called `PythagoreanTriple(N)` which will generate a Pythagorean triplet that sums to `N`. So, we want three numbers, let's call them `x`, `y` and `z`, such that `x + y + z == N` and `z*z == x*x + y*y`. Let's assume `x < y` (it could be the other way round but then we'd just switch `x` and `y`). Now, it's clear that `x < y < N` and so we can get started with a simple `for` loop:

```
1.  def PythagoreanTriple(N): # Generates a pythagorean triple that sums to N
2.      for x in range(1,N+1):
3.          for y in range(x+1,N+1):
```

Now by the time we're inside this second for loop, we know the current values of `x` and `y` so it seems a bit pointless to start again generating *all* possible values of `z` and checking whether `z*z == x*x + y*y`. We can do better: obviously while the square of `z` is less than `x*x + y*y` (which we already know), we can just keep increasing `z`. Only once we know `z*z >= x*x + y*y` do we actually need to check the equality:

```
1.  def PythagoreanTriple(N): # Generates a pythagorean triple that sums to N
2.      for x in range(1,N+1):
3.          for y in range(x+1,N+1):
4.              z = y + 1
5.              while z*z < x*x + y*y:
6.                  z += 1
```

```
            if z*z == x*x + y*y and x+y+z == N:
                return [x,y,z]
    return 'Failed'
```

Of course once we get out our `special` triplet we need the product `x*y*z` (or abc in the question). We could easily do it by hand but let's port in our `prod()` function from the last question and use that:

```
def prod(iterable):
    product = 1
    for i in iterable:
        product *= i
    return product

def PythagoreanTriple(N): # Generates a pythagorean triple that sums to N
    for x in range(1,N+1):
        for y in range(x+1,N+1):
            z = y + 1
            while z*z < x*x + y*y:
                z += 1
            if z*z == x*x + y*y and x+y+z == N:
                return [x,y,z]
    return 'Failed'

print(prod(PythagoreanTriple(1000)))
```

Project Euler problem 10

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

Another prime-related problem:

> The sum of the primes below 10 is $2 + 3 + 5 + 7 = 17$.
>
> Find the sum of all the primes below two million.

From problem 7 we already have a nice function to check if a number is prime or not:

```
import math

def isPrime(number):
    if number == 1:
        return False
    else:
        if number == 2:
            return True
        else:
            for i in range(2,math.ceil(math.sqrt(number))+1):
                if number % i == 0:
                    return False
            return True
```

That means all we need to do is keep checking every number, add any primes to a big list, and stop before we hit 2 million.

```
def sumOfPrimes(N): # Sum of all primes below N
    sum = 0
    for i in range(1,N+1,2):
        if isPrime(i):
            sum += i
    return sum
```

This code does just that. I've told the `range()` function to only return odd numbers because, well – no primes are even except for 2. For that reason I've started with my sum as 2 and initiated the `for` loop starting with `i = 3`.

That's our solution done!

```
import math

def isPrime(number):
    prime = True #Assume prime
    if number == 1:
        return False
    if number == 2:
```

```
            return True
    else:
        for i in range(2,math.ceil(math.sqrt(number))+1):
            if number % i == 0:
                return False
        return True

def sumOfPrimes(N): # Sum of all primes below N
    sum = 2
    for i in range(3,N+1,2):
        if isPrime(i):
            sum += i
    return sum

print(sumOfPrimes(2000000))
```

Project Euler problem 11

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

This next problem reminds us very much of problem 8:

> In the 20×20 grid below, four numbers along a diagonal line have been marked in red.
>
> ```
> 08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 08
> 49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
> 81 49 31 73 55 79 14 29 93 71 40 67 53 88 30 03 49 13 36 65
> 52 70 95 23 04 60 11 42 69 24 68 56 01 32 56 71 37 02 36 91
> 22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
> 24 47 32 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
> 32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
> 67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
> 24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
> 21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
> 78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
> 16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
> 86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
> 19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
> 04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
> 88 36 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
> 04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
> 20 69 36 41 72 30 23 88 34 62 99 69 82 67 59 85 74 04 36 16
> 20 73 35 29 78 31 90 01 74 31 49 71 48 86 81 16 23 57 05 54
> 01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 19 67 48
> ```
>
> The product of these numbers is $26 \times 63 \times 78 \times 14 = 1788696$.
>
> What is the greatest product of four adjacent numbers in the same direction (up, down, left, right, or diagonally) in the 20×20 grid?

Well, this is the same problem as problem 8 but instead of just a long number we're actually dealing with a grid now, so we're going to have to change our data construct to reflect that.

We copy and paste the grid into a text file and open it with Python as before, but instead of putting it all into a string, we want a list where each entry is itself a list of numbers in that line. First we want to have each entry be a different line:

```
1. with open('input.txt') as f:
2.     grid = [line.replace('\n','') for line in f]
```

If you iterate through a text file in Python it automatically returns it one line at a time, which is why we ca just say `for line in f`. Doing `line.replace('\n','')` instead of just `line` gets rid of the new line characters that end up in the string.

Now for *each* entry in `grid` we want to split it up into a list of all the numbers – splitting the string every time there's a space character. This is very simple with the `split()` function:

```
with open('input.txt') as f:
    grid = [line.replace('\n','').split(' ') for line in f]
```

Our last problem is that each entry in each line of `grid` is a short string, not a number – that is, each entry looks something like `'54'` or `'63'` rather than `54` or `63`. So, `for` each entry in each the line we want to convert the entry to an integer using `int()`:

```
with open('input.txt') as f:
    grid = [[int(i) for i in line.replace('\n','').split(' ')] for line in f]
```

Done! Now we're told that the grid is 20 by 20, but just in case we want to use a different grid later on we'll work out the width and height with:

```
width = len(grid[0])
height = len(grid)
```

Remember that `grid` is the list of lines, and `grid[i]` is the list of numbers on the `i` th line.

Okay, so we're told to find the maximum product of four adjacent numbers (let's do `adjacents = 4` for generality) in any direction, which gives us four possible directions to check: up and down, left and right, diagonally top-left to bottom-right and diagonally top-right to bottom-left.

Let's check all groups of vertically adjacent digits first – that is, the direction 'up and down'. Try to visualise this, for each row in the grid we'll take the first number in the row and the three numbers directly below it, then repeat for the next number along in the row, and so on. Then we just drop down a row and start again. In the end we have $20 - 3 = 17$ rows to do this to (since we're always taking 3 numbers *below* the current row).

Using the numbers 20 and 4 as in the question, an implementation of this is as below:

```
maxProduct = 0

# Vertical adjacents
for i in range(17):
    for j in range(20):
        product = prod([grid[i+k][j] for k in range(4)])
        if product > maxProduct:
            maxProduct = product
```

It always helps to imagine you're Python and follow along the first loop or two to understand what's going on. We're still using our `prod()` function that we've defined previously to take the product of all entries in a list.

Now getting rid of `20` and `4` and replacing them with `width`, `height` and `adjacents`, that code becomes:

```
width = len(grid[0])
height = len(grid)

maxProduct = 0
adjacents = 4

# Vertical adjacents
for i in range(height - adjacents + 1):
    for j in range(width):
```

```
        product = prod([grid[i+k][j] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product
```

Now that we've done it for vertically adjacent numbers, we can pretty much use the exact same structure with only small alterations for our other three directions. For horizontally adjacent numbers, the code becomes this:

```
# Horizontal adjacents
for i in range(height):
    for j in range(width - adjacents + 1):
        product = prod([grid[i][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product
```

And it's not hard to carry on this pattern to diagonally adjacent numbers:

```
# Diagonal tl-br adjacents
for i in range(height - adjacents + 1):
    for j in range(width - adjacents + 1):
        product = prod([grid[i+k][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product

# Diagonal tr-bl adjacents
for i in range(adjacents - 1,height):
    for j in range(width - adjacents + 1):
        product = prod([grid[i-k][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product
```

All the while, I was visualising the grid quite a lot while coding this, so visualisation definitely helps in understanding what's going on here.

Finally, once we've checked all four directions (and therefore all possible groups of 4 adjacent numbers) our variable `maxProduct` *must* contain the largest product of any of them. So, we just print it off!

That makes our full program look like this:

```
def prod(iterable):
    product = 1
    for i in iterable:
        product *= i
    return product

with open('input.txt') as f:
    grid = [[int(i) for i in line.replace('\n','').split(' ')] for line in f]

width = len(grid[0])
height = len(grid)

maxProduct = 0
adjacents = 4

# Vertical adjacents
for i in range(height - adjacents + 1):
    for j in range(width):
        product = prod([grid[i+k][j] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product

# Horizontal adjacents
```

```
for i in range(height):
    for j in range(width - adjacents + 1):
        product = prod([grid[i][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product

# Diagonal tl-br adjacents
for i in range(height - adjacents + 1):
    for j in range(width - adjacents + 1):
        product = prod([grid[i+k][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product

# Diagonal tr-bl adjacents
for i in range(adjacents - 1,height):
    for j in range(width - adjacents + 1):
        product = prod([grid[i-k][j+k] for k in range(adjacents)])
        if product > maxProduct:
            maxProduct = product

print(maxProduct)
```

Project Euler problem 12

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

Here goes:

> The sequence of triangle numbers is generated by adding the natural numbers. So the 7^{th} triangle number would be $1 + 2 + 3 + 4 + 5 + 6 + 7 = 28$. The first ten terms would be:
>
> $$1, 3, 6, 10, 15, 21, 28, 36, 45, 55, \ldots$$
>
> Let us list the factors of the first seven triangle numbers:
>
> $$\begin{aligned} \mathbf{3:}&\ 1, 3 \\ \mathbf{6:}&\ 1, 2, 3, 6 \\ \mathbf{10:}&\ 1, 2, 5, 10 \\ \mathbf{15:}&\ 1, 3, 5, 15 \\ \mathbf{21:}&\ 1, 3, 7, 21 \\ \mathbf{28:}&\ 1, 2, 4, 7, 14, 28 \end{aligned} \tag{1}$$
>
> We can see that 28 is the first triangle number to have over five divisors.
>
> What is the value of the first triangle number to have over five hundred divisors?

Ok, so we're going to need to be able to find all the factors of a given number. In problem 3 we wrote a function to find all the *prime* factors:

```
1.  def primeFactors(number):
2.      primefactors = []
3.      while True:
4.          if isPrime(number):
5.              primefactors.append(number)
6.              break
7.          for i in range(1,math.ceil(math.sqrt(number))+1): # Only check up to the square root:
    any other factors can be found from the existing ones
8.              if number % i == 0 and isPrime(i):
9.                  primefactors.append(i)
10.                 number = int(number/i)
11.                 break
12.     return primefactors
```

(We wrote the function `isPrime(number)` to check if `number` is prime or not.) We can do something similar but much simpler if we don't care about primality: once we find a factor `i` of `n`, we just add both `i` and `n/i` (or for best practice, `int(n/i)`) to our list of factors. Once we've got to `sqrt(n)` we know we have them all.

```
def findFactors(n):
    factors = []
    for i in range(1,math.floor(math.sqrt(n))+1):
        if n % i == 0:
            factors.append(i)
            factors.append(int(n/i))
    return factors
```

Next, we make a simple function to find the `n` th triangular number:

```
def triangularNumber(n):
    sum = 0
    for i in range(1,n+1):
        sum += i
    return sum
```

That one's pretty self-explanatory.

So, now we just have to keep finding triangular numbers, check their factors, and if there are more than 500 of them, print off that triangular number and end the program:

```
index = 0

while True:
    if len(findFactors(triangularNumber(index))) >= 500:
        print(triangularNumber(index))
        break
    index += 1
```

Done! Here's the whole thing:

```
import math

def findFactors(n):
    factors = []
    for i in range(1,math.floor(math.sqrt(n))+1):
        if n % i == 0:
            factors.append(i)
            factors.append(int(n/i))
    return factors

def triangularNumber(n):
    sum = 0
    for i in range(1,n+1):
        sum += i
    return sum

index = 0

while True:
    if len(findFactors(triangularNumber(index))) >= 500:
        print(triangularNumber(index))
        break
    index += 1
```

Project Euler problem 13

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

The problem:

> Work out the first ten digits of the sum of the following one-hundred 50-digit numbers.

```
37107287533902102798797998220837590246510135740250
46376937677490009712648124896970078050417018260538
74324986199524741059474233309513058123726617309629
91942213363574161572522430563301811072406154908250
23067588207539346171171980310421047513778063246676
89261670696623633820136378418383684178734361726757
28112879812849979408065481931592621691275889832738
44274228917432520321923589422876796487670272189318
47451445736001306439091167216856844588711603153276
70386486105843025439939619828917593665686757934951
62176457141856560629502157223196586755079324193331
64906352462741904929101432445813822663347944758178
92575867718337217661963751590579239728245598838407
58203565325359399008402633568948830189458628227828
80181199384826282014278194139940567587151170094390
35398664372827112653829987240784473053190104293586
86515506006295864861532075273371959191420517255829
71693888707715466499115593487603532921714970056938
54370070576826684624621495650076471787294438377604
53282654108756828443191190634694037855217779295145
36123272525000296071075082563815656710885258350721
45876576172410976447339110607218265236877223636045
17423706905851860660448207621209813287860733969412
81142660418086830619328460811191061556940512689692
51934325451728388641918047049293215058642563049483
62467221648435076201727918039944693004732956340691
15732444386908125794514089057706229429197107928209
55037687525678773091862540744969844508330393682126
18336384825330154686196124348767681297534375946515
80386287592878490201521685554828717201219257766954
78182833757993103614740356856449095527097864797581
16726320100436897842553539920931837441497806860984
48403098129077791799088218795327364475675590848030
87086987551392711854517078544161852424320693150332
59959406895756536782107074926966537676326235447210
69793950679652694742597709739166693763042633987085
41052684708299085211399427365734116182760315001271
65378607361501080857009149939512557028198746004375
35829035317434717326932123578154982629742552737307
94953759765105305946966067683156574377167401875275
88902802571733229619176668713819931811048770190271
25267680276078003013678680992525463401061632866526
36270218540497705585629946580636237993140746255962
```

```
37074486908231174977792365466257246923322810917141
91430288197103288597806669760892938638285025333403
34413065578016127815921815005561868836468420090470
23053081172816430487623791969842487255036638784583
11487696932154902810424020138335124462181441773470
63783299490636259666498587618221225225512486764533
67720186971698544312419572409913959008952310058822
95548255300263520781532296796249481641953868218774
76085327132285723110424803456124867697064507995236
37774242535411291684276865538926205024910326572967
23701913275725675285653248258265463092207058596522
29798860272258331913126375147341994889534765745501
18495701454879288984856827726077713721403798879715
38298203783031473527721580348144513491373226651381
34829543829199918180278916522431027392251122869539
40957953066405232632538044100059654939159879593635
29746152185502371307642255121183693803580388584903
41698116222072977186158236678424689157993532961922
62467957194401269043877107275048102390895523597457
23189706772547915061505504953922979530901129967519
86188088225875314529584099251203829009407770775672
11306739708304724483816533873502340845647058077308
82959174767140363198008187129011875491310547126581
97623331044818386269515456334926366572897563400500
42846280183517070527831839425882145521227251250327
55121603546981200581762165212827652751691296897789
32238195734329339946437501907836945765883352399886
75506164965184775180738168837861091527357929701337
62177842752192623401942399639168044983993173312731
32924185707147349566916674687634660915035914677504
99518671430235219628894890102423325116913619626622
73267460800591547471830798392868535206946944540724
76841822524674417161514036427982273348055556214818
97142617910342598647204516893989422179826088076852
87783646182799346313767754307809363333018982642090
10848802521674670883215120185883543223812876952786
71329612474782464538636993009049310363619763878039
62184073572399794223406235393808339651327408011116
66627891981488087797941876876144230030984490851411
60661826293682836764744779239180335110989069790714
85786944089552990653640447425576083659976645795096
66024396409905389607120198219976047599490197230297
64913982680032973156037120041377903785566085089252
16730939319872750275468906903707539413042652315011
94809377245048795150954100921645863754710598436791
78639167021187492431995700641917969775599028300699
15368713711936614952811305876380278410754449733078
40789923115535562561142322423255033685442488917353
44889911501440648020369068063960672322193204149535
41503128880339536053299340368006977710650566631954
81234880673210146739058568557934581403627822703280
82616570773948327592232845941706525094512325230608
22918802058777319719839450180888072429661980811197
77158542502016545090413245809786882778948721859617
72107838435069186155435662884062257473692284509516
20849603980134001723930671666823555245252804609722
53503534226472524250874054075591789781264330331690
```

Congratulations for managing to scroll past that enormous list! This is one of those problems that takes a bit of thinking but actually is very simple. If you're adding two enormous numbers (of the same number of digits) and you only want the first 10 digits of the sum, there's a very high chance that just summing the

first 11 or 12 digits of those enormous numbers will give you your answer. It all depends on how big the rest of the digits are and whether adding some of them forces changes to carry over to the left (imagine doing column addition), but we may as well try it and, if we get the wrong answer, just try adding in one more digit of each number.

(Of course, this 'trial and error' approach works in the Project Euler context, but if you're in a real-world situation where you want 100% accuracy then this isn't the best method to use.)

First, we copy and paste the numbers into a text file, open the file and make a list of its lines, as we've done a few times before:

```
with open('input.txt') as f:
    numbers = [line.strip() for line in f]
```

(The `strip()` function just gets rid of any whitespace and new line characters at the end of each line.)

We only want the first 13 (let's start with 13) digits from each line, and we want them as integers not strings, so this code becomes:

```
with open('input.txt') as f:
    numbers = [int(line.strip()[0:13]) for line in f]
```

All we have to do now is sum the entries in `numbers` using the built-in function `sum(numbers)`, and then find the first 10 digits of that sum. To do so, we convert the sum to a string, slice off the first 10 indices of the string, and convert the result back into an integer:

```
summation = sum(numbers)
first10 = int(str(summation)[0:10])
```

Print it all off, and we're done! (It turns out that this does get us the right answer, though depending on the numbers you use there's a very small chance it won't)

The (very short) whole thing is below.

```
with open('input.txt') as f:
    numbers = [int(line.strip()[0:13]) for line in f]

summation = sum(numbers)
first10 = int(str(summation)[0:10])
print(first10)
```

Project Euler problem 14

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

> The following iterative sequence is defined for the set of positive integers:
>
> $$n \to n/2 \ (n \text{ is even})$$
> $$n \to 3n + 1 \ (n \text{ is odd})$$
>
> Using the rule above and starting with 13, we generate the following sequence:
>
> $$13 \to 40 \to 20 \to 10 \to 5 \to 16 \to 8 \to 4 \to 2 \to 1$$
>
> It can be seen that this sequence (starting at 13 and finishing at 1) contains 10 terms. Although it has not been proved yet (Collatz Problem), it is thought that all starting numbers finish at 1.
>
> Which starting number, under one million, produces the longest chain?
>
> **NOTE:** Once the chain starts the terms are allowed to go above one million.

This is a really interesting problem. Before even thinking about what we're going to have to do later on, let's just make a quick function to determine the next number in a Collatz sequence:

```
1.  def nextCollatz(n):
2.      if n % 2 == 0:
3.          return int(n/2)
4.      else:
5.          return 3*n+1
```

This quite literally is nothing more than the rule given in the question. Next, we want to generate the whole sequence given a particular starting number:

```
1.  def CollatzSequence(n):
2.      sequence = [n]
3.      while True:
4.          n = nextCollatz(n)
5.          sequence.append(n)
6.          if n == 1:
7.              break
8.      return sequence
```

We're just appending the next term (using the function `nextCollatz(n)` we just defined) endlessly until we reach 1, and then we return that whole list.

Now we have to find which starting number under a million produces the longest chain. Starting at 999,999 and decreasing the starting number each time rather than the other way round, because it seems

plausible that a higher starting number will result in longer sequences on average, we keep calculating the length of each Collatz sequence and if it's longer than the longest so far, we store both the starting number and the sequence length:

```
def longestCollatz(cap): # Returns the number below 'cap' which generates the longest Collatz sequence, and its length
    longest = 0
    for i in range(cap-1,1,-1):
        length = len(CollatzSequence(i))
        if length > longest:
            longest = length
            longestAt = i
    return longestAt
```

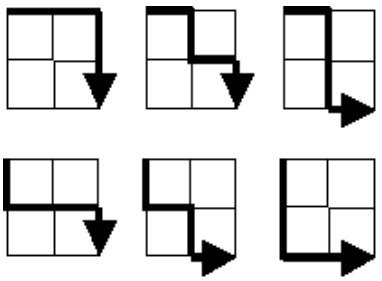
That's all there is to it: we print off the answer and we're done.

```
def nextCollatz(n):
    if n % 2 == 0:
        return int(n/2)
    else:
        return 3*n+1

def CollatzSequence(n):
    sequence = [n]
    while True:
        n = nextCollatz(n)
        sequence.append(n)
        if n == 1:
            break
    return sequence

def longestCollatz(cap): # Returns the number below 'cap' which generates the longest Collatz sequence, and its length
    longest = 0
    for i in range(cap-1,1,-1):
        length = len(CollatzSequence(i))
        if length > longest:
            longest = length
            longestAt = i
    return longestAt

print(longestCollatz(1000000))
```

Project Euler problem 15

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

> Starting in the top left corner of a 2×2 grid, and only being able to move to the right and down, there are exactly 6 routes to the bottom right corner.
>
> 
>
> How many such routes are there through a 20×20 grid?

This looks confusing at first, and it's easy to get lost combinatorially. However, this type of problem is a typical example of where we can use basic dynamic programming.

Dynamic programming is where, instead of trying to do the whole thing at once, you build up gradually piece by piece. In this case, we want to consider very carefully what it means to go from one corner to the next.

Say we label the vertical lines as $i = 0, 1, 2, \ldots, n$ and the horizontal lines as $j = 0, 1, 2, \ldots, n$ for an $n \times n$ grid. Any route from corner $(0, 0)$ to corner (n, n) can only go downwards and rightwards, and so there are only two options for which corner came before corner (i, j): either corner $(i - 1, j)$ or corner $(i, j - 1)$. Therefore, if $f(i, j)$ is the number of routes possible to arrive at corner (i, j), then

$$f(i, j) = f(i - 1, j) + f(i, j - 1). \tag{1}$$

Take a moment to see that that makes sense (it was a sudden realisation for me). Since we know that fact, we can now very easily start at the top-left corner of any grid and gradually work out the number of routes to every corner, eventually getting to the bottom-right corner which will give us our answer.

We start by setting up a matrix – or, in more programming-like-language, an array – which will eventually hold the number of paths to *every* corner on an `x` by `y` grid, but for now we'll just fill with ones. To actually do this in Python, we can just initialise a list inside a list:

```
1.  paths = [[1]*(x+1)]*(y+1)
```

This means `paths` has `y+1` entries, each of which is a list itself with `x+1` entries; if you like, an array wi `y+1` columns and `x+1` rows.

Of course, we already know that there is *one* possible path to the very first corner: that is, just start there and end there. So, we do `paths[0][0] = 1` to set this initial value.

Then, we just want to loop through the array, going through one row at a time, setting each value using the rule in equation (1). Finally, once we've got to the end, we just return the value at corner (x, y), that is, the number of paths to corner (x, y), because that's all we care about:

```
def PathsToPoint(x,y): # This is using basic dynamic programming
    paths = [[1]*(x+1)]*(y+1) # Matrix containing number of paths to each point
    for i in range(1,x+1):
        for j in range(1,y+1):
            paths[i][j] = paths[i-1][j] + paths[i][j-1]
    print(paths)
    return paths[x][y]

print(PathsToPoint(20, 20))
```

And that's the whole program! It's an elegantly simple approach.

If you start to write out the values of this array, you'll see that they match the values of Pascal's triangle, starting at the top-left corner. This makes perfect sense: you write out the values of Pascal's triangle by adding the two numbers directly above the one you're trying to find, which is exactly what this code is doing if you rotate the array clockwise by 45 degrees. In fact, this whole program can be replaced by a single combinatorial calculation. Each possible path to (x, y) contains x horizontal segments and y vertical segments: so, each path will contain $x + y$ segments overall. If you imagine starting from the beginning and repeatedly choosing whether the next segment should be horizontal or vertical, it's easy to see you have to choose a horizontal segment exactly x of these times. Therefore, the problem is reduced to 'how many different ways are there of choosing x objects out of $x + y$?', a problem with a well-established solution:

$$C(x + y, x) = \frac{(x + y)!}{x![(x + y) - x]!} = \frac{(x + y)!}{x!y!}$$

which, if $x = 20$ and $y = 20$ as in the question, reduces to

$$\frac{(20 + 20)!}{20! \cdot 20!} = \frac{40!}{(20!)^2} = 137846528820$$

which is exactly the same answer that our program returns.

Project Euler problem 16

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

> $2^{15} = 32768$ and the sum of its digits is $3 + 2 + 7 + 6 + 8 = 26$.
>
> What is the sum of the digits of the number 2^{1000}?

Easy. Define a quick digits function to return a tuple of the digits of `N`:

```
1.  def digits(N):
2.      return tuple(int(i) for i in str(N))
```

Now just sum the digits of 2^{1000}. Done.

```
1.  def digits(N):
2.      return tuple(int(i) for i in str(N))
3.
4.  print(sum(digits(2**1000)))
```

Project Euler problem 17

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

> If the numbers 1 to 5 are written out in words: one, two, three, four, five, then there are $3 + 3 + 5 + 4 + 4 = 19$ letters used in total.
>
> If all the numbers from 1 to 1000 (one thousand) inclusive were written out in words, how many letters would be used?
>
> **NOTE:** Do not count spaces or hyphens. For example, 342 (three hundred and forty-two) contains 23 letters and 115 (one hundred and fifteen) contains 20 letters. The use of "and" when writing out numbers is in compliance with British usage.

This problem got me thinking about turning numerals into words, and I actually spent quite a while solving this problem much more fully than I had to. I'm not going to explain my full solution: that's for a different post.

Suffice to say, I wrote a script that will print, in British English short-scale words, any integer from 0 to 10^{64} (10^{63} is one *vigintillion,* after which Wikipedia runs out of names). It took quite a lot of effort and made me realise just how subtle our numbering rules actually are. Here's the code:

```
1.  # D Falck. Run this file, the command line interface will ask you for a number to print
    (below 10^64 please)
2.
3.  import math
4.
5.  def digitsOf(N): # Separates N into a list of its digits
6.      N = abs(N) # Make positive
7.      length = math.floor(math.log10(N)) + 1 # How many digits
8.      separated = []
9.      separated.append(int(N % 10)) # We build the list in reverse order, starting with the
    units digit
10.     for i in range(1,length): # Every digit can be found from N modulo something and the
    previous digit
11.         new = int((N % 10**(i+1) - separated[i-1])/(10**i))
12.         separated.append(new)
13.     separated.reverse() # Finally, put the list in the right order
14.     return separated
15.
16. def inWords(N): # Returns the English words for N
17.     digitNames = ['zero','one','two','three','four','five','six','seven','eight','nine']
18.     teenNames =
    ['ten','eleven','twelve','thirteen','fourteen','fifteen','sixteen','seventeen','eighteen','ni
    neteen']
19.     tensNames = ['twenty','thirty','forty','fifty','sixty','seventy','eighty','ninety']
20.     bigNames =
    ['thousand','million','billion','trillion','quadrillion','quintillion','sextillion','septill
    on','octillion','nonillion','decillion','undecillion','duodecillion','tredecillion','quattuor
```

```
decillion','quindecillion','sexdecillion','septendecillion','octodecillion','novemdecillion',
'vigintillion']

    def hundreds(digits): # Deals with triplets of digits
        def units(digit): # Deals with single digits
            return digitNames[digit]

        def tens(digits): # Deals with pairs of digits
            if digits[0] == 0: # If there's a leading zero, pass to the units function
                return units(digits[1])
            elif digits[0] == 1: # If the number is in the teens
                return teenNames[digits[1]]
            else: # For any number higher than 19
                firstWord = tensNames[digits[0]-2]
                if digits[1] == 0: # If a multiple of 10, just return the first word
                    return firstWord
                else: # Otherwise, concatenate the tens digit with the units digit
                    return '{}-{}'.format(firstWord,units(digits[1]))

        if digits[0] == 0: # If a leading zero, pass to the tens function
            return tens(digits[1:])
        elif digits[1:] == [0,0]: # If a multiple of 100, just return 'thingy hundred'
            return '{} hundred'.format(units(digits[0]))
        else: # Otherwise, return 'thingy hundred and thingy'
            return '{} hundred and {}'.format(units(digits[0]),tens(digits[1:]))
    if N == 0:
        return digitNames[0]
    digits = digitsOf(N)
    length = len(digits)
    for dummy in range(40-length): # Add leading zeroes until the number is 40 digits long
        digits.insert(0,0)
    string = hundreds(digits[-3:]) # Start by adding the words for the hundreds digits
    for i in range(len(bigNames)+1): # Now recursively add the words for the thousands,
millions, etc. digits
        if digits[-(6+3*i):-(3+3*i)] == [0,0,0]: # Don't append anything if the digits are
zeroes
            pass
        elif digits[-(3+3*i):] == [0 for dummy in range((6+3*i) - 3)]: # If everything else
is zeroes, make this the only thing in the string
            string = '{} {}'.format(hundreds(digits[-(6+3*i):-(3+3*i)]),bigNames[i])
        elif digits[-(3+3*i):-2] == [0 for dummy in range((6+3*i) - 3 - 2)]: # If no digits
until the tens, add an 'and'
            string = '{} {} and {}'.format(hundreds(digits[-(6+3*i):-
(3+3*i)]),bigNames[i],string)
        else: # Otherwise just add on these digits and a comma
            string = '{} {}, {}'.format(hundreds(digits[-(6+3*i):-
(3+3*i)]),bigNames[i],string)

    if N < 0: # If negative
        string = 'negative {}'.format(string)

    return string

if __name__ == '__main__': # Only run this when executing this file itself
    continuing = True # Runs a basic command line interface
    while continuing:
        request = input('Enter the integer you want printed (or \'s\' to stop):  ')
        if request == 's':
            continuing = False
        else:
            print(inWords(int(request)))
```

It will be a lot easier to code just the first 1000 numbers, which is all this problem calls for.

Anyway, with that out the way, we want to start a new python file and use our functions from this one (which I called `numberprinter.py`). To do that, we just put `from numberprinter import *` at the top along with our usual `import math`. All we do now is iterate through the integers from 1 to 1000 and find how long the string of words for each is.

(We get rid of any hyphens and spaces with `.replace('-','').replace(' ','')` since we're not meant to count these.) The final code for this file is like this:

```
1. import math
2. from numberprinter import *
3. 
4. letterSum = 0
5. for i in range(1,1001):
6.     string = inWords(i)
7.     letterSum += len(string.replace('-','').replace(' ',''))
8. print(letterSum)
```

Project Euler problem 18

August 25, 2017 D Falck Project Euler

Source code available on GitHub.

> By starting at the top of the triangle below and moving to adjacent numbers on the row below, the maximum total from top to bottom is 23.
>
> ```
>    3
>   7 4
>  2 4 6
> 8 5 9 3
> ```
>
> That is, $3 + 7 + 4 + 9 = 23$.
>
> Find the maximum total from top to bottom of the triangle below:
>
> ```
>                75
>               95 64
>              17 47 82
>             18 35 87 10
>            20 04 82 47 65
>           19 01 23 75 03 34
>          88 02 77 73 07 63 67
>         99 65 04 28 06 16 70 92
>        41 41 26 56 83 40 80 70 33
>       41 48 72 33 47 32 37 16 94 29
>      53 71 44 65 25 43 91 52 97 51 14
>     70 11 33 28 77 73 17 78 39 68 17 57
>    91 71 52 38 17 14 91 43 58 50 27 29 48
>   63 66 04 68 89 53 67 30 73 16 69 87 40 31
> 04 62 98 27 23 09 70 98 73 93 38 53 60 04 23
> ```
>
> **NOTE:** As there are only 16384 routes, it is possible to solve this problem by trying every route. However, Problem 67, is the same challenge with a triangle containing one-hundred rows; it cannot be solved by brute force, and requires a clever method! ;o)

This is a really wonderful problem, another great example of dynamic programming. Of course, as the problem tells us, it's perfectly possible to solve this by trying every route possible, but it's ridiculously inefficient. In fact this problem has complexity $O(2^n)$, meaning the time taken increases exponentially with how many rows there are in the triangle (as for each extra row, every route splits into two branches).

No, we're going to have to be clever about this if we want our solution to work for bigger triangles. I encourage you to think about this for a while, because that's how I came across the way to do it.

What works is if, starting at the top, we assign each number a value which is *the maximum sum down to that number*. To find the maximum sum for a particular number, we just add that number to the greater of th two maximum sums directly above it. When we get to the bottom we take the largest of the maximum sums and we have our answer.

