A-level Refresher Course

Keith Ball

Chapter 1. Functions

Functions are some of the most important objects in mathematics. But it was only in the 19^{th} century that mathematicians finally settled upon a clear definition of what they meant by functions. They chose just about the simplest possible concept.

Suppose A and B are sets. A **function** from A to B sends each element of A to something in B. That's all. We don't say anything about **how** the function does its job: the only thing we insist is that it sends each element of the first set, to something in the second.

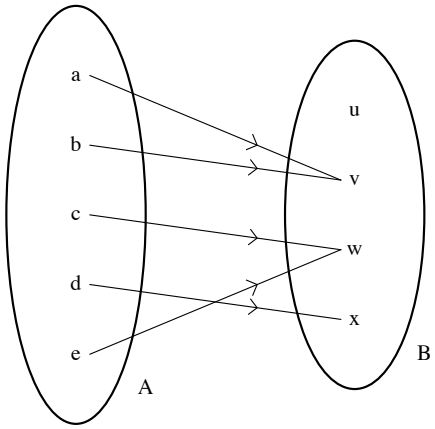
Thus, every function comes "equipped," with two sets: A, the set of points where the function is defined, and B, the set of possible values of the function. If the name of the function is f, we can draw attention to these sets by writing

$$f : A \to B.$$

The set A is called the **domain** of f: B is called its **codomain**. For each element x of the domain we write $f(x)$ for the place to which x is sent: the **image** of x under f.

The figure shows a schematic representation.

There is no stipulation that different elements in A end up at different places in B: two arrows can point to the same place. Nor is there any insistence that every member of B should lie at the point of an arrow: some members of B may be "redundant".



The only rule is that every element of the domain, should be the origin of precisely one arrow: for each x in A, there is an unique image, $f(x)$.

Example. *Let* $s : \mathbf{R} \to \mathbf{R}$ *defined by*

$$s(x) = x^2 \qquad \textit{for each real number } x. \tag{1}$$

What does equation (1) tell us? It says that whatever real number you pick, the image will be the square of that number. The function s, maps each real number to its square. s is the function that squares things.

The last sentence illustrates the real idea: **A function is what it does.** If you want to say what a particular function **is**, you have to say what it **does**, to each point of its domain.

Let's go back, for a moment, to the definition of s:

$$s(x) = x^2, \qquad \text{for each real number } x. \tag{2}$$

As I mentioned, this definition can be written: "s is the function that squares things." When you write the definition this way, you see that the "x" disappears. We don't need an x when we write the definition in words. This brings out the fact that the definition is **not** telling us anything about x: it is telling us about the function s. The statement says that, whatever number you put in to s, you will get out the square of that number. The statement

$$s(w) = w^2, \qquad \text{for each real number } w \tag{3}$$

says exactly the same as (2). Each of them tells us what the function s does: and therefore, what s is.

Now let's look at a rather different function. This time we will take the domain and codomain to be the finite set

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

consisting of just 4 numbers. Let p be the function given by

$$p(1) = 3, \qquad p(2) = 4, \qquad p(3) = 1, \qquad p(4) = 2.$$

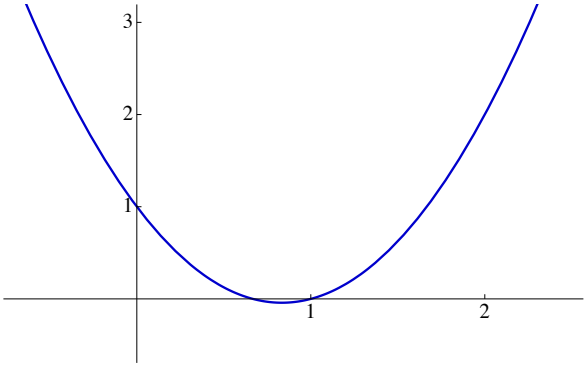
We have not written down any "formula" for the function. But we have certainly specified a perfectly good function: we know what p does to each element of its domain.

It is important to try to rid oneself of the feeling that a function has to be given by a "formula." There is nothing in the definition which requires that it should. All we know is that each point of the domain has to get sent to a point of the codomain. To illustrate this issue let me address the following question.

Find a function f defined for all real numbers with the property that $f(0) = 1$, $f(1) = 0$ and $f(2) = 2$.

So we want

$$\begin{aligned} f(0) &= 1 \\ f(1) &= 0 \\ f(2) &= 2 \end{aligned}$$



You might be tempted to try to find a formula for something like the function on the right. (And clearly I did just that in order to tell the computer how to draw it.)

Here is a simpler possibility. Let $f : \mathbf{R} \to \mathbf{R}$ be defined by

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x = 1 \\ 2 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Once you see this example you understand that the question was a stupid one. Of course there is a function taking the specified values: we just define it to take the specified values. Mathematicians don't spend their time answering stupid questions like this: but sometimes it is important to know that a question **is** stupid and to understand why.

Composition of functions

Functions can be combined in various ways. One of the most important is composition. If you have functions

$$g : A \to B$$

and

$$f : B \to C$$

then it is possible to build a new function which maps from A to C by first applying g and then applying f to the result. This new function sends an element x of A, to the element $f(g(x))$ of C. The new function is called the composition of f and g and is often written $f \circ g$.

So $f \circ g$ is our name for the function which first does g and then does f to the result. Suppose $f : \mathbf{R} \to \mathbf{R}$ and $g : \mathbf{R} \to \mathbf{R}$ are given by

$$\begin{aligned} f(t) &= t^{1/3} \\ g(x) &= 1 + x^2. \end{aligned}$$

What is the composition $f \circ g$? What does $f \circ g$ do to x?

$$f \circ g(x) = f(g(x)) = \left(1 + x^2\right)^{1/3}.$$

Algebra of functions

Suppose $f : \mathbf{R} \to \mathbf{R}$ and $g : \mathbf{R} \to \mathbf{R}$ are functions which, as is indicated, map real numbers to real numbers. We can form a new function

$$f + g$$

in the following way: we define $f + g$ by

$$(f + g)(x) = f(x) + g(x).$$

Remember that in order to say what $f + g$ is, we have to say what it does: and indeed we have. $f + g$ is the function which takes each number x to the sum of the values $f(x)$ and $g(x)$.

This addition of functions is already very familiar to you. You have often worked with polynomial functions

$$x \mapsto 1 + 3x + 2x^2$$

which are built by adding multiples of powers.
The point I want to make here, which you have perhaps glossed over in the past, is that we use the addition of ordinary real numbers,

$$f(x) + g(x)$$

to provide us with a way of adding **functions**.

In a similar way we can multiply functions. We can form a new function $f.g$ from two old ones by defining

$$(f.g)(x) = f(x)g(x).$$

Again you are very familiar with this process. The function $x \mapsto (1 + x)\sqrt{1 + x^2}$ is obtained by multiplying the functions $x \mapsto 1 + x$ and $x \mapsto \sqrt{1 + x^2}$.

Chapter 2. Polynomials

Polynomials make up the simplest class of functions which are varied enough to be interesting. A polynomial is a function such as

$$x \mapsto 1 - 3x + 4/7x^2 + 2\pi x^3$$

which acts on a number (x in this case) by adding up multiples of $1, x, x^2, \ldots$ up to some power x^n. The general example is thus

$$x \mapsto a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

where n is a non-negative integer and $a_0, a_1, \ldots$ are numbers. The highest power with a non-zero coefficient is called the **degree** of the polynomial.

We often use polynomials to approximate more complicated functions: they are varied enough to enable us to approximate pretty accurately, but simple enough for us to be able to calculate them efficiently.

The simplest polynomials are the **linear** functions such as

$$\begin{aligned} x &\mapsto 1 + 4x \\ x &\mapsto 2/7 - x. \end{aligned}$$

A linear polynomial is a function of the form

$$x \mapsto ax + b$$

where a and b are numbers. These functions are **called** linear because if you plot the graph $y = ax + b$ you get a straight line.

One useful property of linear functions is that you can easily solve equations of the form

$$ax + b = 0$$

(where $a \neq 0$): so you can easily find where a linear function takes the value 0. You may have exploited this fact already in Newton's method for approximating the solutions of equations.

The next simplest kind of polynomial is the quadratic polynomial

$$x \mapsto ax^2 + bx + c$$

for numbers a, b and c. If you plot $y = ax^2 + bx + c$ you get a parabola (as long as $a \neq 0$). Again we can solve equations of the form

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

to find x. If $a \neq 0$ and

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

then

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

However, in this case we need square roots to express the solution.

To use the formula in concrete cases, we need an efficient and reliable method for calculating square roots. (Your calculator is equipped with such a method, probably a relative of Newton's method.)

Zeros and factors

The most basic property of polynomials relates their zeros and their factors.

Theorem (**Zeros and factors of polynomials**)**.** *Let p be a polynomial and suppose that $p(\alpha) = 0$: that α is a zero of p. Then we can factorise p as a product of two polynomials*

$$p(x) = (x - \alpha)q(x)$$

for an appropriate polynomial q.

For example, if $p(x) = x^3 - 11x^2 + 7x + 3$ then you can check that $p(1) = 0$ and we can write

$$x^3 - 11x^2 + 7x + 3 = (x - 1)(x^2 - 10x - 3).$$

Once you have factorised the polynomial you can **immediately** see that $p(1) = 0$ because the first factor is 0 when $x = 1$. Putting it another way, if a polynomial is zero at α, it is zero for a very simple reason: there is a linear factor of the polynomial which is

"obviously" zero at α. This is not true in any normal sense for other functions. $\sin x$ is zero at $x = \pi$ but $\sin x$ is not "divisible" by $x - \pi$ in any algebraic sense.

How do we demonstrate that each zero corresponds to a linear factor? The argument we need is very close to what we actually do to find the factor. Suppose we want to factor the polynomial

$$x^3 - 11x^2 + 7x + 3 = (x - 1)q(x).$$

How do we do it? We have in our minds, or on a piece of scrap paper, a tentative product

$$x^3 - 11x^2 + 7x + 3 = (x - 1)(?x^2 + ?x + ?).$$

What do we need to start off with in the second factor if we are going to get x^3 in the product? Clearly we need an x^2:

$$x^3 - 11x^2 + 7x + 3 = (x - 1)(x^2 + ?x + ?).$$

Now what? Our product now looks like $(x - 1)x^2 = x^3 - x^2$ so we need to get an extra $-10x^2$ somehow, in order to end up with $-11x^2$ altogether. We can get this $-10x^2$ by putting in an extra $-10x$ into q:

$$x^3 - 11x^2 + 7x + 3 = (x - 1)(x^2 - 10x + ?).$$

Now we have a product $(x - 1)(x^2 - 10x) = x^3 - 11x^2 + 10x$, so we need a further $-3x$ to get $7x$. So we put -3 into q:

$$x^3 - 11x^2 + 7x + \mathbf{3} = (x - 1)(x^2 - 10x - 3).$$

Now we have no more question marks left to fill, so we just have to hope that the last term of p, the **3**, **automatically** works out right. Sure enough it does. Is this a miracle or can we see why?

If we were to start with any polynomial $p(x)$ and divide by $x - 1$, we could continue to choose terms in $q(x)$ until we had managed to get all the powers of x we wanted except the last one; the constant term. So we would have

$$p(x) = (x - 1).q(x) + r \tag{4}$$

where q is a polynomial and r is a number. Now suppose that $p(x) = 0$ when $x = 1$; ie. $p(1) = 0$. Then, if we put $x = 1$ into equation (4) we get

$$0 = 0.q(1) + r$$

and hence $r = 0$.

In other words the constant term works out automatically. Now if we put $r = 0$ back into (4) we get

$$p(x) = (x - 1).q(x)$$

which is the factorisation that we wanted. The division process just described is a special case of a more general fact.

Theorem (**Division of polynomials**)**.** *If p and d are polynomials then we can divide p by d in the following sense. We can write*

$$p(x) = d(x)q(x) + r(x)$$

where q, the "quotient" is a polynomial, (possibly 0) and r, the "remainder," is a polynomial whose degree is lower than the degree of d, the divisor.

For example if $p(x) = x^4 + x^3 + 2x^2 + 2x + 2$ and $d(x) = x^2 + x + 1$ then we can write,

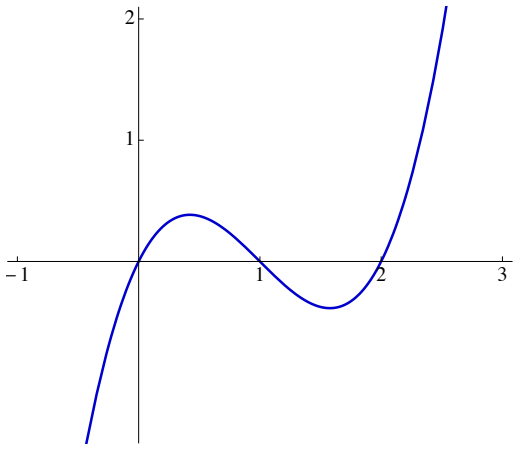
$$x^4 + x^3 + 2x^2 + 2x + 2 \quad = \quad (x^2 + x + 1)(x^2 + 1) + (x + 1).$$

If d is a **linear** polynomial, as in the $(x - 1)$ example, then the remainder r is a constant.

Consequences of factorisation

Once you have factored a polynomial (if you can), you have a pretty good idea of how the polynomial behaves. For example, if $p(x) = x(x - 1)(x - 2)$ then it's pretty easy to see that the graph of $y = p(x)$ looks something like this:

The function is zero at 0, 1 and 2, negative between 1 and 2, positive to the right of 2, and so on.



One of the most important consequences of factorisation is a bit more theoretical. If you multiply together some linear factors, let's say five of them,

$$(x-\alpha_1)(x-\alpha_2)(x-\alpha_3)(x-\alpha_4)(x-\alpha_5)$$

you get a polynomial of degree 5. So we can immediately see that a polynomial of degree 4 cannot have 5 **different** zeros since this would make it a product of 5 or more factors. In general:

Theorem (**Zeros and degree of a polynomial**)**.** *A non-zero polynomial of degree n cannot have more than n different zeros.*

This principle can be reinterpreted. Suppose that f and g are two polynomials of degree at most 5 and suppose we know of 6 different places where f and g take the same value: they agree at 6 different places. In other words I can find $x_0, x_1, \ldots, x_5$ with

$$f(x_0) = g(x_0), \quad f(x_1) = g(x_1), \quad \ldots, \quad f(x_5) = g(x_5).$$

Then the polynomial $f - g$ also has degree at most 5 but is zero at 6 different places. $(f-g)(x_0) = f(x_0) - g(x_0) = 0$ and so on.

So this polynomial must be "identically zero": it must be the zero polynomial. This in turn means that f and g must be the **same** as one another.

Thus we end up with the so-called uniqueness principle.

Theorem (**Polynomial uniqueness**)**.** *If two polynomials of degree at most n, agree at $n+1$ different points, then they are the same polynomial.*

It wasn't too hard to show that polynomials can't have **too many** zeros: we just observed that they obviously can't have too many factors. The question of whether they have **any** zeros is much more difficult. For a start, in order to guarantee that you can factor polynomials, you need to introduce complex numbers. This was not really done until the 18^{th} century. The first arguments which clearly showed that you **can** always factorise polynomials, did not appear until the 19^{th} century. This stunning fact is called the The Fundamental Theorem of Algebra. We shall discuss this remarkable fact later on.

