Mathematics as whole is composed of three parts

We tend to split mathematics into three parts, algebra, analysis and the other stuff. To my mind algebra is the stuff we make up, we define some rules, like a *group* is a set with a map satisfying such and such a list of properties and then we consider the consequences of those rules. Algebra is much loved by computer scientists because computer science is, in some ways, the study of a system of this sort and the logic of programming langauges can be framed in algebraic terms. Computer science was also born during a time of great progress in our understanding of the philosophy and mathematics of axiomatic systems and still bears the historical imprint of that time.

This is a beautiful and profound subject, and we are not here to criticize it. We are, however, here to study a different set of mathematical ideas often bundled together as analysis. If algebra is the stuff we make up, analysis is the mathematics we discover; it is the collection of mathematical ideas we develop in response to some of the mysteries of our world, the fact that objects have entent, the fact that they move. It is not as central to our subject, in and of itself, as algebra but it is important to many of the applications that have driven forward computer-based computations, the simulation of physical systems in service of engineering and science, the stochastic dynamics of stock markets, the complex modelling of climate, the performance of physics-like dynamics in games and, of course recently, the optimization of large neural network in machine learning. We, as computer science, should learn analysis because it is useful in areas we often end up working, because we are interested in machine learning and, of course, because it is a beautiful subject.

This distinction is, of course, too simple; there is axiomatic ideas in analysis and, as we'll allude to, ideas about proof have been important in analysis; algebra in turn, is not solely a mental exercise, it gains inspiration from the real world. Most importantly, however, is that the greatest progress has always come from bringing analysis and algebra together, for example, mathematics in the last fifty years has seen incredible progress in geometry which depends on a combination of ideas from analysis and algebra. This is all beyond this unit, for now we want to revise differentiation.

The law of fluxations?

We will start with the original description of calculus due to the two scientists who discovered it in the seventeenth century: Newton and Leibniz. They based their version of calculus on "infinitessimals": an *infinitessimal* was a very small number, a number so small that it could be regarded as zero. The difficulty in making sense of what this meant, and Leibniz who was also a philosopher invented a weird set of ideas about reality to do just that, does not mean it is not a useful way of thinking about differentiating.

The idea is to work out how quickly a function $f(x)$ is changing. Consider adding a small amount, δx[1] to x then the change in $f(x)$ will be

$$\delta f(x) = f(x + \delta x) - f(x) \tag{1}$$

This means that change in f per change in x is

$$\frac{\delta f(x)}{\delta x} = \frac{f(x + \delta x) - f(x)}{\delta x} \tag{2}$$

This is fine, but $f(x)$ could have any sort of shape and it might change how quickly it is changing from x to $x + \delta x$, we want to know how much it is changing "right now", exactly at x. The solution is to make δx really small and the idea Newton and Leibniz had was to define dx as a number so small it is zero once you have finished with any dividing and so forth. That this might cause problems later on is now obvious, it is a slippery sort of notion, but it sort of works, we define the *derivative* as

$$\frac{df}{dx} = \frac{f(x + dx) - f(x)}{dx} \tag{3}$$

and this is a measure of the rate $f(x)$ is changing as x changes.

How does this work in practice, lets consider a simple example:

$$f(x) = x^3 \tag{4}$$

Now

$$\frac{df}{dx} = \frac{(x + dx)^3 - x^3}{dx} \tag{5}$$

[1] The Greek letter δ, called delta and equivalent in pronounciation to the Latin letter d is often used for small changes.

We can expand out

$$(x+dx)^3 = x^3 + 3x^2dx + 3x(dx)^2 + (dx)^3 \tag{6}$$

and so

$$\frac{df}{dx} = 3x^2 + 3xdx + (dx)^2 \tag{7}$$

and then using the rule that dx is so small it is zero

$$\frac{df}{dx} = 3x^2 \tag{8}$$

and all is good. This works for any power of x since we know

$$(x+a)^n = \sum_{r=0}^{n} \left(\begin{array}{c} n \\ r \end{array} \right) x^{n-r}a^r \tag{9}$$

where I am using the bracket notation for the binomial symbol

$$\left(\begin{array}{c} n \\ r \end{array} \right) = {}_nC_r = \frac{n!}{r!(n-r)!} \tag{10}$$