Let's start by pasting the triangle into `triangle.txt` and opening it with Python. We'll make a list called `triangle` contain as each of its entries a list of the numbers in that row of the triangle.

```
triangle = [] # Put input triangle into a list of lists
with open('input.txt') as input:
    for line in input:
        triangle.append(line.strip().split(' '))
```

Next, we'll initialise our array `maxSum` that will hold all the maximum sums we just talked about:

```
maxsum = [[0 for entry in row] for row in triangle] # Zeroes same size as triangle
```

Now we want to iterate through the whole triangle, one row at a time, using two nested `for` loops. When we get to each number, we want to compare the two entries above it in `maxSum` and add the greater of the two to the number itself, then store that value in the current entry in `maxSum`. I'm going to use the useful `enumerate(iterable)` function which on every iteration returns both the index and value of the current position in `iterable`.

```
for i, row in enumerate(triangle):
    for j, entry in enumerate(row):
        # Compare the two numbers directly above entry and add the largest to the maxsum here
            if maxsum[i-1][j-1] > maxsum[i-1][j]:
                maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
            else:
                maxsum[i][j] = int(entry) + maxsum[i-1][j]
```

Of course, if we're in the first row of the triangle that number won't *have* any entries above it, so we just set the `maxSum` there to be the entry itself.

```
for i, row in enumerate(triangle):
    for j, entry in enumerate(row):
        if i == 0:
            maxsum[i][j] = int(entry)
        else: # Normally, compare the two numbers directly above entry and add the largest to the maxsum here
            if maxsum[i-1][j-1] > maxsum[i-1][j]:
                maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
            else:
                maxsum[i][j] = int(entry) + maxsum[i-1][j]
```

The only thing we haven't accounted for is numbers at the end of rows. These numbers will only have one entry directly above them, so we have no choice about what we do with the maximum sum there:

```
for i, row in enumerate(triangle):
    for j, entry in enumerate(row):
        if i == 0:
            maxsum[i][j] = int(entry)
        elif j == 0: # If at the start of the line, only one option for max sum
            maxsum[i][j] = int(entry) + maxsum[i-1][j]
        elif j == len(row) - 1: # If at the end of the line, only one option
            maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
        else: # Normally, compare the two numbers directly above entry and add the largest to the maxsum here
            if maxsum[i-1][j-1] > maxsum[i-1][j]:
                maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
            else:
                maxsum[i][j] = int(entry) + maxsum[i-1][j]
```

Now once we've got to the end we just want to take the maximum value of the last row of `maxSum`. Put the whole program together and it looks like this:

```
triangle = [] # Put input triangle into a list of lists
with open('input.txt') as input:
    for line in input:
        triangle.append(line.strip().split(' '))

maxsum = [[0 for entry in row] for row in triangle] # Zeroes same size as triangle

for i, row in enumerate(triangle):
    for j, entry in enumerate(row):
        if i == 0:
            maxsum[i][j] = int(entry)
        elif j == 0: # If at the start of the line, only one option for max sum
            maxsum[i][j] = int(entry) + maxsum[i-1][j]
        elif j == len(row) - 1: # If at the end of the line, only one option
            maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
        else: # Normally, compare the two numbers directly above entry and add the largest to the maxsum here
            if maxsum[i-1][j-1] > maxsum[i-1][j]:
                maxsum[i][j] = int(entry) + maxsum[i-1][j-1]
            else:
                maxsum[i][j] = int(entry) + maxsum[i-1][j]

print(max(maxsum[len(triangle)-1]))
```

Chapter 27

Firework problem (projectile loci)

This was a mechanics problem given by Dr Kwasigroch to us in a Year 12 projectile motion lesson. It was interesting and I've since seen variants of this problem come up in loads of different places!

Firework problem (projectile loci)

Damon Falck

December 2016

Question:

A shell explodes at the origin and the resulting debris is projected with an initial speed u and arbitrary angle θ above the horizontal. Find the locus of points reached by debris as θ is varied. What is the locus of points that are not reached by debris as θ is varied, i.e. where is it safe for a bird to fly? Find the equation of the curve that separates these two loci.

We can draw a diagram to represent the situation as follows:

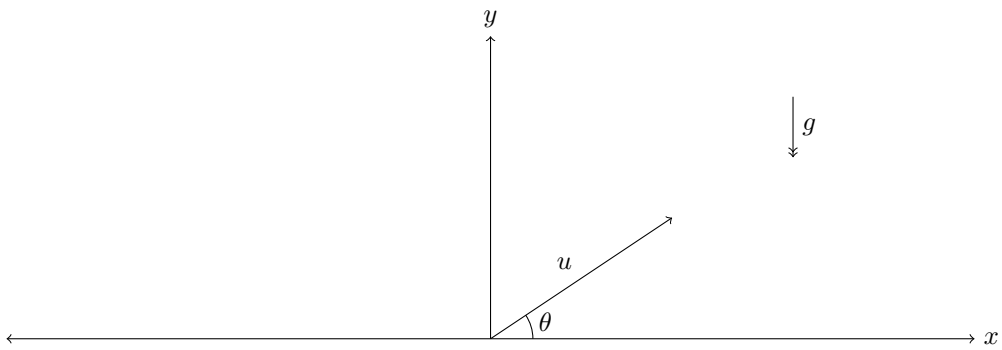


Figure 1: Debris is projected from the origin with initial velocity u and angle θ

We will first find the locus of points which the debris can reach. Let us resolve the initial velocity u into its horizontal and vertical components.

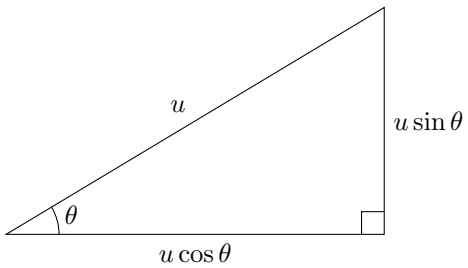


Figure 2: A right triangle to show the components of initial velocity u

We start by deriving an equation for the path of debris.

Applying $s = ut + \frac{1}{2}at^2$ vertically,

$$y = u\sin\theta t - \frac{1}{2}gt^2 \tag{1}$$

where t is the time passed since launch. Now applying $s = ut + \frac{1}{2}at^2 horizontally$,

$$x = u\cos\theta t + 0$$

which gives

$$t = \frac{x}{u\cos\theta}. \tag{2}$$

We can substitute this into (1) to eliminate t:

$$y = u\sin\theta \cdot \frac{x}{u\cos\theta} - \frac{1}{2}g \cdot \frac{x^2}{u^2\cos^2\theta}$$
$$y = \tan\theta x - \frac{g}{2u^2\cos^2\theta}x^2.$$

Now we want to rearrange this to find a quadratic with coefficients of x and y only. We can start by rewriting it as

$$\frac{gx^2}{2u^2} \cdot \frac{1}{\cos^2\theta} - x\tan\theta + y = 0. \tag{3}$$

Using the trigonometric identity

$$\sin^2\theta + \cos^2\theta = 1,$$

by multiplying by $\cos^2\theta$ we get

$$\tan^2\theta + 1 = \frac{1}{\cos^2\theta}.$$

We can substitute this into (3) to give us

$$\frac{gx^2}{2u^2}(\tan^2\theta + 1) - x\tan\theta + y = 0$$

which simplifies to

$$\frac{gx^2}{2u^2}\tan^2\theta - x\tan\theta + \left(y + \frac{gx^2}{2u^2}\right), \tag{4}$$

a quadratic in $\tan\theta$.

What we're interested in is for which points in the xy-plane there exists a real solution for $\tan\theta$ and hence for θ. If at a given point there exists a solution, it means there is some θ such that the path of the debris will pass through this point.

We know that there exists a real solution for $\tan\theta$ if and only if the discriminant $\Delta \geq 0$, so

$$(-x)^2 - 4 \cdot \frac{gx^2}{2u^2} \cdot \left(y + \frac{gx^2}{2u^2}\right) \geq 0$$
$$x^2 - \frac{4gx^2y}{2u^2} - \frac{4g^2x^4}{4u^4} \geq 0$$
$$\frac{u^4x^2 - 2gu^2x^2y - g^2x^4}{u^4} \geq 0$$
$$u^4 - 2gu^2y - g^2x^2 \geq 0. \tag{5}$$

This is the locus of points which the debris can reach, and its shape is shown below.

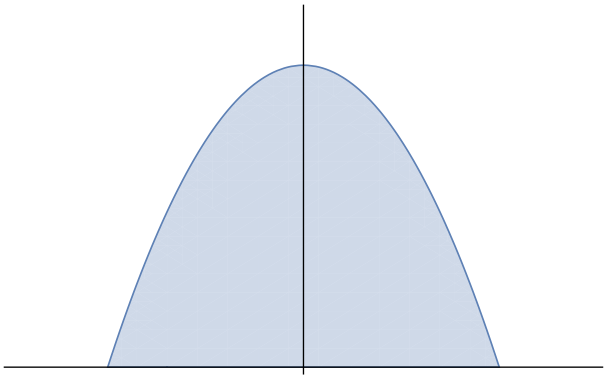


Figure 3: The shape of the locus of points which the debris can reach, plotted in Mathematica

The locus of points which are safe for a bird to fly in is therefore

$$u^4 - 2gu^2y - g^2x^2 < 0, \tag{6}$$

because the two loci are mutually exclusive.

The curve separating these two loci is therefore given by the equation

$$\begin{aligned} u^4 - 2gu^2y - g^2x^2 &= 0 \\ 2gu^2y &= u^4 - g^2x^2 \\ y &= \frac{u^4 - g^2x^2}{2gu^2} \\ y &= -\frac{gx^2}{2u^2} + \frac{u^2}{2g} \end{aligned} \tag{7}$$

which is interestingly the curve of points for which there is only one real solution for θ, meaning that there is only one possible path of the debris that will reach each point in this outer boundary.

Because g and u are constants, this is simply a quadratic in x; in fact it's a just a vertical linear transformation of $y = x^2$. With $u = 10$ ms^{-1} and $g = 9.8$ ms^{-2}, we end up with the equation

$$\begin{aligned} y &= -\frac{9.8x^2}{200} + \frac{100}{19.6} \\ y &= -0.049x^2 + 5.10. \end{aligned} \tag{8}$$

This parabola is plotted below.

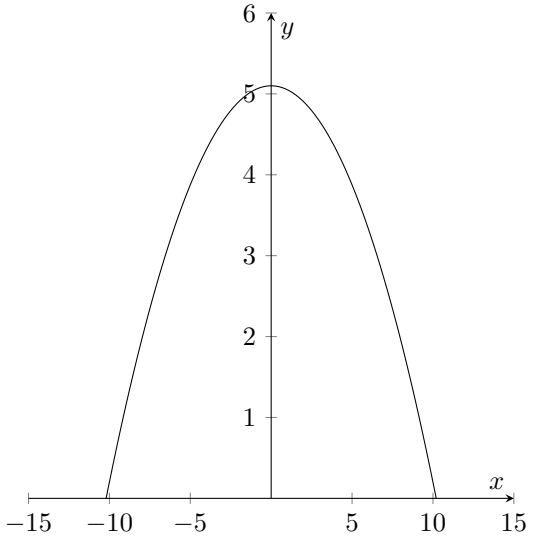


Figure 4: Graph of $y = -0.049x^2 + 5.10$

Chapter 28

Kisses at a party

One of Mr Vaccaro's sheets he used at Year 12 mathematics extension, 'Reasoning and Proof III', was difficult and quite interesting (like usual), and I started my solutions here. I never got around to doing the whole sheet.

Reasoning and Proof III — Solutions

Damon Falck

June 30, 2018

1. (a)

Theorem 1.1. *Let ω_n be the number of people who've had an odd number of kisses after a total of $n \in \mathbb{Z}^*$ kisses in the history of the world. Then ω_n is always even.*

Proof. Because at the beginning of time no one has kissed yet, we can say

$$\omega_0 = 0$$

and after the first kiss, we must have

$$\omega_1 = 2.$$

Now, assume ω_k is odd. Then there are three cases regarding the $(k+1)$st kiss:

Case 1. Both kissers have previously had an odd number of kisses, so after the kiss they both now have an even number of kisses. Hence $\omega_{k+1} = \omega_k - 2$.

Case 2. Both kissers have previously had an even number of kisses, so after the kiss they both now have an odd number of kisses. Hence $\omega_{k+1} = \omega_k + 2$.

Case 3. One kisser has had an odd number of kisses and one an even number. Therefore after the kiss their roles are switched, so $\omega_{k+1} = \omega_k + 1 - 1 = \omega_k$.

In all three cases if ω_k is even then ω_{k+1} is even. Since we know ω_1 is even, by induction ω_n must be even for all n. □

(b)

Theorem 1.2. *During the course of any party there are always at least two people who have kissed the same number of other people at a party.*

Proof. Let there be n people at a party. Therefore the number m_i of other people some individual i has kissed is limited by

$$m_i \in [0, n-1]$$

since it is impossible to kiss oneself. Suppose that every person has kissed a unique number of other people. Then since there are n people and n possible values of m_i, for every integer in the interval $[0, n-1]$ there must be someone who has kissed a corresponding number of people. However, it is impossible for simultaneously one person to have kissed $n-1$ people and another to have kissed no-one. Therefore we reach a contradiction, and so there must always be at least two people who have kissed the same number of other people. □

Chapter 29

An orgy of integration — solutions

An infamous Dr Cheung sheet. I thought it would be beyond me but after doing quite a lot of integration with Miss Brownlee I had a go at these and they were easier than I expected.

An Orgy of Integration - Solutions

Damon Falck

June 30, 2018

1. $$\int x \sin(x^2+3)\,\mathrm{d}x = -\frac{1}{2}\cos(x^2+3) + c.$$

2. $$\begin{aligned}\int \frac{\sec^2 x}{\sqrt{\tan x}}\,\mathrm{d}x &= \int \sec^2 x \tan^{-\frac{1}{2}} x\,\mathrm{d}x \\ &= 2\tan^{\frac{1}{2}} x + c \\ &= 2\sqrt{\tan x} + c.\end{aligned}$$

3. $$\int \cos x \cos(\sin x)\,\mathrm{d}x = \sin(\sin x) + c.$$

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

5. $$\int \frac{2x-\sin x}{x^2+\cos x}\,\mathrm{d}x = \ln\left|x^2+\cos x\right| + c.$$

6. Integrating by parts,

$$\begin{aligned}\int \frac{\ln x}{x}\,\mathrm{d}x &= \ln^2 x - \int \frac{\ln x}{x}\,\mathrm{d}x + c \\ \implies 2\int \frac{\ln x}{x}\,\mathrm{d}x &= \ln^2 x + c \\ \implies \int \frac{\ln x}{x}\,\mathrm{d}x &= \frac{\ln^2 x}{2} + c.\end{aligned}$$

7. $$\begin{aligned}\int \frac{1}{\sqrt{x}(1+\sqrt{x})}\,\mathrm{d}x &= \int \frac{x^{-\frac{1}{2}}}{1+x^{\frac{1}{2}}}\,\mathrm{d}x \\ &= 2\ln\left|1+\sqrt{x}\right| + c.\end{aligned}$$

8. $$\int x\mathrm{e}^{-2x^2}\,\mathrm{d}x = -\frac{1}{4}\mathrm{e}^{-2x^2} + c.$$

9. $$\int \csc x\cot x\mathrm{e}^{\csc x}\,\mathrm{d}x = -\mathrm{e}^{\csc x} + c.$$

10. Integrating by parts,

$$\begin{aligned}\int x(1+x)^{19}\,\mathrm{d}x &= x\cdot\frac{(1+x)^{20}}{20} - \int \frac{(1+x)^{20}}{20}\,\mathrm{d}x + c \\ &= \frac{x(1+x)^{20}}{20} - \frac{(1+x)^{21}}{420} + c.\end{aligned}$$

11. Integrating by parts,

$$\begin{aligned}\int x^2\ln x\,\mathrm{d}x = \frac{1}{3}x^3\ln x &= \int \frac{1}{3}x^2\,\mathrm{d}x + c \\ &= \frac{1}{3}x^3\ln x - \frac{1}{9}x^3 + c.\end{aligned}$$

12. Integrating by parts twice,

$$\begin{aligned}\int x^2\sin x\,\mathrm{d}x &= -x^2\cos x + \int 2x\cos x\,\mathrm{d}x + c \\ &= -x^2\cos x + \left(2x\sin x - \int 2\sin x\,\mathrm{d}x\right) + c \\ &= -x^2\cos x + 2x\sin x + 2\cos x + c.\end{aligned}$$

13. $$\frac{\mathrm{d}}{\mathrm{d}x}(x\ln x) = \ln x + 1.$$

14. Therefore,

$$\int \ln x\,\mathrm{d}x = x\ln x - x + c.$$

15. Integrating twice by parts,

$$\begin{aligned}\int \ln^2 x\,\mathrm{d}x &= x\ln^2 x - \int \frac{2x\ln x}{x}\,\mathrm{d}x + x \\ &= x\ln^2 x - 2x\ln x + 2x + c.\end{aligned}$$

16. As $\sin^2 x + \cos^2 x = 1$, dividing through by $\cos^2 x$ gives $\tan^2 x + 1 = \sec^2 x$. So,

$$\begin{aligned}\int \tan^2 x \,\mathrm{d}x &= \int \Big(\sec^2 x - 1\Big)\,\mathrm{d}x \\ &= \tan x - x + c.\end{aligned}$$

17. Integrating twice by parts,

$$\begin{aligned}\int \mathrm{e}^x \sin x \,\mathrm{d}x &= \mathrm{e}^x \sin x - \int \mathrm{e}^x \cos x \,\mathrm{d}x + c \\ &= \mathrm{e}^x \sin x - \left(\mathrm{e}^x \cos x + \int \mathrm{e}^x \sin x \,\mathrm{d}x\right) + c \\ \implies 2\int \mathrm{e}^x \sin x \,\mathrm{d}x &= \mathrm{e}^x \sin x - \mathrm{e}^x \cos x + c \\ \implies \int \mathrm{e}^x \sin x \,\mathrm{d}x &= \frac{\mathrm{e}^x(\sin x - \cos x)}{2} + c.\end{aligned}$$

18. We know that

$$\begin{aligned}\cos 2x &= \cos^2 x - \sin^2 x \\ &= \cos^2 x - (1 - \cos^2 x) \\ &= 2\cos^2 x - 1 \\ \implies \cos^2 x &= \frac{1 - \cos 2x}{2}.\end{aligned}$$

Hence,

$$\begin{aligned}\int \cos^2 x \,\mathrm{d}x &= \int \left(\frac{1}{2}\cos 2x + \frac{1}{2}\right)\mathrm{d}x \\ &= \frac{1}{4}\sin 2x + \frac{1}{2}x + c\end{aligned}$$

19.
$$\begin{aligned}\int \tan^3 x \,\mathrm{d}x &= \int (\sec^2 x - 1)\tan x \,\mathrm{d}x \\ &= \int \Big(\tan x \sec^2 x - \tan x\Big)\,\mathrm{d}x \\ &= \frac{1}{2}\tan^2 x + \ln|\cos x| + c.\end{aligned}$$

20.
$$\begin{aligned}\int \sin^2 x \cos^2 x \,\mathrm{d}x &= \int \frac{1}{4}\sin^2 2x \,\mathrm{d}x \\ &= \int \frac{1}{4}\left(\frac{1 - \cos 4x}{2}\right)\mathrm{d}x \\ &= \int \left(\frac{1}{8} - \frac{1}{8}\cos 4x\right)\mathrm{d}x \\ &= \frac{1}{8}x - \frac{1}{32}\sin 4x + c.\end{aligned}$$

21. Using the trig factor formulae,

$$\begin{aligned}\int \cos 2x \cos 4x \,\mathrm{d}x &= \int \left(\frac{\cos(2x+4x)}{2} + \frac{\cos(4x-2x)}{2}\right) \mathrm{d}x \\ &= \int \left(\frac{1}{2}\cos 6x + \frac{1}{2}\cos 2x\right) \mathrm{d}x \\ &= \frac{1}{12}\sin 6x + \frac{1}{4}\sin 2x + c.\end{aligned}$$

22.

$$\begin{aligned}\int \sec x \,\mathrm{d}x &= \int \sec x \left(\frac{\sec x + \tan x}{\sec x + \tan x}\right) \mathrm{d}x \\ &= \int \frac{\sec^2 x + \sec x \tan x}{\tan x + \sec x} \,\mathrm{d}x \\ &= \ln|\tan x + \sec x| + c.\end{aligned}$$

23. Integrating twice by parts,

$$\begin{aligned}\int \sec^3 x \,\mathrm{d}x &= \int \sec^2 x \cdot \sec x \,\mathrm{d}x \\ &= \tan x \sec x - \int \tan^2 x \sec x \,\mathrm{d}x + c \\ &= \tan x \sec x - \int (\sec^2 x - 1)\sec x \,\mathrm{d}x + c \\ &= \tan x \sec x - \int \sec^3 x \,\mathrm{d}x + \int \sec x \,\mathrm{d}x + c \\ \implies 2\int \sec^3 x \,\mathrm{d}x &= \tan x \sec x + \int \sec x \,\mathrm{d}x + c \\ \implies \int \sec^3 x \,\mathrm{d}x &= \frac{1}{2}\tan x \sec x + \frac{1}{2}\ln|\tan x + \sec x| + c.\end{aligned}$$

24. Letting $x = 2\tan\theta$, so $\frac{\mathrm{d}x}{\mathrm{d}\theta} = 2\sec^2\theta$ and $\theta = \arctan\left(\frac{x}{2}\right)$,

$$\begin{aligned}\int \frac{3}{x^2+4} \,\mathrm{d}x &= \int \frac{3}{4(tan^2\theta + 1)} \frac{\mathrm{d}x}{\mathrm{d}\theta} \,\mathrm{d}\theta \\ &= \int \frac{3}{4\sec^2\theta} \cdot 2\sec^2\theta \,\mathrm{d}\theta \\ &= \int \frac{3}{2} \,\mathrm{d}\theta \\ &= \frac{3}{2}\theta + c \\ &= \frac{3}{2}\arctan\left(\frac{x}{2}\right) + c.\end{aligned}$$

25. Letting $x = 2\sin\theta$, so $\frac{\mathrm{d}x}{\mathrm{d}\theta} = 2\cos\theta$ and $\theta = \arcsin\left(\frac{x}{2}\right)$,

$$\begin{aligned}\int \frac{1}{\sqrt{4-x^2}}\,\mathrm{d}x &= \int \frac{1}{2\sqrt{1-\sin^2\theta}}\,\frac{\mathrm{d}x}{\mathrm{d}\theta}\,\mathrm{d}\theta \\ &= \int \frac{1}{2\cos\theta}\cdot 2\cos\theta\,\mathrm{d}\theta \\ &= \int 1\,\mathrm{d}\theta \\ &= \theta + c \\ &= \arcsin\left(\frac{x}{2}\right) + c.\end{aligned}$$

26. Letting $t = \sqrt{x^2-4}$, so $x = \sqrt{t^2+4}$ and $\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1}{2}(t^2+4)^{-\frac{1}{2}}\cdot 2t$,

$$\begin{aligned}\int \frac{1}{x\sqrt{x^2-4}}\,\mathrm{d}x &= \int \frac{1}{t\sqrt{t^2+4}}\,\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t \\ &= \int \frac{1}{t\sqrt{t^2+4}}\cdot\frac{1}{2}(t^2+4)^{-\frac{1}{2}}\cdot 2t\,\mathrm{d}t \\ &= \int \frac{1}{t^2+4}\,\mathrm{d}t.\end{aligned}$$

Now letting $t = 2\tan\theta$, so $\frac{\mathrm{d}t}{\mathrm{d}\theta} = 2\sec^2\theta$ and $\theta = \arctan\left(\frac{t}{2}\right)$,

$$\begin{aligned}\int \frac{1}{t^2+4}\,\mathrm{d}t &= \int \frac{1}{4(\tan^2\theta+1)}\,\frac{\mathrm{d}t}{\mathrm{d}\theta}\,\mathrm{d}\theta \\ &= \int \frac{1}{4\sec^2\theta}\cdot 2\sec^2\theta\,\mathrm{d}\theta \\ &= \int \frac{1}{2}\,\mathrm{d}\theta \\ &= \frac{1}{2}\theta + c.\end{aligned}$$

Back-substituting, $\theta = \arctan\left(\frac{t}{2}\right) = \arctan\left(\frac{\sqrt{x^2-4}}{2}\right)$, so

$$\int \frac{1}{x\sqrt{x^2-4}}\,\mathrm{d}x = \frac{1}{2}\arctan\left(\frac{\sqrt{x^2-4}}{2}\right) + c.$$

27. Using the Weierstrass substitution, let $t = \tan\left(\frac{x}{2}\right)$. So,

$$\sin x = \frac{2t}{1+t^2}$$

and

$$\cos x = \frac{1-t^2}{1+t^2}.$$

Also, $\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{2}{1+t^2}$. So,

$$\begin{aligned}
\int \frac{1}{\sin x - 3\cos x - 1}\,\mathrm{d}x &= \int \frac{1}{\frac{2t}{1+t^2} - 3\frac{1-t^2}{1+t^2} - 1}\,\frac{\mathrm{d}x}{\mathrm{d}t}\,\mathrm{d}t \\
&= \int \frac{1+t^2}{2t - 3 + 3t^2 - 1 - t^2} \cdot \frac{2}{1+t^2}\,\mathrm{d}t \\
&= \int \frac{2}{2t^2 + 2t - 4}\,\mathrm{d}t \\
&= \int \frac{1}{t^2 + t - 2}\,\mathrm{d}t \\
&= \int \frac{1}{\left(t + \frac{1}{2}\right)^2 - \frac{9}{4}}\,\mathrm{d}t \\
&= -\int \frac{1}{\frac{9}{4}\left(1 - \left[\frac{2}{3}\left(t + \frac{1}{2}\right)\right]^2\right)}\,\mathrm{d}t.
\end{aligned}$$

Now we let $\sin\theta = \frac{2}{3}\left(t + \frac{1}{2}\right)$, so that $t = \frac{3}{2}\sin\theta - \frac{1}{2}$ and $\frac{\mathrm{d}t}{\mathrm{d}\theta} = \frac{3}{2}\cos\theta$. So, our integral becomes

$$\begin{aligned}
-\int \frac{1}{\frac{9}{4}\left(1 - \sin^2\theta\right)}\,\frac{\mathrm{d}t}{\mathrm{d}\theta}\,\mathrm{d}\theta &= -\int \frac{1}{\frac{9}{4}\cos^2\theta} \cdot \frac{3}{2}\cos\theta\,\mathrm{d}\theta \\
&= -\int \frac{2}{3} \cdot \frac{1}{\cos\theta}\,\mathrm{d}\theta \\
&= -\frac{2}{3}\int \sec\theta\,\mathrm{d}\theta.
\end{aligned}$$

Using our result from question 22, this is $-\frac{2}{3}\ln|\tan\theta + \sec\theta| + c$. Since

$$\theta = \arcsin\left(\frac{2}{3}\left(t + \frac{1}{2}\right)\right) = \arcsin\left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right),$$

we come to

$$\begin{aligned}
\int \frac{1}{\sin x - 3\cos x - 1}\,\mathrm{d}x &= -\frac{2}{3}\ln\left|\tan\left(\arcsin\left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right)\right) + \sec\left(\arcsin\left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right)\right)\right| + c \\
&= -\frac{2}{3}\ln\left|\frac{\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}}{\sqrt{1 - \left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right)^2}} + \frac{1}{\sqrt{1 - \left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right)^2}}\right| + c \\
&= -\frac{2}{3}\ln\left|\frac{2\tan\frac{x}{2} + 4}{3\sqrt{1 - \left(\frac{2}{3}\tan\frac{x}{2} + \frac{1}{3}\right)^2}}\right| + c.
\end{aligned}$$

28. Letting $u = \frac{1}{x}$, so that $x = \frac{1}{u}$ and $\frac{\mathrm{d}x}{\mathrm{d}u} = -\frac{1}{u^2}$,

$$\begin{aligned}\int \frac{1}{x\sqrt{x^2-1}}\,\mathrm{d}x &= \int \frac{1}{\frac{1}{u}\sqrt{\frac{1}{u^2}-1}}\cdot\left(-\frac{1}{u^2}\right)\mathrm{d}u \\ &= -\int \frac{1}{u^2\cdot\frac{1}{u}\sqrt{\frac{1}{u^2}-1}}\,\mathrm{d}u \\ &= -\int \frac{1}{u\sqrt{\frac{1}{u^2}-1}}\,\mathrm{d}u \\ &= -\int \frac{1}{\sqrt{1-u^2}}\,\mathrm{d}u.\end{aligned}$$

Now letting $u = \sin\theta$ so that $\frac{\mathrm{d}u}{\mathrm{d}\theta} = \cos\theta$, we get

$$\begin{aligned}-\int \frac{1}{\sqrt{1-u^2}}\,\mathrm{d}u &= -\int \frac{1}{\sqrt{1-\sin^2\theta}}\cos\theta\,\mathrm{d}\theta \\ &= -\int \frac{\cos\theta}{\cos\theta}\,\mathrm{d}\theta \\ &= -\theta + c \\ &= -\arcsin u + c \\ &= -\arcsin\left(\frac{1}{x}\right) + c.\end{aligned}$$

29.
$$\begin{aligned}\int \frac{4\sin x + 3\cos x}{\sin x + 2\cos x}\,\mathrm{d}x &= \int \left(\frac{-\cos x + 2\sin x}{\sin x + 2\cos x} + \frac{2\sin x + 4\cos x}{\sin x + 2\cos x}\right)\mathrm{d}x \\ &= \int \left(-\frac{\cos x - 2\sin x}{\sin x + 2\cos x} + 2\right)\mathrm{d}x \\ &= -\ln|\sin x + 2\cos x| + 2x + c.\end{aligned}$$

30. Letting $u = x - 1$, so $x = u + 1$ and $\frac{\mathrm{d}x}{\mathrm{d}u} = 1$,

$$\begin{aligned}\int \sqrt{\frac{x-1}{2-x}}\,\mathrm{d}x &= \int \sqrt{\frac{u}{2-u-1}}\cdot 1\cdot\mathrm{d}u \\ &= \int \frac{\sqrt{u}}{\sqrt{1-u}}\,\mathrm{d}u.\end{aligned}$$

We now let $u = \sin^2\theta$, so that $\sqrt{u} = \sin\theta$ and $\frac{\mathrm{d}u}{\mathrm{d}\theta} = 2\sin\theta\cos\theta$. Thus,

$$\begin{aligned}\int \frac{\sqrt{u}}{\sqrt{1-u}}\,\mathrm{d}u &= \int \frac{\sin\theta}{\sqrt{1-\sin^2\theta}}\cdot 2\sin\theta\cos\theta\,\mathrm{d}\theta \\ &= \int \frac{2\sin^2\theta\cos\theta}{\cos\theta}\,\mathrm{d}\theta \\ &= \int 2\sin^2\theta\,\mathrm{d}\theta \\ &= \int (1-\cos 2\theta)\,\mathrm{d}\theta \\ &= \theta - \frac{1}{2}\sin 2\theta + c \\ &= \theta - \sin\theta\cos\theta + c.\end{aligned}$$

So, since $\theta = \arcsin(\sqrt{u} = \arcsin\sqrt{x-1}$, our solution is

$$\begin{aligned}\int \sqrt{\frac{x-1}{2-x}}\,\mathrm{d}x &= \arcsin\left(\sqrt{x-1}\right) - \sqrt{x-1}\cdot\sqrt{1-\left(\sqrt{x-1}\right)^2} + c \\ &= \arcsin\left(\sqrt{x-1}\right) - \sqrt{x-1}\sqrt{2-x} + c.\end{aligned}$$

Chapter 30

A few STEP problems

These were solutions to a few STEP problems I did for Miss Brownlee about halfway through Year 12.

A few STEP problems...

Damon Falck

February 2017

STEP 1 2009 Question 2

We are given a curve with equation

$$y^3 = x^3 + a^3 + b^3$$

where a and b are positive constants. Differentiating,

$$\frac{\mathrm{d}y}{\mathrm{d}x} = x^2(x^3 + a^3 + b^3)^{-2/3}$$

and so at the point $(-a, b)$, the tangent line

$$y = mx + c$$

has gradient

$$\begin{aligned} m = \left.\frac{\mathrm{d}y}{\mathrm{d}x}\right|_{x=-a} &= (-a)^2\left[(-a)^3 + a^3 + b^3\right]^{-2/3} \\ &= a^2(b^3)^{-2/3} \\ &= \frac{a^2}{b^2}. \end{aligned}$$

Now since we know this line passes through point $(-a, b)$ (as it is tangent to the curve there), we can find c:

$$\begin{aligned} b &= \frac{a^2}{b^2}(-a) + c \\ \implies c &= b + \frac{a^3}{b^2}. \end{aligned}$$

Thus, the equation of the tangent line is

$$\begin{aligned} y &= \frac{a^2}{b^2}x + b + \frac{a^3}{b^2} \\ \implies b^2y - a^2x &= a^3 + b^3 \end{aligned}$$

as desired.