Chapter 3. Summation

It often happens in mathematics that we need to refer to the sum of a large number of terms or perhaps of an indeterminate number of terms

$$1 + 2 + 3 + \ldots + n.$$

In these situations it is convenient to use the $\sum$ notation. Thus, the expression could be written as

$$\sum_{i=1}^{n} i.$$

In a similar way we use the expression

$$\sum_{i=1}^{n} x_i \tag{5}$$

to mean

$$x_1 + x_2 + x_3 + \ldots + x_n. \tag{6}$$

Before moving on to calculations, I want to draw your attention to the role of the letter i in each of these expressions. Whereas i appears in the expression (5) it does not appear in the expanded version (6). The letter i is merely a "dummy variable" which is being used to give us an instruction: "Add up the numbers x_1, x_2, $\ldots x_n$." This remark makes it clear that

$$\sum_{i=1}^{13} x_i$$

and

$$\sum_{k=1}^{13} x_k$$

are the same: they are different ways to write the same expression

$$x_1 + x_2 + x_3 + \ldots + x_{13}.$$

It is important to bear in mind the meaning of expressions such as these when using them.

In addition to being a shorthand, sigma notation helps to clarify sums when the terms are complicated.

$$\frac{1}{6}+\frac{1}{24}+\frac{1}{60}+\frac{1}{120}+\frac{1}{210}$$

means the same as

$$\sum_{n=1}^{5}\frac{1}{n(n+1)(n+2)}$$

but it is much easier to see what's going on in the second expression.

Geometric series

The most important example of summation is one that you have met. Suppose r is a number and we consider the sequence

$$(1, r, r^2, r^3, \ldots)$$

in which each number is r times the previous one. Such sequences occur naturally in the calculation of interest repayments and the study of radioactive decay for example. Can we find a simple expression for the sum of the terms of such a sequence

$$\sum_{k=0}^{n} r^k?$$

As we increase n, these sums look more and more complicated and it is less and less easy to see how they behave:

$$\begin{aligned}&1\\&1+r\\&1+r+r^2\\&1+r+r^2+r^3\\&\quad\vdots\end{aligned}$$

However there is a simple trick which enables us to rewrite these sums. If you multiply one of these expressions by $1-r$, almost everything cancels: for example,

$$\begin{aligned}(1-r)(1+r+r^2+r^3) &= 1 + r + r^2 + r^3\\ &\quad - r - r^2 - r^3 - r^4\\ &= 1 - r^4\end{aligned}$$

Now, as long as r is **not** equal to 1, you can divide by $1-r$ and get

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

In the same way, for each integer n,

$$1 + r + r^2 + \ldots + r^n = \frac{1 - r^{n+1}}{1 - r}.$$

It is clear that the method will not work if $r = 1$, but in this case, it is easy to write down the sum anyway:

$$1 + 1 + 1^2 + \ldots + 1^n = n + 1.$$

In all cases we have obtained an expression for the sum

$$\sum_{k=0}^{n} r^k$$

which has the advantage that it does not become more complicated as n increases.

Theorem (**Summation of a geometric progression**)**.** *If r is a number other than 1, and n is a positive integer, then,*

$$1 + r + r^2 + \ldots + r^n = \frac{1 - r^{n+1}}{1 - r}.$$

Notice that we could interpret this theorem as a statement about the factorisation of a polynomial:

$$1 - x^{n+1} = (1 - x)(1 + x + x^2 + \cdots + x^n).$$

Thus, the formula for the sum of a geometric progression is just a statement about factorising a special family of polynomials.

Infinite series

An extremely significant role is played in mathematics by infinite sums. Important functions like the exponential and trigonometric functions can be expressed as so-called power series

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots.$$

Power series are supposed to be like polynomials, but with infinitely many terms. It comes as no surprise that we have to be rather careful when we talk about infinite sums: they aren't quite as simple as finite sums.

Consider the sum,

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots.$$

Let's set about adding these terms one at a time. We get successively

$$1, \quad 1\tfrac{1}{2}, \quad 1\tfrac{3}{4}, \quad 1\tfrac{7}{8}, \dots$$

and it's pretty clear that this sequence of numbers is approaching 2. So it seems reasonable to say that the infinite sum is equal to 2.

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2. \tag{7}$$

On the other hand, suppose we look at the sum

$$1 + 2 + 4 + 8 + 16 + \dots.$$

If we keep adding more terms, the result just shoots off out of sight, so we have no way to make sense of this infinite sum.

Experience has shown us that the most convenient way to make sense of an infinite sum corresponds to this idea of watching what happens as we add successive terms. If we have a sequence of numbers

$$a_1, \quad a_2, \quad a_3, \quad a_4, \quad a_5, \dots$$

and if the sequence of **partial sums**

$$\begin{aligned} &a_1 \\ &a_1 + a_2 \\ &a_1 + a_2 + a_3 \\ &a_1 + a_2 + a_3 + a_4 \\ &\quad\vdots \end{aligned}$$

approaches a fixed number A then we say that the series $\sum_1^\infty a_k$ converges and

$$\sum_{k=1}^{\infty} a_k = A.$$

Using sigma notation we can rewrite the earlier equation (7) as

$$\sum_{k=0}^{\infty} 2^{-k} = 2.$$

It is a special case of a more general statement about geometric series. Suppose that $-1 < r < 1$. Recall that

$$1 + r + r^2 + \ldots + r^n = \frac{1 - r^{n+1}}{1 - r}.$$

Thus our formula for the sum of a geometric progression tells us the size of the partial sums of the infinite series

$$1 + r + r^2 + \ldots.$$

As n gets larger, the right hand side approaches

$$\frac{1}{1 - r}$$

because $r^{n+1} \to 0$. (This depends upon the fact that $-1 < r < 1$.)

So, according to our definition

$$1 + r + r^2 + r^3 + \ldots = \frac{1}{1 - r}. \tag{8}$$

Using $\sum$ notation, we can write this as

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1 - r}.$$

Equation (8) tells us how to sum infinite geometric series. These are some of the most important infinite series in mathematics. They crop up all over the place.

Theorem (**Geometric series**). *If* $-1 < r < 1$ *then*

$$1 + r + r^2 + r^3 + \ldots = \frac{1}{1 - r}.$$

Chapter 4. The binomial Theorem.

The Binomial Theorem concerns the expansions of the powers of a sum of two numbers (hence "binomial"). The powers 2 and 3 give rise to the familiar expansions

$$(x+y)^2 = x^2 + 2xy + y^2$$

and

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3.$$

More generally $(x+y)^n$ has an expansion of the form

$$x^n + ?x^{n-1}y + ?x^{n-2}y^2 + \ldots + ?xy^{n-1} + y^n$$

where the question-marks denote coefficients: the so-called binomial coefficients, which appear in what is known as Pascal's Triangle[1]:

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

The coefficients in the expansion of $(x+y)^n$ appear in the n^{th} row of the triangle, provided we count the single 1 at the top, as the zeroth row. Each row of the triangle is obtained from the previous row in a simple way: to obtain a given entry you add together the two entries above it.

[1]The triangle was known hundreds of years before Pascal in India, China, Persia and probably elsewhere.

To see why Pascal's triangle appears let's try to find the coefficients in the expansion of $(x+y)^4$ from those before. We can write $(x+y)^4$ in terms of the previous expansion in an obvious way.

$$(x+y)^4 = (x+y)(x+y)^3.$$

So, assuming we know the third row of the triangle, we get

$$(x+y)^4 = (x+y)(x^3 + 3x^2y + 3xy^2 + y^3).$$

If we multiply the second factor by x and by y in turn, we get two pieces which contribute to the total:

$$x^4 + 3x^3y + 3x^2y^2 + xy^3$$

$$x^3y + 3x^2y^2 + 3xy^3 + y^4$$

Each of these pieces has coefficients 1,3,3,1 like the 3^{rd} row, but the coefficients are attached to different combinations of x and y. When we combine the two pieces, we therefore add shifted copies of the 3^{rd} row to get

$$x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.$$

Schematically, we could represent it like this

$$\begin{array}{ccccc} 1 & 3 & 3 & 1 & \\ & 1 & 3 & 3 & 1 \\ \hline 1 & 4 & 6 & 4 & 1 \end{array}$$

It is no coincidence that the table above looks rather like a long multiplication. Try multiplying 1331 by 11 on paper and see what you write down. The same shifting operation accounts for the way we build all the other rows of Pascal's Triangle from the previous ones.

We denote the coefficients in the n^{th} row of Pascal's Triangle

$$\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}.$$

The expansion can thus be written

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

The coefficient $\binom{n}{k}$ is pronounced "en choose kay" for reasons described later. The Pascal triangle property of the coefficients can be written as follows

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

So far we have examined Pascal's Triangle and explained why it is built up in the way that it is, but we didn't really do anything more than reinterpret multiplication by $x+y$ in terms of addition. If we are to use binomial coefficients, it would be nice to have a simple formula which will enable us to calculate them without going through the whole business of building Pascal's Triangle.

Can we determine the numerical value of

$$\binom{20}{7},$$

for example, without finding 20 rows of the triangle? The answer is provided by the **Binomial Theorem** which tells us that the coefficients can be written in terms of factorials.

Theorem (**Binomial Theorem**)**.** *For any x and y,*

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

where for each n and $0 \leq k \leq n$

$$\binom{n}{k} = \frac{n!}{(n-k)!\, k!}.$$

The simplest way to prove the Binomial Theorem is to check algebraically that the expressions involving factorials do indeed satisfy the Pascal triangle property:

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Once you know that the factorial expressions do satisfy this equation you can deduce that they are the binomial coefficients by induction: assuming that they give the correct values in the n^{th} row we can conclude that they also give the correct values in the $(n+1)^{st}$ (and so we only need to check the first row in order to get the induction started).

This inductive argument gives a perfectly good proof of the Binomial Theorem but it isn't very illuminating. It just seems to work by magic without really explaining why the coefficients have the factorial formula or how someone came up with the formula in the first place. There is a much more instructive way to find the formulae.

If you were to multiply out the product

$$(x+y)^3 = (x+y)(x+y)(x+y)$$

all at once (instead of squaring $x+y$ and then multiplying again) you would write down each possible product made up of one factor from each bracket:

xxx	xxy	xyy	
	xyx	yxy	
	yxx	yyx	yyy

This procedure makes it clear that the coefficient of xy^2 is equal to 3 because there are three different ways of getting a product of one x and two y's.

The binomial coefficient $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ is thus the number of ways of selecting **two** factors from among the three: namely the two factors which contribute y to the product. It is automatically equal to the number of ways of choosing **one** factor from among the three (the factor which contributes the x).

In the same way, $\begin{pmatrix} n \\ k \end{pmatrix}$ is the number of different choices of k objects from a given n objects. This is why we call it "n choose k". For each k we have that

$$\begin{pmatrix} n \\ k \end{pmatrix} = \begin{pmatrix} n \\ n-k \end{pmatrix}$$

simply because, instead of focusing upon the k we choose, we could instead focus upon the $n-k$ that we don't.

Now we're ready to derive the factorial formula. Let's do a particular example, with $n=7$ and $k=4$. We want to calculate how many different 4-somes we can make out of 7 objects.

Suppose that the objects are numbered from 1 to 7. Imagine first that we write down all possible orderings of the 7 objects,

$$1234567$$
$$3146752$$
$$\vdots$$

There are $7! = 5040$ of them. Now from each ordering, select the first 4 objects. So from the second ordering above we would select the foursome $\{1,3,4,6\}$.

How many times will each 4-some get selected? The 4-some $\{1,3,4,6\}$ will be selected each time that our ordering has these four numbers distributed among the first four positions and the numbers 2, 5 and 7 distributed among the last three positions.

There are $4! \times 3!$ ways of doing this, since the numbers 1, 3, 4 and 6 can be ordered in 4! ways and the other three in 3! ways. So from our 7! orderings, each foursome will get selected $4! \times 3!$ times. This means that the number of different foursomes is

$$\frac{7!}{4!\,3!}$$

which is what we wanted to check.

The same argument works just as well for arbitrary n and k.

What does the Binomial Theorem really say?

There are many ways to state the Binomial Theorem. The important thing to understand is that the theorem links two apparently different questions and provides two different interpretations of the binomial coefficients.

Expanding $(x+y)^n$ tells us that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} :$$

selecting k-somes tells us that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} :$$

and the two things are actually the same.

Chapter 5. Linear equations and matrices.

Consider the following simple problem. You have at your disposal two commercially available mixtures of nitrate and phosphate fertilisers: call the mixtures X and Y. 1kg of each of these mixtures contains the following amounts of each fertiliser.

	X	Y
Nitrate	200g	100g
Phosphate	100g	200g

You wish to make up a bag containing 120g of N and 150g of P. Can you do it, and if so how much of each mixture do you need?

To solve the problem you let x be the number of kilos of X and y be the number of kilos of Y. Then you want to arrange that

$$\begin{aligned} 200x + 100y &= 120 \\ 100x + 200y &= 150. \end{aligned}$$

These are simultaneous linear equations which are easily solved to yield

$$\begin{aligned} x &= 0.3 \\ y &= 0.6 \end{aligned}$$

so that you need 300g of X and 600g of Y (and you can indeed achieve your aim).

Let us think for a moment why this problem gave rise to linear equations (which we have no difficulty solving). Why is it that x and y appear only in simple linear combinations? There are two closely related points involved:

- The amount of nitrate contributed by mixture X is proportional to the amount of mixture X present
- The total amount of nitrate is just the sum of the amount of nitrate coming from X and that coming from Y.

When you lump together some of X and some of Y you simply add the amounts of N contributed by each (and similarly the amounts of P).

This "additivity of lumping together" principle, holds in a wide variety of situations. For example, if you put together two lumps of a certain radioactive substance, then the number of atoms which decay in any given period is just the sum of the numbers from the two lumps. As you may know, radioactive decay is governed by a differential equation rather than by algebraic equations like those above. Nevertheless, we still refer to this differential equation as a **linear** differential equation because it exhibits the same additivity principle as the linear equations above.

Linear equations are the most useful in mathematics, for three reasons:

- They turn up naturally in many situations.
- Even when the true equations are non-linear, we can often approximate them by linear ones.
- Linear equations are usually much easier to solve than nonlinear ones.

(The difficulty of solving non-linear equations is well-illustrated by the fact that we cannot solve, precisely, the equations governing the motion of three heavy objects under gravity.)

Matrices

We have developed a useful shorthand for writing systems of linear equations using vectors and matrices. The system

$$\begin{aligned} 2x - y &= 5 \\ x + 2y &= 5. \end{aligned}$$

is written

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

Our choice about how to multiply vectors by matrices is made deliberately so as to correspond to the way in which linear equations are built from their coefficients. We **define**

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x - y \\ x + 2y \end{pmatrix}$$

precisely because we have found it useful in dealing with linear equations.

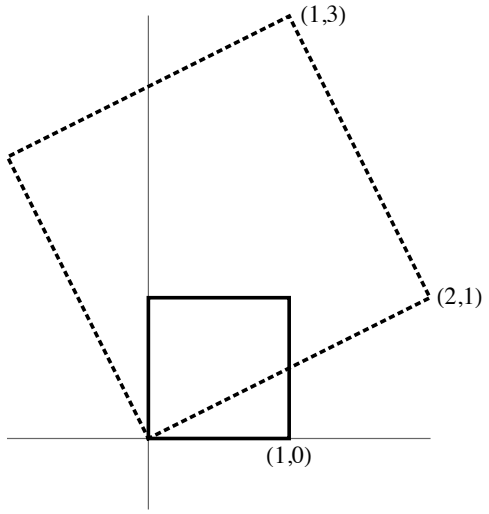
However, once we have decided how to multiply a vector by a matrix, we can study this operation in a slightly different way. Instead of trying to solve some particular set of equations we can think of multiplication by the matrix

$$\begin{pmatrix} 2 & -1 \\ 1 & 2 \end{pmatrix}$$

as giving rise to a transformation of the x, y-plane. The transformation is

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 2x - y \\ x + 2y \end{pmatrix}.$$

The diagram shows what the map does to the unit square. It is a rotation together with an enlargement.



