Introduction to analysis and calculus

This part of the course addresses calculus and analysis[1]. These are fundamental areas of mathematics. They are particularly relevant to computer scientists in the areas of, for example, computational modelling, optimisation and machine learning.

The language of calculus (differentiation and especially integration) will appear in perhaps unexpected areas of your studies, for example when studying expectations in probability theory. In calculus, we study the rate of change. We encounter this in our daily life, for example with economic inflation. The rate of change is also how we model processes on a computer; for example, by looking at the rate of change of how a drug is processed biologically (a set of differential equations), we can extract a description (an equation) of how the drug works. Another application of calculus is to optimisation: by studying the rate of change, we can analyse maxima and minima of functions. This is obviously important in finding out the best/worst outcome of a function, but is also used in optimising the outcomes of functions used in machine learning and AI.

In this section of the course, we'll revise differentiation (in this lecture) and extend our knowledge to multivariate differentiation and partial differentiation. We'll apply this to solving some simple models, and optimisations. Finally some functions are too complicated to be able to solve precisely: we'll look at how calculus appears in helping us find good approximations to solutions.

Differentiation

We will start by 'rediscovering' differentiation. This is an area that will likely be familiar to you from secondary school mathematics, and (unsurprisingly) it's not going to change for this course.

Calculus was 'invented' in the seventeenth century independently and contemporaneously by Newton and Leibniz (and with little cooperation between them; accusations of plagiarism were liberally bandied about – see Wikipedia for a more detailed history of this). It is the study of contin-

[1]This section of the course is based on (or sometimes directly is the) lecture notes by Dr Conor Houghton, who delivered this material in 2022-23. His notes are available at `https://coms10013.github.io/2022-23.html`.

uous change. Both Newton and Leibniz based their version of calculus on "infinitessimals": an *infinitessimal* is a very small number that is 'infinitely close' to zero. So small in fact, that it becomes zero. To study calculus in depth, we would look at sequences, limits of sequences, and notions of convergence. However, this is a deeper study than we will encounter in this course.

Differentiating a straight line

Suppose our function is a straight line of the form $y = mx + c$ (for known values m and c. Our goal is to find out what the rate of change is over an interval, say between x_1 and x_2. This is an overly complicated way of saying: what's the gradient of the function between x_1 and x_2; but since the gradient is constant (it's a line), we're really just asking: "what's the gradient". On the one hand, clearly, it's m, by definition of a straight line. However, another way we could derive this is by evaluating the difference in outputs (the 'y' values) with the difference in inputs:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{(mx_2 + c) - (mx_1 + c)}{x_2 - x_1} = m\,. \tag{1}$$

What we're really interested in though, is finding the rate of change at a single point, say x_1. So for this, we take x_2 to be infinitely close to x_1: let $x_2 = x_1 + h$, with h very close to 0. Because the gradient of a straight line is constant, using (1), we can confirm that the gradient at x_1 is also m.

Differentiating a function

What about if we're using a more complicated function? Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function[2] We'll use the same concepts as we did for a line to find the rate of change at x_1.

Firstly, we'll pick a point that is very close to x_1: $x_1 + h$, for h very small. As we will have picked h to be so small, we'll have 'zoomed in' on the function f so much that it will look like a straight line from our persepective:

[2]To be rigourous, we would need to definie what it means for a function to be continuous. Very loosely, it is enough to understand a continuous function to mean that its pictorial graph has no gaps or 'jumps', as long as we accept that this description is a slightly inaccurate oversimplification.

the rate of change in this interval will then be given by the gradient of this patch of the function:

$$\frac{f(x_1+h)-f(x_1)}{(x_1+h)-x_1} = \frac{f(x_1+h)-f(x_1)}{h}\,.$$

There was nothing special about our choice of x_1, so we can think of it as a variable (rather than a specific value). By taking $h \to 0$, we have rediscovered the derivative of f:

$$\frac{df}{dx} = \lim_{h\to 0} \frac{f(x+h)-f(x)}{h}\,. \tag{2}$$

Here I've used the notation $\lim_{h\to 0}$ as short hand for 'the limit, as h tends to 0'. Dividing by zero, as you well know, is a taboo, and this concept can only work if we have some cancellation in the fraction (that will occur before we take h to zero).

