Math Review for Biologists 2017

Marcelo O. Magnasco
TA: Ana Hočevar
Rockefeller University, New York

February 16, 2017
Course outline.

The course will convene January through March 2017; Lectures, Mondays 9-11AM, plus Tuesdays 3-5 pm for Lab sessions (discussing exercises/hw/Python).

These rough notes are recycled from the previous editions of the course, and will be adjusted to reflect what actually is explained through the course. In particular, we are made a major shift from MATLAB to Python as a programming environment in 2014.

Quantitation has been at the heart of Biology for a long time. Mendel’s work was careful counting. Watson and Crick’s epochal Nature paper, perhaps the most important paper in Biology in the last century, was just a research announcement, the display of the solution to a problem; but the problem was worked out by careful computation of the Fourier transform of a spiral. Similarly, Hodgkin and Huxley’s epochal papers, arguably amongst the most important in Neuroscience last century, were all so detailedly quantitative that they were able to derive, from careful measurement of the initiation of action potentials, how many subunits ion channels had — years before ion channels actually were discovered. Quantitation had a lull for a while, during the first explosion of molecular biology methods; for a while the most quantitation required for a large swath of practicing biologists was to measure the bands in a gel. But with the advent of genomes and high-throughput techniques, quantitation has resurged with a vengeance, as the dimensionality of measurements in Biology has increased dramatically.

The first aim of the course is to teach useful stuff. Except for where the elegance of a proof serves as education of what a formal proof is all about, I entirely avoid proving anything, as opposed to quickly calculating why certain assertions are true. Mathematics, Statistics, Physics and Computation are extraordinarily useful in Biology, to the point that pretty soon they will be indispensible; however, it is not plausible to teach these disciplines in their entirety, so we need to carefully choose the basic and the useful; and as tempting as it is to go into side branches and explain everything at length, we shall try to rein in those urges.

The second aim is to introduce students to ideas that, ideally, every PhD in Science in the 21st century should be familiar with. What exactly is a p-value? (you can’t say why we should believe your thesis if you don’t know). What are computability and decidability? (arguably amongst the most important ideas in Math in the 20th century). What is a Kalman filter? (It landed man on the moon and may underlie much of neural sensory processing). What is the Fast Fourier Transform? (It has saved countless lives through its role enabling all biomedical imaging, and crystallography would not exist without it). What is recursion?

The course starts on precalculus and moves to basic calculus. We rapidly cover limits and derivatives, applications of the derivative, and give a first, superficial view to inversion of the derivative (integration and differential equations) and Taylor expansion.

Take a detour through linear algebra then: vector spaces, subspaces, linear operators. Choice of a basis transforms abstract operations to matrix operators. Examples of linear algebra: the powers of a matrix representing a graph count paths. The solutions of a linear differential equation are a vector space.

We move through computation and computability. Basic computational complexity. Divide and conquer, heapsort and FFT. Dynamic programming, explain Dijkstra and Smith-Waterman-Sellers.
Chapter 1

Getting up to speed: Precalculus.

1.1 Introduction

The first thing you should understand is that you have an incorrect misconception about how math is done. Math is not done with your brain. Math is done with your hands. Let me explain.

I could stand at a blackboard and wave my arms all morning explaining how to perform any routine experimental procedure. From how to pour a gel to how to do a craniotomy. You would leave the classroom convinced you understood and you can do it And then you try and... no, you did not learn it yet. There will be no substitute for trying and failing until you do not fail anymore.

Well, math is exactly the same. I could stand at the blackboard all morning explaining how to take derivatives (in fact, I shall), but there is no substitute for sitting down and taking derivatives yourself. So, while in order to understand what the operations are, what they mean, what are they used for, requires use of your brain, using those operations in a calculation requires you to have internalized them well enough that they are automatic — if it is a struggle to take a derivative, then you will lose your bearings any time you attempt to do a calculation that requires derivation.

A similar misconception pertains to reading a math (or physics or compsci) textbook. It looks like English. It ain’t English. You cannot read a math textbook the way you read a cell biology textbook like Alberts et al. You will read it, you will think you understood, you will close the book and all of those explanations will efervesce away.

Points about developing the basic handicraft. Parts of math that you “do with your hands”. Points about how to read a math textbook.