Now if $a = 1$ and $b = 2$, then the points where the tangent meets the curve can be found by solving simultaneously the equations

$$y^3 = x^3 + 9, \tag{1}$$

$$4y - x = 9 \tag{2}$$

where eq. (1) is the original curve and eq. (2) is its tangent at $(-1, 2)$. Equation (2) gives us

$$y = \frac{x+9}{4}$$

and so substituting this into eq. (2), we know that at all intersection points

$$\begin{aligned}
\left(\frac{x+9}{4}\right)^3 &= x^3 + 9 \\
\implies \frac{(x+9)^3}{64} - x^3 - 9 &= 0 \\
\implies x^3 + 27x^2 + 243x + 729 - 64x^3 - 576 &= 0 \\
\implies -63x^3 + 27x^2 + 243x + 153 &= 0 \\
\implies 7x^3 - 3x^2 - 27x - 17 &= 0 \qquad (3)
\end{aligned}$$

as desired.

Now we want to find a pair of values to satisfy the original equation. Let us take $a = 1$ and $b = 2$ as above (since we already know some information regarding these values). We know that eq. (3) above satisfies the original curve at point $(-1, 2)$; indeed, it is tangent there. So we know two of the roots of the above equation will be $x = -1$, and so $(x+1)^2$ must be a factor.

Factorising out $(x+1)$, we come to

$$(x+1)(7x^2 - 10x - 17) = 0$$

and factorising out the same again,

$$(x+1)^2(7x - 17) = 0.$$

Therefore the tangent and the curve must intersect again at

$$x = \frac{17}{7}.$$

Using the tangent equation 2, we therefore know that at this point

$$y = \frac{x+9}{4} = \frac{\frac{17}{7}+9}{4} = \frac{80}{28} = \frac{20}{7}.$$

So, $(\frac{17}{7}, \frac{20}{7})$ must be a point on the original curve, and so as $a = 1$ and $b = 2$, we know that

$$\left(\frac{20}{7}\right)^3 = \left(\frac{17}{7}\right)^3 + 1^3 + 2^3$$

must hold true.

Multiplying both sides by 7^3, we come to

$$20^3 = 17^3 + 7^3 + 14^3$$

and so the positive integers $p = 20$, $q = 17$, $r = 7$ and $s = 14$ satisfy the equation

$$p^3 = q^3 + r^3 + s^3.$$

STEP 1 2009 Question 3

(i) Looking at the equation

$$x = (a - x)^{\frac{1}{2}},$$

we immediately note that as the square root is defined to be positive only, we require $x \geq 0$ and for x to be real we also require $x \leq a$. Squaring both sides however gives that

$$x^2 = a - x$$

and since squares are always positive, $a - x \geq 0$ and so $x \leq a$.

Hence, we only need to check the first condition. With the exception of these two conditions, the quadratic

$$x^2 + x - a = 0$$

is equivalent, and so the question becomes how many *nonnegative* real roots this quadratic has.

Solving this yields

$$x = \frac{-1 \pm \sqrt{1 + 4a}}{2}.$$

We can see that for the condition $x \geq 0$ to be met, we need

$$-1 \pm \sqrt{1 + 4a} \geq 0.$$

which implies

$$\sqrt{1 + 4a} \leq -1$$

or

$$\sqrt{1 + 4a} \geq 1.$$

The square root is *positive* by definition, so we can never have the first inequality, and so there can at most be one root to this equation. For the second inequality, we require $4a \geq 0$ and so $a \geq 0$.

Thus, there are no real solutions when $a < 0$ and there is one real root when $a \geq 0$, as required.

Let us not consider the equation

$$x = \left[(1 + a)x - a\right]^{\frac{1}{3}}.$$

Unlike above, this is identical in every way to the standard cubic equation

$$x^3 - (1 + a)x + a = 0$$

because cubes can have any sign.

Let $f(x) = x^3 - (1 + a)x + a = 0$. The only factors of a that we know are $1, -1, a, -a$ and by trial and error we find that $f(1) = 0$, and so by the factor theorem $(x - 1)$ must be a factor.

Factorising, we come to

$$(x - 1)(x^2 + x - a) = 0$$

and so we can now consider the quadratic, which being identical to the one discussed previously has discriminant

$$\Delta = 1 + 4a.$$

So, the quadratic has two real roots if

$$\begin{aligned} & \Delta > 0 \\ \implies & 1 + 4a > 0 \\ \implies & a > -\frac{1}{4}, \end{aligned}$$

one real root if

$$a = -\frac{1}{4}$$

and no real roots if

$$a < -\frac{1}{4}.$$

Therefore, considering the cubic in question now, because of the additional real root $x = 1$, we have the following cases:

- Three real roots if $a > -\frac{1}{4}$
- Two real roots if $a = -\frac{1}{4}$
- One real root if $a < -\frac{1}{4}$

(ii) We now consider the equation

$$x = (b + x)^{\frac{1}{2}},$$

which is very similar to that discussed in part (i) with the exception of a minus sign. So, we will tackle it in a somewhat similar way. As before, we require

$$x \geq 0$$

and

$$x \geq -b$$

but, as before, all squares are positive so the second condition is always true. Hence the question becomes how many *nonnegative* real roots there are to the quadratic

$$x^2 - x - b = 0$$

which implies

$$x = \frac{1 \pm \sqrt{1 + 4b}}{2}$$

by the quadratic formula.

This time we will consider the discriminant first. We know

$$\Delta = 1 + 4b$$

and so the number of real roots are:

- Two real roots if $b > -\frac{1}{4}$
- One real root if $b = -\frac{1}{4}$
- No real roots if $b < -\frac{1}{4}$

For nonnegative real roots we now need to consider the numerator of the fraction. Assuming $b > -\frac{1}{4}$so the square root is real and positive, there will be two nonnegative real solutions *only* if

$$\sqrt{1+4b} \leq 1$$
$$b \leq 0.$$

Otherwise, only the + sign yields a positive numerator. If $b \leq -\frac{1}{4}$, the $\pm$ sign becomes irrelevant.

Thus, we have the following cases for the *original* equation in the question:

- Two real roots if $-\frac{1}{4} < b \leq 0$
- One real root if $b > 0$ or $b = -\frac{1}{4}$
- No real roots if $b < -\frac{1}{4}$

STEP 1 2009 Question 5

(i) We're given that

$$V = \frac{1}{3}\pi r^2 h \tag{4}$$

and

$$A = \pi r \ell. \tag{5}$$

Since A is fixed and r is chosen so that V is at its stationary value, we want to find V in terms of r and A only and then differentiate with respect to r.

We know by Pythagoras that

$$h = \sqrt{\ell^2 - r^2}$$

and from eq. (5) that

$$\ell = \frac{A}{\pi r},$$

so substituting both of these into eq. (4) we get

$$\begin{aligned} V &= \frac{1}{3}\pi r^2 \sqrt{\left(\frac{A}{\pi r}\right)^2 - r^2} \\ &= \frac{1}{3}\pi \sqrt{\frac{r^4 A^2}{\pi^2 r^2} - r^6} \\ &= \frac{1}{3}\pi \left[\frac{A^2}{\pi^2} r^2 - r^6\right]^{\frac{1}{2}}. \end{aligned}$$

So, differentiating with respect to r,

$$\frac{\partial V}{\partial r} = \frac{1}{6}\pi \left[\frac{A^2}{\pi^2} r^2 - r^6\right]^{-\frac{1}{2}} \left[\frac{2A^2}{\pi^2} r - 6r^5\right]$$

and so as V is at its maximum value, $\frac{\partial V}{\partial r} = 0$ and so

$$\left[\frac{2A^2}{\pi^2} r - 6r^5\right] = 0.$$

Assuming $r \neq 0$,

$$\begin{aligned} 6r^4 &= \frac{2A^2}{\pi^2} \\ \implies A^2 &= 3\pi^2 r^4 \end{aligned}$$

as desired. Since $\ell = \frac{A}{\pi r}$, the above equation also tells us that

$$\ell = \frac{\sqrt{3\pi^2 r^4}}{\pi r} = \frac{\sqrt{3}\pi r^2}{\pi r} = \sqrt{3}r$$

as desired.

(ii) If instead V is fixed and r is chosen so that A is at its stationary value, we will do something very similar. We have already calculated that

$$V = \frac{1}{3}\pi \left[\frac{A^2}{\pi^2}r^2 - r^6\right]^{\frac{1}{2}},$$

and so with a bit of rearranging to make A the subject,

$$\begin{aligned} V^2 &= \frac{1}{9}\pi^2 \left[\frac{A^2}{\pi^2}r^2 - r^6\right] \\ \implies \frac{1}{9}\pi^2 \frac{A^2}{\pi^2}r^2 &= V^2 + \frac{1}{9}\pi^2 r^6 \\ \implies A^2 &= \frac{9V^2}{r^2} + \frac{9\pi^2 r^6}{9r^2} \\ \implies A &= \left[9V^2 r^{-2} + \pi^2 r^4\right]^{\frac{1}{2}}. \end{aligned}$$

Therefore differentiating with respect to r (with V fixed),

$$\frac{\partial A}{\partial r} = \frac{1}{2}\left[9V^2 r^{-2} + \pi^2 r^4\right]^{-\frac{1}{2}} \left[-18V^2 r^{-3} + 4\pi^2 r^3\right]$$

and so as A is stationary, $\frac{\partial A}{\partial r} = 0$ and so

$$\begin{aligned} \left[-18V^2 r^{-3} + 4\pi^2 r^3\right] &= 0 \\ \implies 2\pi^2 r^3 &= 9V^2 r^{-3} \\ \implies V &= \sqrt{\frac{2\pi^2 r^6}{9}}. \end{aligned}$$

Now we know from eq. (4) that $h = \dfrac{3V}{\pi r^2}$ and so

$$\begin{aligned} h &= \frac{3}{\pi r^2}\sqrt{\frac{2\pi^2 r^6}{9}} \\ &= \sqrt{\frac{18\pi^2 r^6}{9\pi^2 r^4}} \\ &= \sqrt{2r^2} \\ &= \sqrt{2}r. \end{aligned}$$

Thus we have an expression for h in terms of r as desired.

Note: most other questions from this paper done, but not yet written up.

Chapter 31

Differential equations STEP questions

While studying differential equations with Mr Dales I had a go at these two STEP questions he set us. Of course we've done many similar questions since, but at the time they were very satisfying.

Differential Equations STEP Questions

Damon Falck

November 2017

1 STEP 1 2003, Question 8

We're told that the rate at which A converts to B is

$$-\frac{\mathrm{d}}{\mathrm{d}t}(xV) = \frac{\mathrm{d}}{\mathrm{d}t}(yV) = kVxy$$

for a positive constant k. The total volume V is fixed, so the right-hand equality gives

$$\frac{\mathrm{d}y}{\mathrm{d}t} = kxy$$

but we know $x + y = 1$, so

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ky(1-y).$$

Separating the variables and integrating,

$$\int \frac{\mathrm{d}y}{y(1-y)} = \int k\,\mathrm{d}t. \tag{1}$$

The right-hand integral is just $kt + c$ for some constant c, and we can evaluate the left-hand integral using partial fraction decomposition. Let

$$\frac{1}{y(1-y)} \equiv \frac{A}{y} + \frac{B}{1-y}.$$

Then,

$$\begin{aligned} 1 &\equiv A(1-y) + By \\ &\equiv A + (B-A)y \end{aligned}$$

so equating coefficients,

$$A = 1$$

and

$$B - A = 0 \implies B = A = 1.$$

Thus, eq. (1) gives

$$\begin{aligned} \int \frac{\mathrm{d}y}{y} + \int \frac{\mathrm{d}y}{1-y} &= kt + c \\ \implies \ln y - \ln(1-y) &= kt + c \\ \implies \ln\left(\frac{y}{1-y}\right) &= kt + c \\ \implies \frac{y}{1-y} &= \mathrm{e}^{kt+c} = D\mathrm{e}^{kt} \end{aligned}$$

for some constant $D = \mathrm{e}^c$. Rearranging gives

$$\begin{aligned} y &= D\mathrm{e}^{kt} - D\mathrm{e}^{kt}y \\ \implies y(1 + D\mathrm{e}^{kt}) &= D\mathrm{e}^{kt} \\ \implies y &= \frac{D\mathrm{e}^{kt}}{1 + D\mathrm{e}^{kt}} \end{aligned} \tag{2}$$

as desired.

Taking the limit as t tends to infinity, the final value of

$$\lim_{t\to\infty} y = \lim_{t\to\infty} \frac{D\mathrm{e}^{kt}}{1 + D\mathrm{e}^{kt}} = \frac{D\mathrm{e}^{kt}}{D\mathrm{e}^{kt}} = 1.$$

When $y = 1$, $x = 0$, and so the volume of A gradually tends towards zero as time progresses, but A never actually completely converts to B as eq. (2) has no solutions for $y = 1$.

2 STEP 2 2003, Question 8

We are given that

$$\frac{\mathrm{d}y}{\mathrm{d}t} + k\left(\frac{t^2 - 3t + 2}{t + 1}\right) y = 0. \tag{3}$$

Rearranging this, we can separate the variables:

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}t} &= -k\left(\frac{t^2 - 3t + 2}{t + 1}\right) y \\ \implies \frac{\mathrm{d}y}{y} &= -k\left(\frac{t^2 - 3t + 2}{t + 1}\right) \mathrm{d}t. \end{aligned} \tag{4}$$

We can simplify the right hand side using a multiplication grid (or we could have used polynomial long division):

	t	-4
t	t^2	$-4t$
1	t	-4

For this division, we see that the quotient is $t - 4$ and the remainder is 6, and so

$$\frac{t^2 - 3t + 2}{t + 1} = (t - 4) + \frac{6}{t + 1}.$$

Thus, integrating eq. (4), we come to

$$\begin{aligned} \int \frac{\mathrm{d}y}{y} &= \int -k\left(t - 4 + \frac{6}{t + 1}\right) \mathrm{d}t \\ \implies \ln y &= -k\left(\frac{1}{2}t^2 - 4t + 6\ln(t + 1)\right) + \alpha \end{aligned}$$

for some constant α, and so, exponentiating,

$$y = \beta \mathrm{e}^{4kt - \frac{1}{2}kt^2}(t + 1)^{-6k}$$

for some constant $\beta = \mathrm{e}^{\alpha}$. Now we can use the fact that $y = A$ when $t = 0$:

$$A = \beta \mathrm{e}^{0}(0+1)^{-6k} = \beta$$

so that our expression for y in terms of t that we want is

$$y = A\mathrm{e}^{4kt - \frac{1}{2}kt^2}(t+1)^{-6k}. \tag{5}$$

Now, y has a stationary value whenever $\dfrac{\mathrm{d}y}{\mathrm{d}t} = 0$. So, in this situation, eq. (3) gives us

$$\begin{aligned} k\left(\frac{t^2 - 3t + 2}{t+1}\right) y &= 0 \\ \implies (t^2 - 3t + 2)y &= 0 \quad \text{if and only if } t \neq -1 \\ \implies (t-2)(t-1)A\mathrm{e}^{4kt - \frac{1}{2}kt^2}(t+1)^{-6k} &= 0 \end{aligned}$$

which has solutions at $t = -1$, $t = 1$ and $t = 2$, but the original differential equation is asymptotic at $t = -1$ so our two stationary points are at $t = 1$ and $t = 2$.

Using eq. (5), at $t = 1$,

$$y = A\mathrm{e}^{4k - \frac{1}{2}k}(1+1)^{-6k} = A\mathrm{e}^{\frac{7}{2}k} \cdot 2^{-6k}$$

and at $t = 2$,

$$y = A\mathrm{e}^{8k - 2k}(2+1)^{-6k} = A\mathrm{e}^{6k} \cdot 3^{-6k}.$$

The ratio η of these two turning points is therefore

$$\eta = \frac{A\mathrm{e}^{\frac{7}{2}k} \cdot 2^{-6k}}{A\mathrm{e}^{6k} \cdot 3^{-6k}} = \left(\frac{3}{2}\right)^{6k} \mathrm{e}^{-\frac{5}{2}k}$$

as desired.

Again with reference to eq. (5), as $t \to +\infty$ the exponential term becomes $\mathrm{e}^{-\frac{1}{2}kt^2}$, so that if $k > 0$ it will get very small and y will tend to zero, whereas if $k < 0$ it will get very large and y will diverge to infinity. Either way, the effect of the exponential term outweighs the value of $(t+1)^{-6k}$.

Figure 1 shows the graph of y in the case $k > 0$ and fig. 2 shows the case that $k < 0$.

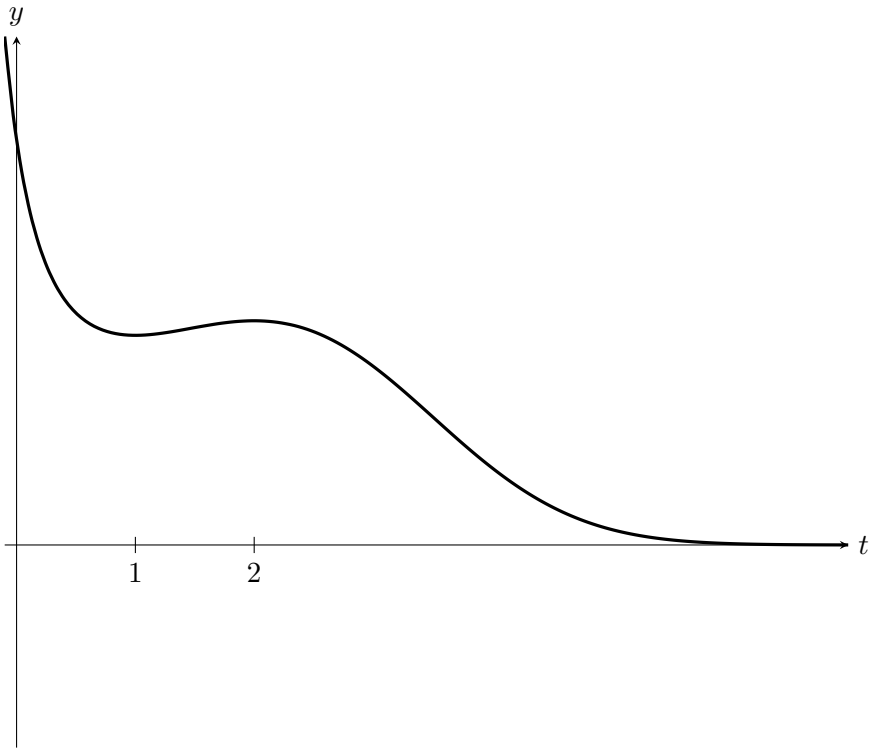


Figure 1: The case when $k > 0$.

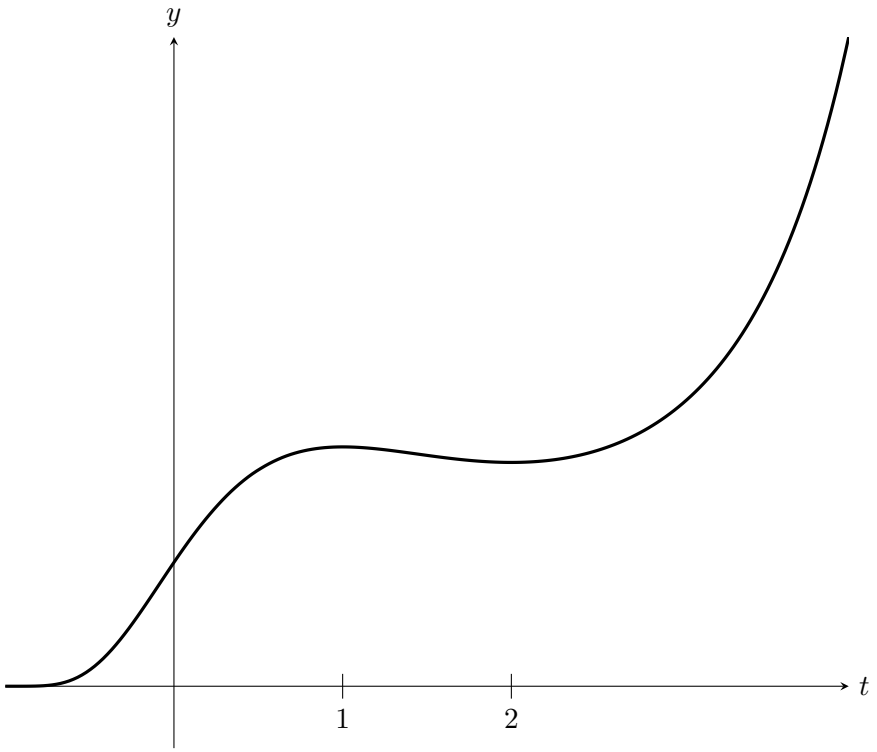


Figure 2: The case when $k < 0$.

Chapter 32

Michaelmas 2016 pure mathematics — half term work

This was my first stab at LaTeX: it sounded useful and cool so I had a go at learning how to use it over the October half term and decided to do my homework and a few other things in it as practice.

Pure - half term work

Damon Falck

October 2016

1 Polynomials (Smedley & Wiseman)

1.1 Exercise 4C

10. We first expand the left hand side of the identity:

$$(x^2 + a)(x - 4) \equiv x^3 + bx^2 + cx - 20$$
$$x^3 - 4x^2 + ax - 4a \equiv x^3 + bx^2 + cx - 20,$$

so by comparing the coefficients we get:

$$-4 = b \tag{1}$$
$$a = c \tag{2}$$
$$-4a = -20. \tag{3}$$

Equation (3) gives $a = \frac{20}{4} = 5$ and so from (2) we know $c = a = 5$. Finally, (1) shows that $b = -4$. Hence, the values of a, b and c are $5, -4$ and 5 respectively. Now we substitute these values into the equation given and solve by factorising:

$$x^3 + bx^2 + cx - 20 = 0$$
$$x^3 - 4x^2 + 5x - 20 = 0$$
$$(x - 4)(x^2 + 5) = 0$$

$$\therefore \quad x = 4 \quad \text{or} \quad x = \pm\sqrt{5}i.$$

Thus there is only one real solution, $x = 4$. □

11. As above, we first expand the left hand side of the identity:

$$(x^2 + b)(x + c) \equiv x^3 - 3x^2 + bx - 15$$
$$x^2 + cx^2 + bx + bc \equiv x^3 - 3x^2 + bx - 15,$$

so by comparing the coefficients,

$$c = -3 \tag{1}$$
$$bc = -15. \tag{2}$$

Hence, by substituting (1) into (2):

$$b \times -3 = -15$$
$$b = 5.$$

Thus we have $b = 5$ and $c = -3$. We can now solve the given equation:

$$\begin{aligned} x^3 - 3x^2 + bx - 15 &= 0 \\ x^3 - 3x^2 + 5x - 15 &= 0 \\ (x-3)(x^2+5) &= 0 \end{aligned}$$

$$\therefore \quad x = 3 \quad \text{or} \quad x = \pm\sqrt{5}i.$$

13. Let $f(x) = (2x-3)(x^2-5x-1)+7$.

Thus by expanding,

$$f(x) = 2x^3 - 13x^2 + 13x + 10.$$

We know[1] that if $f(x)$ has a linear factor $\alpha x + \beta$, then the constant β will be a factor of 10, i.e. one of ±1, ±2, ±5 or ±10.

We start by evaluating $f(1)$ and $f(-1)$. Since $f(1) = 12$ and $f(-1) = -18$, $(x \pm 1)$ are not factors.

Next, we evaluate $f(2)$ and $f(-2)$. By the factor theorem, since $f(2) = 0$, we know $(x-2)$ is a factor of $f(x)$. So,

$$\begin{aligned} f(x) &\equiv (x-2)(ax^2+bx+c) \\ 2x^3 - 13x^2 + 13x + 10 &\equiv (x-2)(ax^2+bx+c). \end{aligned}$$

By inspection[2] we can find that $a = 2$, $b = -9$ and $c = -5$. Now let's substitute these in and factorise further:

$$\begin{aligned} f(x) &\equiv (x-2)(2x^2-9x-5) \\ f(x) &\equiv (x-2)(2x+1)(x-5). \end{aligned}$$

Now, we're given the inequality $(2x-3)(x^2-5x-1) \geq -7$. We can add 7 to both sides of this to simplify the inequality we're trying to solve:

$$\begin{aligned} (2x-3)(x^2-5x-1) &\geq 0 \\ f(x) &\geq 0. \end{aligned}$$

We can see that the solutions to $f(x) = 0$ are $-\frac{1}{2}$, 2 and 5, and so the curve $f(x)$ intersects the x-axis at these points.

As can be seen in figure 1, because $f(x)$ is a cubic function with a positive coefficient of x^3, $f(x) \geq 0$ is true between the first and second x-intercepts and and after the third.

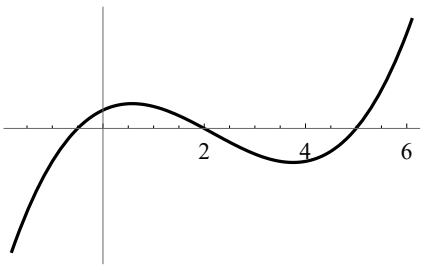


Figure 1: A sketch of $f(x)$

Hence, the solutions to $f(x) \geq 0$ are $-\frac{1}{2} \leq x \leq 2$ and $x \geq 5$.

[1] This is a somewhat lengthy way of finding a factor but it's what Smedley & Wiseman recommend.
[2] Using a farmer's field, or expanding and comparing coefficients.

2 Trigonometry (Smedley & Wiseman)

2.1 Exercise 14A

3. (a)

$$\sin 200^\circ = -\sin(200^\circ - 180^\circ) = -\sin 20^\circ$$

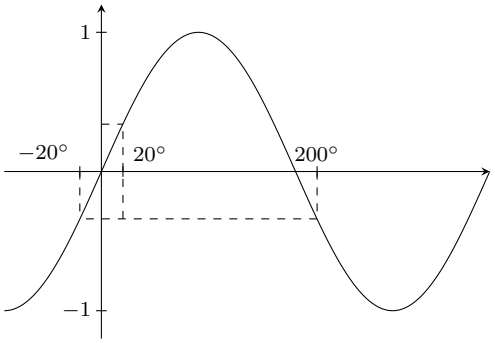


Figure 2: A graph of $\sin x$.

(b)

$$\cos 240^\circ = -\cos(240^\circ - 180^\circ) = -\cos 60^\circ$$

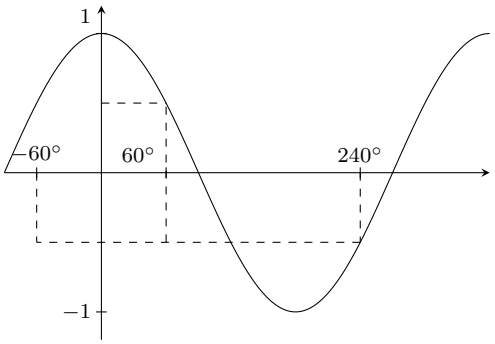


Figure 3: A graph of $\cos x$.

(c)

$$\tan 160^\circ = \tan(160^\circ - 180^\circ) = -\tan 20^\circ$$

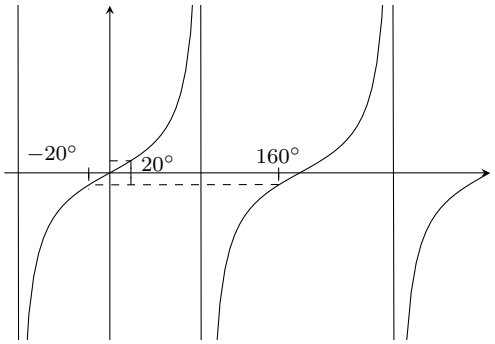


Figure 4: A graph of $\tan x$.

(d)
$$\cos 310^\circ = \cos(310^\circ - 360^\circ) = \cos 50^\circ$$

(e)
$$\tan 220^\circ = \tan(220^\circ - 180^\circ) = \tan 40^\circ$$

(f)
$$\cos 490^\circ = -\cos(490^\circ - 360^\circ - 180^\circ) = -\cos 50^\circ$$

(g)
$$\sin(-20^\circ) = -\sin 20^\circ$$

(h)
$$\cos(-280^\circ) = \cos 280^\circ = \cos(280^\circ - 360^\circ) = \cos 80^\circ$$

4. For $0^\circ \leq \theta \leq 360^\circ$:

(a)
$$\begin{aligned}\sin\theta &= 0.3\\ \theta &= (\arcsin 0.3), (180^\circ - \arcsin 0.3)\\ \theta &= 17.4^\circ, 162.6^\circ\end{aligned}$$

(b)
$$\begin{aligned}\cos\theta &= 0.7\\ \theta &= (\arccos 0.7), (360^\circ - \arccos 0.7)\\ \theta &= 45.6^\circ, 314.4^\circ\end{aligned}$$

(c)
$$\begin{aligned}\tan\theta &= 2\\ \theta &= (\arctan 2), (180^\circ + \arctan 2)\\ \theta &= 63.4^\circ, 243.4^\circ\end{aligned}$$

(d)
$$\begin{aligned}\cos\theta &= -0.5\\ \theta &= (180^\circ - \arccos 0.5), (180^\circ + \arccos 0.5)\\ \theta &= 120^\circ, 240^\circ\end{aligned}$$

(e)
$$\begin{aligned}\sin\theta &= -0.35\\ \theta &= (180^\circ + \arcsin 0.35), (360^\circ - \arcsin 0.35)\\ \theta &= 200.5^\circ, 339.5^\circ\end{aligned}$$

(f)
$$\begin{aligned}\tan\theta &= -7\\ \theta &= (180^\circ - \arctan 7), (360^\circ - \arctan 7)\\ \theta &= 98.1^\circ, 278.1^\circ\end{aligned}$$

(g)

$$\begin{aligned}\sin\theta &= 0.8\\ \theta &= (\arcsin 0.8), (180^\circ - \arcsin 0.8)\\ \theta &= 53.1^\circ, 126.9^\circ\end{aligned}$$

(h)

$$\begin{aligned}\sin\theta &= -1\\ \theta &= 180^\circ + \arcsin 1\\ \theta &= 270^\circ\end{aligned}$$

5. For $-180^\circ \leq \theta \leq 180^\circ$:

(a)

$$\begin{aligned}\csc\theta &= 2\\ \frac{1}{\sin\theta} &= 2\\ \sin\theta &= \frac{1}{2}\\ \theta &= \left(\arcsin\frac{1}{2}\right), \left(180^\circ - \arcsin\frac{1}{2}\right)\\ \theta &= 30^\circ, 150^\circ\end{aligned}$$

(b)

$$\begin{aligned}\sec\theta &= 3\\ \frac{1}{\cos\theta} &= 3\\ \cos\theta &= \frac{1}{3}\\ \theta &= \pm\arccos\frac{1}{3}\\ \theta &= \pm 70.5^\circ\end{aligned}$$

(c)

$$\begin{aligned}\cot\theta &= 0.5\\ \frac{1}{\tan\theta} &= 0.5\\ \tan\theta &= 2\\ \theta &= (\arctan 2), (-180^\circ + \arctan 2)\\ \theta &= 63.4^\circ, -116.6^\circ\end{aligned}$$

(d)

$$\begin{aligned}\cot\theta &= -3\\ \tan\theta &= -\frac{1}{3}\\ \theta &= \left(180^\circ - \arctan\frac{1}{3}\right), \left(-\arctan\frac{1}{3}\right)\\ \theta &= 161.6^\circ, -18.4^\circ\end{aligned}$$

(e)

$$\begin{aligned}\sec\theta &= 6\\ \cos\theta &= \frac{1}{6}\\ \theta &= \pm\arccos\frac{1}{6}\\ \theta &= \pm 80.4^\circ\end{aligned}$$

(f)

$$\begin{aligned}\csc\theta &= 5\\ \sin\theta &= \frac{1}{5}\\ \theta &= \left(\arcsin\frac{1}{5}\right), \left(180^\circ - \arcsin\frac{1}{5}\right)\\ \theta &= 11.5^\circ, 168.5^\circ\end{aligned}$$

(g)

$$\begin{aligned}\sec\theta &= -1\\ \cos\theta &= -1\\ \theta &= \pm 180^\circ\end{aligned}$$

(h)

$$\begin{aligned}\csc\theta &= -10\\ \sin\theta &= -\frac{1}{10}\\ \theta &= \left(-\arcsin\frac{1}{10}\right), \left(-180^\circ + \arcsin\frac{1}{10}\right)\\ \theta &= -5.7^\circ, -174.3^\circ\end{aligned}$$