Putting this into the definition of the derivative gives

$$\frac{dx^n}{dx} = nx^{n-1} \tag{11}$$

This approach also allows us to derive some basic general rules for differentiating. Consider, for example the sum of two functions.

$$\begin{aligned} \frac{d}{dx}[f(x)+g(x)] &= \frac{f(x+dx)+g(x+dx)-f(x)-g(x)}{dx} \\ &= \frac{f(x+dx)-f(x)}{dx} + \frac{g(x+dx)-g(x)}{dx} \\ &= \frac{df}{dx} + \frac{dg}{dx} \end{aligned} \tag{12}$$

or consider multiplying by a constant c

$$\frac{d}{dx}[cf(x)] = \frac{cf(x+dx)-cf(x)}{dx} = c\frac{f(x+dx)-f(x)}{dx} = c\frac{df}{dx} \tag{13}$$

and so differentiation is a linear operation.

A more complicated example is provided by products.

$$\frac{d}{dx}f(x)g(x) = \frac{f(x+dx)g(x+dx) - f(x)g(x)}{dx} \tag{14}$$

Now, this requires a certain amount of nerve, but from the definition

$$f(x+dx) = f(x) + \frac{df}{dx}dx \tag{15}$$

and so

$$\frac{d}{dx}f(x)g(x) = \frac{1}{dx}\left[\left(f(x) + \frac{df}{dx}dx\right)\left(g(x) + \frac{dg}{dx}dx\right) - f(x)g(x)\right] \tag{16}$$

Now, expanding out

$$\frac{df(x)}{dx}f(x)g(x) = \frac{df}{dx}g(x) + f(x)fracdgdx + \frac{df}{dx}\frac{dg}{dx}dx \tag{17}$$

and, in the now familiar way, we set the dx term to zero to get the product rule

$$\frac{d}{dx}f(x)g(x) = \frac{df}{dx}g(x) + f(x)fracdgdx \tag{18}$$

We can differentiate just about anything

In this way we can derive the derivative of common functions:

- polynomials: $dx^n/dx = nx^{n-1}$
- special function:
  1. $d\sin x/dx = \cos x$
  2. $d\cos x/dx = -\sin x$
  3. $d\exp x/dx = \exp x$
  4. $d\log x/dx = 1/x$
- product rule:

$$\frac{d}{dx}uv = \frac{du}{dx}v + u\frac{dv}{dx}$$

- quotient rule:
$$\frac{d}{dx}\frac{u}{v} = \frac{\frac{du}{dv}v - u\frac{dv}{dx}}{v^2}$$

This leaves the most powerful rule of all, the chain rule:

$$\frac{du(v(x))}{dx} = \frac{du}{dv}\frac{dv}{du} \tag{19}$$

This allows us to work out the derivative for function that are writen as a composition, a function of a function. This is the machinery that means we can differente just about anything that has a derivative and, these days, as implemented in autograd in machine learning libraries, allows us to differentiate any calculation made on a computer: any calculation a computer makes is a composition of simple operations, ultimately simple logical operations on bits, and so the chain rule can differentiate the computation, in machine learning libraries this allows the *gradient* to be calculated, the gradient, as we will see is used to optimize, to find maxima and minima of functions, in the case of machine learning, the loss function.

Here is a simple example of the chain rule in action, let

$$f(x) = (2 + x^2)^3 \tag{20}$$

We could do this the hard way by expanding out the bracket:

$$f(x) = 8 + 12x^2 + 6x^4 + x^6 \tag{21}$$

and so

$$\frac{df}{dx} = 24x + 24x^3 + 6x^5 \tag{22}$$

However, we could also write

$$f(v) = v^3 \tag{23}$$

where

$$v = 2 + x^2 \tag{24}$$

So

$$\frac{df}{dv} = 3v^2 \tag{25}$$

and

$$\frac{dv}{dx} = 2x \tag{26}$$

Substituting back in for v and applying the chain rule:

$$\frac{df}{dx} = 6x(2+x^2)^2 \tag{27}$$

which, you can check, is the same as what we got before. In this case there was an alternative, albeit more laborious approach but the chain rule works in cases where there is no alternative. For example

$$f(x) = \exp x^2 \tag{28}$$

so we let $v = x^2$ and $dv/dx = 2x$ while $d\exp v/dv = \exp v = \exp x^2$ and hence

$$\frac{df}{dx} = 2x\exp x^2 \tag{29}$$

A note on notation

Notation is what mathematics is good at, the whole glory of mathematics is that it allows us to write down and discuss difficult abstract ideas. It is surprising then sometimes to reflect on how informal mathematical notation often is, this is particularly frustrating to some computer science folk whose craft requires completely precise notation, in some ways a language like Haskell is what mathematics would look like if we weren't all lazy bags of meat. This is a quick note on that laziness, it is here to help you if you are worried about the notation, but if you weren't worried, don't let it worry you.