Example

To see how this works in practice, let's consider a simple example:

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

Now, from our discovery of how the derivative is calculated in (2), we have

$$\frac{df}{dx} = \lim_{h\to 0} \frac{(x+h)^3 - x^3}{dx}\,. \tag{4}$$

We can expand out the numerator: $(x+h)^3 = x^3 + 3x^2h + 3xh^2 + h^3$ and so

$$\frac{df}{dx} = \lim_{h\to 0} \frac{3x^2h + 3xh^2 + h^3}{h}\,. \tag{5}$$

The h in the denominator cancels out with an h in the numerator:

$$\frac{df}{dx} = \lim_{h\to 0} 3x^2 + 3xh + h^2 \tag{6}$$

and finally, taking $\lim_{h\to 0}$, we see that the second two terms disappear and so

$$\frac{df}{dx} = 3x^2\,. \tag{7}$$

Differentiating powers of x

Our example extends to any power of x. Let's see how this works. We know (e.g. from Maths A) that

$$(x+a)^n = \sum_{r=0}^{n} \binom{n}{r} x^{n-r} a^r \,. \tag{8}$$

Putting this into the definition of the derivative gives

$$\begin{aligned} \frac{dx^n}{dx} &= \lim_{h\to 0} \frac{\sum_{r=0}^{n} \binom{n}{r} x^{n-r} h^r - x^n}{h} \\ &= \lim_{h\to 0} \frac{\sum_{r=1}^{n} \binom{n}{r} x^{n-r} h^r}{h} \\ &= \lim_{h\to 0} \sum_{r=1}^{n} \binom{n}{r} x^{n-r} h^{r-1} \end{aligned}$$

and, as before, all terms involving h disappear as we apply $\lim_{h\to 0}$, so we're left with only the term $r = 1$ in the summation so that

$$\frac{dx^n}{dx} = nx^{n-1} \,. \tag{9}$$

Rules for differentiating

The sum rule tells us that differentiating the sum of two functions is the same as summing their derivatives:

$$\frac{d}{dx}\left[f(x) + g(x)\right] = \frac{d}{dx} f(x) + \frac{d}{dx} g(x) \,. \tag{10}$$

We can extend this to sum any number of functions; by writing $h(x) = -g(x)$, we can also deduce an analogous *difference rule*.

We can also multiply by a constant c:

$$\frac{d}{dx}[cf(x)] = c\frac{df}{dx} \,. \tag{11}$$

This tells us that differentiation is a linear operation.

We also have the product rule:

$$\frac{d}{dx} f(x)g(x) = f(x)\frac{d}{dx} g(x) + \left(\frac{d}{dx} f(x)\right) g(x) \,. \tag{12}$$

The product rule is less straightforward than the previous two rules, so let's see how we'd discover it.

First, notice that we can write $\lim_{h\to 0} f(x+h)$ (very inefficiently) as

$$\lim_{h\to 0} f(x+h) = \frac{h\left(f(x+h) - f(x) + f(x)\right)}{h} = h\frac{df}{dx}(x) + f(x)\,. \tag{13}$$

We can repeat this for $g(x+h)$; substituting these inefficient rewrites into our derivative equation turns out to be a useful thing to do:

$$\frac{d}{dx}f(x)g(x) = \lim_{h\to 0} \frac{\left[h\frac{df}{dx}(x) + f(x)\right]\left[h\frac{dg}{dx}(x) + g(x)\right] - f(x)g(x)}{h}\,. \tag{14}$$

Now, all is left to do is to expand out the numerator and observe the cancellation of the denominator that automatically arises:

$$\frac{d}{dx}f(x)g(x) = \lim_{h\to 0} \frac{df}{dx}g(x) + f(x)\frac{dg}{dx} + h\frac{df}{dx}\frac{dg}{dx}\,. \tag{15}$$

In the now familiar way, the term involving h disappears and we've recovered the product rule as desired.