6. For $0^\circ \leq \theta \leq 360^\circ$:

(a)

$$\begin{aligned}2\sin^2\theta - \sin\theta &= 0\\ \sin\theta(2\sin\theta - 1) &= 0\end{aligned}$$

$$\therefore \quad \textit{either:} \quad \begin{aligned}\sin\theta &= 0\\ \theta &= 0^\circ, 180^\circ, 360^\circ\end{aligned}$$

$$\textit{or:} \quad \begin{aligned}2\sin\theta &= 1\\ \sin\theta &= \frac{1}{2}\\ \theta &= \left(\arcsin\frac{1}{2}\right), \left(180^\circ - \arcsin\frac{1}{2}\right)\\ \theta &= 30^\circ, 150^\circ\end{aligned}$$

$$\therefore \quad \theta = 0^\circ, 30^\circ, 150^\circ, 180^\circ, 360^\circ$$

(b)

$$\begin{aligned} 3\cos^2\theta &= \cos\theta \\ 3\cos^2\theta - \cos\theta &= 0 \\ \cos\theta(3\cos\theta - 1) &= 0 \end{aligned}$$

$$\begin{aligned} \therefore \quad \textit{either:} \quad \cos\theta &= 0 \\ \theta &= 90^\circ, 270^\circ \end{aligned}$$

$$\begin{aligned} \textit{or:} \quad 3\cos\theta &= 1 \\ \cos\theta &= \frac{1}{3} \\ \theta &= \left(\arccos\frac{1}{3}\right), \left(360^\circ - \arccos\frac{1}{3}\right) \\ \theta &= 70.5^\circ, 289.5^\circ \end{aligned}$$

$$\therefore \quad \theta = 70.5^\circ, 90^\circ, 270^\circ, 289.5^\circ$$

(c)

$$\begin{aligned} 5\sin\theta\cos\theta - \sin\theta &= 0 \\ \sin\theta(5\cos\theta - 1) &= 0 \end{aligned}$$

$$\begin{aligned} \therefore \quad \textit{either:} \quad \sin\theta &= 0 \\ \theta &= 0^\circ, 180^\circ, 360^\circ \end{aligned}$$

$$\begin{aligned} \textit{or:} \quad \cos\theta &= \frac{1}{5} \\ \theta &= \left(\arccos\frac{1}{5}\right), \left(360^\circ - \arccos\frac{1}{5}\right) \\ \theta &= 78.5^\circ, 281.5^\circ \end{aligned}$$

$$\therefore \quad \theta = 0^\circ, 78.5^\circ, 180^\circ, 281.5^\circ, 360^\circ$$

(d)

$$\begin{aligned} \tan^2\theta + 4\tan\theta &= 0 \\ \tan\theta(\tan\theta + 4) &= 0 \end{aligned}$$

$$\begin{aligned} \therefore \quad \textit{either:} \quad \tan\theta &= 0 \\ \theta &= 0^\circ, 180^\circ, 360^\circ \end{aligned}$$

$$\begin{aligned} \textit{or:} \quad \tan\theta &= -4 \\ \theta &= (180^\circ - \arctan 4), (360^\circ - \arctan 4) \\ \theta &= 104.0^\circ, 284.0^\circ \end{aligned}$$

$$\therefore \quad \theta = 0^\circ, 104.0^\circ, 180^\circ, 284.0^\circ, 360^\circ$$

(e)

$$6\sin^2\theta - 5\sin\theta + 1 = 0$$
$$(3\sin\theta - 1)(2\sin\theta - 1) = 0$$

$$\therefore \quad \textit{either:} \quad \sin\theta = \frac{1}{3}$$
$$\theta = \left(\arcsin\frac{1}{3}\right), \left(180^\circ - \arcsin\frac{1}{3}\right)$$
$$\theta = 19.5^\circ, 160.5^\circ$$

$$\textit{or:} \quad \sin\theta = \frac{1}{2}$$
$$\theta = \left(\arcsin\frac{1}{2}\right), \left(180^\circ - \arcsin\frac{1}{2}\right)$$
$$\theta = 30^\circ, 150^\circ$$

$$\therefore \quad \theta = 30^\circ, 150^\circ, 19.5^\circ, 160.5^\circ$$

(f)

$$\cot^2\theta - 3\cot\theta + 2 = 0$$
$$(\cot\theta - 1)(\cot\theta - 2) = 0$$

$$\therefore \quad \textit{either:} \quad \cot\theta = 1$$
$$\tan\theta = 1$$
$$\theta = (\arctan 1), (180^\circ + \arctan 1)$$
$$\theta = 45^\circ, 225^\circ$$

$$\textit{or:} \quad \cot\theta = 2$$
$$\tan\theta = \frac{1}{2}$$
$$\theta = \left(\arctan\frac{1}{2}\right), \left(180^\circ + \arctan\frac{1}{2}\right)$$
$$\theta = 26.6^\circ, 206.6^\circ$$

$$\therefore \quad \theta = 26.6^\circ, 45^\circ, 206.6^\circ, 225^\circ$$

(g)

$$\begin{aligned}
\sec^2\theta + 4\sec\theta - 5 &= 0 \\
(\sec\theta + 5)(\sec\theta - 1) &= 0 \\
\therefore \quad \textit{either:} \quad \sec\theta &= 1 \\
\cos\theta &= 1 \\
\theta &= 0^\circ, 360^\circ \\
\textit{or:} \quad \sec\theta &= -5 \\
\cos\theta &= -\frac{1}{5} \\
\theta &= \left(180^\circ - \arccos\frac{1}{5}\right), \left(180^\circ + \arccos\frac{1}{5}\right) \\
\theta &= 101.5^\circ, 258.5^\circ \\
\therefore \quad \theta &= 0^\circ, 101.5^\circ, 258.5^\circ, 360^\circ
\end{aligned}$$

(h)

$$2\cot^2\theta - 7\cot\theta + 6 = 0$$

By applying the quadratic formula[3] with respect to $\cot\theta$:

$$\begin{aligned}
\cot\theta &= \frac{-(-7) \pm \sqrt{(-7)^2 - 4 \times 2 \times 6}}{2 \times 2} \\
&= \frac{7 \pm 1}{4} \\
&= 2, \left(\frac{3}{2}\right) \\
\tan\theta &= \left(\frac{1}{2}\right), \left(\frac{2}{3}\right) \\
\theta &= \left[\left(\arctan\frac{1}{2}\right), \left(180^\circ + \arctan\frac{1}{2}\right)\right], \left[\left(\arctan\frac{2}{3}\right), \left(180^\circ + \arctan\frac{2}{3}\right)\right] \\
\theta &= 26.6^\circ, 206.6^\circ, 33.7^\circ, 213.7^\circ
\end{aligned}$$

[3] I realise now that this could have been factorised very easily.

(i)

$$3\cos\theta + 4\sec\theta = 8$$

Multiply by $\cos\theta$:

$$3\cos^2\theta - 8\cos\theta + 4 = 0$$

Apply quadratic formula:

$$\begin{aligned}\cos\theta &= \frac{-(-8) \pm \sqrt{(-8)^2 - 4\times 3\times 4}}{2\times 3}\\ &= \frac{8\pm 4}{6}\\ &= \left(\frac{2}{3}\right), \left(\frac{5}{3}\right) \qquad \text{N.B. } \sin\theta \text{ cannot be } \frac{5}{3} \text{ as } \frac{5}{3} > 1.\\ \theta &= \left(\arccos\frac{2}{3}\right), \left(360^\circ - \arccos\frac{2}{3}\right)\\ \theta &= 48.2^\circ, 311.8^\circ\end{aligned}$$

(j)

$$4\sin\theta + 1 = 3\csc\theta$$

Multiply by $\sin\theta$:

$$\begin{aligned}4\sin^2\theta + \sin\theta - 3 &= 0\\ (\sin\theta + 1)(4\sin\theta - 3) &= 0\end{aligned}$$

$$\therefore \quad \textit{either:} \quad \sin\theta = -1$$
$$\theta = 270^\circ$$

$$\textit{or:} \quad \sin\theta = \frac{3}{4}$$
$$\theta = \left(\arcsin\frac{3}{4}\right), \left(180^\circ - \arcsin\frac{3}{4}\right)$$
$$\theta = 48.6^\circ, 131.4^\circ$$

$$\therefore \quad \theta = 48.6^\circ, 131.4^\circ, 270^\circ$$

(k)

$$3\sec\theta + 11 = 4\cos\theta$$

Multiply by $\cos\theta$:

$$\begin{aligned}4\cos^2\theta - 11\cos\theta - 3 &= 0\\ (\cos\theta - 3)(4\cos\theta + 1) &= 0\end{aligned}$$

$\cos\theta = 3$ is impossible because $3 > 1$, so:

$$\begin{aligned}\cos\theta &= -\frac{1}{4}\\ \theta &= \left(180^\circ - \arccos\frac{1}{4}\right), \left(180^\circ + \arccos\frac{1}{4}\right)\\ \theta &= 104.5^\circ, 255.5^\circ\end{aligned}$$

(l)

$$\begin{aligned}(\tan\theta+1)^2 &= 9\\ \tan^2\theta+2\tan\theta-8 &= 0\\ (\tan\theta-2)(\tan\theta+4) &= 0\end{aligned}$$

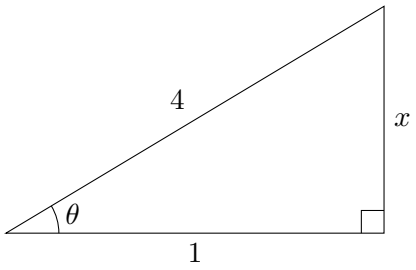
$$\begin{aligned}\therefore \quad \textit{either:} \quad \tan\theta &= 2\\ \theta &= (\arctan 2), (180^\circ+\arctan 2)\\ \theta &= 63.4^\circ, 243.4^\circ\end{aligned}$$

$$\begin{aligned}\textit{or:} \quad \tan\theta &= -4\\ \theta &= (180^\circ-\arctan 4), (360^\circ-\arctan 4)\\ \theta &= 104.0^\circ, 284.0^\circ\end{aligned}$$

$$\therefore \quad \theta = 63.4^\circ, 104.0^\circ, 243.4^\circ, 284.0^\circ$$

2.2 Exercise 14B

4. As we know $\sin\theta = \frac{1}{4}$ and $\theta < 90^\circ$, this question can be answered using a right triangle, as shown.



Thus by Pythagoras,

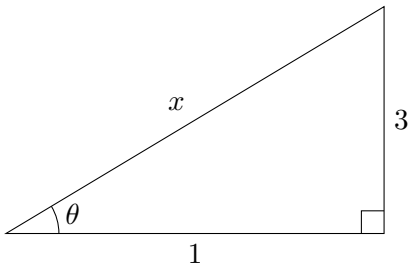
$$\begin{aligned}x^2 &= 4^2-1^2\\ x &= \sqrt{15}\end{aligned}$$

(a)
$$\sin\theta = \frac{\text{opp}}{\text{hyp}} = \frac{x}{4} = \frac{\sqrt{15}}{4}$$

(b)
$$\tan\theta = \frac{\text{opp}}{\text{adj}} = \frac{x}{1} = \sqrt{15}$$

(c)
$$\csc\theta = \frac{\text{hyp}}{\text{opp}} = \frac{4}{x} = \frac{4\sqrt{15}}{15}$$

5. As we know $\tan\theta = 3$ and $\theta < 90°$, this question can similarly be represented using the right triangle shown.



Thus by Pythagoras,

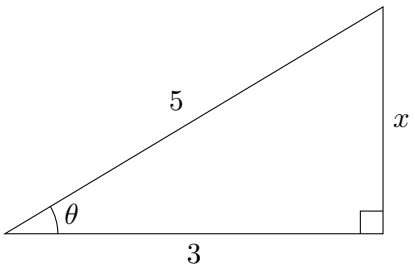
$$x^2 = 1^2 + 3^2$$
$$x = \sqrt{10}$$

(a)
$$\sin\theta = \frac{\text{opp}}{\text{hyp}} = \frac{3}{x} = \frac{3\sqrt{10}}{10}$$

(b)
$$\sec\theta = \frac{\text{hyp}}{\text{adj}} = \frac{x}{1} = \sqrt{10}$$

(c)
$$\csc\theta = \frac{\text{hyp}}{\text{opp}} = \frac{x}{3} = \frac{\sqrt{10}}{3}$$

6. Once again, as we know $\sec\theta = \frac{5}{3}$ and $\theta < 90°$, this question can similarly be represented using the right triangle shown.



Thus by Pythagoras,

$$x^2 = 5^2 - 3^2$$
$$x = 4$$

(a)
$$\sin\theta = \frac{\text{opp}}{\text{hyp}} = \frac{x}{5} = \frac{4}{5}$$

(b)
$$\tan\theta = \frac{\text{opp}}{\text{adj}} = \frac{x}{3} = \frac{4}{3}$$

(c)
$$\cot\theta = \frac{\text{adj}}{\text{opp}} = \frac{3}{x} = \frac{3}{4}$$

2.3 Exercise 15A

4. We are given:

$$\tan(A - B) = \frac{1}{2} \tag{1}$$
$$\tan(A) = 3 \tag{2}$$

We also know the trigonometric identity

$$\tan(\theta + \phi) \equiv \frac{\tan\theta + \tan\phi}{1 - \tan\theta\tan\phi}$$

Hence we can apply this to equation (1):

$$\frac{\tan A + \tan(-B)}{1 - \tan A\tan(-B)} = \frac{1}{2}$$

The tangent function is odd, so

$$\frac{\tan A - \tan B}{1 + \tan A\tan B} = \frac{1}{2}$$

Now we substitute in the value for $\tan A$ given by equation (2):

$$\begin{aligned}\frac{3 - \tan B}{1 + 3\tan B} &= \frac{1}{2}\\ 6 - 2\tan B &= 1 + 3\tan B\\ 5\tan B &= 5\\ \tan B &= 1\end{aligned}$$

5. We know that:

$$\tan(P + Q) = 5 \tag{1}$$
$$\tan Q = 2 \tag{2}$$

We can now apply the trig identity

$$\tan(\theta + \phi) \equiv \frac{\tan\theta + \tan\phi}{1 - \tan\theta\tan\phi},$$

to equation (1):

$$\frac{\tan P + \tan Q}{1 - \tan P\tan Q} = 5$$

We substitute in equation (2) and solve:

$$\begin{aligned}\frac{\tan P + 2}{1 - 2\tan P} &= 5\\ \tan P + 2 &= 5 - 10\tan P\\ 11\tan P &= 3\\ \tan P &= \frac{3}{11}\end{aligned}$$

6. We're given:

$$\tan(\theta - 45^\circ) = 4$$

Applying the trig identity used above,

$$\frac{\tan\theta - \tan 45^\circ}{1 + \tan\theta \tan 45^\circ} = 4$$

$\tan 45^\circ = 1$, so

$$\begin{aligned} \frac{\tan\theta - 1}{1 + \tan\theta} &= 4 \\ \tan\theta - 1 &= 4 + 4\tan\theta \\ 3\tan\theta &= -5 \\ \tan\theta &= -\frac{5}{3} \end{aligned}$$

23. All we know is that, since P, Q and R are angles in a triangle,

$$P + Q + R = 180^\circ$$

Let's try and manipulate this a bit until we have nice stuff with tan in it. First we move R to the right hand side:

$$P + Q = 180^\circ - R$$

Therefore

$$\tan(P + Q) = \tan(180^\circ - R$$

Now we apply the trigonometric identity $\tan(\theta + \phi) = \dfrac{\tan\theta + \tan\phi}{1 - \tan\theta\tan\phi}$

$$\frac{\tan P + \tan Q}{1 - \tan P \tan Q} = \frac{\tan 180^\circ + \tan(-R)}{1 - \tan 180^\circ \tan(-R)}$$

The tangent function is odd, so

$$\frac{\tan P + \tan Q}{1 - \tan P \tan Q} = \frac{\tan 180^\circ - \tan R}{1 + \tan 180^\circ \tan R}$$

We know $\tan 180^\circ = 0$, so

$$\frac{\tan P + \tan Q}{1 - \tan P \tan Q} = -\tan R$$

Expand and simplify

$$\begin{aligned} \tan P + \tan Q &= -\tan R(1 - \tan P \tan Q) \\ \tan P + \tan Q &= -\tan R + \tan P \tan Q \tan R \\ \tan P + \tan Q + \tan R &= \tan P \tan Q \tan R \end{aligned}$$

Hence

$$\frac{\tan P + \tan Q + \tan R}{\tan P \tan Q \tan R} \equiv 1 \qquad \square$$

24. We can split $15°$ into numbers we know about:

$$\tan 15° = \tan(45° - 30°)$$

Apply the trig identity used above

$$= \frac{\tan 45° + \tan(-30°)}{1 - \tan 45° \tan(-30°)}$$

The tangent function is odd, so

$$= \frac{\tan 45° - \tan 30°}{1 + \tan 45° \tan 30°}$$

Substitute in $\tan 45° = 1$ and $\tan 30° = \frac{\sqrt{3}}{3}$

$$= \frac{1 - \frac{\sqrt{3}}{3}}{1 + \frac{\sqrt{3}}{3}}$$

Multiply by the conjugate of the denominator

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

And now we expand and simplify

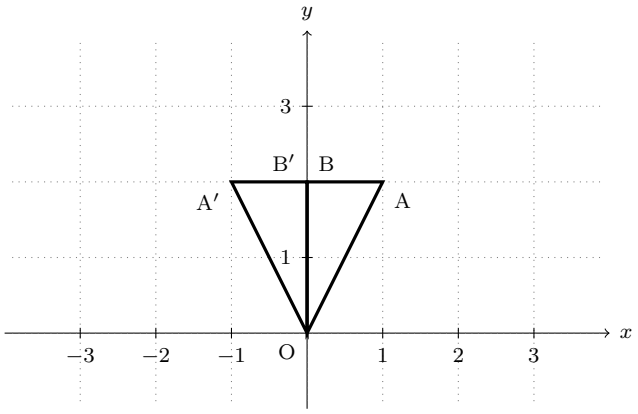
$$\begin{aligned} &= \frac{1 - \frac{2\sqrt{3}}{3} + \frac{1}{3}}{1 - \frac{1}{3}} \\ &= \frac{4 - 2\sqrt{3}}{3} \div \frac{2}{3} \\ &= \frac{4 - 2\sqrt{3}}{2} \\ \tan 15° &= 2 - \sqrt{3} \qquad \square \end{aligned}$$

3 Matrices (FP1)

3.1 Exercise 1B

4. (i) For the matrix $\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$,

(a) The following transformation occurs:

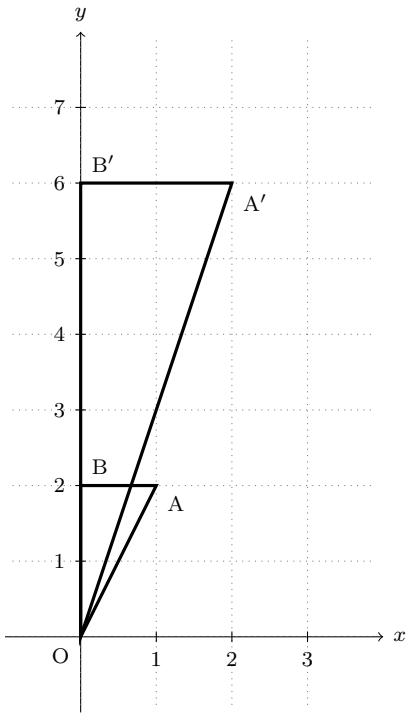


(b) A′ is at $(-1, 2)$ and B′ is at $(0, 2)$.

(c) The effect of the transformation is a reflection in the y-axis.

(ii) For the matrix $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$,

(a) The following transformation occurs:

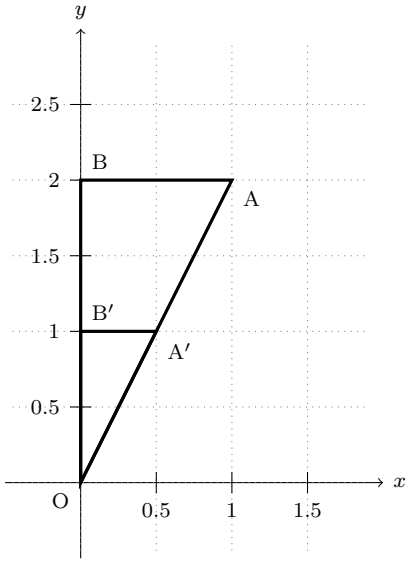


(b) A′ is at $(2, 6)$ and B′ is at $(0, 6)$.

(c) The effect of the transformation is a two-way stretch, with scale factor 2 horizontally and 3 vertically.

(iii) For the matrix $\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$,

(a) The following transformation occurs:

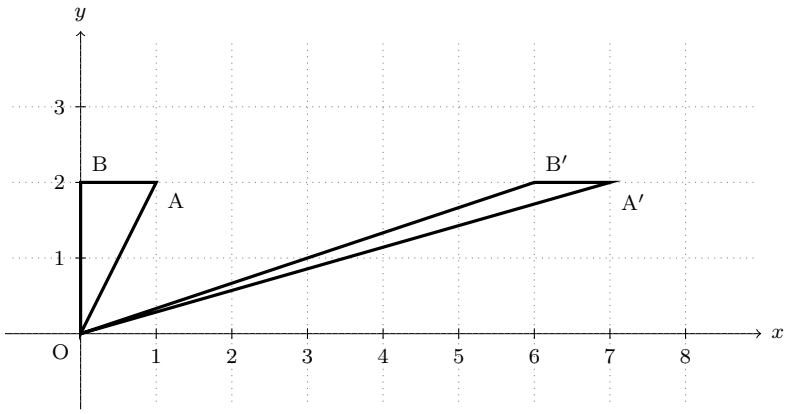


(b) A′ is at $(\frac{1}{2}, 1)$ and B′ is at $(0, 1)$.

(c) The effect of the transformation is a two-way stretch, with scale factor $\frac{1}{2}$ in both directions.

(iv) For the matrix $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$,

(a) The following transformation occurs:

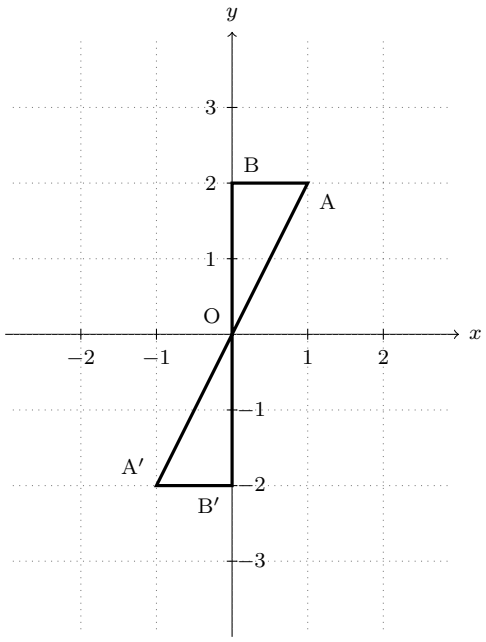


(b) A′ is at $(7, 2)$ and B′ is at $(6, 2)$.

(c) The effect of the transformation is a shear parallel to the x-axis.

(v) For the matrix $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$,

(a) The following transformation occurs:

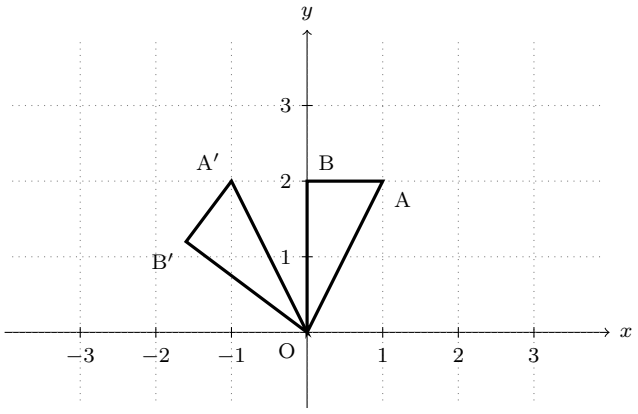


(b) A′ is at $(-1, -2)$ and B′ is at $(0, -2)$.

(c) The effect of the transformation is a rotation of 180° around the origin.

(vi) For the matrix $\begin{pmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{pmatrix}$,

(a) The following transformation occurs:



(b) A′ is at $(-1, 2)$ and B′ is at $(-1.6, 1.2)$.

(c) The effect of the transformation is a rotation of $\left[90° - \arctan\left(\frac{1.2}{1.6}\right)\right] = 53.1°$ anticlockwise around the origin.

5.
$$\begin{pmatrix} 4 & 3 \\ 5 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4x+3y \\ 5x+4y \end{pmatrix}$$

Hence:

$$\begin{aligned} \mathrm{A}' &= (1,0)' = (4,5) \\ \mathrm{B}' &= (1,1)' = (7,9) \\ \mathrm{C}' &= (0,1)' = (3,4) \end{aligned}$$

Using the shoelace formula[4] we can find the area of the transformed shape:

$$\begin{aligned} \text{Area} &= \frac{(x_1y_2 + x_2y_3 + x_3y_4 + x_4y_1) - (y_1x_2 + y_2x_3 + y_3x_4 + y_4x_1)}{2} \\ &= \frac{(0\times 5 + 4\times 9 + 7\times 4 + 3\times 0) - (0\times 4 + 5\times 7 + 9\times 3 + 4\times 0)}{2} \\ &= \frac{4\times 9 + 7\times 4 - 5\times 7 - 9\times 3}{2} \\ &= \frac{36 + 28 - 35 - 27}{2} \\ &= \frac{2}{2} \\ &= 1 \end{aligned}$$

Hence the new quadrilateral has the same area as the unit square.

3.2 Exercise 1C

4. (i) The coordinates of the five points of the flag are as follows:

$$\begin{gathered} (1,0) \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{gathered}$$

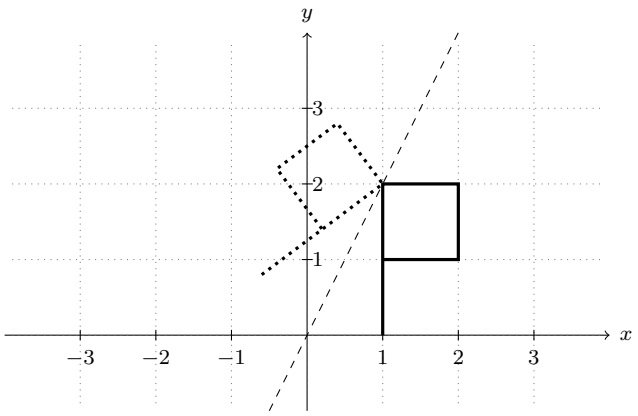
The flag can thus be described by matrix $\begin{matrix} x \\ y \end{matrix}\begin{pmatrix} 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 1 & 2 \end{pmatrix}$.

(ii) The image of the flag can be descibed as:

$$\begin{pmatrix} -0,6 & 0.8 \\ 0.8 & 0.6 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 1 & 2 \end{pmatrix} = \begin{pmatrix} -0.6 & 0.2 & 1 & -0.4 & 0.4 \\ 0.8 & 1.4 & 2 & 2.2 & 2.8 \end{pmatrix}$$

[4] Which apparently exists? Seems a lot easier than trying to draw the shape and split it into triangles and use trig etc.

(iii) The transformation is shown below:



Thus, the transformation is a reflection in the line $y = 2x$.

5. (i) The coordinates of the four points of the rectangle are as follows:

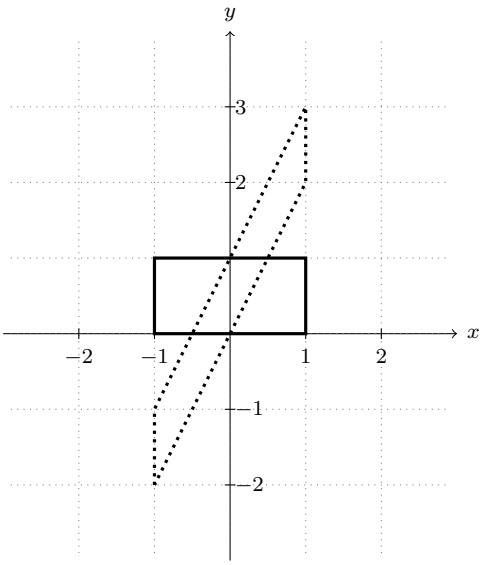
$$(-1, 0)$$
$$(-1, 1)$$
$$(1, 1)$$
$$(1, 0)$$

The flag can thus be described by matrix $\begin{matrix} x \\ y \end{matrix}\begin{pmatrix} -1 & -1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$.

(ii) The image of the flag can be descibed as:

$$\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} -1 & -1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & -1 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

(iii) The transformation is shown below:



Thus, the transformation is a shear parallel to the y-axis.

3.3 Exercise 1D

5. (i)
$$\mathbf{RS} = \begin{pmatrix} 2 & -1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 3 & 0 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 8 & -4 \\ -3 & 12 \end{pmatrix}$$

(ii)
$$\begin{pmatrix} 8 & -4 \\ -3 & 12 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} = \begin{pmatrix} 20 \\ -18 \end{pmatrix}$$

So the image of point $(2, -1)$ under the transformation **RS** is $(20, -18)$.

6. (i) $\mathbf{R}_1\mathbf{R}_2 = \mathbf{R}_2\mathbf{R}_1$ because a rotation of 25° anticlockwise followed by one of 40° anticlockwise has the same effect as a rotation of 40° anticlockwise followed by one of 25° anticlockwise.

(ii)
$$\mathbf{R}_1 = \begin{pmatrix} \cos 25^\circ & -\sin 25^\circ \\ \sin 25^\circ & \cos 25^\circ \end{pmatrix}$$
$$\mathbf{R}_2 = \begin{pmatrix} \cos 40^\circ & -\sin 40^\circ \\ \sin 40^\circ & \cos 40^\circ \end{pmatrix}$$
$$\therefore \quad \mathbf{R}_1\mathbf{R}_2 = \begin{pmatrix} \cos 25^\circ & -\sin 25^\circ \\ \sin 25^\circ & \cos 25^\circ \end{pmatrix} \begin{pmatrix} \cos 40^\circ & -\sin 40^\circ \\ \sin 40^\circ & \cos 40^\circ \end{pmatrix}$$
$$= \begin{pmatrix} \cos 25^\circ \cos 40^\circ - \sin 25^\circ \sin 40^\circ & -\cos 25^\circ \sin 40^\circ - \sin 25^\circ \cos 40^\circ \\ \cos 25^\circ \sin 40^\circ + \sin 25^\circ \cos 40^\circ & \cos 25^\circ \cos 40^\circ - \sin 25^\circ \sin 40^\circ \end{pmatrix}$$
$$= \begin{pmatrix} \cos(25^\circ + 40^\circ) & -\sin(25^\circ + 40^\circ) \\ \sin(25^\circ + 40^\circ) & \cos(25^\circ + 40^\circ) \end{pmatrix}$$
$$= \begin{pmatrix} \cos 65^\circ & -\sin 65^\circ \\ \sin 65^\circ & \cos 65^\circ \end{pmatrix}$$

(iii) $\mathbf{R}_1\mathbf{R}_2$ represents a single rotation of 65° anticlockwise around the origin.

8. As given, the matrix representing reflection in the line $y = mx$ is
$$\frac{1}{1+m^2} \begin{pmatrix} 1-m^2 & 2m \\ 2m & m^2-1 \end{pmatrix}$$
.

(i) Where **P** represents reflection in the line $y = \frac{1}{\sqrt{3}}x$ and **Q** represents reflection in the

line $y = \sqrt{3}x$:

$$\mathbf{P} = \frac{1}{1+\left(\frac{1}{\sqrt{3}}\right)^2}\begin{pmatrix} 1-\left(\frac{1}{\sqrt{3}}\right)^2 & \frac{2}{\sqrt{3}} \\ \frac{2}{\sqrt{3}} & \left(\frac{1}{\sqrt{3}}\right)^2 - 1 \end{pmatrix}$$
$$= \frac{3}{4}\begin{pmatrix} \frac{2}{3} & \frac{2\sqrt{3}}{3} \\ \frac{2\sqrt{3}}{3} & -\frac{2}{3} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$
$$\mathbf{Q} = \frac{1}{1+\sqrt{3}^2}\begin{pmatrix} 1-\sqrt{3}^2 & 2\sqrt{3} \\ 2\sqrt{3} & \sqrt{3}^2 - 1 \end{pmatrix}$$
$$= \frac{1}{4}\begin{pmatrix} -2 & 2\sqrt{3} \\ 2\sqrt{3} & 2 \end{pmatrix}$$
$$= \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

(ii) A reflection in the line $y = \frac{1}{\sqrt{3}}x$ followed by a reflection in the line $y = \sqrt{3}x$:

$$\mathbf{QP} = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

We know $\arccos\left(\frac{1}{2}\right) = 60°$ and $\arcsin\left(\frac{\sqrt{3}}{2}\right) = 60°$, so this transformation is equivalent to a rotation of 60° anticlockwise around the origin.

(iii) A reflection in the line $y = \sqrt{3}x$ followed by a reflection in the line $y = \frac{1}{\sqrt{3}}x$:

$$\mathbf{PQ} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{pmatrix}\begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$
$$= \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

We know $\arccos\left(\frac{1}{2}\right) = 60°$ and $\arcsin\left(\frac{\sqrt{3}}{2}\right) = 60°$, so this transformation is equivalent to a rotation of 60° clockwise around the origin.

9. (i) If matrix $\mathbf{R}$ represents a single rotation of 30° around the origin,

$$\mathbf{R} = \begin{pmatrix} \cos 30° & -\sin 30° \\ \sin 30° & \cos 30° \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}$$

(ii) As shown in the previous question, if matrix $\mathbf{M}$ represents a reflection in the line $y = \sqrt{3}x$,

$$\mathbf{M} = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}$$

(iii)

$$\mathbf{MR} = \begin{pmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

This represents a reflection inl the line $y = x$.