In the same way, every 2×2 matrix gives rise to a transformation of the plane. The matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

takes the point (x, y) to the point $(ax + by, cx + dy)$.

One matrix is especially important: the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

leaves vectors unchanged. We call it the (2×2) identity matrix.

Since every matrix gives rise to a map, the obvious question is "Which maps arise from matrices?" Not all of them do: only those with certain special properties. Suppose I am thinking of a matrix and I tell you what it does to the points $(1, 0)$ and $(0, 1)$. For example,

$$\begin{aligned} (1,0) &\mapsto (3,7) \\ (0,1) &\mapsto (4,5) \end{aligned}$$

What is the matrix? Let's suppose it is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We can see that such a matrix takes the point $(1, 0)$ to the point (a, c). So it must be that, $a = 3$ and $c = 7$. Similarly, $b = 4$ and $d = 5$. So the matrix is

$$\begin{pmatrix} 3 & 4 \\ 7 & 5 \end{pmatrix}.$$

This tells us that any map that is given by a matrix is a very special map: once you know what it does to the points $(1, 0)$ and $(0, 1)$, you know what it does to everything. The reason for this is an additivity property for matrix multiplication. If M is a matrix and u and v are vectors then

$$M.(u + v) = M.u + M.v.$$

Multiplication by M preserves addition of vectors: if you add the vectors you add their images. Similarly, if you double a vector you double its image or if you multiply a vector by the number t, you multiply its image by t.

Every map of the plane that is given by a matrix multiplication has these properties. If M is a matrix, u and v are vectors and λ is a number then:

$$M.(u + v) = M.u + M.v$$

$$M.(\lambda u) = \lambda M.u.$$

A map with this property is called a **linear** map. All maps of the plane given by matrix multiplication are linear maps.

Have we answered our earlier question: "Which maps arise from matrices?" We can see that all matrix maps are linear: is it true that all linear maps are given by matrices? Yes indeed: and we have more or less demonstrated this, already. (See if you can give an explanation of this fact: that any map which is linear is given by a matrix.)

Now we know that the linear maps of the plane to itself are exactly those maps which are given by matrices. This should prompt us to ask the following question. If linear maps are just the same as matrix maps, why have we invented a fancy new name for them: "linear"? The reason is that linear maps turn up in many other situations (not just maps of the plane), where matrices do not seem to be remotely relevant.

Definition (**Linear maps**)**.** *A map M is linear if whenever u and v are vectors and λ is a number,*

$$M.(u+v) = M.u + M.v$$

$$M.(\lambda u) = \lambda M.u.$$

There are at least two operations that you have met, other than matrix multiplication, which possess properties like these: differentiation and integration. For example, when you differentiate the sum of two functions, you can do it by differentiating each of them and then adding the results. Differentiation and integration are linear maps: except that these maps act, not on vectors or points, but on functions.

Chapter 6. Matrix multiplication.

In the last chapter I talked about transformations of the plane which are given by matrix multiplication. I glossed over the question of why these transformations of the plane are interesting or useful. We shall see that among other things, rotations about the origin are linear maps. Rotations are certainly very important: we need to understand rotations in order to be able to relate observations made by one person to those made by someone else, who is facing in a different direction.

For the moment, I want to take for granted that linear maps **are** useful and to talk a bit more about some of their properties. One of the first things we always do when we have thought of some kind of mathematical operation is to see what happens if we do one after another (if we can). Suppose we have a pair of 2×2 matrices M and N. If we transform the plane using M and then transform using N we will end up with some overall transformation.

There is a question which is practically screaming to be asked. Is this new transformation, of the same kind as before: is it given by a matrix? Or have we found a new **kind** of transformation? If the new transformation is given by a matrix, which matrix is it: how does it relate to M and N? Even if you didn't know the answer before, you would probably have guessed what it is.

Let's have an example. Suppose that

$$M = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \text{ and } N = \begin{pmatrix} 3 & 2 \\ 1 & 5 \end{pmatrix}.$$

After multiplication by M, the vector $\begin{pmatrix} x \\ y \end{pmatrix}$ ends up at

$$\begin{pmatrix} 2x + y \\ 3x + 4y \end{pmatrix}.$$

When you multiply this by N you get

$$\begin{pmatrix} 3 & 2 \\ 1 & 5 \end{pmatrix} \cdot \begin{pmatrix} 2x + y \\ 3x + 4y \end{pmatrix} = \begin{pmatrix} 12x + 11y \\ 17x + 21y \end{pmatrix}.$$

This means that the map **is** given by a matrix; namely

$$\begin{pmatrix} 12 & 11 \\ 17 & 21 \end{pmatrix}.$$

How is this matrix related to M and N? As you will have guessed, this new matrix is the product $N.M$.

$$\begin{pmatrix} 3 & 2 \\ 1 & 5 \end{pmatrix} . \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 12 & 11 \\ 17 & 21 \end{pmatrix}.$$

(Note that N is written before M in the product, even though the map it corresponds to consists of "first do M and then do N". This inversion of the order results from the fact that we write maps on the left of the vectors to which they apply.) In general, whenever we apply a matrix map M followed by a matrix map N we get a matrix map given by the matrix product $N.M$.

Theorem (**The reason for matrix multiplication**)**.** *Composition of matrix maps corresponds to multiplication of matrices.*

You might like to check this for yourself by demonstrating it for an arbitrary pair of matrices. Owing to the fact that you have met matrix multiplication before, I chose simply to **tell** you that composition of maps corresponds to matrix multiplication. But this tends to hide the crucial point: the **reason** that we multiply matrices the way we do, (and not some other way) is that we want to talk about combining transformations, one after another. Funny rules for combining bracketed arrangements of numbers are not our real aim. Our **real** aim is to describe what happens when we combine matrix transformations of the plane (or higher-dimensional space). Matrices and matrix multiplication are the tools that do the job.

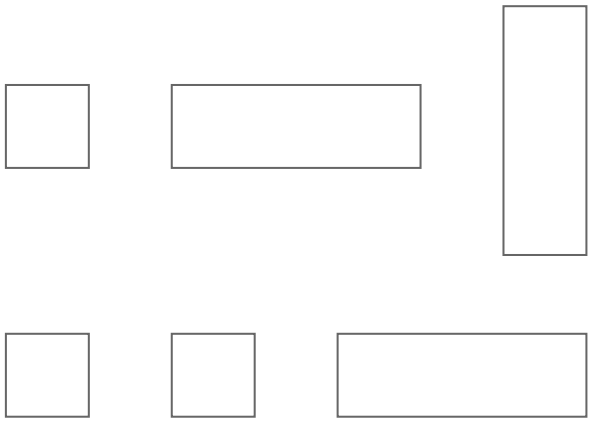
When you first came across matrix multiplication you may have been a bit distressed by the fact that it is not what we call "commutative"; the order of the matrices makes a difference. The two products below do not produce the same result:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} . \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} . \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

The two matrices we are multiplying correspond to two transformations

- A quarter turn anticlockwise
- A stretch by a factor of 3 in the x direction

Think about what happens to a square if you first stretch in the x direction and then rotate through 90^o; or if you first rotate and then stretch **in the x direction**.



You shouldn't be surprised that matrix multiplication doesn't commute: transformations don't commute.

Matrix inverses

Once we know how to combine linear maps, we can ask how to invert (or undo) them. If you give me a linear transformation of the plane, can I find a linear transformation which returns every point to its original position? Let's try it for the matrix

$$M = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix}.$$

Recall the "identity" matrix

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which we saw in the last chapter, and which corresponds to the transformation which leaves all points where they are. Our task is to find a matrix, let's call it

$$\begin{pmatrix} u & v \\ x & y \end{pmatrix}$$

with the property that

$$\begin{pmatrix} u & v \\ x & y \end{pmatrix} . \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

so that the combined transformations yield the identity map which doesn't move anything. (Now we see why something as boring as the identity map plays an important role.)

The matrix on the left is

$$\begin{pmatrix} 2u+3v & u+4v \\ 2x+3y & x+4y \end{pmatrix} .$$

So we would like to satisfy the following linear equations

$$\begin{aligned} 2u + 3v &= 1 \\ u + 4v &= 0 \\ 2x + 3y &= 0 \\ x + 4y &= 1. \end{aligned}$$

Although these are 4 equations in 4 variables, they separate automatically into two sets of 2 equations. When you solve them you get

$$u = \frac{4}{5}, \qquad v = -\frac{1}{5}, \qquad x = -\frac{3}{5}, \qquad y = \frac{2}{5}$$

so that the matrix we are looking for is

$$\begin{pmatrix} \frac{4}{5} & -\frac{1}{5} \\ -\frac{3}{5} & \frac{2}{5} \end{pmatrix} .$$

This matrix is called the inverse (or left inverse) of M and we usually write it M^{-1}. Notice that once you have found the inverse, you can check that $M^{-1}.M = I$ very easily; much more easily than you could **find** M^{-1}.

We constructed the matrix M^{-1} above in order to satisfy $M^{-1}.M = I$. However we get an unexpected (and extremely important) bonus. Since we are talking about matrices, it isn't obvious what will happen when we multiply M and M^{-1} in the opposite order, $M.M^{-1}$. It turns out that we **automatically** get the identity, I. In other words, if M^{-1} is the left inverse of M then it is also the right inverse of M. It is for this reason that we just refer to M^{-1} as **the** inverse of M and we think of M and M^{-1} as inverses of one another.

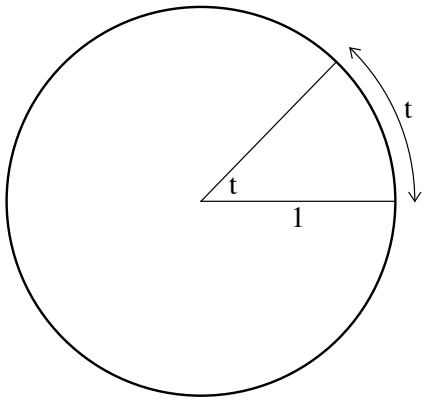
Theorem (**Left and right inverses**)**.** *If two square matrices multiply to give the identity in one order, then they automatically do the same, in the other order.*

For 2×2 matrices you can check this by hand. For general $n \times n$ matrices some theory is needed to handle the same question.

Chapter 7. The trigonometric functions

The Babylonians chose to measure angle in degrees, but this is a very arbitrary measure, which is unsuitable for most mathematical purposes. The most natural way to measure angle is in radians. Let's recall what that means?

We draw the circle of radius 1 and then for a given angle we use the length of the circular arc that it spans as the measure of the angle. If the circular arc bounding the sector has length t, then we say that the angle of the sector is t.

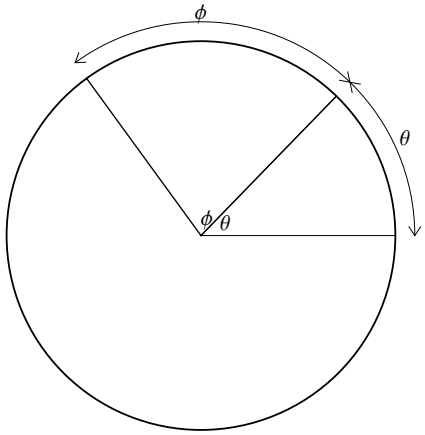


Often when one first meets radians one feels that they have a rather mysterious quality. They don't. We just use the **length** of the circular arc, as a way to measure the **angle**. Nothing could be simpler.

It is immediately clear that a full turn is angle 2π, a half turn is π and so on. Naturally, if we have a circle whose radius is different from 1, we have to calculate angle by taking the **ratio** of the length of the arc, to the radius.

Let us start by noticing that radians have one property, (in common with degrees), which is vital.

When you rotate through one angle and then through another, the total angle you get is the sum of the original ones. This is clear because the same statement holds for lengths: lengths add up in the obvious way.

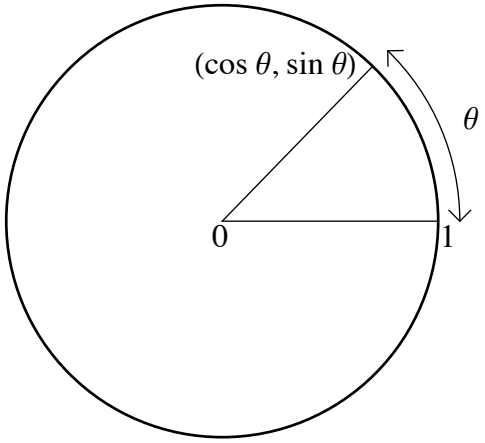


Later we shall discuss the crucial property of radians which is **not** possessed by degrees, (nor by **any** measure of angle other than radians).

Cosine and Sine

Frequently in mathematics, we need to be able to relate circular to linear motion. For this we need the trigonometric functions. Again, draw a circle of radius 1.

Mark a point on the circle at an angle θ, measured from the horizontal. The x and y coordinates of this point are the numbers $\cos\theta$ and $\sin\theta$.



Again, we can express each of these numbers as a **ratio** of lengths, when we look at a circle whose radius is different from 1.

There is a tendency to think of trigonometry as "the study of right-angled triangles." Such a view makes trigonometry seem absurdly specialised and rather arbitrary. In reality, the importance of the trigonometric functions stems from the importance of the circle. Right-angled triangles come into the picture because we use axes that are at right-angles to one another.

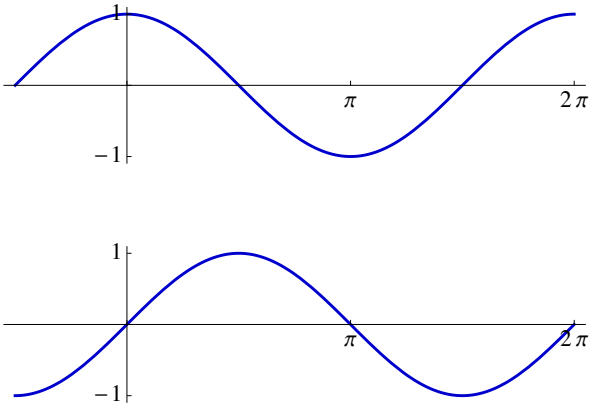
The geometric definition of cos and sin makes the following standard properties obvious. For any θ,

$$\cos^2\theta + \sin^2\theta = 1.$$

(This is Pythagoras' Theorem.) Both functions cos and sin repeat themselves after an angle 2π: they are periodic with period 2π. cos is an even function while sin is an odd function and for any θ,

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

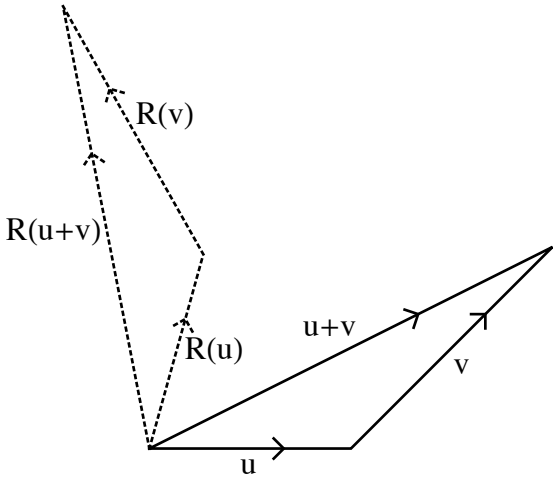
The graphs of cos and sin are the following familiar pictures.