We can differentiate just about anything

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

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

$$\frac{du(v(x))}{dx} = \frac{du}{dv}\frac{dv}{dx} \tag{16}$$

This allows us to work out the derivative for function that are written as a composition, a function of a function. This is the machinery that means we can differentiate 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 optimise, 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

$$f(x) = (2 + x^2)^3 \tag{17}$$

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

$$f(x) = 8 + 12x^2 + 6x^4 + x^6 \tag{18}$$

and so

$$\frac{df}{dx} = 24x + 24x^3 + 6x^5 \tag{19}$$

However, we could also write

$$f(v) = v^3 \tag{20}$$

where

$$v = 2 + x^2 \tag{21}$$

So

$$\frac{df}{dv} = 3v^2 \tag{22}$$

and

$$\frac{dv}{dx} = 2x \tag{23}$$

Substituting back in for v and applying the chain rule:

$$\frac{df}{dx} = 6x(2 + x^2)^2 \tag{24}$$

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

$$f(x) = \exp x^2 \tag{25}$$

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

$$\frac{df}{dx} = 2x \exp x^2 \tag{26}$$

A note on notation

There are lots of different ways to write 'differentiate $f(x)$', and you may have seen or will seen a variety of these already:

1. $\frac{df}{dx}$ – this is the notation we've been using here
2. $f'(x)$ – pronounced *f prime of x*
3. $\dot{f}(x)$ – this notation is used more in physics (where you might have more commonly seen it as $\dot{\mathbf{x}}$).

The uses of different notation is usually due to cultural differences between disciplines. The first notation (which we'll continue to use, and is the most common) is Leibniz's notation; the second notation is known as Lagrange notation (although it was actually first used by Euler); the third is Newton's notation.

What if we want to differentiate a function at a point? Say we'd like to differentiate the function f at the value $x = 3$. Using Lagrange's notation (item 2 above), it's fairly straightforward how to do this: $f'(3)$. Similarly, using Newton's notation, we'd write $\dot{f}(4)$. What about *our* notation (or rather Leibniz's notation, item 1 above)?

What about something like $\frac{df(4)}{dx}$? This is actually slightly ambiguous – is this the differential of $f(x)$ evaluated at $x = 4$ or is it the differential of (the constant) $f(4)$? To avoid this ambiguity we often use the "restrict to" notation

$$\left.\frac{df}{dx}\right|_{x=4}. \tag{27}$$

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

$$\frac{df}{dx} = 3x^2 \tag{28}$$

and

$$\left.\frac{df}{dx}\right|_{x=4} = 48\,.$$

Higher derivatives

There is another convention that it is useful to mention:

$$\frac{d^2 f}{dx^2} = \frac{d}{dx}\frac{df}{dx} \tag{29}$$

so when writing the derivative of a derivative, we use powers. For example

$$\frac{d^3 f}{dx^3} = \frac{d}{dx}\frac{d}{dx}\frac{df}{dx} \tag{30}$$

and so on.

Higher derivatives are calculated in exactly the same way as the initial derivative, by applying (2) as many times as desired. The initial derivative tells us about the *rate of change* of a function; higher derivatives tell us about the rate of change of that rate of change. For example, in economics, inflation tells us how the value of money changes over time; the rate of inflation is a second order derivative, and tells us how quickly inflation is changing.

Finally, whereas f is a *function* (it maps a space to a space), d/dx is an operator: it maps a function to another function:

$$\frac{d}{dx} : (\mathbb{R} \to \mathbb{R}) \to (\mathbb{R} \to \mathbb{R}) \tag{31}$$

but that really is a discussion for another day!

Summary

This set of notes revises basic calculus. We looked at the sum rule, the constant multiple rule and the product rule. Most importantly, we looked at the chain rule. We gave a list of standard derivatives.