4 Complex numbers (FP1)

4.1 Exercise 2A

3. Given $z = 2 + 3j, w = 6 - 4j$:

(i)
$$\Re(z) = 2$$

(ii)
$$\Im(w) = -4$$

(iii)
$$z^\star = 2 - 3j$$

(iv)
$$w^\star = 6 + 4j$$

(v)
$$z^\star + w^\star = (2 - 3j) + (6 + 4j) = 8 + j$$

(vi)

$$\begin{aligned} z^\star - w^\star &= (2-3j) - (6+4j) \\ &= -4 - 7j \end{aligned}$$

(vii)

$$\Im(z + z^\star) = 0$$

(viii)

$$\Re(w - w^\star) = 0$$

(ix)

$$\begin{aligned} zz^\star - ww^\star &= (2+3j)(2-3j) - (6-4j)(6+4j) \\ &= 13 - 52 \\ &= -39 \end{aligned}$$

(x)

$$\begin{aligned} (z^3)^\star &= (-46+9j)^\star \\ &= -46 - 9j \end{aligned}$$

(xi)

$$\begin{aligned} (z^\star)^3 &= (2-3j)^3 \\ &= -46 - 9j \end{aligned}$$

(xii)

$$\begin{aligned} zw^\star - z^\star w &= (2+3j)(6+4j) - (2-3j)(6-4j) \\ &= 26j - (-26)j \\ &= 52j \end{aligned}$$

4.2 Exercise 2B

1. (i)

$$\frac{1}{3+j} = \frac{3-j}{(3+j)(3-j)} = \frac{3}{10} - \frac{1}{10}j$$

(ii)

$$\frac{1}{6-j} = \frac{6+j}{(6-j)(6+j)} = \frac{6}{37} + \frac{1}{37}j$$

(iii)

$$\frac{5j}{6-2j} = \frac{5j(6+2j)}{(6-2j)(6+2j)} = \frac{-10+30j}{40} = -\frac{1}{4} + \frac{3}{4}j$$

(iv)

$$\frac{7+5j}{6-2j} = \frac{(7+5j)(6+2j)}{(6-2j)(6+2j)} = \frac{32+44j}{40} = \frac{4}{5} + \frac{11}{10}j$$

(v)

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

(vi)
$$\frac{47-23j}{6+j} = \frac{(47-23j)(6-j)}{(6+j)(6-j)} = \frac{259-185j}{37} = 7-5j$$

(vii)
$$\frac{2-3j}{3+2j} = \frac{(2-3j)(3-2j)}{(3+2j)(3-2j)} = \frac{-13i}{13} = -j$$

(viii)
$$\frac{5-3j}{4+3j} = \frac{(5-3j)(4-3j)}{(4+3j)(4-3j)} = \frac{11-27j}{25} = \frac{11}{25} - \frac{27}{25}j$$

(ix)
$$\frac{6+j}{2-5j} = \frac{(6+j)(2+5j)}{(2-5j)(2+5j)} = \frac{7+32j}{29} = \frac{7}{29} + \frac{32}{29}j$$

(x)
$$\frac{12-8j}{(2+2j)^2} = \frac{12-8j}{8j} = \frac{j(12-8j)}{-8} = \frac{-8-12j}{8} = -1 - \frac{3}{2}j$$

2. (i) We must find $a, b \in \mathbb{R}$ such that
$$(a+bj)^2 = 21+20j \tag{1}$$
We can expand the left hand side:
$$a^2 + 2abj + b^2j^2 = 21+20j$$
$$a^2 - b^2 + 2abj = 21+20j$$
Now set the real parts equal
$$a^2 - b^2 = 21 \tag{2}$$
and set the imaginary parts equal
$$2ab = 20 \tag{3}$$
We can divide (3) by $2a$ to get $b = \frac{20}{2a} = \frac{10}{a}$, and now substitute this into (2):
$$a^2 - \left(\frac{10}{a}\right)^2 = 21 \tag{4}$$
And now solve (4) for a:
$$a^2 - \frac{100}{a^2} = 21$$
Multiply by a^2
$$a^4 - 21a^2 - 100 = 0$$
Use the quadratic formula with regards to a^2
$$\begin{aligned} a^2 &= \frac{-(-21) \pm \sqrt{(-21)^2 - 4\times 1 \times -100}}{2\times 1} \\ &= \frac{21 \pm 29}{2} \\ &= 25, -4 \\ a &= \sqrt{25} \quad \text{because } a \in \mathbb{R}, a > 0. \\ &= 5 \end{aligned}$$

Now we substitute this back into (3):

$$b = \frac{10}{5} = 2$$

$$\therefore a = 5, b = 2$$

We can check this by substituting these values back into (1):

$$\begin{aligned} \text{LHS} &= (a + bj)^2 \\ &= (5 + 2j)(5 + 2j) \\ &= 25 + 20j - 4 \\ &= 21 + 20j \end{aligned}$$

$$\text{RHS} = 21 + 20j$$

$$\therefore \quad \text{LHS} = \text{RHS}$$

Thus our solutions are correct.

4. (i)

$$\begin{aligned} (1 + j)z &= 3 + j \\ z &= \frac{3 + j}{1 + j} \\ &= \frac{(3 + j)(1 - j)}{(1 + j)(1 - j)} \\ &= \frac{4 - 2j}{2} \\ &= 2 - j \end{aligned}$$

(ii)

$$\begin{aligned} (3 - 4j)(z - 1) &= 10 - 5j \\ z - 1 &= \frac{10 - 5j}{3 - 4j} \\ &= \frac{(10 - 5j)(3 + 4j)}{(3 - 4j)(3 + 4j)} \\ &= \frac{50 + 25j}{25} \\ &= 2 + j \\ z &= 3 + j \end{aligned}$$

(iii)

$$\begin{aligned} (2 + j)(z - 7 + 3j) &= 15 - 10j \\ z - 7 + 3j &= \frac{15 - 10j}{2 + j} \\ &= \frac{(15 - 10j)(2 - j)}{(2 + j)(2 - j)} \\ &= \frac{20 - 35j}{5} \\ &= 4 - 7j \\ z &= 11 - 10j \end{aligned}$$

(iv)

$$\begin{aligned}(3+5j)(z+2-5j) &= 6+3j \\ z+2-5j &= \frac{6+3j}{3+5j} \\ &= \frac{(6+3j)(3-5j)}{(3+5j)(3-5j)} \\ &= \frac{33-21j}{34} \\ z &= \frac{33-21j}{34} - 2 + 5j \\ &= \frac{33-21j+34(-2+5j)}{34} \\ &= \frac{-35+149j}{34}\end{aligned}$$

5.

Find all the complex numbers z for which $z^2 = 2z^\star$.

Let $z = a + bj$ where $z \in \mathbb{C}$ and $a, b \in \mathbb{R}$. Hence we want to solve:

$$(a+bj)^2 = 2(a+bj)^\star \tag{1}$$

Now we expand it out:

$$a^2 - b^2 + 2abj = 2a - 2bj$$

Set the real parts equal

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

And set the imaginary parts equal

$$2ab = -2b \tag{3}$$

We simplify (3) to get $a = \dfrac{-2b}{2b} = -1$ with $b \neq 0$, and substitute this into (2):

$$\begin{aligned}(-1)^2 - b^2 &= 2(-1) \\ 1 - b^2 &= -2 \\ b^2 &= 3 \\ b &= \pm\sqrt{3}\end{aligned}$$

If $b = 0$, that is if $z \in \mathbb{R}$, equation (2) tells us that $a^2 = 2a \Rightarrow a = 0, 2$.

Hence our solutions to (1) are $z = (-1+3j), (-1-3j), 0, 2$.

12. Let $z = a + bj$ and $w = c + dj$ where $z, w \in \mathbb{C}$ and $a, b, c, d \in \mathbb{R}$. We're given:

$$z + jw = 13 \tag{1}$$
$$3z - 4w = 2j \tag{2}$$

First we expand out equation (1):

$$\begin{aligned}(a+bj) + j(c+dj) &= 13 \\ a + bj + cj - d &= 13\end{aligned}$$

And we set the real and imaginary parts equal respectively

$$a - d = 13 \tag{3}$$
$$b + c = 0 \tag{4}$$

Now we repeat this process for equation (2):

$$3(a + bj) - 4(c + dj) = 2j$$
$$3a + 3bj - 4c - 4dj = 2j$$

Setting the real and imaginary parts equal respectively

$$3a - 4c = 0 \tag{5}$$
$$3b - 4d = 2 \tag{6}$$

Thus we can simultaneously solve equations (3), (4), (5), and (6) for variables a, b, c and d.
These can be represented as a matrix equation:

$$\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 3 & 0 & -4 & 0 \\ 0 & 3 & 0 & -4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 13 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

Now we multiply by the inverse

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 3 & 0 & -4 & 0 \\ 0 & 3 & 0 & -4 \end{pmatrix}^{-1} \begin{pmatrix} 13 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

Find the inverse[5] as $M^{-1} = \dfrac{1}{\det(M)} \operatorname{adj}(M)$:

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \frac{1}{25} \begin{pmatrix} 16 & 12 & 3 & -4 \\ -12 & 16 & 4 & 3 \\ 12 & 9 & -4 & -3 \\ -9 & 12 & 3 & -4 \end{pmatrix} \begin{pmatrix} 13 \\ 0 \\ 0 \\ 2 \end{pmatrix}$$

Evaluate the matrix product

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \frac{1}{25} \begin{pmatrix} 200 \\ -150 \\ 150 \\ -125 \end{pmatrix}$$
$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 8 \\ -6 \\ 6 \\ -5 \end{pmatrix}$$

Thus $a = 8$, $b = -6$, $c = 6$ and $d = -5$.
Therefore $z = 8 - 6j$ and $w = 6 - 5j$.

4.3 Exercise 2C

Note[6]

1. (i)

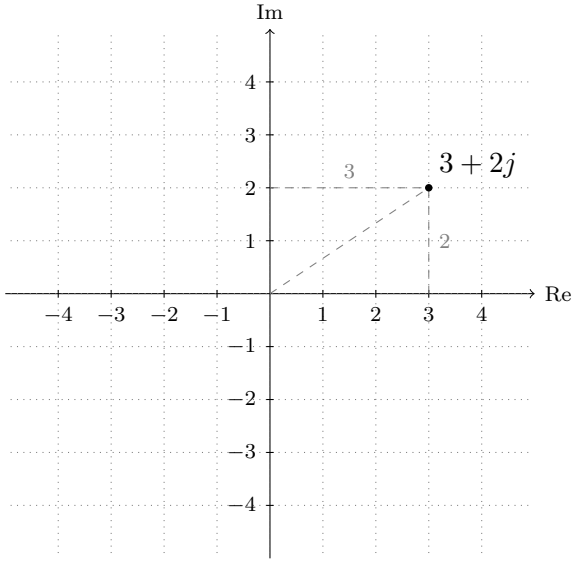


Figure 5: Argand diagram of $3 + 2j$.

$\therefore$ By Pythagoras, $|3 + 2j| = \sqrt{3^2 + 2^2} = \sqrt{13}$.

(ii)

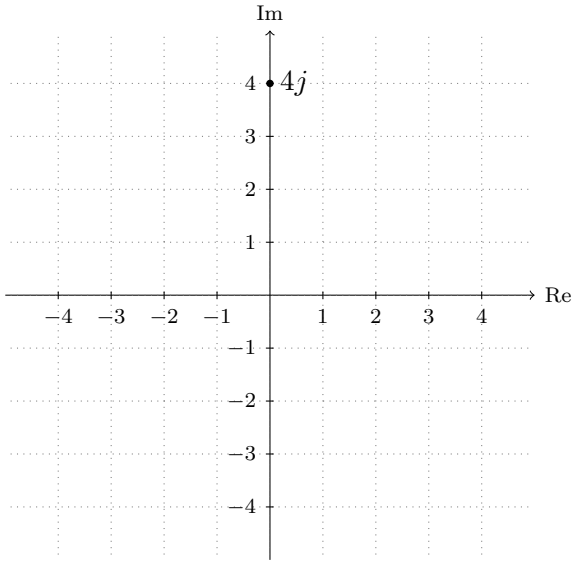


Figure 6: Argand diagram of $4j$.

$\therefore$ It can be seen that $|4j| = 4$.

[6]For question 1 here I didn't see 'single Argand diagram' in the question until afterwards... Anyway I've left them separate for this question.

(iii)

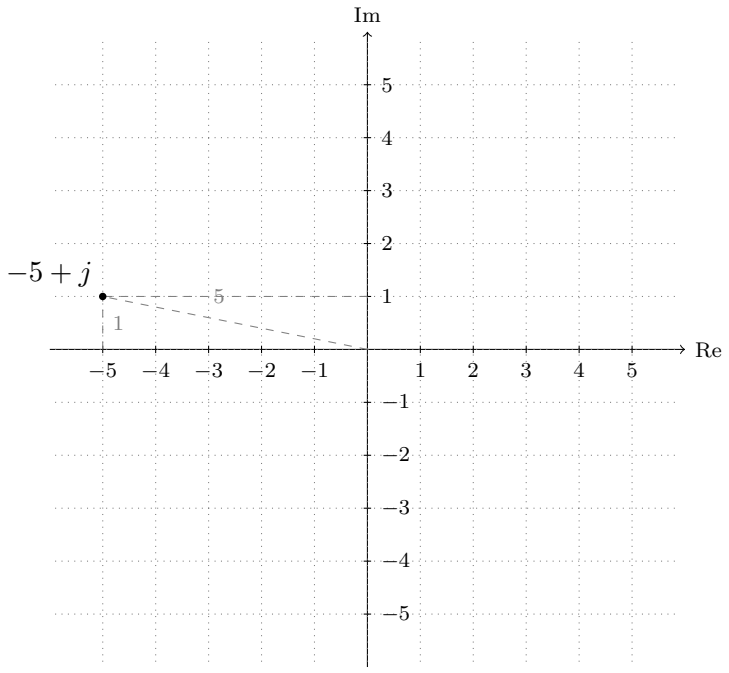


Figure 7: Argand diagram of $-5+j$.

$\therefore$ By Pythagoras, $|-5+j| = \sqrt{5^2+1^2} = \sqrt{26}$.

(iv)

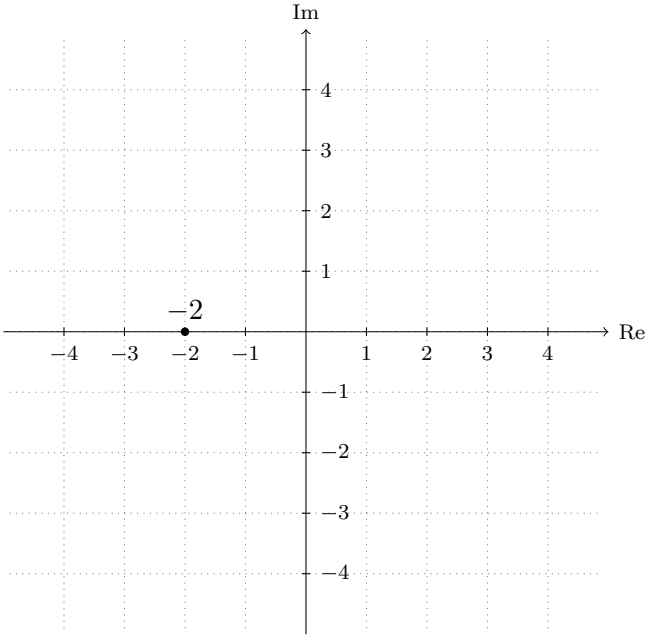


Figure 8: Argand diagram of -2.

$\therefore$ It can be seen that $|-2| = 2$.

(v)

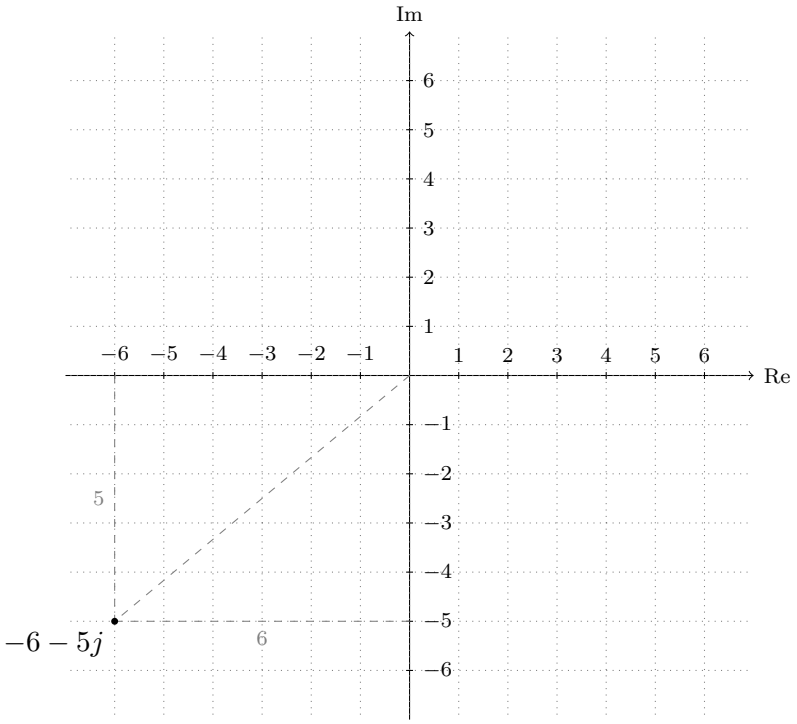


Figure 9: Argand diagram of $-6-5j$.

$\therefore$ By Pythagoras, $|-6-5j| = \sqrt{6^2+5^2} = \sqrt{61}$.

(vi)

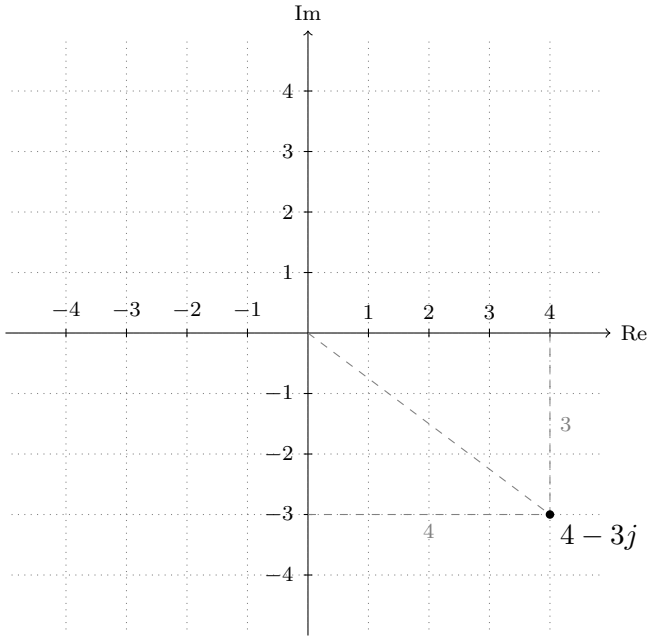


Figure 10: Argand diagram of $4-3j$.

$\therefore$ By Pythagoras, $|4-3j| = \sqrt{4^2+3^2} = \sqrt{25} = 5$.

2. Given $z = 2 - 4j$, the points (i) through (viii) are represented on the Argand diagram in figure 11.

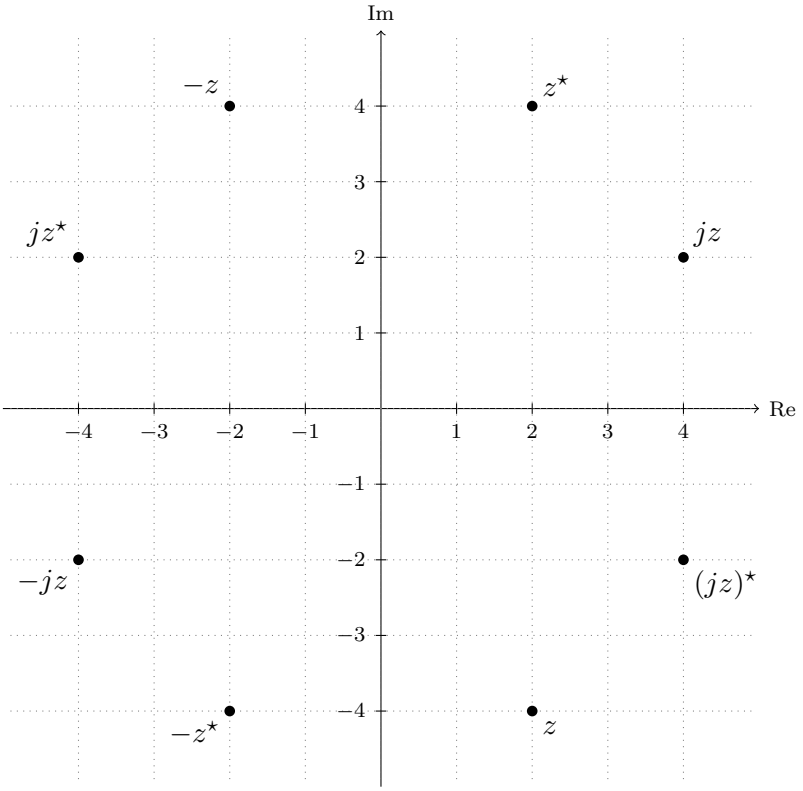


Figure 11: Argand diagram for question 2.

3. Given $z = 10 + 5j$ and $w = 1 + 2j$, the points (i) through (v) are represented on the Argand diagram in figure 12.

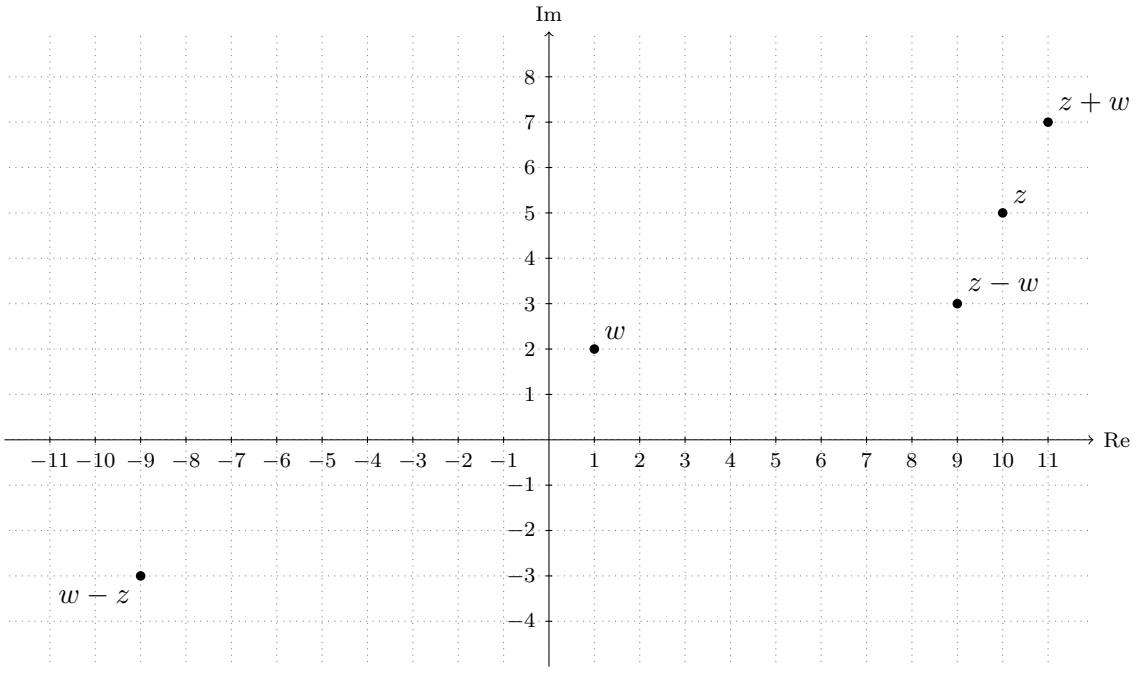


Figure 12: Argand diagram for question 3.

4. Given $z = 3 + 4j$ and $w = 5 - 12j$:

(i)
$$|z| = |3 + 4j| = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$$

(ii)
$$|w| = |5 - 12j| = \sqrt{5^2 + 12^2} = \sqrt{169} = 13$$

(iii)
$$|zw| = |(3 + 4j)(5 - 12j)| = |63 - 16j| = \sqrt{63^2 + 16^2} = \sqrt{4225} = 65$$

(iv)
$$\left|\frac{z}{w}\right| = \left|\frac{3 + 4j}{5 - 12j}\right| = \left|\frac{(3 + 4j)(5 + 12j)}{(5 - 12j)(5 + 12j)}\right| = \left|\frac{-33 + 56j}{169}\right| = \left|-\frac{33}{169} + \frac{56}{169}j\right|$$
$$= \sqrt{\left(\frac{33}{169}\right)^2 + \left(\frac{56}{169}\right)^2} = \sqrt{\frac{1089}{28561} + \frac{3136}{28561}} = \sqrt{\frac{25}{169}} = \frac{5}{13}$$

(v)
$$\left|\frac{w}{z}\right| = \left|\frac{5 - 12j}{3 + 4j}\right| = \left|\frac{(5 - 12j)(3 - 4j)}{(3 + 4j)(3 - 4j)}\right| = \left|\frac{-33 - 56j}{25}\right| = \left|-\frac{33}{25} - \frac{56}{25}j\right|$$
$$= \sqrt{\left(\frac{33}{25}\right)^2 + \left(\frac{56}{25}\right)^2} = \sqrt{\frac{1089}{625} + \frac{3136}{625}} = \sqrt{\frac{169}{25}} = \frac{13}{5}$$

It can be seen that $|zw| = |z||w|$, $\left|\frac{z}{w}\right| = \frac{|z|}{|w|}$ and $\left|\frac{w}{z}\right| = \frac{|w|}{|z|}$.

4.4 Exercise 2G

1.
$$z^3 - z^2 - 7z + 15 = 0 \tag{1}$$

If $2 + j$ is a root, then $z - (2 + j)$ must be a factor[7]:

$$[z - (2 + j)](az^2 + bz + c) = z^3 - z^2 - 7z + 15 \tag{2}$$
$$az^3 + bz^2 + cz - (2 + j)az^2 - (2 + j)bz - (2 + j)c = z^3 - z^2 - 7z + 15$$
$$az^3 + (b - (2 + j)a)z^2 + (c - (2 + j)b)z - (2 + j)c = z^3 - z^2 - 7z + 15$$
$$az^3 + (b - 2a - aj)z^2 + (c - 2b - bj)z - (2c + cj) = z^3 - z^2 - 7z + 15$$

where $a, b, c \in \mathbb{C}$. Comparing coefficients gives

$$a = 1 \tag{3}$$
$$b - 2a - aj = -1 \tag{4}$$
$$c - 2b - bj = -7 \tag{5}$$
$$-(2c + cj) = 15 \tag{6}$$

Equation (3) gives $a = 1$, equation (4) gives $b = -1+2a+aj = 1+j$, and equation (5) gives $c = -7 + 2b + bj = -7 + 2(1 + j) + (1 + j)j = -6 + 3j$. Hence,

$$[z - (2 + j)][(z^2 + (1 + j)z + (-6 + 3j)] = 0$$

We can solve the second factor here as a quadratic for the other roots:

$$\begin{aligned} z &= \frac{-(1+j) \pm \sqrt{(1+j)^2 - 4 \times 1 \times (-6+3j)}}{2 \times 1} \\ &= \frac{-1-j \pm \sqrt{24-10j}}{2} \\ &= -3, 2-j \end{aligned}$$

So our final roots to equation (1) are $z = 2 \pm j, -3$.

2. Let $f(z) = z^3 - 15z^2 + 76z - 140$. We are solving:

$$z^3 - 15z^2 + 76z - 140 = 0 \tag{1}$$

so $f(z) = 0$. Because we know one root of this equation is an integer, $(z + \alpha)$ must be a factor, where $\alpha \in \mathbb{Z}$. Thus, where $a, b, c \in \mathbb{C}$, we can express the following equality:

$$(z+\alpha)(az^2 + bz + c) = z^3 - 15z^2 + 76z - 140 \tag{2}$$

We know that α must be a factor of -140, namely one of
$\pm 1, \pm 2, \pm 4, \pm 5, \pm 7, \pm 10, \pm 14, \pm 20, \pm 28, \pm 35, \pm 70, \pm 140$.

The factor theorem states that if $f(-\alpha) = 0$ then $z + \alpha$ is a factor. $f(1) = -78$ and $f(-1) = -232$, so $(z \pm 1)$ are not factors. $f(2) = -40$ and $f(-2) = -360$ so $(z \pm 2)$ are not factors. $f(4) = -12)$ and $f(-4) = -748$, so $(z \pm 4)$ are not factors. $f(5) = -10$ and $f(-5) = -1020$, so $(z \pm 5)$ are not factore. $f(7) = 0$ so by the factor theorem, $(z - 7)$ is a factor. Hence:

$$(z-7)(az^2 + bz + c) = z^3 - 15z^2 + 76z - 140$$

We expand the LHS:

$$\begin{aligned} az^3 + bz^2 + cz - 7az^2 - 7bz - 7c &= z^3 - 15z^2 + 76z - 140 \\ az^3 + (b-7a)z^2 + (c-7b)z - 7c &= z^3 - 15z^2 + 76z - 140 \end{aligned}$$

By comparing coefficients, we get:

$$a = 1 \tag{3}$$
$$b - 7a = -15 \tag{4}$$
$$c - 7b = 76 \tag{5}$$
$$-7c = -140 \tag{6}$$

Equation (3) gives $a = 1$, equation (4) gives $b = 7a - 15 = -8$ and equation (5) gives $c = 7b + 76 = 20$. Hence we're solving the equation:

$$(z-7)(z^2 - 8z + 20) = 0$$

Thus one of our roots is $z = 7$ and the others can be found by solving the quadratic

$$\begin{aligned} z^2 - 8z + 20 &= 0 \\ z &= \frac{8 \pm \sqrt{(-8)^2 - 4 \times 1 \times 20}}{2 \times 1} \\ &= \frac{8 \pm 4j}{2} \\ &= 4 \pm 2j \end{aligned}$$

Hence the solutions to $f(z) = 0$ are $z = 7, 4 \pm 2j$.

3. We must solve
$$z^3 + pz^2 + qz + 12 = 0 \quad (1)$$
Where $p, q \in \mathbb{R}$. We're given that $1 - j$ is a root, so we know that one possible value of z is as follows:
$$\begin{aligned} z &= 1 - j \\ z^2 &= (1 - j)^2 = -2j \\ z^3 &= -2j(1 - j) = -2 - 2j \end{aligned}$$
We can substitute these into equation (1):
$$\begin{aligned} (-2 - 2j) + p(-2j) + q(1 - j) + 12 &= 0 \\ -2 - 2j - 2pj + q - qj + 12 &= 0 \\ (q + 10) + (-2 - 2p - q)j &= 0 \end{aligned}$$
Equating the real and imaginary parts gives
$$q + 10 = 0 \quad (2)$$
$$-2 - 2p - q = 0 \quad (3)$$
Equation (2) gives $q = -10$ and equation (3) gives $p = \dfrac{-q - 2}{2} = \dfrac{10 - 2}{2} = 4$. Substitute these back into equation (1):
$$z^3 + 4z^2 - 10z + 12 = 0 \quad (4)$$
Because $1 - j$ is a root and the coefficients are real, its conjugate $1 + j$ must be a root too. Thus $[z - (1 - j)]$ and $[z - (1 + j)]$ are factors, which means that $(z - 1 + j)(z - 1 - j)$ is a factor. We expand this factor:
$$\begin{aligned} (z - 1 + j)(z - 1 - j) &= [(z - 1) - j][(z - 1) + j] \\ &= (z - 1)^2 + 1 \\ &= z^2 - 2z + 2 \text{ is a factor.} \end{aligned}$$
The other factor can be found to be $(z + 6)$ by looking at the coefficient of z^3 and the constant term in equation (4). So:
$$(z + 6)(z^2 - 2z + 2) = 0$$
Hence the roots are $z = -6, 1 \pm j$.

4.
$$z^4 - 10z^3 + 42z^2 - 82z + 65 = 0 \quad (1)$$
We're given that $3 + 2j$ is a root, and so because the coefficients are all real, its conjugate $3 - 2j$ must also be a root.

Hence $[z - (3 + 2j)]$ and $[z - (3 - 2j)]$ are factors, meaning (z-3-2j)(z-3+2j) is a factor. Let's expand this factor:
$$\begin{aligned} (z - 3 - 2j)(z - 3 + 2j) &= [(z - 3) - 2j][(z - 3) + 2j] \\ &= (z - 3)^2 + 4 \\ &= z^2 - 6z + 13 \text{ is a factor.} \end{aligned}$$

Using a farmer's field[8] allows us to find the other factor to be $z^2 - 4z + 5$. Hence:

$$(z^2 - 6z + 13)(z^2 - 4z + 5) = 0$$

We already know roots $3 \pm 2j$, so our other two roots are the solutions of the quadratic equation $z^2 - 4z + 5 = 0$:

$$\begin{aligned} z &= \frac{4 \pm \sqrt{(-4)^2 - 4 \times 1 \times 5}}{2 \times 1} \\ &= \frac{4 \pm 2j}{2} \\ &= 2 \pm j \end{aligned}$$

Thus our final roots are $z = 3 \pm 2j, 2 \pm j$.