1.2 Numbers

Explain N, Z, Q, R, C, \( R^n \otimes m \)

1.3 Basic Algebra

Basic algebra is like chess, it has “moves”. You start by noticing that if you do the same thing to both sides of an equality it stays an equality, and use that to learn to move things from side to side and top to bottom. Then you learn how to gather and spread things like common factors. Once you learn the moves, like chess, you learn sequences of moves that allow you to accomplish more complex objectives. Solving an equation then is the planning of a set of moves that will take you from here to there.
Go through the detailed motions of expanding \((x + 1)^2\) →

\[(x + 1) \cdot (x + 1) = x(x + 1) + 1(x + 1) = xx + x1 + 1x + 11 = x^2 + 2x + 1\]

Moving terms right to left, etc, take \(x + xy = 3y\) and solve for \(x\) and then for \(y\).

Exercises:
- Solve for \(y\): \(x = \frac{ay + b}{cy + d}\),
- Expand \((x + 1)^4\)...

(Construct and explain Pascal’s triangle)

\[
\begin{array}{cccccc}
1 & & & & & \\
1 & 1 & & & & \\
1 & 2 & 1 & & & \\
1 & 3 & 3 & 1 & & \\
1 & 4 & 6 & 4 & 1 & \\
1 & 5 & 10 & 10 & 5 & 1 \\
1 & 6 & 15 & 20 & 15 & 6 & 1
\end{array}
\]

### 1.4 Absolute value

The absolute value is the notion of distance on the line. Given Pythagoras’ thm, the distance from a point \((x, y)\) to the origin is given by \(\sqrt{x^2 + y^2}\), which on the line \(y = 0\) becomes \(\sqrt{x^2} = |x|\) (The symbol \(\equiv\) means “equal, by definition”; this is meant to remind you that there is nothing deep about the equality, but, rather, that it is used to define the terms in question). Usually useful to understand absolute value in geometrical terms (by imagining disks), but if all else fails it’s ok to do brute force from the definition.

Properties of the absolute value: \(|2| = 2, \ |−2| = 2\). From there we get \(|x| = 2\) to mean the set \(x = \pm 2\). Then, \(|x| < 2\) is the “one dimensional disk” of points which are closer to the origin than to the number 2: \(x \in (−2, 2)\) where the parenthesis notation means “all points between”, and the round parenthesis do not include endpoints (an open set) while the square parenthesis include endpoints [a closed set]. So \(|x| \leq 2 \rightarrow x \in [−2, 2]\) since 2 and -2 both satisfy the equation.

Go through: \(|x| + x/2 < 3\)

### 1.5 The exponential and the log.

Innumerable examples in biology of exponentials. (Go through some; bacteria, Fechner, lungs).

Take white and black putty. Flatten them into

The exponential is defined through “raising to a power”, the shorthand for multiplying by the same stuff a given number of times. Because the full \(x^y\) function for all real \(x\) and \(y\) is defined as an “extension” of the kindergarten definition of \(a^b\) for positive integers \(a\) and \(b\), many times the properties of the exponential can be understood simply by thinking of the integer powers, and then trying to do the appropriate extensions. So

\[2^7 = \underbrace{2222222}_4 = 2^4 \cdot 2^3\]

(just count!!!) and hence the formula

\[n^{a+b} = n^a \cdot n^b\]

comes from just counting how many times we’ve multiplied by \(n\). Most, if not all, properties of the exponential can be understood in these terms without recourse to higher whatever.
CHAPTER 1. GETTING UP TO SPEED: PRECALCULUS

The extensions of the exponential from the integer “power” definition go as follows. If the product formula is to hold for negative exponents, then at the very least \( x^1x^{-1} = x^0 \) needs to hold. Hence the product formula cannot hold unless we define \( x^{-1} \equiv \frac{1}{x} \). From whence \( 2 \equiv \frac{1}{2} \). (Again, the dot on the equal means this is the definition of a negative power). From there, \( 2^{-3} = \frac{1}{2} \frac{1}{2} \) and it is easy to verify that with this simple definition the formula holds for arbitrary negative integer powers. Then we do know how to multiply reals, so we can do \( 2.732^3 \) so it’s trivial to see the formula extended to arbitrary integer powers of real numbers.

Finally, at least for a couple of chapters, \( 2.732^{-3.178} \) will be carried out by pressing the \( x^y \) key on the Casio: all that we care for now is that it a well-defined extension, which agrees with the basic definition on the integers, and does nothing crazy outside them. The holdout will be trying to understand what a negative number raised to a fractional power looks like — in other words, \(-1^{0.5}\), the imaginary unit. For the time being let’s avoid raising negative numbers to fractional powers.