There is no simple way to calculate cos and sin for a general angle. In order to approximate them, we need to use the techniques of calculus, just as we do for the exponential and logarithmic functions. However, for certain special angles, we can find the values of cos and sin using simple geometric arguments. For integral multiples of $\frac{\pi}{2}$ we can easily see that both functions are ± 1 or 0 and can easily check which. The cosines and sines of $\frac{\pi}{3}$ and $\frac{\pi}{4}$ are also not too hard to calculate. In the examples, you are asked to find $\cos\frac{\pi}{8}$, $\cos\frac{\pi}{12}$ and $\cos\frac{2\pi}{5}$. You could continue to think up new angles for which you can obtain exact expressions, but this is not an efficient way to calculate for a general angle. We shall discuss a more systematic approach in later chapters. I want to devote the rest of this chapter to the addition formulae for cosine and sine.

The addition formulae

In an earlier chapter I mentioned that rotations about the origin are linear maps: they are given by matrices. Let's quickly convince ourselves of that. What we want to know is that if u and v are vectors and R is a rotation, then $R(u+v) = R(u) + R(v)$. The question is, do we get the same thing if we rotate the sum of $u+v$, as if we add together the rotated vectors $R(u)$ and $R(v)$. The picture tells the story:



As we rotate the vectors u and v we rotate their sum because we can represent the sum as the third side of a triangle with u and v on the other sides.

So now we know that rotations about the origin are linear maps. What are their matrices? Let's find the matrix that gives the rotation through an angle θ anticlockwise? Suppose it is the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We want to find the numbers a, b, c and d. As we did in the last chapter, we can look at what the matrix does to a special choice of vectors:

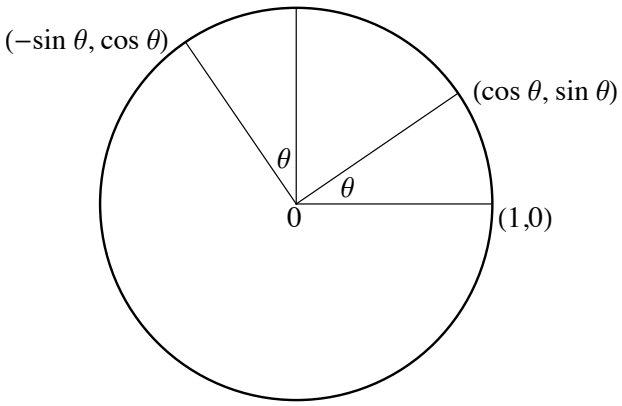
$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The images of these two vectors are

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ c \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix}.\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}$$

respectively.

What are the images **supposed** to be?



The picture shows that after rotation through an angle θ,

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

So the matrix of a rotation about the origin through the angle θ is

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

These matrices (for different θ) are used in many ways throughout physics and engineering: whenever you programme a computer to handle rotations you need to give it these matrices. We don't have any other quantitative description of a rotation.

The first use to which we will put these matrices is to discover the addition formulae for cos and sin. You have met these formulae and begun to realise that they turn up a lot. The reason is that addition of angle is a natural thing to do: as we saw, it corresponds to following one rotation by another. Rotation through ϕ, followed by rotation through θ produces rotation through $\theta + \phi$.
The matrices for rotations through θ and ϕ are

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}.$$

The matrix for the sum is of course

$$\begin{pmatrix} \cos(\theta+\phi) & -\sin(\theta+\phi) \\ \sin(\theta+\phi) & \cos(\theta+\phi) \end{pmatrix}.$$

But we saw yesterday that the matrix for the composition, one map followed by another, is obtained by multiplying the matrices of the separate maps. So the last matrix is equal to the product

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} . \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$$

which is

$$\begin{pmatrix} \cos\theta\cos\phi - \sin\theta\sin\phi & -\cos\theta\sin\phi - \sin\theta\cos\phi \\ \sin\theta\cos\phi + \cos\theta\sin\phi & -\sin\theta\sin\phi + \cos\theta\cos\phi \end{pmatrix} .$$

By equating the two expressions for the "$\theta + \phi$" matrix we can immediately read off the standard addition formulae.

Theorem (**The addition formulae**)**.**

$$\cos(\theta + \phi) = \cos\theta\cos\phi - \sin\theta\sin\phi$$

$$\sin(\theta + \phi) = \sin\theta\cos\phi + \cos\theta\sin\phi.$$

If you haven't ever tried it, you might find it instructive to derive the addition formulae using "old-fashioned" constructive geometry: it isn't **too** hard: but it isn't too pleasant either.

Once we have the addition formulae for cos and sin we can derive the formula for tan: try it. Naturally we can also derive the double angle formulae: for example

$$\begin{aligned} \cos 2\theta &= \cos(\theta + \theta) \\ &= \cos\theta\cos\theta - \sin\theta\sin\theta \\ &= \cos^2\theta - \sin^2\theta \\ &= 2\cos^2\theta - 1. \end{aligned}$$

In a few chapters' time, we will look at the addition formulae in the context of complex numbers and see an alternative way to understand (and hence remember) them.

Chapter 8. Differentiation.

Ancient Greek mathematicians devoted enormous effort to calculating the volumes or areas of different solids or shapes. Antiquity's greatest genius, Archimedes, was so delighted by his discovery of the areas of segments of the sphere that he is said to have had a diagram of it inscribed on his tomb. Today, such problems are, almost literally, child's play.

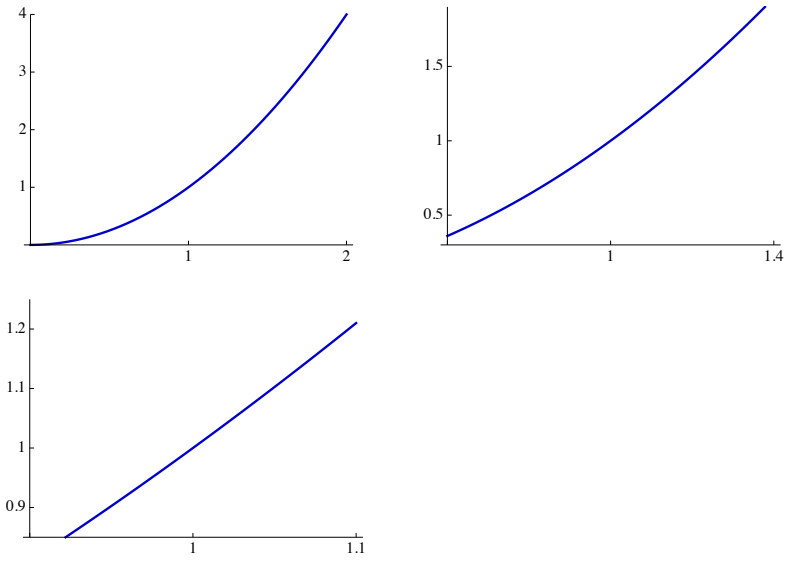
In the middle of the seventeenth century, with an extraordinary burst of activity, Isaac Newton not only explained the motion of the planets and the behaviour of falling bodies on the basis of a single principle of gravitation, but also created the mathematical tool known as calculus. By relating the two operations that we now call differentiation and integration, and by inventing a systematic method for carrying out the former, Newton practically restarted mathematics. Ever since Newton, mathematical knowledge has grown at a rate incomparably greater than it did at any time before him. In this and the next two chapters I want to recall the major ideas involved in differentiation; in finding derivatives. The aim is to study rates of change of numerical quantities, (relative to one another).

The concept of the derivative

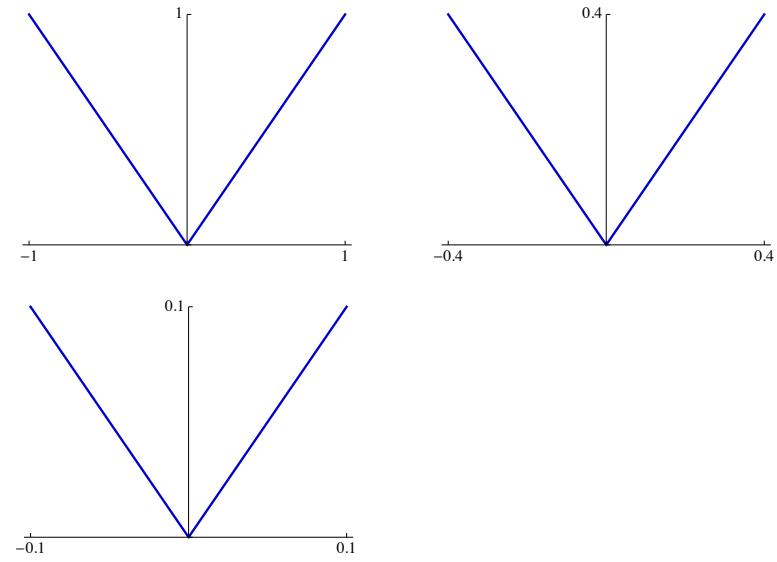
As you know the earth is round (more or less). However, if you look out of the window, it looks flat. The more closely you look at a curve, such as the curve of the earth's surface, the flatter it looks.

Let's look at a mathematical curve which is easier to describe than the earth's surface; for example the curve $y = s(x) = x^2$. I have drawn three graphs of this function for different ranges of the x-coordinate, near to the point $(1, 1)$ on the curve. The ranges are,

$$\begin{array}{rcccl} 0 & < & x & < & 2 \\ 0.6 & < & x & < & 1.4 \quad \text{and} \\ 0.9 & < & x & < & 1.1 \end{array}$$

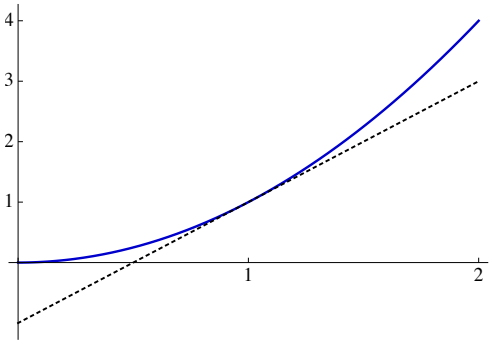


The graphs for the shorter ranges have been blown up to fill the same width of picture. What you notice is that smaller pieces of the curve look flatter **even when you blow them up to the same size**. Contrast this with what happens when you focus onto a sharp corner:



In this case the corners look exactly the same as you blow them up: there is no flattening. However in the case of the smooth curve, as you focus on smaller and smaller pieces, your graph looks more and more like a straight line, even when you scale up the picture.

Not only does the curve look more and more straight as you focus on smaller pieces: there is a particular straight line that the curve is copying: the tangent line. This straight line just brushes the curve at a particular point.



How can we find which line is the tangent? One thing we know about the tangent is obvious: it passes through the point. The problem is to find its slope. The tangent line to a graph, at a point, is supposed to provide a good description of how the function is behaving near that point. Can we understand **algebraically** how the function $x \mapsto x^2$ is behaving near $x = 1$?

If x is close to 1, then I can write x as $1 + h$ for some small number h. The value of the function at x is thus

$$x^2 = (1 + h)^2 = 1 + 2h + h^2.$$

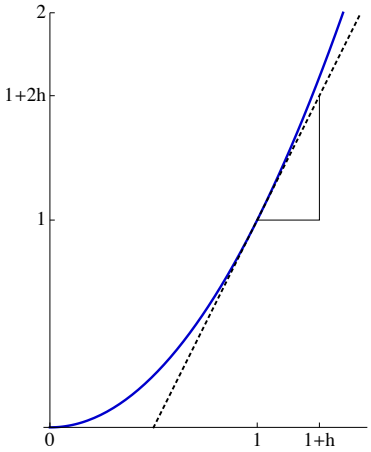
The first thing you notice about this expression is that if h is small, then the function is fairly close to 1. If x is close to 1 then x^2 is close to 1. That hardly comes as a surprise. Much more important however, is that if h is a small number, then h^2 is extremely small. For example, if $h = 0.01$ then $h^2 = 0.0001$ which is much smaller.

Thus, if x is close to 1, so that h is a small number, the value of $x^2 = 1 + 2h + h^2$ is very close to

$$1 + 2h.$$

So we can see that if we increase x just a bit, from 1 to $1 + h$, then we increase the value of x^2 by about twice as much; from 1 to $1 + 2h$. The instantaneous rate of change of the function, at $x = 1$, is 2: as we increase x from 1 to $1 + h$ the function increases twice as fast.

This ratio 2, is the slope of the tangent line that we wanted to find. As you know we call it the derivative of the function at $x = 1$.



Let's try to find the derivative of the function $x \mapsto x^2$ at other places. Let's try an arbitrary value $x = c$. How does the function behave near c? Put $x = c + h$ and evaluate the function at x:

$$x^2 = (c + h)^2 = c^2 + 2ch + h^2.$$

Near c we can see that the function is well approximated by the linear function $c + h \mapsto c^2 + 2ch$. As we increase x from c to $c + h$, the value of x^2 increases by about $2c$ times as much; from c^2 to $c^2 + 2ch$. So the derivative of the function at c is $2c$.

Let's do another example. Suppose $q(x) = x^3$ and we wish to calculate the derivative of q at an arbitrary number x. We choose a number nearby $x + h$ and evaluate $q(x + h)$.

$$q(x + h) = (x + h)^3 = x^3 + 3x^2h + 3xh^2 + h^3$$

If h is very small, the last expression differs from $q(x) = x^3$ by about $3x^2h$. So the slope of the curve at the point (x, x^3) is $3x^2$.

Can we give a more systematic description of what we just did? We evaluated q at x and $x + h$ and looked at the difference between them:

$$\begin{aligned} q(x + h) - q(x) &= x^3 + 3x^2h + 3xh^2 + h^3 - x^3 \\ &= 3x^2h + 3xh^2 + h^3 \end{aligned}$$

Then we picked out "the coefficient of h" in this difference; namely $3x^2$. We can do the "picking out" algebraically: divide the difference by h to get $3x^2 + 3xh + h^2$ and now see what happens to this as h approaches zero. Clearly as h approaches 0, the expression $3x^2 + 3xh + h^2$ approaches $3x^2$.

Our procedure thus consisted of forming the quotient

$$\frac{q(x+h) - q(x)}{h}$$

and then asking what happens to it, as h approaches zero. We can try to repeat this procedure for any function. Let's call it f. We evaluate the function at c and at $c + h$. We take the difference, and divide by h:

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

Now we ask, does this ratio approach a limiting value, as h approaches 0?

Let's do one more example. Suppose $r(x) = \frac{1}{x}$. The value of r at x is

$$r(x) = \frac{1}{x}.$$

The value at $x + h$ is

$$\frac{1}{x+h}.$$

Now it isn't so easy to see what to do with this. We can't "expand" this reciprocal in the same way as a square or a cube. We have to use our more formal description of the procedure. We write down the ratio

$$\frac{r(x+h) - r(x)}{h} = \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \frac{1}{h}\left(\frac{1}{x+h} - \frac{1}{x}\right).$$

In order to see what happens to this expression as h approaches 0, our only hope is to simplify it, by combining the fractions in the bracket. We get

$$\begin{aligned} \frac{1}{h}\left(\frac{x - (x+h)}{(x+h)x}\right) &= \frac{1}{h}\left(\frac{-h}{(x+h)x}\right) \\ &= \frac{-1}{(x+h)x}. \end{aligned}$$

Thus

$$\frac{r(x+h)-r(x)}{h} = \frac{-1}{(x+h)x}.$$

Now we are home. As h approaches 0, $x+h$ approaches x and so this expression approaches

$$\frac{-1}{x^2}$$

(at least provided $x \neq 0$). So, the derivative of the function $x \mapsto \frac{1}{x}$ at x is $\frac{-1}{x^2}$.

Let us summarise. Let f be a function defined near x. If the quotient

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

approaches some limiting value as h approaches 0, we say that f has derivative at x, equal to this value. We denote it

$$f'(x).$$

Near c the function f is approximated by a linear function

$$f(x+h) \approx f(x) + f'(x).h$$

whose slope is equal to the derivative.