6. Given $w = 1 - j$:

   (i)

$$\begin{aligned} w^2 &= (1 - j)^2 = -2j \\ w^3 &= -2j(1 - j) = -2 - 2j \\ w^4 &= (-2j)^2 = -4 \end{aligned}$$

   (ii) We can substitute in the values of w we just found:

$$\begin{aligned} w^4 + 3w^3 + pw^2 + qw + 8 &= 0 \\ (-4) + 3(-2 - 2j) + p(-2j) + q(1 - j) + 8 &= 0 \\ -4 - 6 - 6j - 2pj + q - qj + 8 &= 0 \\ (q - 2) + (-2p - q - 6)j &= 0 \end{aligned}$$

   By equating the real and imaginary parts, we get

$$q - 2 = 0 \quad (1)$$
$$-2p - q - 6 = 0 \quad (2)$$

   Equation (1) gives $q = 2$ and equation (2) gives $p = -\frac{q + 6}{2} = -4$.

   (iii) We're given

$$z^4 + 3z^3 + pz^2 + qz + 8 = 0$$

   We observe that this is identical to the quartic equation with respect to w given in part (ii), and so we know that w is a root of z, namely that $1 - j$ is a root. Hence, because all of the coefficients are real, the conjugate $1 + j$ is a root too.

   Thus two roots of this quartic equation are $z = 1 \pm j$.

[8] Or a multiplication grid? What should I say in formal proofs?

5 Graph sketching (FP1)

5.1 Exercise 3A

Note[9]

2.

$$y = \frac{2}{(x-3)^2}$$

When $x = 0$, $y = \frac{2}{(-3)^2} = \frac{2}{9}$.

When $y = 0$, $\frac{2}{(x-3)^2} = 0$ so there are no x-intercepts.

There is a vertical asymptote at $x = 3$, because $\frac{2}{(3-3)^2} = \frac{2}{0}$ is undefined.

At this asymptote: as $x \to 3^-$, $y \to \infty$ and as $x \to 3^+$, $y \to \infty$.

As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^+$.

Hence:

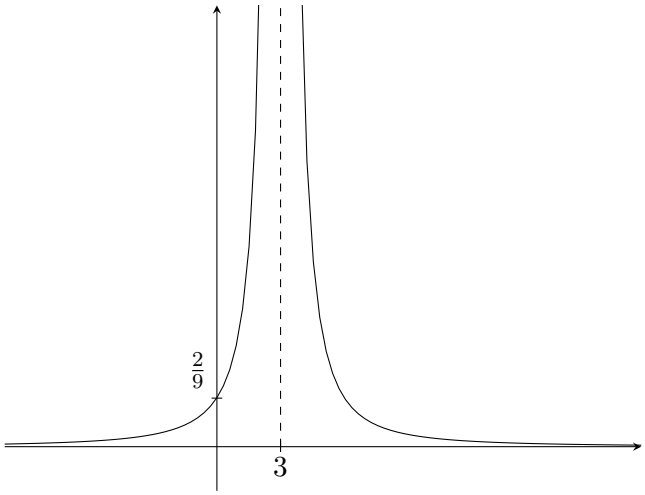


Figure 13: Graph of $y = \frac{2}{(x-3)^2}$.

4.

$$y = \frac{x}{x^2 - 4}$$

When $x = 0$, $y = \frac{0}{0^2 - 4} = 0$.

When $y = 0$, $\frac{x}{x^2 - 4} = 0$ so $x = 0$.

There are vertical asymptotes at $x = 2$ and $x = -2$, because $\frac{2}{2^2 - 4} = \frac{2}{0}$ is undefined, as is $\frac{-2}{(-2)^2 - 4)} = \frac{-2}{0}$.

[9] I realise it's completely counter-intuitive to 'sketch' a graph on a computer. I've worked all the points out and done it on paper first, and these weren't automatically generated.

At these asymptotes: as $x \to 2^-$, $y \to -\infty$ and as $x \to 2^+$, $y \to \infty$,
as $x \to (-2)^-$, $y \to -\infty$ and as $x \to (-2)^+$, $y \to \infty$.
As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^-$.
Hence:

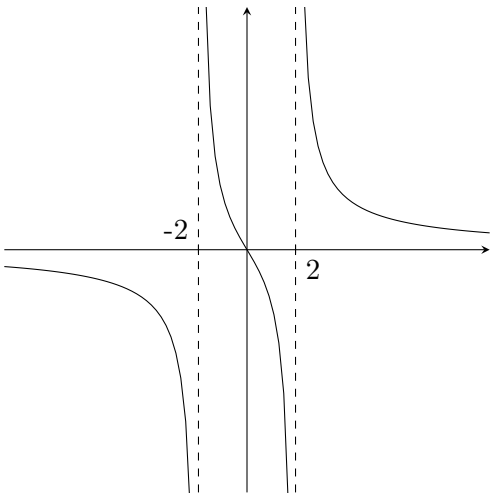


Figure 14: Graph of $y = \dfrac{x}{x^2 - 4}$.

6.

$$y = \frac{x - 5}{(x + 2)(x - 3)}$$

When $x = 0$, $y = \dfrac{-5}{2 \times -3} = \dfrac{5}{6}$.

When $y = 0$, $\dfrac{x - 5}{(x + 2)(x - 3)} = 0$ so $x = 5$.

There are vertical asymptotes at $x = -2$ and $x = 3$, because $\dfrac{-2}{(-2 + 2)(-2 - 3)} = \dfrac{-2}{0}$ is undefined, as is $\dfrac{3}{(3 + 2)(3 - 3)} = \dfrac{3}{0}$.

At these asymptotes: as $x \to (-2)^-$, $y \to -\infty$ and as $x \to (-2)^+$, $y \to \infty$,
as $x \to 3^-$, $y \to \infty$ and as $x \to 3^+$, $y \to -\infty$.
As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^-$.
Hence:

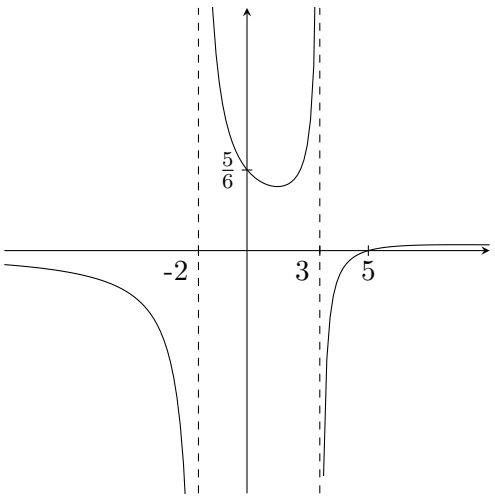


Figure 15: Graph of $y = \dfrac{x-5}{(x+2)(x-3)}$.

8.

$$y = \frac{x}{x^2+3}$$

When $x = 0$, $y = \dfrac{0}{3} = 0$.

When $y = 0$, $\dfrac{x}{x^2+3} = 0$ so $x = 0$.

There are no vertical asymptotes as the denominator, $x^2 + 3$, is never equal to 0 for real values of x.

As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^-$.

Hence:

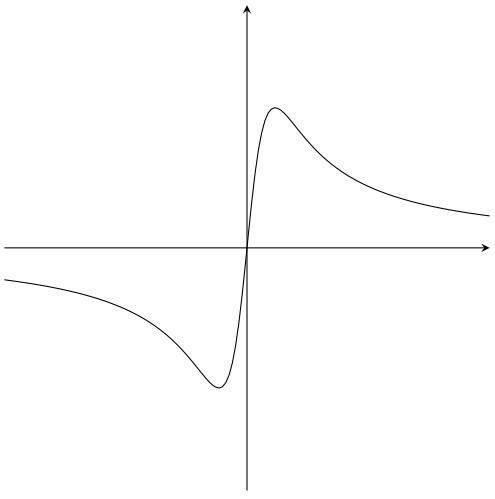


Figure 16: Graph of $y = \dfrac{x}{x^2+3}$. Of course we don't know anything about the shape except from what side the curve approaches zero on the left and the right.

14. (i)

$$y = \frac{1}{(x+1)(3-x)}$$

When $x = 0$, $y = \dfrac{1}{1 \times 3} = \dfrac{1}{3}$.

When $y = 0$, $\dfrac{1}{(x+1)(3-x)} = 0$ so there are no x-intercepts.

There are vertical asymptotes at $x = -1$ and $x = 3$, because $\dfrac{-1}{(-1+1)(3+1)} = \dfrac{-1}{0}$ is undefined, as is $\dfrac{3}{(3+1)(3-3)} = \dfrac{3}{0}$.

At these asymptotes: as $x \to (-1)^-$, $y \to -\infty$ and as $x \to (-1)^+$, $y \to \infty$, as $x \to 3^-$, $y \to \infty$ and as $x \to 3^+$, $y \to -\infty$.

As $x \to \infty$, $y \to 0^-$ and as $x \to -\infty$, $y \to 0^-$.

Hence:

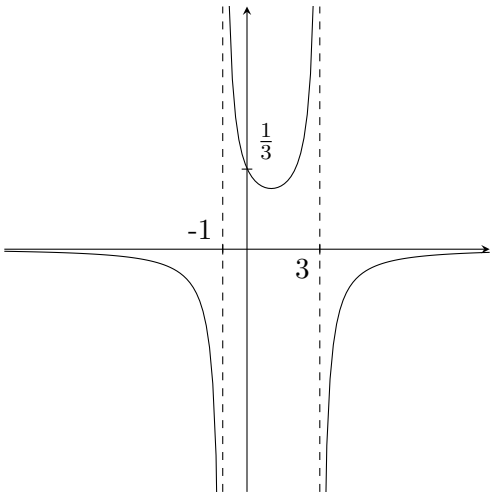


Figure 17: Graph of $y = \dfrac{1}{(x+1)(3-x)}$.

(ii) The line of symmetry is halfway between the two vertical asymptotes, so $x = 1$.

Therefore at the local minimum here, $y = \dfrac{1}{(1+1)(3-1)} = \dfrac{1}{4}$.

Hence the local minimum here is at point $(1, \frac{1}{4})$.

(iii) The equation $\dfrac{1}{(x+1)(3-x)} = k$ is equivalent to the function on the graph above, where k is represented by y. Thus,

(a) There are two distinct real solutions for x whenever there are two x-values for each y-value. As can be seen on the graph, this is true for y-values above $y = \dfrac{1}{4}$, the y-value of the local minimum found in part (ii), and all y-values below 0.

Hence, $k > \dfrac{1}{4}, k < 0$.

(b) There is one real solution for x whenever there is one x-value for every y value and vice versa. This is true only at $y = \dfrac{1}{4}$.

Hence, $k = \dfrac{1}{4}$.

(c) There are no real solutions for x at all y-values that the curve never passes through. It can be seen from the graph that this is true between $y = 0$ and $y = \frac{1}{4}$.
Hence, $0 \leq k < \frac{1}{4}$.

5.2 Exercise 3B

1. (i)

$$y = (x+3)(x-1)(2x-7)$$

At $x = 0$, $y = 3 \times -1 \times -7 = 21$.
At $y = 0$, $(x+3)(x-1)(2x-7) = 0$ so $x = -3, 1, \frac{7}{2}$.
The coefficient of x^3 is positive, so as $x \to \infty$, $y \to \infty$ and as $x \to -\infty$, $y \to -\infty$.
Hence:

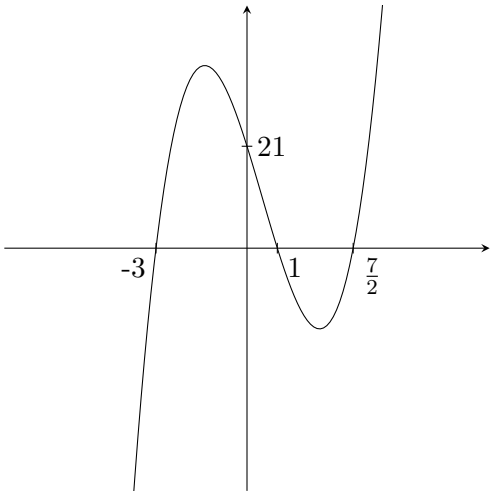


Figure 18: Graph of $y = (x+3)(x-1)(2x-7)$.

(ii) By looking at the graph, it can be seen that $(x+3)(x-1)(2x-7) > 0$ when $-3 < x < 1$ or $x > \frac{7}{2}$.

2. (i)

$$y = \frac{x+2}{x-1}$$

At $x = 0$, $y = \frac{2}{-1} = -2$.
At $y = 0$, $\frac{x+2}{x-1} = 0$ so $x = -2$.
There is a vertical asymptote at $x = 1$, because $\frac{1+2}{1-1} = \frac{3}{0}$ is undefined.
At this asymptote, as $x \to 1^-$, $y \to -\infty$ and as $x \to 1^+$, $y \to \infty$.
As $x \to \infty$, $y \to 1^+$ and as $x \to -\infty$, $y \to 1^-$.
Hence:

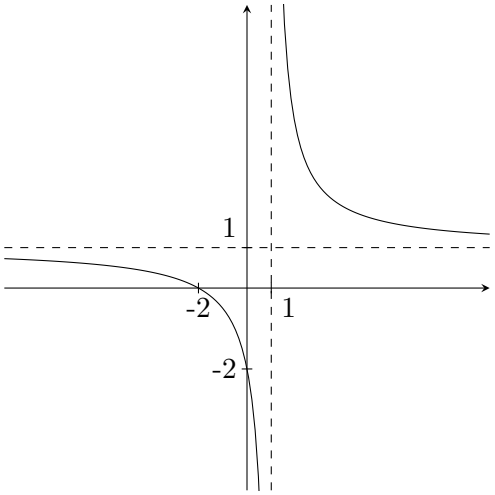


Figure 19: Graph of $y = \dfrac{x+2}{x-1}$.

(ii) By looking at the graph, it can be seen that $\dfrac{x+2}{x-1} \geq 0$ when $x \leq -2$ or $x > 1$.

3. (i)

$$y = x^2$$

Parabola through $(0, 0)$ with gradient $2x$.

$$y = 2x + 3$$

Line through $(0, 3)$ with gradient 2.

x-intercept is at $-\dfrac{3}{2}$.

Hence:

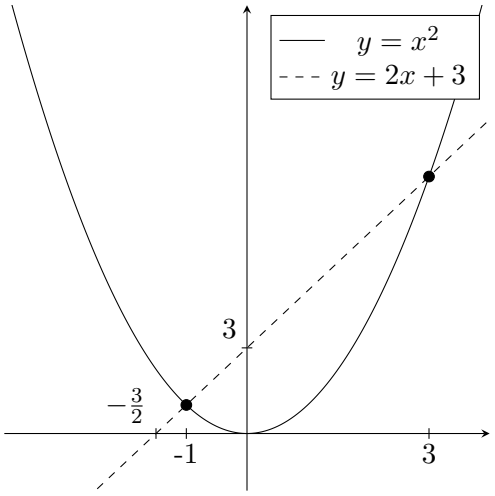


Figure 20: Graph of $y = x^2$ and $y = 2x + 3$.

(ii) By looking at the graph, it can be seen that $x^2 < 2x + 3$ between the x-values at the two intersections of the two lines.

Hence we'll solve the two equations simultaneously.

$$y = x^2 \quad (1)$$

$$y = 2x + 3 \quad (2)$$

Hence

$$\begin{aligned} x^2 &= 2x + 3 \\ x^2 - 2x - 3 &= 0 \\ x &= \frac{2 \pm \sqrt{(-2)^2 - 4 \times 1 \times -3}}{2 \times 1} \\ &= \frac{2 \pm 4}{2} \\ &= 3, -1 \end{aligned}$$

(These are now marked onto the graph above.)
Therefore, the inequality is true for $-1 < x < 3$.

4. (i)

$$y = \frac{8}{x}$$

No y or x-intercepts.
Vertical asymptote at $x = 0$, as $\frac{8}{0}$ is undefined.
At this asymptote: as $x \to 0^-$, $y \to -\infty$ and as $x \to 0^+$, $y \to \infty$.
As $x \to \infty$, $y \to 0^+$ and as $x \to -\infty$, $y \to 0^-$.

$$y = x^2$$

Parabola through $(0, 0)$ with gradient $2x$.
Hence:

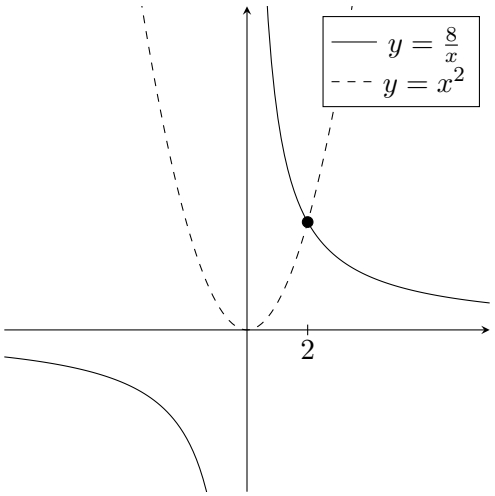


Figure 21: Graph of $y = \frac{8}{x}$ and $y = x^2$.

(ii) It can be seen from the graph that $x^2 \geq \frac{8}{x}$ both to the left of the y-axis, and to the right of their intersection point.
Thus let us find the intersection point:

$$\begin{aligned} \frac{8}{x} &= x^2 \\ x &= \sqrt[3]{8} \\ x &= 2 \end{aligned}$$

(This point is now marked onto the graph above.)
Therefore, the inequality is true for $x < 0$ or $x \geq 2$.

Part V

Incomplete work

Chapter 33

Quaternions (DJV)

This was one of the two 'extension' Year 12 question packs from Mr Vaccaro. It was incredibly interesting and while its density meant I never got very far with it, I plan to come back soon and finish it off.

Some summer pure

Damon Falck

August 2017

Contents

1 DJV Quaternions

Throughout this section, i is the imaginary unit and $\mathbf{1}$ is the 2×2 identity matrix.

1.1 Quaternions as Matrices

1. Using the definitions given:

   (a)

$$\boldsymbol{i}^2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{1}$$

$$\boldsymbol{ij} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \boldsymbol{k}$$

$$\boldsymbol{ik} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -\boldsymbol{j}$$

(b)

$$\boldsymbol{ji} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = -\boldsymbol{k}$$
$$\boldsymbol{j}^2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{1}$$
$$\boldsymbol{jk} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \boldsymbol{i}$$

(c)

$$\boldsymbol{ki} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \boldsymbol{j}$$
$$\boldsymbol{kj} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = -\boldsymbol{i}$$
$$\boldsymbol{k}^2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathbf{1}$$

(d)

$$\boldsymbol{ijk} = \boldsymbol{k}^2 = -\mathbf{1}$$
$$\boldsymbol{jki} = \boldsymbol{i}^2 = -\mathbf{1}$$
$$\boldsymbol{kij} = \boldsymbol{j}^2 = -\mathbf{1}$$

(e)

$$\boldsymbol{ikj} = -\boldsymbol{j}^2 = \mathbf{1}$$
$$\boldsymbol{jik} = -\boldsymbol{k}^2 = \mathbf{1}$$
$$\boldsymbol{kji} = -\boldsymbol{i}^2 = \mathbf{1}$$

(f)
- $\boldsymbol{i}^2$, $\boldsymbol{j}^2$ and $\boldsymbol{k}^2$ are all equal to $-\mathbf{1}$.
- Depending on the order, the product of all three matrices $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$ is $\pm\mathbf{1}$.
- Changing the position of any two adjacent matrices when multiplying switches the sign of the product (e.g. $\boldsymbol{ij} = \boldsymbol{k}$ whereas $\boldsymbol{ji} = -\boldsymbol{k}$, and $\boldsymbol{ijk} = -\mathbf{1}$ whereas $\boldsymbol{jik} = \mathbf{1}$).
- The product of any two distinct matrices (out of $\boldsymbol{i}$, $\boldsymbol{j}$ and $\boldsymbol{k}$), depending on their order, is either the positive or negative of the third, remaining matrix.

(g)

$$\boldsymbol{i}^{-1} = \frac{1}{i \cdot -i - 0 \cdot 0} \begin{pmatrix} -i & -0 \\ -0 & i \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = -\boldsymbol{i}$$
$$\boldsymbol{j}^{-1} = \frac{1}{0 \cdot 0 - 1 \cdot -1} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -\boldsymbol{j}$$
$$\boldsymbol{k}^{-1} = \frac{1}{0 \cdot 0 - i \cdot i} \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} = -\boldsymbol{k}$$

2. (a) Every member $\boldsymbol{q} \in \mathbb{H}$ can be written as

$$\begin{aligned}
\boldsymbol{q} &= a\mathbf{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k} \\
&= a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} + c\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + d\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \\
&= \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} bi & 0 \\ 0 & -bi \end{pmatrix} + \begin{pmatrix} 0 & c \\ -c & 0 \end{pmatrix} + \begin{pmatrix} 0 & di \\ di & 0 \end{pmatrix} \\
&= \begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix} \\
&= \begin{pmatrix} (a+bi) & (c+di) \\ -(c+di)^\star & (a+bi)^\star \end{pmatrix}.
\end{aligned}$$

Let $z = a + bi$ and $w = c + di$ where $z, w \in \mathbb{C}$. Then,

$$\boldsymbol{q} = \begin{pmatrix} z & w \\ -w^\star & z^\star \end{pmatrix}. \qquad \square$$

(b)

$$\boldsymbol{q}^{-1} = \frac{1}{zz^\star + ww^\star}\begin{pmatrix} z^\star & -w \\ w^\star & z \end{pmatrix} = \frac{1}{|z|^2 + |w|^2}\begin{pmatrix} z* & -w \\ w* & z \end{pmatrix}$$

(c)

$$\begin{aligned}
\therefore \quad \boldsymbol{q}^{-1} &= \frac{1}{(a^2+b^2)+(c^2+d^2)}\begin{pmatrix} a-bi & -c-di \\ c-di & a+bi \end{pmatrix} \\
&= \frac{1}{a^2+b^2+c^2+d^2} \cdot \left[\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} -bi & 0 \\ 0 & bi \end{pmatrix} + \begin{pmatrix} 0 & -c \\ c & 0 \end{pmatrix} + \begin{pmatrix} 0 & -di \\ -di & 0 \end{pmatrix}\right] \\
&= \frac{1}{a^2+b^2+c^2+d^2} \cdot \left[a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - b\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} - c\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} - d\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}\right] \\
&= \frac{a\mathbf{1} - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k}}{a^2+b^2+c^2+d^2}
\end{aligned}$$

(d) If $\boldsymbol{q}^\star = a\mathbf{1} - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k}$, then

$$\begin{aligned}
\boldsymbol{q}\boldsymbol{q}^\star &= (a\mathbf{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k})(a\mathbf{1} - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k}) \\
&= a\mathbf{1}a\mathbf{1} - a\mathbf{1}b\boldsymbol{i} - a\mathbf{1}c\boldsymbol{j} - a\mathbf{1}d\boldsymbol{k} + b\boldsymbol{i}a\mathbf{1} - b\boldsymbol{i}b\boldsymbol{i} - b\boldsymbol{i}c\boldsymbol{j} - b\boldsymbol{i}d\boldsymbol{k} \\
&\quad + c\boldsymbol{j}a\mathbf{1} - c\boldsymbol{j}b\boldsymbol{i} - c\boldsymbol{j}c\boldsymbol{j} - c\boldsymbol{j}d\boldsymbol{k} + d\boldsymbol{k}a\mathbf{1} - d\boldsymbol{k}b\boldsymbol{i} - d\boldsymbol{k}c\boldsymbol{j} - d\boldsymbol{k}d\boldsymbol{k} \\
&= a^2\mathbf{1} - ab\boldsymbol{i} - ac\boldsymbol{j} - ad\boldsymbol{k} + ab\boldsymbol{i} + b^2\mathbf{1} - bc\boldsymbol{k} + bd\boldsymbol{j} \\
&\quad + ac\boldsymbol{j} + bc\boldsymbol{k} + c^2\mathbf{1} - cd\boldsymbol{i} + ad\boldsymbol{k} - bd\boldsymbol{j} + cd\boldsymbol{i} + d^2\mathbf{1} \\
&= (a^2+b^2+c^2+d^2)\mathbf{1}.
\end{aligned}$$

(e) $|\boldsymbol{q}|^2 = a^2 + b^2 + c^2 + d^2$ is the determinant of matrix $\boldsymbol{q}$:

$$\det(\boldsymbol{q}) = \det\begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix} = (a+bi)(a-bi)-(c+di)(-c+di) = a^2+b^2+c^2+d^2$$

(f)
$$\boldsymbol{q}^{-1} = \frac{a\mathbf{1} - b\boldsymbol{i} - c\boldsymbol{j} - d\boldsymbol{k}}{a^2 + b^2 + c^2 + d^2} = \frac{\boldsymbol{q}^\star}{|\boldsymbol{q}|^2}$$

(g)
- Quaternion multiplication is non-commutative, whereas complex number multiplication is commutative, but both are associative and distribute over addition.
- Family $\mathbb{H}$ has 4 dimensions (one real and three imaginary), whereas family $\mathbb{C}$ has 2 dimensions (one real and one imaginary).
-

(h) Writing $\boldsymbol{s} = \dfrac{\boldsymbol{q}}{\boldsymbol{r}}$ (where $\boldsymbol{q}, \boldsymbol{r}, \boldsymbol{s} \in \mathbb{H}$ and $\boldsymbol{q} \neq \boldsymbol{r} \neq \boldsymbol{s}$) either implies that $\boldsymbol{rs} = \boldsymbol{q}$ or that $\boldsymbol{sr} = \boldsymbol{q}$, and because quaternion multiplication is non-commutative, $\boldsymbol{sr} \neq \boldsymbol{rs}$ so both can't be true.
In other words, either $\dfrac{\boldsymbol{q}}{\boldsymbol{r}} = \boldsymbol{q}\boldsymbol{r}^{-1}$ or $\dfrac{\boldsymbol{q}}{\boldsymbol{r}} = \boldsymbol{r}^{-1}\boldsymbol{q}$, and we know $\boldsymbol{q}\boldsymbol{r}^{-1} \neq \boldsymbol{r}^{-1}\boldsymbol{q}$ so the notation $\dfrac{\boldsymbol{q}}{\boldsymbol{r}}$ is ambiguous.

3. (a)
$$\begin{aligned}\boldsymbol{qr} &= \begin{pmatrix} a & b \\ -b^\star & a^\star \end{pmatrix}\begin{pmatrix} c & d \\ -d^\star & c^\star \end{pmatrix} \\ &= \begin{pmatrix} ac - bd^\star & ad + bc^\star \\ -a^\star d^\star - b^\star c & a^\star c^\star - b^\star d \end{pmatrix} \\ &= \begin{pmatrix} (ac - bd^\star) & (ad + bc^\star) \\ -(ac - bd^\star)^\star & (ad + bc^\star)^\star \end{pmatrix}\end{aligned}$$
and so if $z = ac - bd^\star$ and $w = ad + bc^\star$, then $\boldsymbol{qr} = \begin{pmatrix} z & w \\ -w^\star & z^\star \end{pmatrix}$.

(b) (Assuming here that the question is asking to prove $(a, b) + (c, d) = (a + c, b + d)$.)
Where $a, b, c, d \in \mathbb{C}$ and $(a, b), (c, d) \in \mathbb{H}$,
$$\begin{aligned}(a, b) + (c, d) &= \begin{pmatrix} a & b \\ -b^\star & a^\star \end{pmatrix} + \begin{pmatrix} c & d \\ -d^\star & c^\star \end{pmatrix} \\ &= \begin{pmatrix} a + c & b + d \\ -(b + d)^\star & (a + c)^\star \end{pmatrix} \\ &= (a + c, b + d)\end{aligned}$$
as the complex conjugate is distributive across addition. □

(c) As shown in part (a),
$$(a, b)(c, d) = (ac - bd^\star, ad + bc^\star)$$

(d)
$$\boldsymbol{r}^{-1} = \frac{1}{cc^\star + dd^\star}\begin{pmatrix} c^\star & -d \\ d^\star & c \end{pmatrix} = \frac{1}{|c|^2 + |d|^2}\begin{pmatrix} c^\star & -d \\ d^\star & c \end{pmatrix}$$

$$\begin{aligned}\therefore \quad \boldsymbol{qr}^{-1} &= \frac{1}{|c|^2+|d|^2}\begin{pmatrix} a & b \\ -b^\star & a^\star \end{pmatrix}\begin{pmatrix} c^\star & -d \\ d^\star & c \end{pmatrix} \\ &= \frac{1}{|c|^2+|d|^2}\begin{pmatrix} ac^\star + bd^\star & bc - ad \\ -(b^\star c^\star + a^\star d^\star) & a^\star c - b^\star d \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{ac^\star + bd^\star}{|c|^2+|d|^2} & \dfrac{bc-ad}{|c|^2+|d|^2} \\ -\left(\dfrac{bc-ad}{|c|^2+|d|^2}\right)^\star & \left(\dfrac{a^\star c + b^\star d}{|c|^2+|d|^2}\right)^\star \end{pmatrix} \qquad \square\end{aligned}$$

(e) Hence as $\boldsymbol{qr}^{-1}$ can be written as $(a,b) \div (c,d)$,

$$(a,b) \div (c,d) = \left(\frac{a^\star c + b^\star d}{|c|^2+|d|^2}, \frac{bc-ad}{|c|^2+|d|^2}\right). \qquad \square$$

4. For simplicity we assume $\mathbb{N}$ is the set of positive integers.

Lemma 1. *The set of integral quaternions (Lipschitz integers)*

$$L = \{\boldsymbol{q} \in \mathbb{H} : \boldsymbol{q} = a\mathbf{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}, \{a, b, c, d\} \subseteq \mathbb{Z}$$

is closed under multiplication.

Proof. Let $\boldsymbol{q} = a\mathbf{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}$ and $\boldsymbol{r} = e\mathbf{1} + f\boldsymbol{i} + g\boldsymbol{j} + h\boldsymbol{k}$ be two integer quaternions, where $a, b, c, d, e, f, g, h \subseteq \mathbb{Z}$.

So, tediously multiplying through, using the unit quaternion identities established in question 1:

$$\begin{aligned}
\boldsymbol{qr} &= (a\mathbf{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k})(e\mathbf{1} + f\boldsymbol{i} + g\boldsymbol{j} + h\boldsymbol{k}) \\
&= ae\mathbf{1}^2 + af\mathbf{1}\boldsymbol{i} + ag\mathbf{1}\boldsymbol{j} + ah\mathbf{1}\boldsymbol{k} \\
&\quad + be\boldsymbol{i}\mathbf{1} + bf\boldsymbol{i}^2 + bg\boldsymbol{i}\boldsymbol{j} + bh\boldsymbol{i}\boldsymbol{k} \\
&\quad + ce\boldsymbol{j}\mathbf{1} + cf\boldsymbol{j}\boldsymbol{i} + cg\boldsymbol{j}^2 + ch\boldsymbol{j}\boldsymbol{k} \\
&\quad + de\boldsymbol{k}\mathbf{1} + df\boldsymbol{k}\boldsymbol{i} + dg\boldsymbol{k}\boldsymbol{j} + dh\boldsymbol{k}^2 \\
&= ae\mathbf{1} + af\boldsymbol{i} + ag\boldsymbol{j} + ah\boldsymbol{k} \\
&\quad + be\boldsymbol{i} - bf\mathbf{1} + bg\boldsymbol{k} - bh\boldsymbol{j} \\
&\quad + ce\boldsymbol{j} - cf\boldsymbol{k} - cg\mathbf{1} + ch\boldsymbol{i} \\
&\quad + de\boldsymbol{k} + df\boldsymbol{j} - dg\boldsymbol{i} - dh\mathbf{1} \\
&= (ae - bf - cg - dh)\mathbf{1} + (af + be - cg - dg)\boldsymbol{i} \\
&\quad + (ag - bh + ce + df)\boldsymbol{j} + (ah + bg - cf + de)\boldsymbol{i}.
\end{aligned}$$

Now the product of two integers is an integer and the sum of two integers is an integer, so each of these terms is an integer: thus $\boldsymbol{qr}$ is itself an integral quaternion, and so the integral quaternions are closed under multiplication. □

Lemma 2. *For any invertible matrices A and B,*

$$\det(A)\det(B) = \det(AB).$$

Proof. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$. So,

$$\begin{aligned}
\det(A) &= ad - bc, \\
\det(B) &= eh - fg \\
\implies \det(A)\det(B) &= (ad - bc)(eh - fg) \\
&= adeh - adfg - bceh + bcfg.
\end{aligned}$$

However,

$$\begin{aligned}
AB &= \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix} \\
&= \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix} \\
\implies \det(AB) &= (ae + bg)(cf + dh) - (af + bh)(ce + dg) \\
&= acef + adeh + bcfg + bdgh - acef - adfg - bceh - bdgh \\
&= adeh + bcfg - adfg - bceh \\
&= \det(A)\det(B)
\end{aligned}$$

and we're done. □

Lemma 3. *The set of determinants of integral quaternions*

$$D = \{\det(\boldsymbol{q}) : \boldsymbol{q} \in L\}$$

is closed under multiplication.