For instance, the exponential has important *monotonicity* properties:

- \( x^n \) is monotonically increasing in \( x \) for \( x > 0 \).
- \( x^n \) is a monotonically increasing function of \( n \) for \( x > 1 \), and monotonically decreasing in \( n \) for \( 0 < x < 1 \).

Some properties have to do with obvious issues of preserving the properties.

- multiplying by 1 is the same as nothing, so \( 1^n \equiv 1 \)(the triple equal means “always equal”),
- multiplying anything by zero is zero, so \( 0^n \equiv 0 \) \( \forall n > 0 \) (the inverted A means “for all”)
- \( e^0 \) means “a piece of blank paper”, so we define it as \( \equiv 1 \). Since \( 17 + 0 = 17 \) for all possible values of 17, in order for the formula \( e^{a+b} = e^ae^b \) to be extensible to \( a = 0 \), we need \( e^0 = 1 \). See for yourself.
- Then \( e^{-1} = 1/e \) can be seen as required by the extension of the sum formula to negative \( a \): \( e^0 = e^1e^{-1} \). Please work it through and convince yourselves that there is no other choice. When there is no other choice, we say the extension is “natural”.

Exercise: solve for \( x \) in \( e^{2x} - 10e^x + 25 = 0 \).

It’s important to understand that the exponential function grows extremely rapidly. Demonstration: kneading together two pieces of putty until they are mixed down to molecular dimensions requires 27 kneads since \( 2^{-27} \) cm is about 1 angstrom.

Exponentials in biology: go through recursive levels of division in lung, trees, rabbit growth, bacterial log phase, exponential damping (photobleaching, lysing...), decibels...

From here we get to the logarithm, which is defined as the function which “undoes” the exponential:

\[ \ln e^x \equiv x \]

which means: \( \bullet = \ln \bullet \) is given by finding \( \bullet \) satisfying \( e^\bullet = \bullet \).

Again, the logarithm increases quite slowly, as slowly as the exp is fast.

Log to different bases, to invert the more general equation \( \log_a(a^x) \equiv x \)

### 1.6 Functions and their properties

Functions are correspondences between a *domain* and a *range*. Usual notation is \( f : A \rightarrow B \), which means that \( f \) takes an argument which is an element of the domain \( A \) and returns a value within the range \( B \). The function has to depend exclusively upon \( A \) in order to be a function. For example, if \( A \) is a set of 7 coins, tossing them and counting how many heads is not a function, because it will not give us the same number each time we try, since it depends on something other than the set of seven coins—the toss. On the other hand, we
can define a set $A$ of all possible configurations of a set of seven coins: there are 128 such, for example $↑↑↓↑↓↓↑$ etc. On this set $A$ we can define a function that counts how many heads, so $B$ would be the set of integers from 0 to 7 inclusive; then $f(↑↑↓↑↓↓↑) = 4$ and so on and so forth.

Draw domain and range.

In calculus books thorny issues of the exact domain and range are usually brushed aside since we only consider functions $f : \mathbb{R} \to \mathbb{R}$ (where $\mathbb{R}$ is the set of real numbers).

Functions can be combined to make new functions. Addition of functions proceeds in the “natural” fashion: $(f + g) : \mathbb{R} \to \mathbb{R}$ is a function defined by pointwise addition:

$$(f + g)(x) \doteq f(x) + g(x)$$

Please note! The first $+$ sign there denotes addition of functions, while the second $+$ denotes addition of real numbers. While this is an abuse of notation, $(f + g)(x)$ is cumbersome and strictly for wimps. Because the two $+$s are different things, and the first one is being defined through this equation, the equal has the customary dot.

Example: we can have a $\#$heads function and a $\#$tails function, their sum is $N$. We do not need to have to sum the domains, just the ranges. But we do need to have the same domains to add.

In similar fashion we define substraction, multiplication and division. Since dividing by zero is “bad”, we have to be careful when dividing by a function whose values include zero. Brief aside: the set of values $x$ satisfying $f(x) = 0$ are called the “zeroes” of $f$. Strictly speaking not to abuse notation we would have to write something like $\text{zeros}(f) \doteq \{x \ni f(x) = 0\}$ but.