We now understand what we mean by the derivative and we know how to calculate derivatives for some special functions. We could continue, adding to our repertoire of differentiable functions: but we would go nuts. Instead, we need a machine to do the work for us. The point is that the functions we use in mathematics are built up in fairly simple ways from a few basic functions. The machine tells us how to differentiate complicated functions, once we already know how to differentiate the pieces out of which they are built.

Chapter 9. The differentiation machine: exponentials.

The machine has three basic pieces:

- the sum rule
- the product rule
- the chain rule

and special attachments to differentiate particular types of function:

- polynomials
- exponentials
- trigonometric functions.

The sum rule expresses the derivative of a sum of functions in terms of the derivatives of those functions. The product rule tells us how to differentiate a product of functions as long as we already know how to differentiate the factors. If f and g are differentiable functions then

$$(f+g)' = f' + g'$$

and

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

Once we know the derivatives of the two very simple functions $x \mapsto 1$ and $x \mapsto x$ we can use the product rule to find the derivatives of all the monomial functions

$$\begin{array}{ccc} x & \mapsto & x^2 \\ x & \mapsto & x^3 \\ & \vdots & \end{array}$$

In other words we can find the derivatives of all functions $x \mapsto x^n$ where n is a positive integer, since each of these is built up from the function $x \mapsto x$ by multiplication. Using the sum rule as well, we can now find the derivatives of all polynomials.

The third part of the differentiation machine is the chain rule: it tells us how to express the derivative of a composition of functions in terms of the derivatives of those functions. Suppose f and g are functions and we form the composition

$$x \mapsto f(g(x)).$$

(We usually write this function as $f \circ g$.) Choose a number x where we want to calculate the rate of change of $f \circ g$. The value of the function at x depends upon the value of g at x and the value of f at $g(x)$. Similarly, the values of $f \circ g$ **near** x depend upon the values of g near x and the values of f near $g(x)$. So it is not surprising that the rate of change of the composition, depends upon the derivative $g'(x)$ of g at cx, and also the derivative $f'(g(cx))$ of f at the point $g(x)$.

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

Let's have an example. Let g and f be given by $g(x) = 1 + x^2$ and $f(t) = \frac{1}{t}$ respectively. Then

$$g'(x) = 2x$$

while

$$f'(t) = \frac{-1}{t^2}$$

Hence

$$f'(g(x)) \quad = \quad \frac{-1}{g(x)^2} = \frac{-1}{(1+x^2)^2}$$

So

$$\frac{d}{dx}\frac{1}{1+x^2} = \frac{-2x}{(1+x^2)^2}.$$

Derivatives of inverses

Inverse functions play a rather important role in mathematics: the logarithm as the inverse of the exponential, the inverse sine function and square roots for example. So it natural to ask whether we can find the derivative of the inverse of a function whose derivative we already know. In order to find out, we need some sort of idea.

Suppose f is differentiable and g is its inverse. Since f and g are inverses of one another, if you do g followed by f you get back where you started:

$$f(g(x)) = x$$

for each x in the domain of g. The right hand side of this equation, we know how to differentiate explicitly: we get 1. The left hand side, we can differentiate using the chain rule, to get

$$f'(g(x)).g'(x).$$

Thus we have found that for each x

$$f'(g(x)).g'(x) = 1$$

and hence

$$g'(x) = \frac{1}{f'(g(x))}.$$

The upshot is the following principle If f and g are inverses then

$$g'(x) = \frac{1}{f'(g(x))}.$$

Let's have an example: let's look at the function g given by

$$g(x) = x^{\frac{1}{2}} = \sqrt{x}.$$

The thing we know about g is that its inverse is the function

$$t \mapsto t^2$$

that squares things. As above we'll call this inverse function f. So $f(t) = t^2$.

Now, since $f(t) = t^2$ we know that $f'(t) = 2t$. So

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{2g(x)} = \frac{1}{2x^{1/2}} = 1/2x^{-1/2}.$$

Thus we get the familiar derivative for the function $x \mapsto x^{1/2}$ by using our knowledge of the derivative of its inverse.

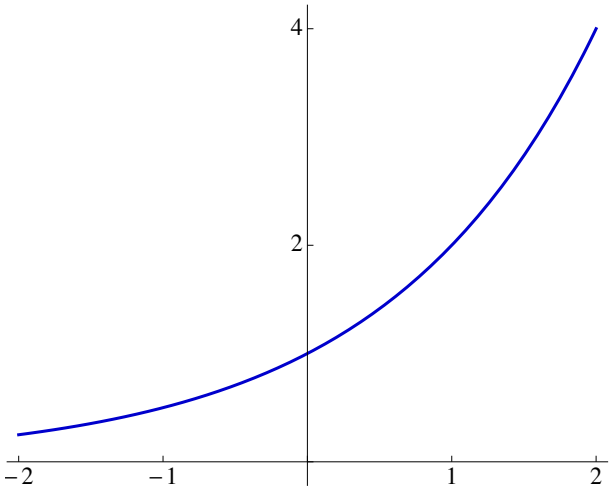
The exponential function

There are many functions in which the variable is the exponent: $x \mapsto 2^x$, $x \mapsto (5.267)^x$ and so on. These functions have certain things in common: most especially, if f is an exponential function then

$$f(x + y) = f(x).f(y)$$

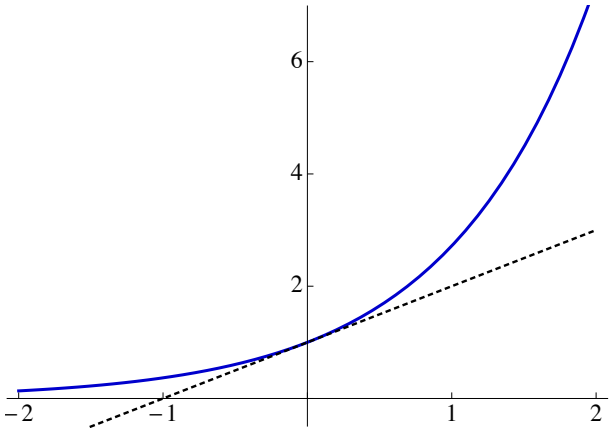
Exponentials turn addition into multiplication.

But there is one exponential function which is special: and because it is so special we call it **the** exponential function. Let's look at the graph of $y = 2^x$.



If you try to estimate the slope of this curve at the point $(0, 1)$ you will find that it comes out to be about 0.693... This is not terribly convenient. If you try the same thing with the graph of $y = 3^x$ you get a slope of about 1.099... Again this is not very helpful.

However, there is a number, which we call e, between 2 and 3 with the property that the curve $y = e^x$ has a slope equal to 1 at $(0, 1)$. The curve and its tangent are shown below.



Why is it vital that we should have a special exponential function whose derivative at 0

is 1? As you know

$$\frac{d}{dx}e^x = e^x$$

the exponential function is its own derivative. For this to be true we need that the slope at 0 is equal to 1: because $e^0 = 1$.

Our choice $f(x) = e^x$ of the exponential function is intended to guarantee that the exponential is its own derivative. In the first instance it just guarantees that e^x has the correct slope at $x = 0$. What this means is that as $h \to 0$

$$\frac{f(0+h) - f(0)}{h} = \frac{e^h - 1}{h} \to 1. \tag{9}$$

Now let's calculate the derivative everywhere else. We want to know what happens to the ratio

$$\frac{f(x+h) - f(x)}{h} = \frac{e^{x+h} - e^x}{h}$$

as $h \to 0$. But this is

$$\frac{e^x.e^h - e^x}{h} = e^x\ \frac{e^h - 1}{h} \to e^x.$$

Hence if $f(x) = e^x$ then $f'(x) = e^x$. The exponential function is thus its own derivative. This is what makes the exponential function useful in calculus.

The above treatment of the exponential function was a bit cavalier: I made no attempt to justify my claim that the number e exists, although this claim is intuitively very reasonable. I certainly made very little attempt to calculate e: could it be $\frac{19}{7}$ perhaps? We shall return to this later.

Chapter 10. The differentiation machine: the trigonometric functions.

In Chapter 7 we recalled that we measure angle in radians and found the addition formulae for sine and cosine. My aim in this chapter is to find the derivatives of the trigonometric functions and in doing so explain why radians are so important.

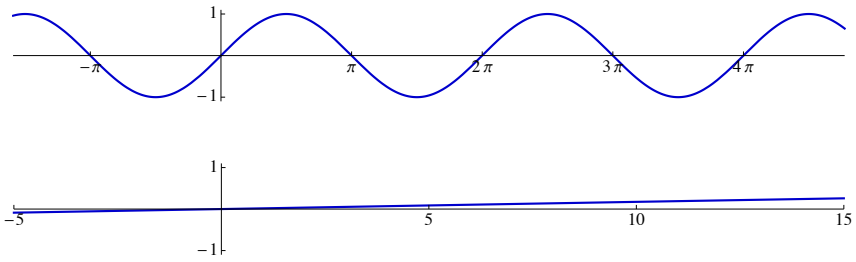
The two graphs below show the two functions

$$x \mapsto \sin x$$

and

$$x \mapsto \sin x^\circ.$$

The two graphs are drawn on the same sized axes and in each case the horizontal and vertical axes have the same scales.



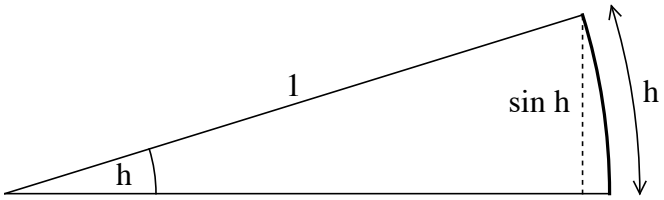
Notice that for the first function I did not specify the units of angle: radians are taken for granted. For the second function, angle is measured in degrees. Clearly the two graphs look very different. The characteristic wiggly behaviour of sine doesn't show up at all on the second graph because the horizontal scale only goes up to 15°. More significantly, the first graph has a slope at the origin that looks like a reasonably sized number; something like 1; not something very large or very small, like 1000 or 0.021.

In fact the slope of the second graph at the origin is 0.0174.... whereas the slope of the first graph is 1: the graph is instantaneously rising at 45°, at the origin. Let's see why we get a slope of 1 for the graph of $y = \sin x$ at 0, **as long as we measure angle in radians**.

We want to show that the curve $y = \sin x$ looks like the line $y = x$ when x is a small number: that the tangent line to to the sine graph at $x = 0$ is the line $y = x$. So we want to verify that when x is small, the ratio

$$\frac{\sin x}{x}$$

is close to 1. For this we need to consider a small sector of a circle of radius 1.



It is intuitively clear that for small angles the ratio of the arc to the height is very close to 1. Remember that the length of the arc is the relevant quantity, precisely because we are using radians to measure angle. (You may object that this "intuitive" argument is not very precise, even though it is pretty convincing. In fact the only real obstacle to our making it precise, is that we haven't been **absolutely** precise about what we mean by the length of a curve.)

Theorem (**The slope of sine at 0**)**.** *As x approaches 0, the ratio*

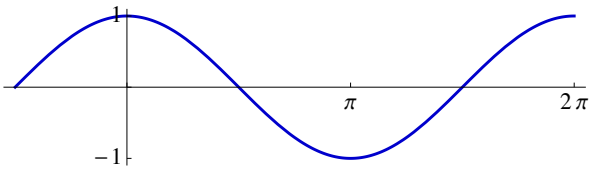
$$\frac{\sin x}{x}$$

approaches 1, **provided we measure angle in radians**. *Consequently, at 0, the slope of the curve $y = \sin x$, is 1.*

Why is it such a good thing to have slope equal to 1 at the origin? As you know, and as we shall see in detail next week, the derivative of sine is cosine. In other words the slope of the curve $y = \sin x$ at $x = t$ is equal to $\cos t$; not $5.3 \cos t$: not $0.017 \cos t$ but $\cos t$. Clearly, if the slope were not equal to 1 at the origin we wouldn't get cosine; because $\cos 0 = 1$.

The reason we choose to measure angle in radians is that when we look at the functions sine and cosine applied to angles in radians, we get the simplest possible derivatives: the derivative of sin is cos and the derivative of cos is -sin. To finish this section let's remark that the slope of cosine at 0 is easy to determine.

Theorem (**The slope of cosine at 0**). *At 0, the slope of the curve $y = \cos x$, is 0.*



The function is even so if it has a tangent line at $x = 0$ that line will be the graph of an even function and hence horizontal.

The derivatives of the trigonometric functions

In the previous section we saw that the slope of the curve $y = \sin x$ at 0, is 1 and that the slope of the curve $y = \cos x$ at 0 is 0. The first amounted to showing that as h approaches 0,

$$\frac{\sin h}{h} \to 1.$$

The second means that

$$\frac{\cos h - 1}{h} \to 0$$

because to find the slope of $y = f(x)$ at x we look at

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

and ask what happens to it as h approaches 0. In this case f is the function cos and $x = 0$ so we look at

$$\frac{\cos(h) - \cos 0}{h} = \frac{\cos h - 1}{h}.$$

Our aim now is to find the derivatives of these functions at all other places. We want to calculate, in particular, what happens as h approaches 0 to

$$\frac{\sin(x+h) - \sin x}{h}.$$

To do this we use the addition formula which tells us that

$$\sin(x+h) = \sin x.\cos h + \cos x.\sin h.$$

Hence

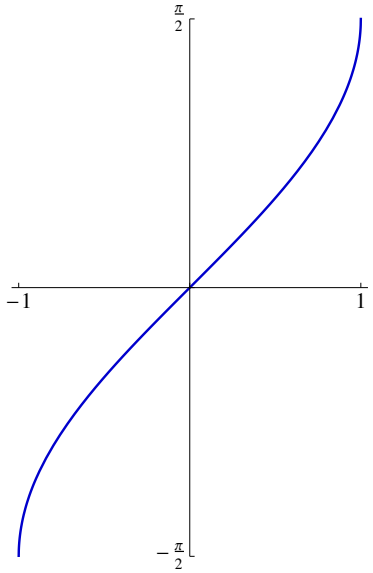
$$\begin{aligned}\frac{\sin(x+h)-\sin x}{h} &= \frac{\sin x.\cos h+\cos x.\sin h-\sin x}{h}\\ &= \sin x\,\frac{\cos h-1}{h}+\cos x\,\frac{\sin h}{h}\\ &\to (\sin x).0+(\cos x).1=\cos x.\end{aligned}$$

As h approaches 0, the first term approaches 0 while the second term approaches $\cos x$. Thus, the derivative of sin is cos.

You can find the derivative of cos in much the same way.

The derivative of $\sin^{-1}$

As you know the inverse sin function which maps the interval $\{x : -1 \le x \le 1\}$ onto the interval $\{y : -\pi/2 \le y \le \pi/2\}$ has a slightly surprising derivative. The graph of the function looks like the figure on the right.



We can see that the slope is 1 at 0 and is infinite at ± 1. But it is not obvious what the slope is elsewhere. To find it we need to use the method discussed in Chapter 9.

To use the chain rule we set

$$f(t)=\sin t \quad \text{and} \quad g(x)=\sin^{-1} x$$

and observe that

$$f(g(x))=x.$$

Consequently

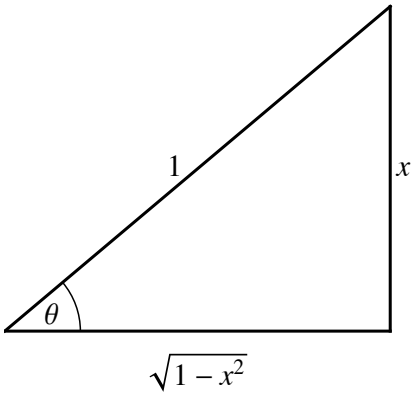
$$f'(g(x))g'(x)=1$$

and hence

$$\begin{aligned} g'(x) &= \frac{1}{f'(g(x))} \\ &= \frac{1}{\cos g(x)} \\ &= \frac{1}{\cos(\sin^{-1} x)}. \end{aligned}$$

This is not the expression you are accustomed to. The expression $\cos(\sin^{-1} x)$ simplifies. Suppose x is $\sin\theta$ so that $\theta = \sin^{-1} x$. Our aim is to find $\cos(\sin^{-1} x)$ which is $\cos\theta$.