Anyway, the problems often start with functions, a function is a map between spaces, so we might say

$$f : \mathbb{R} \to \mathbb{R} \tag{30}$$

and then, if the function is mapping x to x^3+3 we could right

$$f : x \mapsto x^3 + 3 \tag{31}$$

Now, say we wanted to know what 4 maps to, we would write $f(4) = 67$, the trouble starts when we write $f(x)$; do we mean the value x goes to under the map, or do we mean the function itself, is $f(x) \in \mathbb{R}$ or is $f(x) \in (\mathbb{R} \to \mathbb{R})$. Frustratingly for people who think of mathematics as an exact and careful process, we sort of assume this will be clear by the context.

This can be a bit awkward with differentiation, no one is every really sure what $df(x)/dx$ means, is it the differential of a function called $f(x)$ or is it the differentiation of f evaluated at x, is it

$$\frac{df(x)}{dx} \text{ or } \frac{df}{dx}(x) \tag{32}$$

This is a problem if you write something like

$$\frac{df(2)}{dx} \tag{33}$$

Is this the differential of $f(x)$ evaluated at $x = 2$ or is it the differential of $f(2)$; it looks like the latter, but that would be silly since $f(2)$ doesn't depend on x and so its differential is zero! To avoid this ambiguity we often use the "restrict to" notation

$$\left.\frac{df}{dx}\right|_{x=2} \tag{34}$$

so, if $f(x) = x^3 + 3$ then

$$\frac{df}{dx} = 3x^2 \tag{35}$$

and

$$\left.\frac{df}{dx}\right|_{x=2} = 12 \tag{36}$$

This hasn't really cleared up what d/dx is, it is in fact an operator, it maps a function to another function:

$$\frac{d}{dx} : (\mathbb{R} \to \mathbb{R}) \to (\mathbb{R} \to \mathbb{R}) \tag{37}$$

but that really is a discussion for another day!

There is another convention that it is useful to mention:

$$\frac{d^2 f}{dx^2} = \frac{d}{dx}\frac{df}{dx} \tag{38}$$

so when writing the derivative of a derivative, we use powers; it isn't worth worry about the precise relationship between taking a power and taking a sequence of derivative, just think of it as a piece of notation, so the thing on the left hand side is just a convenient way to write the thing on the right hand side,

$$\frac{d^3 f}{dx^3} = \frac{d}{dx}\frac{d}{dx}\frac{df}{dx} \tag{39}$$

and so on.

While we are talking about notation, there is another thing to clear up; the use of dots and primes. The d/dx notation makes a lot of sense and provided you don't think it entitles you to make cancellations you can't, it is useful that it makes the chain rule look like you are cancelling the dv's. However, to save space we sometimes use the following two notations, for functions of x we write

$$f'(x) = \frac{df}{dx} \tag{40}$$

and for functions of t

$$\dot{f}(t) = \frac{df}{dt} \tag{41}$$

There is no rule that says you can't use a dot for functions of x or a prime for functions of t, these are just conventions and in a paper you should really make it clear what you are doing, but in informal mathematical discussion we often switch between d/dt or d/dx and a dot or prime without even really noticing we're doing it. The prime notation is also used for higher derivatives,

$$f''(x) = \frac{d^2 f}{dx^2} \tag{42}$$

and so on, when it gets inconvenient to write lots of primes, we use a number in brackets:

$$f^{(4)}(x) = \frac{d^4 f}{dx^4} \tag{43}$$

or even, in the general case

$$f^{(n)}(x) = \frac{d^n f}{dx^n} \tag{44}$$

Summary

This set of notes revises basic calculus using the old-fashioned notion of infinitessimals. We gave a list of standard derivatives and looked at the chain rule.