A new and interesting operation between functions is called “composition”, and is defined by applying one after the other. The image being pressing two function keys after another on the Casio.

$$(g \circ f)(x) \doteq f(g(x))$$

Care has to be exercised with the order. Nobody uses the $f \circ g$ notation except when making it clear that functional composition is intended as opposed to multiplication, and so it is rarely seen outside of textbooks, and so there may be disagreement on whether $f(g(x))$ or $g(f(x))$ is intended; the problem is that the guy on the right is the first one to act, so $f \circ g$ should be read “g composed with f” since the first key pressed on the Casio is $g$ and then $f$ is pressed.

In any event functional composition is a wonderful and tricky thing. First thing to note is that it is not symmetric: $f \circ g$ and $g \circ f$ are not at all the same thing; $\sin \exp x \neq \exp \sin x$.

An important use of functional composition is to define inverses. Given $f$, there might be a function $g$ satisfying the equation

$$(f \circ g)(x) \equiv x$$

(the $\equiv$ means this has to be equal “for all values of $x$”). There is not always a functional inverse, in particular when a given value of the function is obtained for several different values of the argument. Also, the range of the inverse must be the domain of the original function and viceversa, which can pose problems of exactly what is the domain: $\sqrt{x^2} = (\sqrt{x})^2$ for $x > 0$, but they differ for negative values of $x$: $\sqrt{x^2} = |x|$, while $(\sqrt{x})^2$ is not defined for negative $x$ (the Casio says “error” upon trying). An even worse case is the exponential/log pair: $\log(\exp(x))$ is defined $\forall x \in \mathbb{R}$ because any real, whether positive or negative, can be exponentiated. OTOH, $\exp(\log(x))$ is only defined for $x > 0$ since the log of a negative number is an error.


Chapter 2

Limits and derivatives.

2.1 Distance on the plane

The distance formula between two points on the plane: since \( c^2 = a^2 + b^2 \), we get \( d^2 = (x_1 - x_0)^2 + (y_1 - y_0)^2 \).

![Distance formula between two points on the plane](image)

Proof of Pythagoras’ thm: cut triangles and move them to form a figure made of two squares whose sides are the sides of the original triangle...

![Proof of Pythagoras’ thm](image)

2.2 The linear equation

Graphs of functions: vertical lines must intersect only once.

Linear equation is \( y = mx + n \). How much does the line rise vertically when we displace the horizontal by \( \Delta x \) ?

\[
\begin{align*}
 x & \rightarrow mx + n \\
 x + \Delta x & \rightarrow m(x + \Delta x) + n = mx + n + m\Delta x
\end{align*}
\]
so
\[ \Delta y = m \Delta x \rightarrow m = \frac{\Delta y}{\Delta x} \]

### 2.3 Limits

The notion of limit embodies the idea of “where would the function go if it did the sensible thing”. Proper definition requires “\( \forall \varepsilon > 0 \exists \delta > 0 \ni \) ...” (“for every epsilon bigger than zero there exists a delta bigger than zero such that ...”). We shall not take this approach since you’ve already done this before.

Limits may or may not exist. \( \lim_{x \to 0} \left| \frac{x}{x} \right| \) does not exist, because the function is \( \equiv 1 \) on the right and \( \equiv -1 \) on the left, so the “right limit” and the “left limit” do not agree. Similarly, \( \lim_{x \to -2} \sqrt{x} \) does not exist because there’s nothing on the immediate sides of \(-2\) to take the limit from: the function ceased existing at \(0\). Finally, \( \lim_{x \to 0} \frac{1}{x} \) does not exist because the function “blows up”. So, in order for the limit to exist, the following conditions must be satisfied.

1. the function has to be defined in the neighbourhood of the limit point, existance of limits
2. the function has to stay bounded within this neighbourhood, and
3. the limit has to be the same from every direction (just both sides in the case of one real variable).

If the limit does exist and it agrees with the function’s value, then the function is continuous at that point.

Example:

\[
\lim_{x \to -3} \frac{x^2 - 9}{x + 3} = \lim_{x \to -3} \frac{(x + 3)(x - 3)}{x + 3} = \lim_{x \to -3} (x - 3) = -6
\]

where we were able to simplify the \(x + 3\) above and below because within the limit sign, \(x\) is supposed to be \(very\) close to \(-3\), but \(not\) \(actually\) \(-3\), where we would not have been able to simplify.