In other words we are told that $\sin\theta = x$ and we want to find $\cos\theta$. A picture will do it:



If $\sin\theta = x$ then $\cos\theta = \sqrt{1-x^2}$.

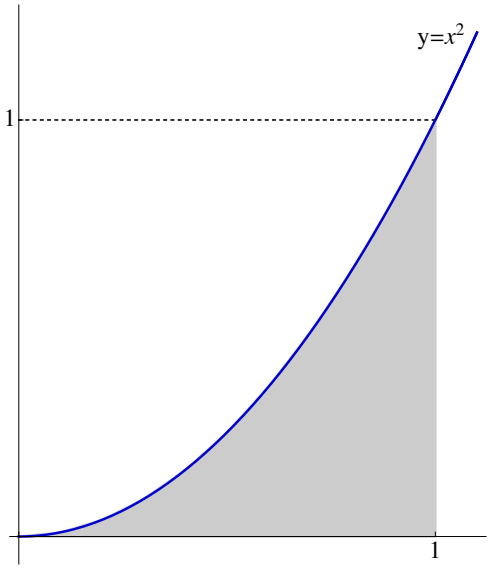
Thus we end up with

$$g'(x) = \frac{1}{\sqrt{1-x^2}}.$$

You might object that this is a rather complicated way to find the derivative. Can you find another way?

Chapter 11. Integration

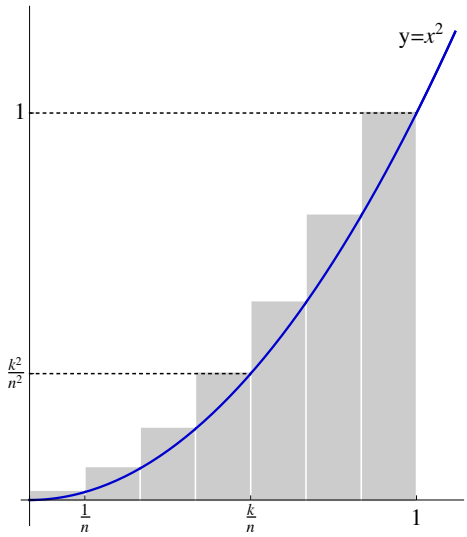
Earlier we talked about differentiation: the first part of calculus. Now we are going to look at the second part: integration. Integrals appear all over mathematics and have many different interpretations and uses. They originate in the basic problem of calculating area. Consider the function $x \mapsto x^2$ on the interval $\{x : 0 \leq x \leq 1\}$.



Archimedes discovered how to calculate the area shown shaded.

Let us calculate this area just as Archimedes did.

We divide up the interval into many equal pieces, and upon each piece we draw a rectangle whose top touches the curve.



We can calculate the sum of all the areas of the rectangles, and we would expect this to be a good estimate for the area we want. Let's do the calculation for a general number, n, of equal pieces. The rectangles are numbered from 1 to n. The first rectangle is based on the interval $\{x : 0 \leq x \leq \frac{1}{n}\}$. The second on the interval $\{x : \frac{1}{n} \leq x \leq \frac{2}{n}\}$. In general, the k^{th} rectangle is based on the interval from $\frac{k-1}{n}$ to $\frac{k}{n}$. The height of the k^{th} rectangle is thus

$$\left(\frac{k}{n}\right)^2 = \frac{k^2}{n^2}.$$

The width of each rectangle is $\frac{1}{n}$ so the area of the k^{th} rectangle is

$$\frac{k^2}{n^3}.$$

The total area (of all the rectangles) is thus

$$\frac{1^2}{n^3} + \frac{2^2}{n^3} + \frac{3^2}{n^3} + \ldots + \frac{n^2}{n^3} = \frac{1}{n^3}\sum_{j=1}^{n} j^2.$$

It is not too hard to find a formula for the sum of the square numbers from 1 to n. Using this we can write the total area as a single expression involving n. It is

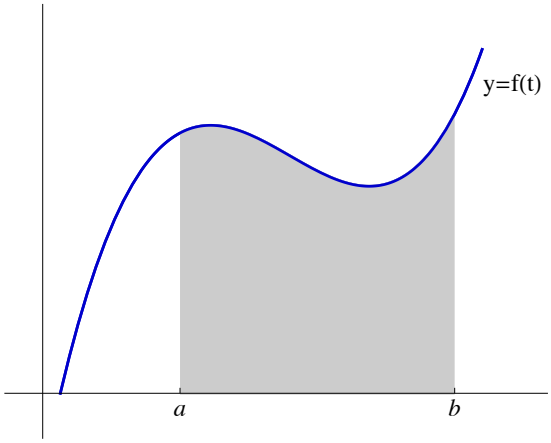
$$\frac{1}{n^3}\,\frac{n(n+1)(2n+1)}{6} = \frac{2n^2+3n+1}{6n^2}.$$

The total area of the blocks is

$$\frac{2n^2+3n+1}{6n^2} = \frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2}.$$

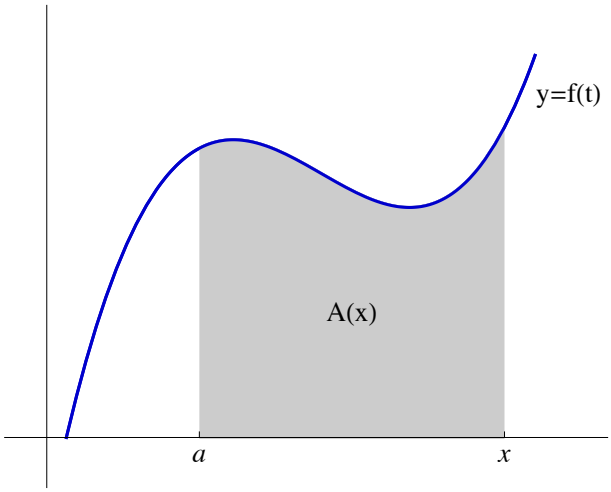
As the number of rectangles, n, gets larger, the fractions $\frac{1}{2n}$ and $\frac{1}{6n^2}$ approach zero. So, the total area of the rectangles approaches $\frac{1}{3}$. So we can see that the actual area under the curve is equal to $\frac{1}{3}$.

We can in principle do this for any nice continuous function. This is a way to construct the integral formally but it is an extremely cumbersome way to calculate integrals. We need something more powerful. Suppose we have a function f and we look at its graph.



Instead of trying to calculate just one area underneath f, for example the area between $t = a$ and $t = b$, we become far more adventurous.

We decide to try to calculate infinitely many areas all at once: we try to calculate the area between $t = a$ and $t = x$ for every number x. The area we get will certainly depend upon the value of x: it will be some function of x. Let's call it $A(x)$ as shown below.



What can we say about the function A? The crucial discovery is that we can immediately write down the **derivative** of the area function A.

$$A' = f.$$

The rate at which we add on new area as we increase x, is equal to the height of the curve: $f(x)$. This principle is so important that we call it the Fundamental Theorem of Calculus.

Theorem (**The Fundamental Theorem of Calculus**). *If f is a continuous function and we set*

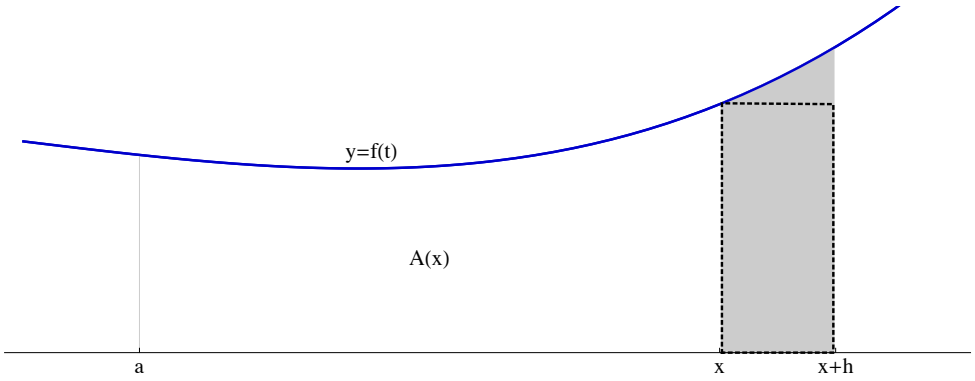
$$A(x) = \int_a^x f(t)\,dt$$

for each x, then

$$A'(x) = f(x)$$

for each x.

Let's see roughly why the fundamental theorem holds. Consider the graph $y = f(t)$ from $t = a$ to $t = x$ and a little further to $t = x + h$.



In order to calculate the derivative of $A(x)$ we need to look at the ratio

$$\frac{A(x+h) - A(x)}{h}$$

and ask what happens to it as h approaches 0. The difference $A(x + h) - A(x)$ is the shaded area. It is very similar to the rectangle outlined with a dashed line whose base has length h and whose height is $f(x)$. So $A(x + h) - A(x)$ is approximately $hf(x)$.

Therefore the ratio

$$\frac{A(x+h) - A(x)}{h}$$

is approximately $f(x)$. This approximation gets better as the width of the rectangle shrinks to zero. So as h approaches 0

$$\frac{A(x+h) - A(x)}{h}$$

approaches $f(x)$. Consequently

$$A'(x) = f(x).$$

Theorem (**The Fundamental Theorem of Calculus**). *If f is continuous and we set*

$$A(x) = \int_a^x f(t)\,dt$$

for each x, then

$$A'(x) = f(x)$$

for each x.

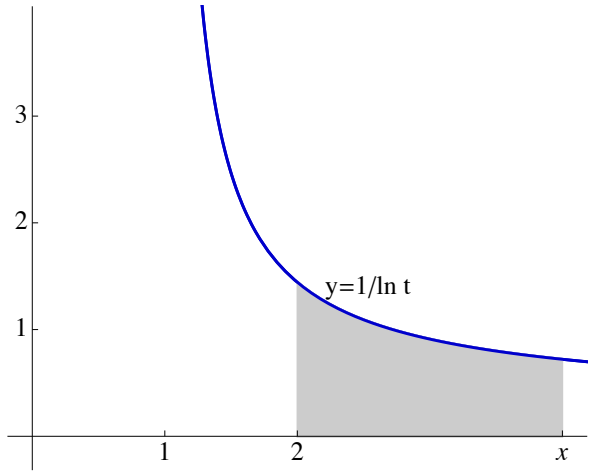
The fundamental theorem really tells us two things. On the one hand, it tells us how to differentiate a function which is given as a definite integral. On the other hand it gives us a faster way to calculate the integrals of certain functions.

The second point is more familiar to you, even though the first point is really the simpler of the two. Let's have an example of the first point. Suppose we know that the function F is given by

$$F(x) = \int_2^x \frac{1}{\ln t}\,dt$$

for each $x > 2$.

F is a perfectly good function: it tells you the area under the curve $y = \frac{1}{\ln t}$ between $t = 2$ and $t = x$.



However, it turns out that you can't express F using the normal functions of mathematics, combined in the usual algebraic ways. You can't write down any formula for F which is "more explicit" than this

$$F(x) = \int_2^x \frac{1}{\ln t}\,dt.$$

However, there are things you can say about F. You know its derivative: $F'(x) = \frac{1}{\ln x}$. It is extremely important to be comfortable with the idea that we can define a function using a definite integral, because such definitions occur all the time when you solve differential equations.

Chapter 12. Integration II

Let's return to what I called the second point about the Fundamental Theorem of Calculus and recall what you already know. If you want to evaluate an integral like $\int_0^1 t^2\,dt$ what you really do is to say: I shall evaluate **all** the integrals $\int_0^x t^2\,dt$, for all possible values of x. Suppose that, as above, you write

$$A(x) = \int_0^x t^2\,dt.$$

You don't yet know what the function A is but you know its derivative: $A'(x) = x^2$.

Now you ask yourself a question. Can I think of **any** function whose derivative is x^2? The answer is yes. The obvious one is

$$x \mapsto \frac{1}{3}x^3.$$

But this is not the only such function. If you add a constant to this function you don't change its derivative. If you just shift a graph upwards, you don't change its slope. So there are many functions whose derivative is x^2: for example

$$\begin{aligned} x &\mapsto \frac{1}{3}x^3 \\ x &\mapsto \frac{1}{3}x^3 + 2 \\ x &\mapsto \frac{1}{3}x^3 + \pi \\ &\vdots \end{aligned}$$

The obvious question is, are these the only ones? Is it true that if you know the slope of a function at every point, then you know what the function is, apart from a possible constant?

The answer is yes. I shan't attempt to justify this but I think it is at least intuitively reasonable. So now we know that our function $A(x)$ is a function of the form

$$x \mapsto \frac{1}{3}x^3 + C$$

for some constant C. It only remains to decide which of these functions: to find the value of C. We could do this if we knew the value of A at **just one** point.

We do: we know $A(0)$.

$$A(0) = \int_0^0 t^2\,dt = 0$$

because if you integrate over a range with no thickness you get no area. So $A(x) = \frac{1}{3}x^3 + C$ and $A(0) = 0$. That tells us that $C = 0$ and so the function we want is

$$x \;\mapsto\; \frac{1}{3}x^3.$$

So we have found that for every number x,

$$\int_0^x t^2\,dt = \frac{1}{3}x^3.$$

We can now go back and find the integral that we originally wanted. We just put $x = 1$ and we get

$$\int_0^1 t^2\,dt = \frac{1}{3}.$$

That's how we find definite integrals using the fundamental theorem. But, as you know, there is short cut. Instead of finding all functions whose derivative is x^2 and then choosing the right one, we can use a trick. We just think of one such function, say

$$F(x) = \frac{1}{3}x^3$$

and then we calculate $F(1) - F(0)$. By subtracting off $F(0)$ we automatically modify the function to get the right constant C.

I have explained the long-winded process for two reasons: partly to show why the trick works and partly to explain something much more important. Frequently at school you are asked to find **indefinite** integrals, and you were told to put in the constant C. It is important to understand what you are doing. Try not to think of the indefinite integral as a **thing** (a mathematical thing like a number) but as a question. The indefinite integral

$$\int t^2\,dt$$

is asking the question, "Which functions have x^2 as their derivative? Tell me all of them." The answer is "All functions of the form $x \mapsto 1/3x^3 + C$". And now you see why that is a good question.

Sums and integrals

A famous series, studied by Euler in the 18^{th} century is the following

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots$$

whose terms are the reciprocals of the squares. If you begin to calculate the partial sums

$$\begin{aligned} 1 &= 1 \\ 1 + \tfrac{1}{4} &= \tfrac{5}{4} \\ 1 + \tfrac{1}{4} + \tfrac{1}{9} &= \tfrac{49}{36} \\ 1 + \tfrac{1}{4} + \tfrac{1}{9} + \tfrac{1}{16} &= \tfrac{205}{144} \\ &\vdots \end{aligned}$$

you quickly despair of finding a pattern. It is highly unlikely that there is any simple formula for these partial sums: (we shall see some evidence in a moment). So it is not at all clear whether this series converges: whether these sums approach a limit. However, one thing **is** obvious about the partial sums listed above: they get larger as we add more terms, because the terms that we are adding are positive.