Proof. Let $\boldsymbol{q}, \boldsymbol{r}$ be two integral quaternions. So, $\det(\boldsymbol{q}), \det(\boldsymbol{r}) \in D$. By lemma 2,

$$\det(\boldsymbol{q}) \det(\boldsymbol{r}) = \det(\boldsymbol{qr}),$$

but lemma 1 guarantees that $\boldsymbol{qr} \in L$ and so $\det(\boldsymbol{qr}) \in D$. Hence, D is closed under multiplication. □

Theorem. *If*

$$\mathfrak{F} = \{n \in \mathbb{N} : n = a^2 + b^2 + c^2 + d^2, \{a, b, c, d\} \subseteq \mathbb{N}\}$$

then $\mathfrak{F}$ is closed under multiplication.

Proof. As shown in question 2 (e), the determinant of the quaternion $\boldsymbol{q} = a\boldsymbol{1} + b\boldsymbol{i} + c\boldsymbol{j} + d\boldsymbol{k}$ is $\det(\boldsymbol{q}) = a^2 + b^2 + c^2 + d^2$. Therefore, $\boldsymbol{q} \in L \iff \det(\boldsymbol{q}) \in D$ if and only if $a, b, c, d \in \mathbb{Z}$. However, any number expressible as the sum of the squares of four integers is expressible as the sum of the squares of four naturals, as $x^2 \equiv (-x)^2$, and so every element of D is in fact an element of F, and vice versa: the sets D and $\mathfrak{F}$ are equal.

Hence since $D = \mathfrak{F}$ and since D is closed under multiplication by lemma 3, we conclude that $\mathfrak{F}$ is indeed closed under multiplication. □

1.2 Quaternions as Numbers

5. Here the identities are shown in an order such that each identity relies only on ones previously shown.[1]

$$\begin{aligned}
i^2 = j^2 = k^2 = ijk = -1 \quad &\text{is given} \\
ij = -ij \cdot -1 = -ijk^2 = (-ijk)k = k \\
jk = -jk \cdot -1 = -i^2jk = i(-ijk) = i \\
jki = (i)i = -1 \\
ki = -ki \cdot -1 = -j^2ki = j(-jki) = j \\
i^{-1} = -i^{-1} \cdot -1 = -i^{-1}ijk = -jk = -i \\
k^{-1} = -k^{-1} \cdot -1 = -ijkk^{-1} = -ij = -k \\
j^{-1} = -j^{-1} \cdot -1 = -j^{-1}jki = -ki = -j \\
ik = (-i)(-k) = i^{-1}k^{-1} = (ki)^{-1} = j^{-1} = -j \\
ji = (-j)(-i) = j^{-1}i^{-1} = (ij)^{-1} = k^{-1} = -k \\
kj = (-k)(-j) = k^{-1}j^{-1} = (jk)^{-1} = i^{-1} = -i
\end{aligned}$$

The others were shown without matrices in question 1.

[1] The latter half of these do rely on thinking of the quaternion units as matrices; perhaps later I'll try to find a purely algebraic way.

1.2.1 Division by Quaternions

6. (a) LHS:

$$\left(\frac{1-2i-2j-k}{10}\right)(1+2i+2j+k) = \frac{(1-2i-2j-k)(1+2i+2j+k)}{10}$$
$$= \frac{1^2+2^2+2^2+1^2}{10}$$
$$= \frac{10}{10} = 1$$

RHS:

$$(1+2i+2j+k)\left(\frac{1-2i-2j-k}{10}\right) = \frac{(1+2i+2j+k)(1-2i-2j-k)}{10}$$
$$= \frac{1^2+2^2+2^2+1^2}{10}$$
$$= \frac{10}{10} = 1$$

$$\therefore \quad \text{LHS} = \text{RHS} = 1 \qquad \square$$

(b) Let $q = a + bi + cj + dk$.

$$\begin{aligned} qq^\star &= (a+bi+cj+dk)(a-bi-cj-dk) \\ &= a^2 - abi - acj - adk + abi - b^2i^2 - bcij - cdik \\ &\quad + acj - bcji - c^2j^2 - cdjk + dka - bdki - cdkj - d^2k^2 \\ &= a^2 - abi - acj - adk + abi + b^2 - bck + cdj \\ &\quad + acj + bck + c^2 - cdi + adk - bdj + cdi + d^2 \\ &= a^2 + b^2 + c^2 + d^2 \end{aligned}$$

And now

$$\begin{aligned} q^\star q &= (a-bi-cj-dk)(a+bi+cj+dk) \\ &= a^2 + abi + acj + adk - abi - b^2i^2 - bcij - bdik \\ &\quad - cja - bcji - c^2j^2 - cdjk - dka - bdki - cdkj - d^2k^2 \\ &= a^2 + b^2 + c^2 + d^2 \end{aligned}$$

So we can see that

$$qq^{-1} = q\frac{q^\star}{|q|^2} = \frac{qq^\star}{|q|^2} = \frac{|q|^2}{|q|^2} = 1$$
$$q^{-1}q = \frac{q^\star}{|q|^2}q = \frac{q^\star q}{|q|^2} = \frac{qq^\star}{|q|^2} = \frac{|q|^2}{|q|^2} = 1$$

$$\therefore \quad qq^{-1} = q^{-1}q \qquad \square$$

1.2.2 Quaternion Arithmetic

7. (a)

$$(1+3i-j+k) + (2-4i+j-k) = 3-i$$

(b)

$$\begin{aligned}(2+3i-2j+k)(1+2i-j+k) &= 2+4i-2j+2k+3i+6i^2-3ij+3ik\\ &\quad -2j-4ji+2j^2-2jk+k+2ki-kj+k^2\\ &= 2+4i-2j+2k+3i-6-3k-3j\\ &\quad -2j+2k-2-2i+k+2j+i-1\\ &= -7+6i-5j+2k\end{aligned}$$

(c)

$$\begin{aligned}(1+2i-j+k)(2+3i-2j+k) &= 2+3i-2j+k+4i+6i^2-4ij+2ik\\ &\quad -2j-2ji+2j^2-jk+2k+3ki-2kj+k^2\\ &= 2+3i-2j+k+4i-6-4k-2j\\ &\quad -2j+2k-2-i+2k+3j+2i-1\\ &= -7+8i-3j+k\end{aligned}$$

(d)

$$\begin{aligned}(2+3i-2j+k)^{-1} &= \frac{2-3i+2j-k}{2^2+3^2+2^2+1^2}\\ &= \frac{2-3i+2j-k}{18}\end{aligned}$$

(e)

$$\begin{aligned}(1+i)(2-3j)^{-1} &= (1+i)\frac{2+3j}{2^2+3^2}\\ &= \frac{1}{13}(1+i)(2+3j)\\ &= \frac{1}{13}(2+3j+2i+3ij)\\ &= \frac{2+2i+3j+3k}{13}\end{aligned}$$

(f)

$$\begin{aligned}(2+j+k)^{-1}(1+j) &= \frac{2-j-k}{2^2+1^2+1^2}(1+j)\\ &= \frac{1}{6}(2-j-k)(1+j)\\ &= \frac{1}{6}(2+2j-j-j^2-k-kj)\\ &= \frac{3+i+j-k}{6}\end{aligned}$$

(g)

$$\begin{aligned}(1+i+j+k)^2 &= 1+i+j+k+i+i^2+ij+ik\\ &\quad +j+ji+j^2+jk+k+ki+kj+k^2\\ &= 1+i+j+k+i-1+k-j\\ &\quad +j-k-1+i+k+j-i-1\\ &= -2+2i+2j+2k\end{aligned}$$

(h)

$$\begin{aligned}(1-i+j+k)^2 &= 1-i+j+k-i+i^2-ij-ik\\ &\quad +j-ji+j^2+jk+k-ki+kj+k^2\\ &= 1-i+j+k-i-1-k+j\\ &\quad +j+k-1+i+k-j-i-1\\ &= -2-2i+2j+2k = 2(-1-i+j+k)\\ \therefore (1-i+j+k)^3 &= 2(-1-i+j+k)(1-i+j+k)\\ &= 2(-1+i-j-k-i+i^2-ij-ik\\ &\quad +j-ji+j^2+jk+k-ki+kj+k^2)\\ &= 2(-1+i-j-k-i-1-k+j\\ &\quad +j+k-1+i+k-j-i-1)\\ &= 2(-4) = -8\end{aligned}$$

1.2.3 Quaternion Algebra

8. In general,

$$\begin{aligned}(a_1i+a_2j+a_3k)(b_1i+b_2j+b_3k) &= a_1b_1i^2+a_1b_2ij+a_1b_3ik\\ &\quad +a_2b_1ji+a_2b_2j^2+a_2b_3jk\\ &\quad +a_3b_1ki+a_3b_2kj+a_3b_3k^2\\ &= -a_1b_1+a_1b_2k-a_1b_3j\\ &\quad -a_2b_1k-a_2b_2+a_2b_3i\\ &\quad +a_3b_1j-a_3b_2i-a_3b_3\\ &= -(a_1b_1+a_2b_2+a_3b_3)+(a_2b_3-a_3b_2)i\\ &\quad +(a_3b_1-a_1b_3)j+(a_1b_2-a_2b_1)k.\end{aligned}$$

9. Using the expression derived above,

$$\begin{aligned}(a_1i+a_2j+a_3k)^2 &= -(a_1^2+a_2^2+a_3^2)+(a_2a_3-a_3a_2)i+(a_3a_1-a_1a_3)j+(a_1a_2-a_2a_1)k\\ &= -(a_1^2+a_2^2+a_3^2)+0i+0j+0k\\ &= -a_1^2-a_2^2-a_3^2\end{aligned}$$

which is evidently real.

10. (a) For any quaternions $q = a+bi+cj+dk$ and $r = e+fi+gj+hk$,

$$\begin{aligned}qr &= (a+bi+cj+dk)(e+fi+gj+hk)\\ &= ae+afi+agj+ahk+bei+bfi^2+bgij+bhik\\ &\quad +cej+cfji+cgj^2+chjk+dek+dfki+dgkj+dhk^2\\ &= ae+afi+agj+ahk+bei-bf+bgk-bhj\\ &\quad +cej-cfk-cg+chi+dek+dfj-dgi-dh\\ &= (ae-bf-cg-dh)+(af+be+ch-dg)i+(ag-bh+ce+df)j+(ah+bg-cf+de)k\end{aligned}$$

and so,

$$|qr|^2 = (ae-bf-cg-dh)^2+(af+be+ch-dg)^2+(ag-bh+ce+df)^2+(ah+bg-cf+de)^2$$

which is the same[2] as

$$\begin{aligned}|qr|^2 &= (a^2+b^2+c^2+d^2)(e^2+f^2+g^2+h^2)\\ &= |q|^2|r|^2\end{aligned}$$

as we were hoping for.

(b) A matrix-based proof is included in lemma 2 of question 4.

[2] Because Mathematica says so.

Chapter 34

Möbius transformations (DJV)

This was the second of the two packs from Mr Vaccaro just mentioned. Again, I didn't get far at all, but it was thought-provoking to say the least.

Möbius Transformations

Damon Falck

January 2017

1 Preliminaries

1. (a) The matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{1}$$

produces a rotation of θ anticlockwise around the origin.

(b) Let us compute $R_\theta R_\phi$:

$$\begin{aligned} R_\theta R_\phi &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -(\sin\theta\cos\phi + \sin\phi\cos\theta) \\ \sin\theta\cos\phi + \sin\phi\cos\theta & \cos\theta\cos\phi - \sin\theta\sin\phi \end{pmatrix}. \end{aligned}$$

Now applying the compound angle identities, this simplifies to

$$R_\theta R_\phi = \begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix},$$

which by comparison with eq. (1) can be seen to represent a rotation of $\theta + \phi$ anticlockwise around the origin. This is what we'd expect, as first R_ϕ produces a rotation of ϕ, and then R_θ produces a further rotation of θ.

(c) We want to prove that

$$\tan(a+b) \equiv \frac{\tan a + \tan b}{1 - \tan a \tan b}.$$

First, we can express the LHS as

$$\tan(a+b) \equiv \frac{\sin(a+b)}{\cos(a+b)},$$

which using the compound angle identities

$$\sin(\theta + \phi) \equiv \sin\theta\cos\phi + \sin\phi\cos\theta$$

and

$$\cos(\theta + \phi) \equiv \cos\theta\cos\phi - \sin\theta\sin\phi),$$

we can simplify to

$$\tan(a + b) \equiv \frac{\sin a\cos b + \sin b\cos a}{\cos a\cos b - \sin a\sin b}.$$

Now dividing the numerator and denominator by $\cos a\cos b$, we get

$$\tan(a + b) \equiv \frac{\frac{\sin a}{\cos a} + \frac{\sin b}{\cos b}}{1 - \frac{\sin a\sin b}{\cos a\cos b}}$$

which becomes

$$\tan(a + b) \equiv \frac{\tan a + \tan b}{1 - \tan a\tan b}$$

as desired. □

(d) We can construct a right triangle with angle θ and opposite side length 1 as below:

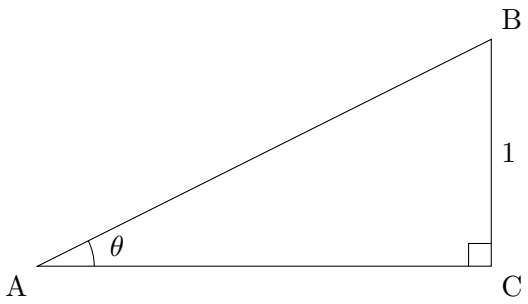


Figure 1: We're given that $BC = 1$ and $\angle BAC = \theta$.

Therefore,

$$AC = \frac{BC}{\tan\theta} = \frac{1}{\tan\theta}.$$

We also know that

$$\begin{aligned}\tan(\angle ABC) &= \frac{AC}{BC} \\ &= \frac{\frac{1}{\tan\theta}}{1} \\ &= \frac{1}{\tan\theta}.\end{aligned}$$

Finally, $ABC = \dfrac{\pi}{2} - \theta$ (because it's a right triangle) and so

$$\tan\left(\frac{\pi}{2} - \theta\right) = \frac{1}{\tan\theta}$$

as desired. □
(Shown only for acute θ.)

2. It can be seen that

$$NC = \tan\phi$$

and

$$NB = \tan\theta.$$

Therefore,

$$AC = \tan\theta + \tan\phi. \tag{2}$$

Assuming point M is placed so that BM is perpendicular to NB (and *not* necessarily tangent to the circle), $MBNC$ forms a rectangle. Therefore

$$BM = NC = \tan\phi \tag{3}$$

and

$$CM = NC = 1. \tag{4}$$

Because $DBAC$ is a cyclic quadrilateral, $\angle DCA$ and $\angle DBA$ sum to π, so as $\angle DCA$ is a right angle,

$$\frac{\pi}{2} + \angle DBA = \pi$$
$$\angle DBA = \frac{\pi}{2}.$$

Thus, as can be seen from the diagram,

$$\angle MBA = \angle MBN + \theta = \angle DBA + \angle MBD$$

and so because $\angle DBA$ and $\angle MBN$ are both $\frac{\pi}{2}$,

$$\frac{\pi}{2} + \theta = \frac{\pi}{2} + \angle MBD$$
$$\angle MBD = \theta. \tag{5}$$

Hence, now considering $\triangle MBD$, we can see that

$$DM = BM \cdot \tan(\angle MBD) = \tan\phi\tan\theta$$

from eqs. (3) and (5). So, because of eq. (4),

$$DC = 1 - DM = 1 - \tan\theta\tan\phi. \tag{6}$$

We're given that $\angle NDC = \phi$ and $\angle NDA = \theta$, so $\angle CDA = \theta + \phi$. Thus, because of eqs. (2) and (6),

$$\tan(\theta + \phi) = \frac{AC}{DC}$$
$$\tan(\theta + \phi) = \frac{\tan\theta + \tan\phi}{1 - \tan\theta\tan\phi}$$

as desired. □

2 Stereographic Projection

3. (a) *(Assuming $C \subset \mathbb{C}$.)*
Let $t \in \mathbb{R}$ be the input to σ (the real part of P) and let $a + bi \in \mathbb{C}$ be the output, so that our map is

$$\sigma : \mathbb{R} \to C : t \mapsto a + bi.$$

The circle C has centre $(0, \frac{1}{2}$ and radius $\frac{1}{2}$, so

$$C = \{x + yi \in \mathbb{C} : x^2 + (y - \frac{1}{2})^2 = \frac{1}{4}.\}.$$

Similarly, the line NP with gradient $-\frac{ON}{OP} = -\frac{1}{t}$ passing through point $N(0, 1)$ is given by

$$\{\text{line}NP\} = \{x + yi \in \mathbb{C} : x = t - ty\}.$$

So, at their intersection (point P'), both

$$a^2 + (b - \frac{1}{2})^2 = \frac{1}{4} \tag{7}$$

and

$$a = t - tb \tag{8}$$

are true.

$$\therefore \quad (t - tb)^2 + (b - \frac{1}{2})^2 = \frac{1}{4}$$
$$t^2 - 2t^2b + t^2b^2 + b^2 - b + \frac{1}{4} = \frac{1}{4}$$
$$(t^2 + 1)b^2 - (2t^2 + 1)b + t^2 = 0$$

So we have a quadratic in b:

$$\begin{aligned} b &= \frac{(2t^2+1) \pm \sqrt{(2t^2+1)^2 - 4(t^2+1)t^2}}{2(t^2+1)} \\ &= \frac{2t^2+1 \pm 1}{2t^2+1} \\ &= 1, \frac{t^2}{t^2+1}. \end{aligned}$$

Discarding $b = 1$ (this is point N), we have

$$b = \frac{t^2}{t^2+1}.$$

Hence by eq. (8),

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

so

$$\begin{aligned} a &= \frac{b}{t} \\ t &= \frac{b}{a}. \end{aligned}$$

So we have that our forwards map is

$$\sigma : \mathbb{R} \to C : t \mapsto \frac{t + t^2 i}{t^2+1} \tag{9}$$

and our inverse map is

$$\sigma^{-1} : C \setminus \{N\} \to \mathbb{R} : a + bi \mapsto \frac{b}{a}.$$

The only case where σ^{-1} isn't defined is when

$$a = 0, b = 1$$

i.e. at point N. Thus, as σ is invertible everywhere else, σ is a bijection between $\mathbb{R}$ and $C \setminus \{N\}$. □

It can be seen that

$$\tan(\angle ONP') = \frac{OP}{ON} = \frac{t}{1} = t$$

(where $t = \Re(P)$).

(b) It can be seen from eq. (9) that if $t = \infty$ then

$$\sigma(t) = \lim_{t\to\infty} \left(\frac{t + t^2 i}{t^2 + 1} \right) = i,$$

which is at point N. As σ is injective, the only value that could be mapped onto N is $t = \infty$. Because $\infty \notin \mathbb{R}$, point N is not in the range of σ.

Alternatively, for N to be mapped onto, the line PN would have to make an angle of $\frac{\pi}{2}$ (or $-\frac{\pi}{2}$) with ON, meaning it would be parallel to the real line, so they would never intersect. The same result follows.

A simple way to get around this would be to add a point to the real line at ∞, which would then map to point N.

(c) The function $\arctan : \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ maps the real part of P to the angle $\angle ONP$.

4. (a) The function f has domain and range $\mathbb{R} \setminus \{0\}$, because $f(0)$ is not defined and there is no real x for which $f(x) = 0$.

(b) The function f' has domain and range $C \setminus \{O, N\}$, because $f'(\sigma(x)) = \sigma(f(x))$; $f(x)$ cannot map to or from 0 or ∞, so $\sigma(f(x))$ cannot map to or from O or N.

(c) We know that by definition,

$$f'(\sigma(x)) = \sigma(f(x)) = \sigma\left(\frac{1}{x}\right),$$

so from def. (9),

$$f'(\sigma(x)) = \frac{\dfrac{1}{x} + \dfrac{1}{x)^2} i}{\dfrac{1}{(x)^2} + 1} = \frac{x + i}{x^2 + 1}.$$

By comparison with def. (9), it can be seen that

$$f'(\sigma(x)) = \Re(\sigma(x)) + \frac{\Im(\sigma(x))}{x^2}$$

so $\Re(f'(z)) = \Re(z)$.

We now know that the range of f' is $C \setminus \{O, N\}$, and that the transformation $f' \mapsto C$ conserves the real component (it's a vertical transformation). Any two points on C with the same real part must have a midpoint with imaginary part $\frac{1}{2}$ (because it's a circle with centre $(0, \frac{1}{2})$), so wherever it's defined f' must be a reflection of the circle C in the line $\Im(z) = \dfrac{1}{2}$. □

The function f' isn't defined for points O and N, so they are not affected by this transformation. Were the function definitions extended to include these points (and still produce the same reflection), O and N would switch position.

(d) The function $f \circ f$ is the identity function with domain $\mathbb{R} \setminus \{0\}$.
The function $f' \circ f'$ is the identity function with domain $C \setminus \{O, N\}$.

5. (a) We can show this with simple geometry.

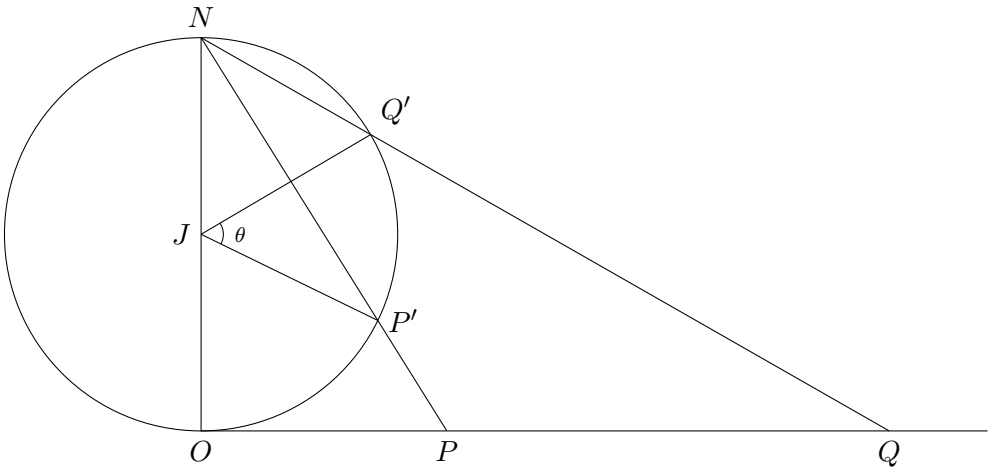


Figure 2: Point $J(0, \frac{1}{2})$ has been added at the centre of circle C. Note that P and Q have been swapped so that θ is acute.

Let $\angle P'JQ'$ be θ, so that r'_θ moves P' to Q'. The angle subtended at the centre of a circle is double the angle subtended at the circumference, so

$$\angle P'NQ' = \frac{\theta}{2}.$$

Therefore,

$$\angle ONQ' = \angle ONP' + \frac{\theta}{2}. \quad \square$$

(b) Let $OP = x$. Therefore $r_\theta(x) = OQ$. We can see that

$$\begin{aligned} OQ &= ON \cdot \tan \angle ONQ \\ r_\theta(x) &= 1 \cdot \tan(\angle ONP' + \angle P'NQ') \\ &= \tan\left(\arctan\frac{x}{1} + \frac{\theta}{2}\right). \end{aligned}$$

Applying the tangent compound angle identity,

$$\begin{aligned} r_\theta(x) &= \frac{\tan(\arctan x) + \tan\frac{\theta}{2}}{1 - \tan(\arctan x) \cdot \tan\frac{\theta}{2}} \\ &= \frac{x + a}{1 - ax} \end{aligned}$$

where $a = \tan\frac{\theta}{2}$. $\square$

(c) The denominator of $r_\theta\left(\frac{1}{a}\right)$ is $1 - a \cdot \frac{1}{a} = 0$, resulting in a division by zero. Hence, $r_\theta\left(\frac{1}{a}\right)$ is undefined.

(d) Let us find r_θ^{-1}. We know that

$$\begin{aligned} r_\theta(x) &= \frac{x+a}{1-ax} \\ \therefore \quad r_\theta(x) - r_\theta(x)ax &= x + a \\ x &= \frac{r_\theta(x) - a}{r_\theta(x)a + 1}. \end{aligned}$$

Hence,

$$r_\theta^{-1} = \frac{x-a}{1+ax}.$$

The only time that r_θ^{-1} is undefined is when the denominator is zero, so $x = -\frac{1}{a}$. Thus, r_θ maps to all points in $\mathbb{R}$ except for $-\frac{1}{a}$. □

6. If $g(x) = -\frac{1}{x}$ then, with reference to def. (9),

$$\begin{aligned} \sigma(g(x)) &= \sigma\left(-\frac{1}{x}\right) \\ \therefore \quad \sigma(g(x)) &= \frac{-\frac{1}{x} + \left(\frac{1}{x}\right)^2 i}{\left(-\frac{1}{x}\right)^2 + 1} = \frac{\frac{i}{x^2} - \frac{1}{x}}{\frac{1}{x^2} + 1} = \frac{i - x}{x^2 + 1}. \end{aligned}$$

Therefore by comparison with def. (9), it can be seen that

$$\Re(\sigma(g(x))) = -\Re(\sigma(x))$$

and so by a similar argument to question 4 (c), the circle must be reflected first in the line $\Im(z) = \frac{1}{2}$ and then in the line $\Re(z) = 0$. Hence, the real transformation $g(x) = -\frac{1}{x}$ is induced by the mapping $\sigma(g(\sigma^{-1}(x)))$, which is a rotation of π radians around the point $(0, \frac{1}{2})$.

Chapter 35

Cavendish quantum mechanics

These were the beginnings of my solutions to Dr Cheung's book on quantum mechanics that I worked on near the start of the 2017 summer holiday. These are incomplete and unchecked but I'm planning on pressing on with them when I next get a chance.

A Cavendish Quantum Mechanics Primer

Solutions

June 30, 2018

Preface

Note: throughout these solutions, equation numbers enclosed in parentheses — for example, eq. (3.15) — always refer to equations printed in the original text. All equations in these solutions themselves are referred to by numbers enclosed in square brackets — for example, eq. [3.15].

Contents

Chapter 1

Preliminaries — some underlying quantum ideas and mathematical tools

Exercise 1.1

Derive from the electric, gravitational and harmonic potentials their force laws. Explain the sign of the forces — is it what you expect? Take care over the definition of the zero of potential. Does the position where the potential is zero matter?

Solution

Remembering the definition of a force in terms of potential, for each of these instances the equation $f = \frac{\mathrm{d}V}{\mathrm{d}x}$ will give us our force law. That is, we just have to differentiate the potential to get our force.

In the first case, the electric potential is $V(x) = \frac{Q_1 Q_2}{4\pi\epsilon_0 x}$ (where ϵ_0 is the permittivity of free space) and so differentiating, our force law is

$$f = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{Q_1 Q_2}{4\pi\epsilon_0 x}\right) \tag{1.1}$$

$$= -\frac{Q_1 Q_2}{4\pi\epsilon_0 x^2}. \tag{1.2}$$

The force is negative because it is of course repulsive; on either charge the direction of the force is opposite to the direction of the other charge (and f is positive only when the force is in same direction as the separation x).

We'll do exactly the same thing for the two other cases. The potential associated with a gravitational field is $V(x) = -\frac{Gm_1 m_2}{x}$ where G is the gravitational constant, and so differentiating,

$$f = \frac{\mathrm{d}}{\mathrm{d}x}\left(-\frac{Gm_1 m_2}{x}\right) \tag{1.3}$$

$$= \frac{Gm_1 m_2}{x^2}. \tag{1.4}$$

Here the force is positive because it is an attractive force; the same explanation as before applies.

Finally, the harmonic potential is $V(x) = \frac{1}{2}qx^2$ and hence in the same manner, the associated force is

$$f = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{2}qx^2\right) \quad [1.5]$$

$$= qx. \quad [1.6]$$

Again, the force f required to effect an extension x is always in the same direction as that extension, and so f has a positive sign.

All of these results have similar forms, and we recognise them as the classical laws more commonly stated than their potential-involving counterparts.

Regarding the position of the zero of potential, for both electric and gravitational fields the potential will become zero only when the two objects separate to infinity (as the separation x is in the denominator). However, the harmonic potential is zero if and only if there is no extension at all — that is, $x = 0$.

Exercise 1.2

Consider a particle of mass m passing a potential well of width a, as shown in Fig. 1.3. The particle has total energy $E > V_0$, the depth of the well. Calculate the time taken by the particle to traverse the figure.

Solution

Solution given in the text.

Exercise 1.3

A particle of mass m slides down, under gravity, a smooth ramp which is inclined at an angle θ to the horizontal. At the bottom, it is joined smoothly to a similar ramp rising at the same angle θ to the horizontal to form a V-shaped surface. If the particle slides smoothly around the join, determine the period of oscillation, T, in terms of the initial horizontal displacement x_0 from the centre join. Note the shape of the potential well.

Hint

We see that the potential well appears as a sloping line similar to the one along which the particle is constrained to move. It is only this linear slope at angle θ to the horizontal, that happens to resemble the potential energy graph of the same shape, which misleads us into thinking that we can see the potential energy. The potential energy is a concept, represented pictorially by a graph and the shape of the graph happens, in some cases, to resemble the mechanical system.

Solution

The particle is acting under gravity and so its potential is given by $V(h) = mgh$ where h is the height above the bottom of the ramps. (Here we take this bottom point as our reference frame.) A simple bit

of trigonometry gives us that $h = x\tan\theta$ where x is horizontal distance from the centre join, and so our potential is $V(x) = mgx\tan\theta$. This leads to the potential well shown below.

Assuming the particle starts from rest, its initial energy is *just* its initial potential energy, that is $V(x_0) = mgx_0\tan\theta$. We know the total energy (always the sum of the kinetic and potential energies) must be constant, and so

$$\frac{1}{2}mv^2 + mgx\tan\theta = mgx_0\tan\theta \tag{1.7}$$

which means that

$$v(x) = \sqrt{2g(x_0 - x)\tan\theta}. \tag{1.8}$$

Taking the particle's descent down the first ramp only (of length $\frac{x_0}{\cos\theta}$), our initial velocity is $v(x_0) = 0$ and our final velocity is $v(0) = \sqrt{2gx_0\tan\theta}$, so using the constant acceleration formula $s = \frac{u+v}{2}t$ we find the time taken to reach the centre join:

$$t = \frac{2\frac{x_0}{\cos\theta}}{0 + \sqrt{2gx_0\tan\theta}} = 2\sqrt{\frac{x_0}{g\sin 2\theta}}. \tag{1.9}$$

Therefore by symmetry, the total period of oscillation T is just twice this time, so

$$T = 4\sqrt{\frac{x_0}{g\sin 2\theta}}. \tag{1.10}$$

(Note that in calculating this answer we use the identity $\sin 2\theta = 2\sin\theta\cos\theta$.)

Exercise 1.4

A particle moves in a potential $V(x) = \frac{1}{2}qx^2$. If it has a total energy $E = E_0$ give an expression for its velocity as a function of position $v(x)$. What is the amplitude of its motion?

Solution

We know the total energy E is always the sum of the kinetic and potential energies, that is

$$\frac{1}{2}mv^2 + \frac{1}{2}qx^2 = E. \tag{1.11}$$

So, making v the subject, our velocity function is

$$v(x) = \sqrt{\frac{2E - qx^2}{m}}. \tag{1.12}$$

The maximum displacement in either direction will occur when $v = 0$, so

$$\sqrt{\frac{2E - qx^2}{m}} = 0 \tag{1.13}$$

$$\implies 2E - qx^2 = 0 \tag{1.14}$$

$$\implies x = \pm\sqrt{\frac{2E}{q}}. \tag{1.15}$$

Therefore the amplitude A of its motion is the difference between these two maxima, that is

$$A = \sqrt{\frac{2E}{q}} - \left(-\sqrt{\frac{2E}{q}}\right) = \sqrt{\frac{8E}{q}}. \tag{1.16}$$

Exercise 1.5

The potential energy of a particle of mass m as a function of its position along the x axis is shown in Fig. 1.4.

(a) Sketch a graph of the force versus position in the x direction which acts on a particle moving in this potential well with its vertical steps. Why is this potential unphysical?

(b) Sketch a more realistic force versus position curve for a particle in this potential well. For a particle moving from $x = 0$ to $x = \frac{3a}{2}$, which way does the force act on the particle? If the particle was moving in the opposite direction, which way would the force be acting on the particle?

Hint

Take care over the physical meaning of the potential energy. It can look misleadingly like the physical picture of a particle sliding off a high shelf, down a very steep slope and then sliding along the floor, reflecting off the left hand wall and then back up the slope. This is too literal an interpretation since, for example, the potential change might be due to an electrostatic effect rather than a gravitational one, and the time spent moving up or down the slope is due to artificially putting in an extra vertical dimension in a problem which is simply about motion in only one dimension. An example of where there is literally motion vertically as well as horizontally, is that of a frictionless bead threaded on a parabolic wire. The motion is not the same as in the one-dimensional simple harmonic motion of Ex. 1.4. Although the potential energy is expressible in the form $\frac{1}{2}qx^2$ due to the constraint of the wire, the kinetic energy involves both the x and y variables.