A note on the equality of functions: in order to be equal, functions must agree not just numerically, but also on their domains and ranges. If the domains are not equal, the functions are not equal, even though they might agree everywhere they are defined. So, for instance, consider the functions

\[
f : \mathbb{R} - \{1\} \to \mathbb{R} \quad f(x) = \frac{(x^2 - 1)}{(x - 1)} \\
g : \mathbb{R} \to \mathbb{R} \quad g(x) = x + 1 \\
h : \mathbb{R} - \{1\} \to \mathbb{R} \quad h(x) = x + 1
\]

then \(f \neq g\) (since the domains are different) but \(f = h\) (since the domains are the same, and the numerical value of the function happens to be identical for every point of the domain). Usual confusion arises because in practice we do not specify the domain of the function explicitly, and a lot of gnashing of teeth would be saved if we did, as above. What must be born in mind is that, by default, if we write \((x^2 - 1)/(x - 1)\) we mean \(f\), since the numerator explicitly forbids \(\{1\}\) from the domain, while if we write \(x + 1\) we mean \(g\), not \(h\). By convention, the domain is the largest one conceivable from the way the function is written. If two functions agree exactly on their values whenever both are defined, but one continuation has a strictly larger domain (as \(f \) and \(g\)) we say the larger one is a continuation of the smaller one: \(g\) continues \(f\).
2.4 Derivatives

The derivative of a function \( f \) is another function, usually called \( f' \) or \( \frac{df}{dx} \), which, at every point \( x \), gives us the slope of the function \( f \) at that point. This slope is obtained as a limit of the local change in the function \( \Delta f \) (or, if we are thinking in terms of \( y = f(x) \), the customary \( \Delta y \)) divided by \( \Delta x \) as the latter tends to zero.

\[
\frac{df}{dx} = f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}
\]

Let us check that our definition, which is based on the notion that the slope is a derivative of a function and gives you back another function, its derivative. It is called an operator. We do it as

\[
(f + g)' = \lim_{\Delta x \to 0} \frac{(f + g)(x + \Delta x) - (f + g)(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) + g(x + \Delta x) - f(x) - g(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x) + g(x + \Delta x) - g(x)}{\Delta x} = f'(x) + g'(x)
\]

\[
(\alpha f)' = \lim_{\Delta x \to 0} \frac{\alpha f(x + \Delta x) - \alpha f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\alpha f(x + \Delta x) - \alpha f(x)}{\Delta x} = \alpha f'(x)
\]

\[
\frac{d}{dx} x^2 = 2x:
\]

\[
\lim_{\Delta x \to 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x} = \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + \Delta x^2 - x^2}{\Delta x} = \lim_{\Delta x \to 0} 2x + \Delta x = 2x
\]

where please note that the third term in the expansion of the square dies out altogether in the limit since it is overall proportional to \( \Delta x \). This is important.
CHAPTER 2. LIMITS AND DERIVATIVES.

• \( \frac{d}{dx} x^n = nx^{n-1} \):

\[
\lim_{\Delta x \to 0} \frac{(x + \Delta x)^n - x^n}{\Delta x} = \lim_{\Delta x \to 0} x^n + \binom{n}{1} x^{n-1} \Delta x + \frac{1}{2} \binom{n}{2} x^{n-2} \Delta x^2 + \cdots + \Delta x = \lim_{\Delta x \to 0} \frac{x^n + n x^{n-1} \Delta x + \frac{n(n-1)}{2} x^{n-2} \Delta x^2 + \cdots}{\Delta x} = nx^{n-1} + \frac{n(n-1)}{2} x^{n-2} \Delta x + \cdots = nx^{n-1}
\]

where we have used the binomial expansion formula (remember the first exercise?); the subsequent terms all are proportional to \( \Delta x \) to some power and thus die out just like above. If you are annoyed by the binomials you can try to derive this by induction, or using the chain rule...