If we are adding up a series of positive terms like this, there are only two possible kinds of behaviour: either the partial sums approach a finite limit, or they increase without bound. Either the sums eventually surpass any value we can think of or they are trapped below some number. In the second case they get squeezed up against a ceiling and are forced to approach a limit.

If we could show that all the finite sums of the series

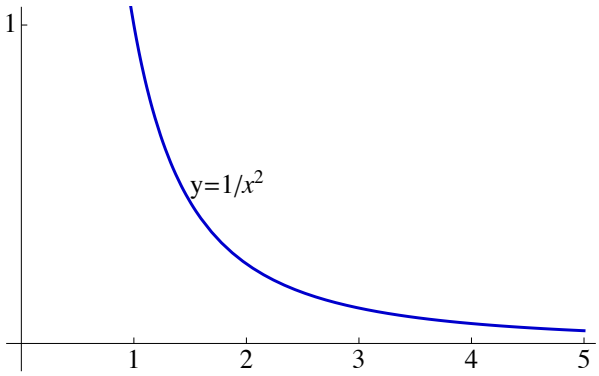
$$\sum_{k=1}^{\infty} \frac{1}{k^2}$$

are less than 2 (for example), then we could conclude that they approach some limit, even though we would not know what limit. Note that the bound we find, 2, need not be the

limit. It is just some value that traps the sums. The limit can't be larger than 2 but it could be smaller.

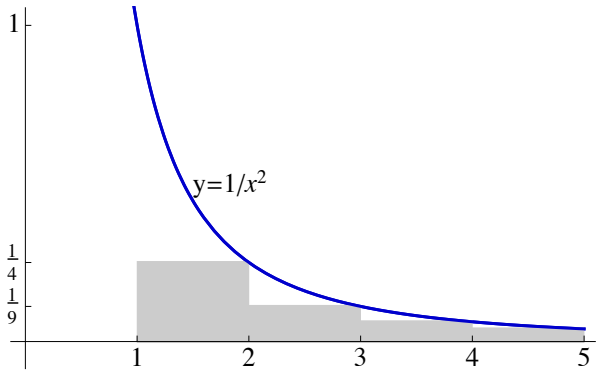
The idea will be to compare the sums with certain integrals which we are able to evaluate. Look at a graph of the function $x \mapsto \frac{1}{x^2}$.

The graph decreases. So, as long as $1 \leq x \leq 2$ the graph stays above the height $\frac{1}{4}$.



Thus, there is a rectangle of height $\frac{1}{4}$ and width 1 underneath the curve between $x = 1$ and $x = 2$. The area of this rectangle is $\frac{1}{4}$.

In a similar way, there are rectangles of area $\frac{1}{9}$, $\frac{1}{16}$ and so on, beneath the curve, between successive integers.



From this we can see that

$$\sum_2^n \frac{1}{k^2} \leq \int_1^n \frac{1}{x^2}\,dx = 1 - \frac{1}{n} \leq 1.$$

So, for each n,

$$\frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ldots + \frac{1}{n^2} \leq 1.$$

If we add in the extra 1 at the beginning we obtain the advertised estimate

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ldots + \frac{1}{n^2} \leq 2.$$

Hence, all the partial sums of this series are at most 2, and the series converges to something. What Euler managed to do, three hundred years ago, was to find out what that "something" is. He discovered the extraordinary and beautiful fact that

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \ldots = \frac{\pi^2}{6}.$$

Chapter 13. Taylor Approximation

Approximation of one kind or another plays a vital role in mathematics: in some ways it is the central problem in mathematics. Most of the equations that arise in physics cannot be solved exactly and we need to find ways to approximate their solutions. The advent of computers has made it possible to use approximation methods that would otherwise be too complex. But these methods have to be invented before they can be programmed. Even in cases where we are accustomed to thinking that we can solve equations, it turns out, when we look closely, that some approximation is going on. For example, we normally feel that we have no difficulty in solving equations like

$$x^2 = 2.$$

But the answer, $\sqrt{2}$ is a number which needs to be approximated if we wish to use it on a computer. We need a way of calculating the first 10 decimal places for $\sqrt{2}$ (or the first 15 or ...).

Similarly, we are in the habit of thinking that the integral

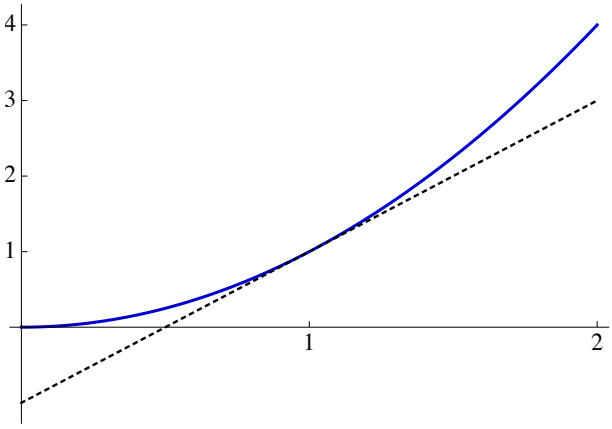
$$\int_1^2 \frac{1}{x}\, dx$$

is an integral that we can "do exactly." The value is ln2. But if I ask you "How do you **calculate** ln2?" you would have trouble giving me a much better answer than

$$\ln 2 = \int_1^2 \frac{1}{x}\, dx$$

and finding some way to estimate this integral. In mathematics, we need to be able to approximate particular numbers, like ln2, and also to approximate functions.

It turns out that the first problem is often best tackled via the second. In this chapter I want to discuss the most fundamental and important way to approximate functions: Taylor approximation. As we saw in an earlier chapter, the derivative of a function at a point tells us the slope of the tangent to the graph of the function at that point.



The derivative tells us how to approximate a function near a particular point, by a linear function. Although linear approximations are extremely important for many reasons, it is often necessary to approximate by other kinds of functions: for example, by polynomials of degree larger than one. How can we approximate a function be a quadratic polynomial?

Just as a linear function is determined by its value at a point together with its slope so a quadratic function is determined by its value at a point, its derivative at that point and its second derivative there. Indeed, suppose I have a quadratic polynomial p given by

$$p(x) = a + bx + cx^2.$$

Then

$$p'(x) = b + 2cx$$

and

$$p''(x) = 2c.$$

Once I tell you the second derivative $p''(0)$, you can find the coefficient

$$c = \frac{p''(0)}{2}.$$

From the derivative at 0, you can find $b = p'(0)$. The value of the function itself, $p(0)$, is a.

To put it another way, if p is a **quadratic** polynomial then for every x,

$$p(x) = p(0) + p'(0)x + \frac{p''(0)}{2}x^2.$$

This suggests a natural way to approximate functions by quadratic polynomials, in the neighbourhood of a particular point. We approximate a function f, near the point c, by the **unique** quadratic polynomial whose value at c, whose derivative at c and whose second derivative at c, are the same as those of f. From what we said above, we can see immediately that if c is 0, the quadratic polynomial in question is given by

$$p(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2.$$

Let's have an example. Suppose $f(x) = \frac{1}{1-x}$. The first two derivatives of f are

$$f'(x) = \frac{1}{(1-x)^2}$$

and

$$f''(x) = \frac{2}{(1-x)^3}.$$

Substituting 0 for x we obtain

$$f(0) = 1, \quad f'(0) = 1, \quad f''(0) = 2.$$

Hence, the quadratic Taylor approximation to f at 0 is

$$x \mapsto f(0) + f'(0)x + \frac{f''(0)}{2}x^2 = 1 + x + x^2.$$

Notice again the sequence of operations in the above calculations. We are given a function:

$$x \mapsto \frac{1}{1-x}.$$

1. We use the machine to differentiate this function an appropriate number of times.
2. We substitute 0 into each of these derivatives, to obtain the successive derivatives **at 0**. These numbers do not depend upon x.
3. We use these numbers to build the Taylor approximation.

We have seen how to approximate a function near 0 by a quadratic polynomial. Needless to say we can approximate near a different point by using the derivatives at that point.

For a general point a we get that the quadratic Taylor approximation to a function f near a is given by

$$p(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2.$$

Now let's move on to higher order approximations. The n^{th} order Taylor approximation to f near a should be a polynomial of degree n whose value, and whose first n derivatives, at a are the same as those of f at a. The polynomial is given by

$$p(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \frac{f^{(3)}(a)}{6}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n.$$

(Try differentiating this a few times to see that it really does have the correct derivatives.)

The geometric series

Let's use the Taylor approximation formula to find further approximations to the function

$$x \mapsto \frac{1}{1-x}.$$

The successive derivatives of this function are

$$\frac{1}{1-x}, \quad \frac{1}{(1-x)^2}, \quad \frac{2}{(1-x)^3}, \quad \frac{6}{(1-x)^4}, \ \ldots, \ \frac{n!}{(1-x)^{n+1}}, \ \ldots$$

The values of these derivatives at 0 are

$$1, \quad 1, \quad 2, \quad 6, \quad 24, \ \ldots, \ n!, \ \ldots$$

This makes it easy to see that the n^{th} order Taylor approximation at zero, to the function

$$x \mapsto \frac{1}{1-x}$$

is

$$1 + x + x^2 + x^3 + x^4 + \ldots + x^n.$$

In consequence we expect that if x is close to 0, and n is large,

$$\frac{1}{1-x} \approx 1 + x + x^2 + x^3 + \ldots + x^n.$$

This should not come as a great surprise. When we talked about geometric series, we found that as long as $|x| < 1$,

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

In other words, the infinite sum converges to the value $\frac{1}{1-x}$. This is the same thing as saying that we get good approximations to $\frac{1}{1-x}$ by taking a large enough number of terms of the series. For the function $x \mapsto \frac{1}{1-x}$, the Taylor approximations at zero approach the value of the function, as long as $|x| < 1$.

This kind of situation occurs quite frequently: there are many functions for which the Taylor approximations at a point converge to the correct value, at least near the point. When this happens we get an infinite series that represents the function (at least near a point). Such a series is called the Taylor series for the function at that point. We can thus state the above remarks in the following way. The Taylor series for the function $x \mapsto \frac{1}{1-x}$ at zero, is

$$1 + x + x^2 + x^3 + x^4 + \dots .$$

In the next chapter I will talk about perhaps the most important Taylor series of all.

Chapter 14. The exponential and logarithm

In this chapter I want to talk about the Taylor series for the exponential function $x \mapsto e^x$. This function is its own derivative: so **all** its derivatives are the same function. Therefore, at the point 0, all the derivatives of this function are equal to 1. Hence, the Taylor series of the function, at 0, is

$$1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots = \sum_0^\infty \frac{x^n}{n!}.$$

It turns out that this series makes sense for any value of x. The series converges, however large x may be. The reason (roughly) is that the denominators $n!$ grow extremely rapidly as n increases, so that the terms of the series eventually become very small, even if x is large. So we can write

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

for every number x.

If you remember, I stated earlier that there is a number e for which the function $x \mapsto e^x$ has derivative equal to 1 at 0. I did not demonstrate the existence of this number e. We can now say what this number is:

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots.$$

Hence, for example, we can estimate e, as accurately as we wish, in a fairly simple way. If you prefer, you could take this series to be the **definition** of e.

In fact, you could go a bit further, and say that you wouldn't mind having a slightly clearer definition of e^x. After all, it isn't entirely obvious what we mean by $e^{\sqrt{2}}$. How do we multiply the number e by itself $\sqrt{2}$ times? If you wish, you could take the Taylor series

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

as the **definition** of e^x. What I am more interested in today is to see how to relate the series to the properties with which you are more familiar.

The exponential function has two crucial properties both of which have already been mentioned.

- $$\frac{d}{dx}e^x = e^x.$$
- $$e^{x+y} = e^x.e^y$$

  for all numbers x and y.

Remember that the first of these was the property that we used to find the Taylor series. It is not too surprising that we can go back.

Let us make a fairly bold assumption: that we can differentiate the series

$$1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \dots$$

"term by term," just as if it were a **finite** sum. What do we get?

$$0 + 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots.$$

So, when we differentiate the series we get exactly the same series back again. The 0 at the front makes no difference to the value.

What about the multiplicative property? When we multiply e^x by e^y we have to expand the product of two brackets

$$\left(1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots\right)\left(1 + y + \frac{y^2}{2} + \frac{y^3}{6} + \cdots\right).$$

So we have to form all possible products of one term from the first bracket and one term from the second bracket (and then add all of these products together).

When you are multiplying two infinite sums, you have to be a bit careful about how you write down the products so as not to miss any. In order to find a systematic way of doing

it, the best thing is to draw a grid.

	1	x	$x^2/2$	$x^3/6$	$\dots$
1	1	x	$x^2/2$	$x^3/6$	$\dots$
y	y	xy	$x^2y/2$	$\dots$	
$y^2/2$	$y^2/2$	$xy^2/2$	$\dots$		
$y^3/6$	$y^3/6$	$\dots$			
$\vdots$	$\vdots$				

Along the top are the terms of the first sum. Down the side are the terms of the second. In the grid are the products of all possible pairs, located in the obvious positions. Our job is to add up all the products in the grid.

You might be tempted to add up all the products in the first row, then move on to the second and so on. But with infinite sums this is not a very good idea. You will never get to the end of the first row: so you will never pick up **any** terms from **any** of the other rows.

One way to make sure that we include everything is to add them diagonally. We start with the top left corner 1. Then we take the next NE-SW diagonal down: the short diagonal containing x and y. That gives us the sum $x+y$. Then the next diagonal down gives us $x^2/2+xy+y^2/2$. Continuing like this we pick up all the products in the grid.

$$1+(x+y)+(x^2/2+xy+y^2/2)+(x^3/6+x^2y/2+xy^2/2+y^3/6)+\cdots$$

Now let us collect each diagonal sum over a common denominator. The first is $x+y$. The second is

$$\frac{x^2+2xy+y^2}{2}$$

and the third is

$$\frac{x^3+3x^2y+3xy^2+y^3}{6}.$$

We immediately see that we have the binomial expansions of $(x+y)^2$ and $(x+y)^3$. So the whole sum looks like

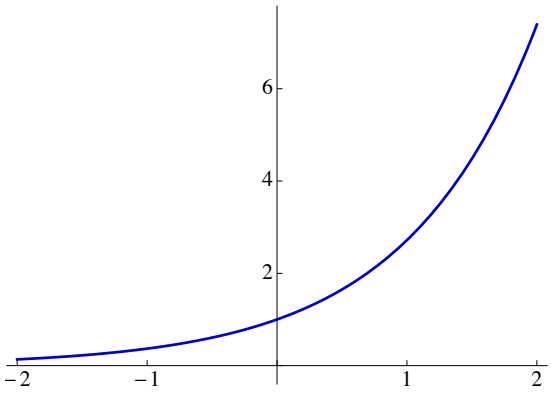
$$1 + (x+y) + \frac{(x+y)^2}{2} + \frac{(x+y)^3}{6} + \cdots$$

and this is $\exp(x+y)$.

What this argument shows is that the characteristic property of the exponential function $e^{x+y} = e^x.e^y$ is actually the same as the binomial theorem.

The logarithm

The exponential function is strictly increasing: so it never takes the same value twice.



As x moves out to ∞ on the right e^x increases to infinity while as x moves out to $-\infty$ the function decreases to 0. So for every positive y there is one and only one value of x for which

$$e^x = y.$$

This x is called $\ln y$: the natural logarithm of y. The number $\ln y$ is the answer to the question

$$e^{?} = y.$$

The upshot is that there is a function $y \mapsto \ln y$ defined for positive y satisfying

$$e^{\ln y} = y \text{ for all positive } y$$

$$\ln(e^x) = x \text{ for all real } x.$$

The natural logarithm is the inverse of the exponential function.