Solution

(a) We remember that force is simply the negative derivative of potential (with respect to position), and so we see that sketching the force graph is very similar to sketching an acceleration-time graph given a velocity-time graph. This leads to the following sketch.

We clearly see that the infinite force spikes cannot be physical; thus the potential is unphysical as it does not change smoothly.

(b) A more realistic force versus position curve can be achieved simply by smoothing over the spikes slightly so that we have short but not infinitesimal forces acting on the particle.

When the particle is moving from $x = 0$ to $x = \frac{3a}{2}$, its potential increases which means that it kinetic energy must decrease, and so the force is acting against the particle's motion. Similarly, when the particle is moving from $x = \frac{3a}{2}$ to $x = 0$, its potential decreases and thus its kinetic energy increases, so the force is acting with the particle's motion. Either way, the force acts in the negative x direction.

Exercise 1.6

Consider again the particle in Ex. 1.5. If it has a total mechanical energy E equal to $3V_0$, calculate the period for a complete oscillation.

Solution

For this exercise we will follow a very similar approach to Exs. 1.2 and 1.3. From $x = 0$ to $x = a$, the particle has zero potential and therefore a kinetic energy of $\frac{1}{2}mv^2 = 3V_0$ which gives us a speed of

$$v = \sqrt{\frac{6V_0}{m}}. \tag{1.17}$$

From $x = a$ to $x = \frac{3a}{2}$, the particle has a potential of $2V_0$ and therefore a kinetic energy of $\frac{1}{2}m(v')^2 = 3V_0 - 2V_0$ giving a speed of

$$v' = \sqrt{\frac{2V_0}{m}}. \tag{1.18}$$

The particle cannot escape the well between $x = 0$ and $x = \frac{3a}{2}$ because it does not have enough total energy. Making use of the definition of speed again, the total time taken to travel from $x = 0$ to $x = \frac{3a}{2}$ (a complete oscillation) is

$$T = a\sqrt{\frac{m}{6V_0}} + \frac{a}{2}\sqrt{\frac{m}{2V_0}} = a\sqrt{\frac{m}{2V_0}}\left(\frac{\sqrt{3}}{3} + \frac{1}{2}\right). \tag{1.19}$$

Exercise 1.7

The variance σ^2 in the values of x is the average of the square of the deviations of x from its mean, that is,

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle. \tag{1.20}$$

Prove the above agrees with the standard result of $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$ for both discrete and continuous x.

Solution

We'll prove this separately for discrete and continuous x.

For discrete x, we know $\langle f(x) \rangle = \sum_i p_i f(x_i)$ and so expanding the form given, the variance is

$$\sigma^2 = \sum_i p_i (x_i^2 - 2x_i\langle x \rangle + \langle x \rangle^2). \tag{1.21}$$

Making use of the fact that summation is commutative, we split this up into three separate sums,

$$\sigma^2 = \sum_i p_i x_i^2 - \sum_i 2p_i x_i \langle x \rangle + \sum_i p_i \langle x \rangle^2, \tag{1.22}$$

from each of which we can factor out any constant $\langle x \rangle$, giving

$$\sigma^2 = \sum_i p_i x_i^2 - 2\langle x \rangle \sum_i p_i x_i + \langle x \rangle^2 \sum_i p_i. \tag{1.23}$$

It is immediately apparent that we can rewrite this in terms of averages again, leading to

$$\sigma^2 = \langle x^2 \rangle - 2\langle x \rangle \langle x \rangle + \langle x \rangle^2 \cdot 1 \tag{1.24}$$

$$= \langle x^2 \rangle - \langle x \rangle^2 \tag{1.25}$$

as desired.

The process is almost identical for continuous x: using $\langle f(x)\rangle = \int P(x)x\,\mathrm{d}x$ instead where $P(x)$ is the probability density function, eq. [1.21] becomes

$$\sigma^2 = \int P(x)(x^2 - 2x\langle x\rangle + \langle x\rangle^2)\,\mathrm{d}x \tag{1.26}$$

but we happen to know integration can be split up just like addition, and so

$$\begin{aligned}\sigma^2 &= \int P(x)x^2\,\mathrm{d}x - \int P(x)2x\langle x\rangle\,\mathrm{d}x + \int P(x)\langle x\rangle^2\,\mathrm{d}x &[1.27]\\ &= \int P(x)x^2\,\mathrm{d}x - 2\langle x\rangle\int P(x)x\,\mathrm{d}x + \langle x\rangle^2\int P(x)\,\mathrm{d}x &[1.28]\\ &= \langle x^2\rangle - 2\langle x\rangle\langle x\rangle + \langle x\rangle^2\cdot 1 &[1.29]\\ &= \langle x^2\rangle - \langle x\rangle^2 &[1.30]\end{aligned}$$

just like before.

It can also be proven that $\langle x + y\rangle = \langle x\rangle + \langle y\rangle$ for any x and y; this is a slightly more general version of the above.

Exercise 1.8

Prove that $\tan 2\theta = 2\tan\theta/(1-\tan^2\theta)$ and further that $\tan 4\theta = 4\tan\theta(1-\tan^2\theta)/(1-6\tan^2\theta+\tan^4\theta)$.

If $t = \tan(\theta/2)$, then show that $\sin\theta = 2t/(2+t^2)$ and $\cos\theta = (1-t^2)/(1+t^2)$, while $\tan\theta = 2t/(1-t^2)$ is also a special form of the $\tan 2\theta$ identity.

Prove that $1 + \tan^2\theta = \sec^2\theta$ where $\sec\theta = 1/\cos\theta$.

These relations are useful in integration by substitution.

Solution

We'll tackle these proofs one step at a time. Firstly, by the definition of the tangent function we know that $\tan 2\theta = \frac{\sin 2\theta}{\cos 2\theta}$ and so using identities (1.9) and (1.10), we know

$$\tan 2\theta = \frac{2\sin\theta\cos\theta}{\cos^2\theta - \sin^2\theta} \tag{1.31}$$

as $2\cos^2\theta - 1 = \cos^2\theta - \sin^2\theta$ by the Pythagorean identity (1.8). Factoring out $\cos^2\theta$ from both the numerator and the denominator, we end up with

$$\tan 2\theta = \frac{2\tan\theta}{1 - \tan^2\theta} \tag{1.32}$$

as desired.

Using this new identity with our angle as 2θ instead of θ, it is clear that

$$\tan 4\theta = \frac{2\tan 2\theta}{1 - \tan^2 2\theta}. \tag{1.33}$$

Now, we just substitute in eq. [1.32] and simplify it down a bit:

$$\tan 4\theta = \frac{2\left(\frac{2\tan\theta}{1-\tan^2\theta}\right)}{1-\left(\frac{2\tan\theta}{1-\tan^2\theta}\right)^2} \tag{1.34}$$

$$= \frac{4\tan\theta(1-\tan^2\theta)}{(1-\tan^2\theta)^2-(2\tan\theta)^2} \tag{1.35}$$

$$= \frac{4\tan\theta(1-\tan^2\theta)}{1-2\tan^2\theta+\tan^4\theta-4\tan^2\theta} \tag{1.36}$$

$$= \frac{4\tan\theta(1-\tan^2\theta)}{1-6\tan^2\theta+\tan^4\theta} \tag{1.37}$$

just as we hoped for. Note that to get from eq. [1.34] to eq. [1.35] we just multiply the whole fraction by $\frac{(1-\tan^2\theta)^2}{(1-\tan^2\theta)^2}$, and then we expand the denominator.

Next, we are asked to consider $t = \tan(\theta/2)$. A simple triangle will get us where we need:

(insert figure here) [1.38]

It is clear that $\cos(\theta/2) = \frac{1}{\sqrt{1+t^2}}$ and $\sin(\theta/2) = \frac{t}{\sqrt{1+t^2}}$. Now, making use of this, the sine double angle identity (1.10) gives

$$\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} \tag{1.39}$$

$$= 2\left(\frac{t}{\sqrt{1+t^2}}\right)\left(\frac{1}{\sqrt{1+t^2}}\right) \tag{1.40}$$

$$= \frac{2t}{1+t^2} \tag{1.41}$$

as hoped for. Very similarly, the cosine double angle identity (1.9) gives

$$\cos\theta = \cos^2\theta - \sin^2\theta \tag{1.42}$$

$$= 2\cos^2\theta - 1 \tag{1.43}$$

$$= 2\left(\frac{1}{\sqrt{1+t^2}}\right)^2 - 1 \tag{1.44}$$

$$= \frac{2}{1+t^2} - 1 \tag{1.45}$$

$$= \frac{1-t^2}{1+t^2} \tag{1.46}$$

just as we were looking for. To get from eq. [1.42] to eq. [1.43] we just use the Pythagorean identity $\sin^2\theta + \cos^2\theta = 1$.

Finally, applying the $\tan 2\theta$ identity directly to this situation gives us

$$\tan\theta = \frac{2\tan\frac{\theta}{2}}{1-\tan^2\frac{\theta}{2}} = \frac{2t}{1-t^2}. \tag{1.47}$$

The last identity is very simple to show: take the Pythagorean identity $\sin^2\theta + \cos^2\theta = 1$ which is derived from a right triangle of unit hypoteneuse, and divide through by $\cos^2\theta$:

$$\frac{\cos^2\theta}{\cos^2\theta} + \frac{\sin^2\theta}{\cos^2\theta} = \frac{1}{\cos^2\theta} \implies 1 + \tan^2\theta = \sec^2\theta. \tag{1.48}$$

Exercise 1.9

Plot $e^{-x^2/2\sigma^2}$ for a range of positive and negative x. Label important points on the x axis (including where the function is $1/e$) and the y axis. Pay special attention to $x = 0$, including slope and curvature there. What is the effect on the graph of varying σ?

Solution

SOLUTION PLACEHOLDER (to be plotted)

Exercise 1.10

Show that $\frac{d}{dx}(\tan x) = \sec^2 x$.

Solution

We know $\tan x = \frac{\sin x}{\cos x}$. Using the product rule, this is $\sin x \cdot (\cos x)^{-1}$ and so

$$\frac{d}{dx}\tan x = \sin x \cdot \frac{d}{dx}(\cos x)^{-1} + (\cos x)^{-1} \cdot \frac{d}{dx}\sin x \quad [1.49]$$

$$= \sin x \cdot [-(\cos x)^{-2}] \sin x + (\cos x)^{-1} \cdot \cos x \quad [1.50]$$

$$= -\frac{\sin^2 x}{\cos^2 x} + 1 \quad [1.51]$$

$$= 1 - \tan^2 x \quad [1.52]$$

$$= \sec^2 x. \quad [1.53]$$

To get to the last line we used the identity $1 + \tan^2\theta = \sec^2\theta$ that we proved in Ex. 1.8.

If you've learned the quotient rule, it is perfectly possible (and slightly easier) to use this instead of the product rule here.

Exercise 1.11

Plot $\sin(kx)$, $\cos(kx)$ and $e^{\pm kx}$ for positive and negative x, and plot $\ln(kx)$ for positive x. Label important points (e.g. intersections with axes, maxima and minima) on the x and y axes. What happens to these points and the graph if you change k? Revise elementary properties of the exponential and logarithmic functions. What are $(e^x)^2$, e^x/e^y, $a \ln x$ and $\ln x + \ln y$?

Solution

The plots asked for are as follows:

(insert figures here) [1.54]

For each of these functions, increasing k will cause a scaling of factor $1/k$ in the x direction — this can be verified by trying out a few values.

Elementary properties of the exponential and logarithmic functions give the following identities:

$$(e^x)^2 = e^{2x}, \tag{1.55}$$

$$\frac{e^x}{e^y} = e^{x-y}, \tag{1.56}$$

$$a \ln x = \ln(x^a), \tag{1.57}$$

$$\ln x + \ln y = \ln(xy). \tag{1.58}$$

Part VI

Programmes, handouts, cheat sheets, etc.

Chapter 36

The Vaccaro Society

Some of the most fun I had at Highgate was in running the Vaccaro Society — we had loads of talks from teachers and students and it was always exciting to go to, for me at least. I've included most of the programmes I printed off for them as well as a handout I made for our one-off lunchtime proof of the Basel problem.

MathSci Forum

Programme

Thursday 11th May 2017

I am the pope — Oliver Smouha [1 min]

Drawing the impossible — Leon Galli [25 min]

A mathematical proof that your mother doesn't love you — Saul Austin [10 min]

A tangent on pigeonholes — Damon Falck [5 min]

Half a revolution in cubics — Gianmarco Luppi [5 min]

Going round in circles — Thalia Seale [10 min]

Total running time: 56 min

Next time will be on Thursday 25th May.

MathSci Forum

Programme

Thursday 25th May 2017

Turning code into maths — Jack Saville [10 min]

Cubics (round 2) — Miss Brownlee [7 min]

Generating functions — Dr Dessain [10 min]

Predicting when the world will end — Joel Gottlieb [5 min]

Estimating! — Mr Vaccaro [10 min]

A second proof that $\sin(\cos x) < \cos(\sin x)$ — Dr Cheung [5 min]

Euler's solution to the Basel problem — Damon Falck [8 min]

Your maths is worth £1 million — Mr Wright [2 min]

Total running time: 57 min

Next time will be on Thursday 15th June.

MathSci Forum

Programme

Thursday 15th June 2017

Your maths is worth £1 million — Mr Wright [2 min]

The geometry of tricky quadratics — Mr Vaccaro [2 min]

Lemma 1: de Moivre's identity — Annie Kaissides [5 min]

Lemma 2: the binomial expansion — Thalia Seale [8 min]

Lemma 3 — Damon Falck [10 min]

Cubics(ish) round 3 / Lemma 4: Vieta's formulas — Gianmarco Luppi [11 min]

Lemma 5 — Damon Falck [6 min]

Lemma 6 — Thalia Seale [4 min]

Theorem: $\boldsymbol{\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}}$ — Damon Falck [7 min]

Total running time: 55 min

Next time will be on Thursday 29th June.

MathSci Forum

Programme

Thursday 6th July 2017

A neat little proof — Mr Wright [2 min]

More on Stirling's approximation! — Dr Cheung [10 min]

Differentiators and integrators — Roch Briscoe [4 min]

e is irrational — Gianmarco Luppi [8 min]

The world's hardest logic puzzle — Oliver Smouha [5 min]

Some musings on friends and strangers — Miss Brownlee [5 min]

The velocities of gas molecules — Damon Falck [10 min]

$\exists \{a, b\} \subset (\mathbb{R} \setminus \mathbb{Q}) : a^b \in \mathbb{Q}$ — Gianmarco Luppi [2 min]

Infinite thanks to DJV — Thalia Seale [5 min]

Dividing by zero and the lightbulb moment — Dr Strangeway [5 min]

Goodbye — Mr Vaccaro [1 min]

Total running time: 57 minutes

Today's talks are being filmed for internal use only.
If you do not wish to appear on camera, just say so.

This is the last session this academic year. Thanks so much to everyone for making this possible: we'll be back in September.

Filmed talks will be made available on HERO shortly.

THE VACCARO SOCIETY

Short talks in mathematics, theoretical physics and computer science.

Programme

Thursday 21st September 2017

Choose it or lose it — Oliver Gottlieb [5 min]

The Maxwell-Boltzmann distribution — Yasmin Yazdani [5 min]

Squaring the circle — Damon Falck [11 min]

Deranged ramblings — Gianmarco Luppi [11 min]

AM–GM — Dr Cheung [10 min]

Solving the cubic — Chandrasekhar Iyengar [6 min]

$\int \sqrt{\tan x}\,\mathrm{d}x$ — Thalia Seale [5 min]

Total running time: 58 minutes

Today's talks are being filmed for internal use only.
If you do not wish to appear on camera, just say so.

The next session will be on Thursday 5th October. **Anyone** can do a talk — just let me know by Sunday 1st October. Thanks for coming! DF

Today's filmed talks will be made available on HERO shortly.

THE VACCARO SOCIETY

Short talks in mathematics, theoretical physics and computer science.

Programme

Thursday 12th October 2017

Proofs without words — Leon Galli [10 min]

AM–GM (part 2) — Dr Cheung [10 min]

The Game of Life — Thalia Seale [15 min]

The unfortunate truth of arc lengths — Bruno Edwards [4 min]

Optimus primes — Gianmarco Luppi [16 min]

Total running time: 55 minutes

Today's talks are being filmed for internal use only.
If you do not wish to appear on camera, just say so.

The next session will be on Thursday 9th November if all goes to plan. **Anyone** can do a talk — just let me know by Sunday 5th November. Thanks for coming! DF

Today's filmed talks will be made available on HERO shortly.

THE VACCARO SOCIETY

Short talks in mathematics, theoretical physics and computer science.

Programme

Thursday 9th November 2017

Tupper's self-referential formula — Oliver Smouha [5 min]

Some Bayesian statistics — Mr Galdal Gibbs [20 min]

Solving $\cos x = 2$ — Gianmarco Luppi [15 min]

As the crow flies — Damon Falck, Thalia Seale [16 min]

Total running time: 56 minutes

Today's talks may be filmed for internal use only.
If you do not wish to appear on camera, just say so.

The next session will be on Thursday 23rd November if all goes to plan. **Anyone** can do a talk — just let me know by Sunday 19th November. Thanks for coming! DF

THE VACCARO SOCIETY

Short talks in mathematics, theoretical physics and computer science.

Programme

Thursday 18th January 2018

An n-dimensional everything formula — Jake Saville [5 min]

The Calculus done geometrically — Dr Cheung [5 min]

The devil's game — Oliver Smouha [2 min]

The kangaroo, the flea and the pizza delivery — Damon Falck [12 min]

The geometry of Koch island — Archie Campbell [10 min]

The catenary curve — Mr Dales [8 min]

Maths taken to the limit — Gianmarco Luppi [10 min]

Total running time: 52 minutes

We will be hosting an all-day mathematics conference on Friday 2nd February with Professor Imre Leader; if you would like to give a short talk there (whether new or recycled) please let me know as soon as possible. Aside from that, I'll be in touch with dates for the next Vaccaro society.

Thanks for coming.

DF

A rigorous collaborative proof for the Basel problem

Gianmarco Luppi, Annie Kaissides, Damon Falck, Thalia Seale

Thursday 15th June 2017

This proof first appeared in Augustin-Louis Cauchy's 1821 seminal textbook *Cours d'Analyse*.

Lemma 1 (de Moivre's identity). For any complex x and integer n,

$$(\cos x + \mathrm{i}\sin x)^n = \cos(nx) + \mathrm{i}\sin(nx).$$

Lemma 2 (the binomial expansion). For any complex a, b and positive integer n,

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + b^n$$

where $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

Lemma 3. The distinct roots of the mth degree polynomial

$$p(t) = \binom{2m+1}{1}t^m - \binom{2m+1}{3}t^{m-1} \pm \cdots + (-1)^m\binom{2m+1}{2m+1}$$

are $t = \cot^2\left(\dfrac{r\pi}{2m+1}\right)$ for $r = 1, 2, \ldots, m$.

Lemma 4 (Vieta's formulas). For any general polynomial

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with distinct roots $r_1, r_2, \ldots, r_n$, the sum of the roots is

$$r_1 + r_2 + \cdots + r_{n-1} + r_n = -\frac{a_{n-1}}{a_n}$$

and the product of the roots is

$$r_1 r_2 r \cdots r_{n-1} r_n = (-1)^n \frac{a_0}{a_n}.$$

Lemma 5. If $x_r = \dfrac{r\pi}{2m+1}$, then

$$\sum_{r=1}^{m} \cot^2 x_r = \frac{2m(2m-1)}{6}$$

and

$$\sum_{r=1}^{m} \csc^2 x_r = \frac{2m(2m+2)}{6}.$$

Lemma 6. For any real x in radians such that $0 < x < \frac{\pi}{2}$,

$$\cot^2 x < \frac{1}{x^2} < \csc^2 x.$$

Theorem.

$$\sum_{r=1}^{\infty} \frac{1}{r^2} = \frac{\pi^2}{6}.$$

Cheat sheet

- The imaginary unit i is defined as $i = \sqrt{-1}$.
- The binomial coefficient, or choose function, $\binom{n}{k}$ represents the number of ways of choosing k objects out of a total of n.
- "n factorial" is defined as $n! = n(n-1)(n-2)\cdots 2 \cdot 1$.
- Writing $\sum_{k=0}^{n} x_k$ means $x_0 + x_1 + \cdots + x_n$: that is, summing over k from 0 to n.
- We use radians for our angle measures; π radians is equivalent to 180°.
- Some more trigonometric functions are defined as follows:

$$\sec x = \frac{1}{\cos x}, \qquad \operatorname{cosec} x = \csc x = \frac{1}{\cos x}, \qquad \operatorname{cotan} x = \cot x = \frac{1}{\tan x} = \frac{\cos x}{\sin x}.$$

- The limit as x tends to infinity of some function $f(x)$ is written as $\lim_{x\to\infty} f(x)$. This is the value the function approaches for large x.

Chapter 37

The Maths Bash

This interschool maths conference was originally my friend's idea, and it was a brilliant day. Here is the programme as well as the booklet of starter problems I compiled from all the problems sent to us by speakers.

The first

HIGHGATE MATHS BASH

Welcome to the Highgate Maths Bash, a day-long celebration of the beauty of mathematics.

The day will revolve around twenty-two short talks from students and teachers at London Academy of Excellence Tottenham, King's College London Mathematics School and Highgate School. We are also very grateful to Professor Imre Leader for joining us from Cambridge to provide our keynote lecture.

Please do have a look at the attached starter problems to get you thinking about the upcoming talks!

Today's programme:

10:30am–10:55am

Mingle and refreshments in the Sir Martin Gilbert Library.

10:55am–11:15am

Introductory problem-solving session in Dyne House Auditorium.

11:15am–12:20pm

Short talks:

Simple brainteasers
Leon Galli (Highgate)
5 minutes

Benford's law
Ruby Gray (Highgate)
5 minutes

A real turning point
Thalia Seale (Highgate)
20 minutes

Some probability
Sae Koyama (KCLMS)
8 minutes

The problem of quickest descent
Damon Falck (Highgate)
18 minutes

Calculus and its modern-day applications
Rigon Tahiraj (KCLMS)
8 minutes

12:20pm–1:05pm

Lunch in the Dining Hall.

1:10pm–2:10pm

Keynote lecture: Professor Imre Leader, University of Cambridge

2:10pm–2:45pm

Short talks:

How winning games is helping robots take over the world

Rufus Walkden (Highgate)
8 minutes

The Josephus problem

Haroon Aftab, Muhammad Hussain (LAET)
5 minutes

Graph theory: the bridges of Königsberg

Nada Baessa (KCLMS)
8 minutes

Graph theory: the four-colour map theorem

Manuj Mishra (KCLMS)
8 minutes

Proof by chance

Mr A Bottomley (Highgate teacher)
5 minutes

2:45pm–2:55pm

A short break and refreshments in Dyne House Foyer and Terrace.

2:55pm–4:30pm

Short talks:

Pirates!

Oliver Smouha (Highgate)
5 minutes

Complex logarithms

Gianmarco Luppi (Highgate)
15 minutes

The Mandelbrot set

Tabs Goldman (KCLMS)
8 minutes

Proof that Christmas = Halloween

Morgan Saville (Highgate)
2 minutes

Paradoxes: Hilbert's hotel and the hanging problem

Natasha Goldman (KCLMS)
8 minutes

A counterexample to Fermat's last theorem?

Devam Savjani (LAET)
5 minutes

Just a little theorem

Sarah Henderson (Highgate)
5 minutes

Irrationality

Miss P Brownlee (Highgate teacher)
5 minutes

Diffie-Hellman

Evan Quiney (KCLMS)
8 minutes

The many proofs of Pythagoras' theorem

Miles Keat (KCLMS)
8 minutes

The Kepler problem

Dr A Cheung (Highgate teacher)
20 minutes

Timing will be extremely tight so unfortunately we cannot allow talks to go over their time limit. We'd also appreciate it it talk changeovers are as quick as possible. Thanks!

Motivational problems:

These problems have been provided by our speakers to get you thinking about areas of mathematics related to their talk.

(Since there are so many, the best thing to do is probably just pick a couple you're interested in and have a go. Not all of these problems are easily solvable!)

Find three distinct integers x, y and z such that $x^3 + y^3$ is as close to z^3 as possible.

Suspend your disbelief for a moment and pretend that you are the manager of an infinitely booked infinite hotel (that is, a hotel with infinitely many rooms — from 1, 2, 3 to infinity — and all of these rooms are occupied). How could you make room for a new guest (if that's even possible)? How about infinitely many guests? And if you were able to make room for new guests, would the total number of guests staying at the hotel change?

How do you find the derivative of a function?

How do you rotate a point in 2D space?

How many integers less than p^2 are coprime with p^2 when p is prime?

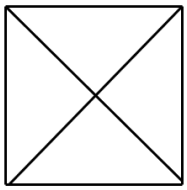
What is the remainder when a^6 is divided by 9, given that a is an integer which shares no prime factors with 9?

Can you draw 5 shapes such that every single shape shares at least part of an edge (not just a corner!) with every other shape?

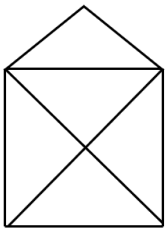
What curve would you choose between two points in the same vertical plane such that a frictionless object sliding down the curve released from rest at the first point will reach the second point in the shortest time?

A *unicursal* shape is one which you can draw without lifting your pen off the paper and without retracing any lines.

Look at the shape below. Is it unicursal?



How about this shape?



Can you think of why? If not, come up with more shapes which you think are or aren't unicursal.

Korma: Prove $\sqrt{2}$ is irrational.

Bhuna: Prove $\sqrt{6}$ is irrational.

Dopiaza: Prove e is irrational.

Vindaloo: Prove e^2 is irrational.

Phall: Prove e^4 is irrational.

Show that any complex number $a + b\mathrm{i}$ can be written in the form $re^{i\theta}$ given Euler's identity $e^{i\theta} \equiv \cos\theta + i\sin\theta$.

Twenty people stand in a circle. Each person will kill the person directly to their left starting with the first person. What is the position of the last person left alive?

Define the function $P_c(z)$ as follows:

$$P_c : \mathbb{C} \to \mathbb{C}, P_c(z) = z^2 + c$$

where $c \in \mathbb{C}$ is a constant. Then $P_c^n(0)$ means that $P_c(z)$ has been composed with itself n times, and we've set $z = 0$. In other words, we can think of this as a recursive relationship.

$$z_(n + 1) = z_n^2 + c$$

Define a set M as follows:

$$M = \{c \in C \mid \exists\ S \in R : \forall\ n \in N,\ \left|P_c^n(0)\right| \leqslant s\}.$$

We're going to investigate the elements of this set. Put more simply, we're going to consider what happens when we change the value of c in our recurrence relation. To find numbers in the Mandelbrot set, for the first iteration, z always equals 0 and c varies.

An example of a number in the set is 0. We begin with $z = 0$ and $c = 0$ and then get every successive term being equal to 0, so the sequence converges and 0 is in the set.

If we wanted to test if 1 was in the set we would make $c = 1$.

Investigate what happens if we let $c = -1$, $\frac{1}{2}$ or 1 and see if you can find any other numbers that are in the set. What about imaginary numbers?

A group of 5 pirates has 100 gold coins. They have to decide amongst themselves how to divide the treasure, but must abide by pirate rules:

- The most senior pirate proposes the division.
- All of the pirates (including the most senior) vote on the division. If half or more vote for the division, it stands. If less than half vote for it, they throw the most senior pirate overboard and start again.
- The pirates are perfectly logical. In order of priority they care a)

about not being thrown overboard, b) about maximising their share of the gold and c) about throwing another pirate overboard if they can.

So, what division should the most senior pirate suggest to the other four?

What is the link between an ancient Greek mathematician and a former president of the United States?

Let $f(x)$ be an increasing function through the points $(0, 0)$ and (x_0, y_0) where $x_0, y_0 > 0$. Find the value a that minimises the area given by

$$\int_0^a f(x)\,\mathrm{d}x + \int_a^{x_0} (y_0 - f(x))\,\mathrm{d}x.$$

Given that $2^x = 16$, what is x? 4, obviously.

But, consider $2^x \equiv 1 \pmod 7$; what values of x satisfy this equation?

Can you find a singular solution for this equation, perhaps using mod?

Suppose a fair coin is flipped 3 times.

What is the probability that heads will come up every time?

What is the probability that heads will come up exactly twice? Once? No times?

Draw an equilateral triangle. Pick a random point in the interior and draw lines from this point to each corner of the triangle.

Label these three lines a, b and c and label two of the angles around the middle point x and y. Is there an easy way we can construct a triangle with sides a, b and c to find the third angle?

Chapter 38

Mechanics cheat sheet

An irrelevant document but quite fun I thought; I made this for a mechanics test near the beginning of Year 12 where Dr Kwasigroch let us bring in a cheat sheet to the exam.

MECHANICS CHEAT-SHEET

DAF, 08-01-2017

FORMULAE OF MOTION

Newton II:

$$F = \frac{\mathrm{d}\,\vec{p}}{\mathrm{d}t} = ma.$$

Principle of impulse:

$$m\,\mathrm{d}v = F\,\mathrm{d}t.$$

Principle of work done:

$$s\sum F = \frac{1}{2}m(v^2 - u^2).$$

CONSTANT ACCELERATION FORMULAE

$$\begin{aligned} v &= u + at, \\ s &= ut + \frac{1}{2}at^2, \\ s &= vt - \frac{1}{2}at^2, \\ v^2 &= u^2 + 2as, \\ s &= \frac{1}{2}(u+v)t. \end{aligned}$$

PROJECTILE FORMULAE

Time of flight:

$$T = \frac{2u\sin\theta}{g}.$$

(To derive, apply $s = ut + \frac{1}{2}at^2$ vertically where $s = 0, T \neq 0$.) Projectile range:

$$R = \frac{u^2\sin 2\theta}{g}.$$

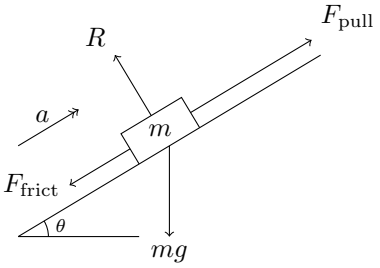
(To derive, apply $s = ut + \frac{1}{2}at^2$ horizontally, substitute in the time of flight eq. and simplify with trig identity.) Path equation:

$$y = \tan\theta x - \frac{gx^2}{u^2\cos^2\theta}.$$

(To derive, apply $s = ut + \frac{1}{2}at^2$ in both directions, then eliminate t.)

BLOCKS ON SLOPES

Draw a free-body diagram of the block:



Resolve mg parallel and perpendicular to slope, then apply Newton II in both axes.

For connected blocks on slopes, repeat and set up simultaneous equations for tension.

MULTIPHASE CONSTANT ACCELERATION PROBLEMS

Draw a table for s, u, v, a and t for every object/path and set up all the simultaneous constant acceleration equations you can using the information given. Solve simultaneously.

COLLISIONS

Total momentum is conserved in a closed system:

$$m_1\vec{u}_1 + m_2\vec{u}_2 = m_1\vec{v}_1 + m_2\vec{v}_2.$$

(Can be proven using Newton II and III between two colliding particles.) In a perfectly elastic collision, total kinetic energy is conserved:

$$\frac{1}{2}m_1u_1^2 + \frac{1}{2}m_2u_2^2 = \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2.$$

In a perfectly inelastic collision, the particles coalesce, so

$$\vec{v}_1 = \vec{v}_2 = \frac{P_i}{m_1 + m_2}$$

where P is the total initial momentum, and therefore

$$E_{kf} = \frac{P^2}{2(m_1+m_2)^2}.$$

Hence for any collision,

$$\frac{P^2}{2(m_1+m_2)^2} \leq E_{kf} \leq E_{ki}.$$

In elastic collisions, it is also true that

$$v_2 - v_1 = u_1 - u_2.$$

This can be derived by combining conservation of momentum and conservation of K.E. by collecting terms of mass, completing the square and then dividing the equations.

Collision questions may involve extensive use of the constant acceleration formulae too.

LAW OF RESTITUTION

The coefficient of restitution e is defined as

$$e = \frac{\text{speed of separation}}{\text{speed of approach}}.$$

So, $e = \frac{v_2 - v_1}{u_1 - u_2}$ if u_1, u_2, v_1 and v_2 are all *in the same direction.*

If $e = 0$ then $v_2 = v_1$ so the collision is perfectly inelastic. If $e = 1$ then $v_2 - v_1 = u_1 - u_2$ so the collision is perfectly elastic.

So, in non-explosive collisions,

$$0 \leq e \leq 1.$$

COEFFICIENT OF STATIC FRICTION

For a *stationary* object on a rough surface, pull F cannot exceed a certain force before the object starts moving:

$$F \leq \mu R$$

where μ is the coefficient of static friction and R is the normal reaction force on the object from the surface.