As you know for all positive u and v,

$$\ln(uv) = \ln u + \ln v.$$

Let's prove this from the definition. We know that $\ln(uv)$ is the number whose exponential is uv:

$$e^{\ln(uv)} = uv.$$

But we know that the number in question is $\ln u + \ln v$ because

$$uv = e^{\ln u}.e^{\ln v} = e^{\ln u + \ln v}.$$

Hence

$$\ln(uv) = \ln u + \ln v.$$

The derivative of the log

Let us use the chain rule to compute the derivative of ln from that of the exponential. Let f be the exponential function and g the logarithm.

$$f(g(x)) = x$$

and so

$$f'(g(x))g'(x) = 1.$$

In the present case

$$f'(x) = e^x$$

and hence

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\exp(\ln x)} = \frac{1}{x}.$$

Chapter 15. Complex numbers

For many reasons, it is important to be able to solve algebraic (or polynomial) equations: equations like

$$x^3 - 3x^2 + x - 5 = 0.$$

But the system of real numbers does not provide us with solutions, even to simple quadratic equations, because negative numbers do not have real square roots. To solve this problem, we adjoin to the real number system, a further number, which we call i (or sometimes j), with the property that $i^2 = -1$. We then study the set of all numbers of the form $x + iy$ where x and y are real numbers. We call these expressions, **complex numbers**.

You might at once wonder whether this i is **really** a number. What is i? Where is it? Mathematicians have learnt from painful experience that to ponder such philosophical questions is invariably fruitless. During the 15^{th} century, there were misgivings about the existence of negative numbers: such misgivings now look absurd. The only things that matter are whether or not we can build a sensible arithmetic, using complex numbers, and whether this arithmetic gives us useful mathematical tools?

How do we do arithmetic with complex numbers? I have to tell you how to add and multiply complex numbers. The sum of $x + iy$ and $u + iv$ is the number

$$(x + u) + i(y + v).$$

We add complex numbers by adding their real parts and adding their imaginary parts, just as if these were ordinary algebraic expressions.

The rule for multiplication looks more complicated: the product of $x + iy$ and $u + iv$ is the number

$$(xu - yv) + i(xv + yu).$$

But the rule is really very simple. We multiply the numbers as if they were ordinary algebraic expressions:

$$(x + iy)(u + iv) = xu + xiv + iyu + i^2yv$$

and then we replace i^2 by -1.

So, for complex arithmetic, we need just one new rule:

$$i^2 = -1.$$

However, there are certain things we need to check before we rush off and do complex arithmetic. The most important thing is to check that each complex number other than $0 + 0.i$ has a reciprocal. For example, is there a complex number equal to

$$\frac{1}{2+3i}?$$

Can we find a number $x + iy$ for which

$$(2+3i)(x+iy) = 1 + 0.i?$$

Since the left side of this is

$$(2x - 3y) + i(3x + 2y)$$

we really have two simultaneous equations here:

$$\begin{aligned} 2x - 3y &= 1 \\ 3x + 2y &= 0. \end{aligned}$$

These have the unique solution

$$x = 2/13, \qquad y = -3/13.$$

So the answer is yes:

$$\frac{1}{2+3i} = \frac{2}{13} - \frac{3}{13}i.$$

As you know, there is a way to streamline the calculation above. You write the expression

$$\frac{1}{2+3i}$$

in the form

$$\frac{2-3i}{(2+3i)(2-3i)}$$

by "multiplying top and bottom by $2-3i$." You now observe that the bottom is $2^2-(3i)^2 = 4 - 9i^2 = 4 + 9 = 13$.

$$\frac{1}{2+3i} = \frac{2-3i}{(2+3i)(2-3i)} = \frac{2-3i}{13}.$$

Since the bottom is now a real number we know how to break the number up into real and imaginary parts.

$$\frac{2-3i}{13} = \frac{2}{13} - \frac{3}{13}i.$$

Once you have reciprocals, you have no problem dividing one complex number by another (which isn't zero). So we now have a system of arithmetic for complex numbers.

The fundamental theorem of algebra

Once we have introduced complex numbers we can find a square root for each negative number: in fact two square roots. For example, each of the numbers i and $-i$ has -1 as its square. Similarly, -4 has the square roots $2i$ and $-2i$. We can then solve all quadratic equations

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

It would be natural to expect that if you want to solve cubic equations, you have to throw in some more numbers, and to solve quartics, yet more.... However, it turns out that **to solve polynomial equations, you only ever need the complex numbers**. Even if you wish to solve equations with complex coefficients (not real ones), you can do it with complex numbers.

For example, we know that there is at least one complex number z which satisfies

$$z^8 + (2+i)z^5 + iz - 3 = 0.$$

This astonishing fact is so important that we call it "The fundamental theorem of algebra."

Now suppose that p is a polynomial with complex coefficients. Once you have found a zero of p, $p(\alpha) = 0$ say, you can factorise: $p(x) = (x-\alpha)q(x)$ where q is a polynomial of degree one less than the degree of p. So you can continue the process, by finding a zero of q. Eventually, you end up with the original polynomial p, expressed as a product of linear factors with complex coefficients. We can thus state the fundamental theorem of algebra in the following way.

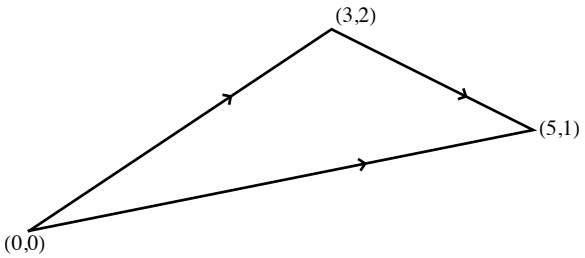
Theorem (The fundamental theorem of algebra)**.** *Every polynomial with complex coefficients can be written as a product of linear factors with complex coefficients.*

The fundamental theorem of algebra is **not** easy to prove. I certainly shan't attempt to prove it here. The first really convincing demonstration was found by Gauss, at the beginning of the 19^{th} century.

The complex plane

Since each complex number $x+iy$ is built out of two real components we can represent the complex numbers by the points of the plane: the number $x+iy$ corresponds to the point (x,y). Such a representation might be nothing more than an artificial correspondence. However, it turns out to be anything but artificial: the arithmetic of the complex numbers is intimately related to the geometry of the plane. Let us begin gently.

The simplest observation is that addition of complex numbers corresponds to addition of vectors in the plane. For example, the statement $(3+2i)+(2-i)=5+i$ can be represented by the diagram:



Thus, the map which takes a point $x+iy$ to the point $(x+iy)+(2-i)$ acts on the plane as a translation, by the displacement $(2,-1)$.

Multiplication is a bit more complicated. Let's try multiplication by $2+3i$. This takes a number $x+iy$ to the number

$$(2x-3y)+i(3x+2y).$$

Thus it corresponds to the map

$$(x,y)\mapsto(2x-3y,3x+2y).$$

This map is a linear map: it is given by a matrix mutiplication:

$$\begin{pmatrix} 2x-3y \\ 3x+2y \end{pmatrix} = \begin{pmatrix} 2 & -3 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

It is a combination of a rotation and an enlargement. The scale factor of the enlargement is $\sqrt{13}$, which should ring a bell. The number 13 appeared when we were calculating the reciprocal of $2+3i$.

More generally, the map corresponding to multiplication by $a+ib$ is the linear map given by the matrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

This map is a combination of an enlargement and rotation. The scale factor of the enlargement is the number

$$\sqrt{a^2+b^2}.$$

The real number $\sqrt{a^2+b^2}$ is thus related to the algebraic behaviour of the complex number $a+ib$. Of course, it is also the distance of the point (a,b) from $(0,0)$. The appearance of the Euclidean (or Pythagorean) distance, lies at the heart of the relationship between the complex numbers and the geometry of the plane. We call the number $\sqrt{a^2+b^2}$, the absolute value of the complex number $a+ib$ and denote it with vertical bars

$$|a+ib|$$

just as for real numbers.

Thus addition of complex numbers corresponds to translation of the plane and multiplication of complex numbers corresponds to matrix operations on the plane. The arithmetic of complex numbers **fits** with the geometry of the plane. When we talk about the complex plane we don't just mean that complex numbers are pairs of real numbers: the complex numbers really **do** form a plane.

One final remark is appropriate for this chapter. We frequently abbreviate our notation $x+iy$ for a complex number. Unless we are explicitly interested in the real and imaginary parts of a number, we usually represent it by a single letter instead of 2 (or 3 if you include the letter i). Thus, we quite happily write w or z to mean a complex number.

Chapter 16. The complex exponential function

In the last chapter I introduced the complex numbers and mentioned the fantastic fact that all polynomial equations have complex solutions. This fact alone would be enough to make the complex numbers important in mathematics. But their importance in physics and engineering stems from another, equally astonishing discovery. In this chapter I will explain how we extend the exponential function to the complex numbers: how we define the exponential of a complex number.

I pointed out that we have to be a bit careful what we mean by $e^{\sqrt{2}}$. It isn't clear what it means to multiply a number by itself $\sqrt{2}$ times. However, at a pinch, I bet you could have come up with a sensible definition. The situation looks much more difficult when we move into the complex plane. What on earth is 7^i? That's what this chapter will be about.

If things are going to work out nicely, one thing certainly ought to be true. Since

$$7 = e^{\ln 7},$$

we had better have

$$7^i = (e^{\ln 7})^i = e^{i \ln 7}.$$

So we can hope that the only thing we really need to do, is to find a way to calculate

$$e^z,$$

for each complex number z.

At this point we make an inspired guess: (or Euler and his contempories made one). Remember that for a real number x we had an expression for e^x as an infinite series:

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots.$$

Can we use the same formula to tell us the value for e^z when z is a complex number? What could go wrong? The first thing we have to do is to make sense of the infinite sum.

In the case of real numbers, this involved determining whether the successive sums approach a limiting value. Can we make sense of complex numbers approaching a limit?

Yes, because we can measure how far apart they are: we can measure the distance between complex numbers because they are points in the plane. So, we can at least **say** "We define

$$e^z = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \dots."$$

The second important thing we need to do is to check that our new function has the right property for an exponential:

$$e^{w+z} = e^w.e^z$$

for any two complex numbers w and z. We can do this, just as we did for real numbers, because the binomial theorem works just as well for complex numbers. Once we have the multiplicative property for the exponential, we feel a bit happier about writing it e^z.

Now comes the really important question. Is our complex exponential function useful, or is it just a mathematical trick with no real point? To answer this we try to understand a bit about it. If you wanted to calculate e^{s+it} for a complex number $s + it$ written in terms of its real and imaginary parts, the natural thing to do would be say

$$e^{s+it} = e^s.e^{it}.$$

Since s is a real number, we already understand e^s. The problem is to understand e^{it}, for each real number t.

When we put it into the series, we get

$$1 + it + \frac{(it)^2}{2} + \frac{(it)^3}{6} + \dots.$$

Because $i^2 = -1$ the second term of this series can be simplified to $-\frac{t^2}{2}$. The third term also simplifies because $i^3 = -i$. If we continue removing even powers of i we can see that the series splits easily into its real and imaginary parts. The odd numbered terms give the imaginary part and the even numbered terms, the real part.

We have

$$e^{it} = \left(1 - \frac{t^2}{2} + \frac{t^4}{24} - \dots\right) + i\left(t - \frac{t^3}{6} + \frac{t^5}{5!} - \dots\right).$$

The two series for the real and imaginary parts of e^{it} are very familiar. They are the series for $\cos t$ and $\sin t$ respectively. What we have found is that

$$e^{it} = \cos t + i \sin t.$$

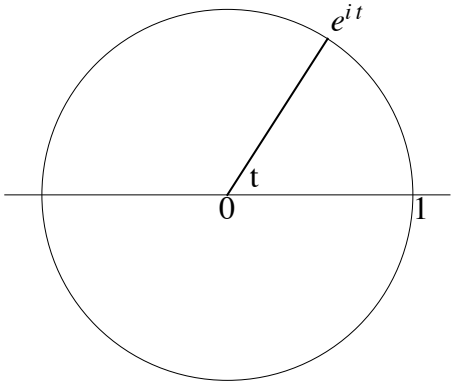
This remarkable formula links together the exponential series, which begins as a purely **analytic** construction, and the trigonometric functions, which are **geometric** constructions.

Some of the most powerful mathematical ideas are those which link together algebra or analysis and geometry. The physicist Richard Feynman described the formula

$$e^{it} = \cos t + i \sin t.$$

as "our jewel."

Going back to our picture of the complex numbers as points in the plane, we can now draw in the point e^{it} for a real number t.



Points of this form are exactly the points which lie on the circle of radius 1, centre 0, because the points have coordinates $(\cos t, \sin t)$. We now have a way to define the exponential of a complex number and we are able to give a geometric meaning to the values we get. This, finally, justifies our view of the complex numbers as a plane. We have an algebraic way to describe the circle of radius 1, sitting in the complex numbers. We have a map

$$t \mapsto e^{it}$$

from the real line to the complex plane, which winds the real line around the circle of radius one, infinitely many times.

The rest of this chapter is devoted to applications of the jewel formula. The first is a quick derivation of the addition formulae for the trigonometric functions. The second relates the derivatives of the exponential and trigonometric functions.

We know from the magic formula of the last section that for any real numbers θ and ϕ,

$$\cos(\theta+\phi) + i\sin(\theta+\phi) = e^{i(\theta+\phi)}.$$

But as soon as we see the exponential of a sum, we can immediately write it as a product, and each of the factors can then be put back in terms of trigonometric functions:

$$\begin{aligned} e^{i(\theta+\phi)} &= e^{i\theta}e^{i\phi} \\ &= (\cos\theta + i\sin\theta)(\cos\phi + i\sin\phi) \\ &= (\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\sin\theta\cos\phi + \cos\theta\sin\phi). \end{aligned}$$

So we get

$$\cos(\theta+\phi) + i\sin(\theta+\phi) = (\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\sin\theta\cos\phi + \cos\theta\sin\phi).$$

This equation involving complex numbers can be regarded as two equations involving real numbers. By equating real and imaginary parts we can read off the addition formulae.

$$\begin{aligned} \cos(\theta+\phi) &= \cos\theta\cos\phi - \sin\theta\sin\phi \\ \sin(\theta+\phi) &= \sin\theta\cos\phi + \cos\theta\sin\phi. \end{aligned}$$

Once you have seen this, you can never again dare to admit that you can't remember the addition formulae for the trigonometric functions.

We can also use the jewel to link together the derivatives of the exponential and trigonometric functions.

$$\cos t + i\sin t = e^{it}.$$

If we differentiate e^{it} with respect to t we get ie^{it}. Hence

$$\frac{d}{dt}(\cos t + i\sin t) = \frac{d}{dt}e^{it} = ie^{it} = i(\cos t + i\sin t) = i\cos t - \sin t.$$

By comparing real and imaginary parts we can confirm that

$$\frac{d}{dt}\cos t = -\sin t \quad \text{and} \quad \frac{d}{dt}\sin t = \cos t.$$

The formula

$$e^{it} = \cos t + i\sin t.$$

only makes sense if we measure the angle in radians. Recall that we chose to use **the** exponential function rather than **an** exponential function ($x \mapsto 2^x$ for example) in order to make the derivative of the function as simple as possible:

$$\frac{d}{dx}e^x = e^x.$$

We chose to measure angle in radians in order to make the derivatives of the trigonometric functions as simple as possible

$$\frac{d}{dx}\cos x = -\sin x \quad \text{and} \quad \frac{d}{dx}\sin x = \cos x.$$

We now see that these two choices were in fact the **same** choice. Together they lead to one of the most remarkable human achievements of all time

$$e^{ix} = \cos x + i \sin x.$$