• \( \frac{d}{dx} \sqrt{x} = \frac{1}{2\sqrt{x}} \): This is just a special case of the one above, but since we derived the case above only for integer \( n \) we will show this now to make it plausible that the formula above works for non-integer values of \( n \):

\[
\lim_{\Delta x \to 0} \frac{\sqrt{x + \Delta x} - \sqrt{x}}{\Delta x} = \lim_{\Delta x \to 0} \frac{\sqrt{x + \Delta x} - \sqrt{x}}{\Delta x} \frac{\sqrt{x + \Delta x} + \sqrt{x}}{\sqrt{x + \Delta x} + \sqrt{x}} = \lim_{\Delta x \to 0} \frac{x + \Delta x - x}{\Delta x (\sqrt{x + \Delta x} + \sqrt{x})} = \lim_{\Delta x \to 0} \frac{1}{\sqrt{x + \Delta x} + \sqrt{x}} = \frac{1}{2\sqrt{x}}
\]

We now turn to some central tricks.

_Last class we saw the basic rules of differentiation. Today we shall review them and try to use derivatives for something useful._

### 2.6 Handy Summary of Rules and Formulas

The **three fundamental rules** are the ones allowing us to derive sums, products, and functional composition.

Derivative of linear combinations: \( (af + bg)' = af' + bg' \)

(Where \( a \) and \( b \) are numbers)

Derivative of functional composition: \( \frac{d}{dx} f(g(x)) = f'(g(x))g'(x) \)

Or also \( y = f(x), z = g(y), \frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx} \). Derivative of products: \( (fg)' = f'g + fg' \)

Then we have some formulas:

\[
\frac{d}{dx} x^n = nx^{n-1} \text{ (valid for any } n \text{, in particular } \frac{d}{dx} \sqrt{x} = \frac{1}{2\sqrt{x}} \)
\]

\[
\frac{d}{dx} e^x = e^x \quad \frac{d}{dx} \ln x = \frac{1}{x} \quad \left( \frac{1}{x} \right)' = -\frac{1}{x^2} \text{ (obtained by applying the chain rule with } f = g \text{ and } g = 1/y) \]

We now turn to some central tricks.
2.7 The chain rule

The chain rule states how to take the derivative of a functional composition. Much like several chained gears,

\[ f(x) \rightarrow g(y) \rightarrow z = g(f(x)) \]

then if we want to take a derivative we need to see how an increment propagates through the chain

\[ f(x + \Delta x) \rightarrow g(y + \Delta y) \rightarrow z + \Delta z = g(f(x + \Delta x)) \]

so when we want to compute the derivative of the whole we write out the limit and then multiply and divide by \( \Delta y \), and swap \( \frac{d}{dx}g(f(x)) \).

\[
\frac{\Delta z}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta z}{\Delta y} \frac{\Delta y}{\Delta x}
\]

and then by sleigh of hand we separate the latter onto two different limits

\[
= \lim_{\Delta y \to 0} \frac{\Delta z}{\Delta y} \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}
\]

The only step offering some difficulty in the proof is the latter, because in order to show the single limit equals the product of the two limits, you will need to use the fact that both limits in the product exist: hence the chain rule requires that both derivatives exist.

Let’s do this in standard notation to know exactly what the arguments to both the derivatives are:

\[
\frac{d}{dx}g(f(x)) \equiv \lim_{\Delta x \to 0} \frac{g(f(x + \Delta x)) - g(f(x))}{\Delta x}
\]

we see that if we define

\[ y + \Delta y = g(x + \Delta x) \]

the right hand side then has a derivativoid shape:

\[
= \lim_{\Delta x \to 0} \frac{g(y + \Delta y) - g(y)}{\Delta x}
\]

So please multiply and divide by \( \Delta y \) and note that the limit \( \Delta x \to 0 \) implies \( \Delta y \to 0 \) if the limit of the quotient exists (thus, if the derivative of \( f \) exists!). Then use the original fact that \( \Delta y = g(x + \Delta x) - g(x) \) and ...

\[
= \lim_{\Delta y \to 0} \frac{g(y + \Delta y) - g(y)}{\Delta y} \lim_{\Delta x \to 0} \frac{g(y + \Delta y) - g(y)}{\Delta x} \lim_{\Delta x \to 0} \frac{y + \Delta y - y}{\Delta x}
\]

or in words, the derivative of the outer function evaluated at the inner function, times the derivative of the inner function (evaluated at \( x \)).

I used to derive the product rule too – can’t find it in the notes?
2.8 Exponentials galore

Let us compute the derivative of the exponential:

\[
\frac{d}{dx} e^x = \lim_{\Delta x \to 0} \frac{e^{x+\Delta x} - e^x}{\Delta x} = e^x \lim_{\Delta x \to 0} \frac{e^{\Delta x} - 1}{\Delta x}
\]

and the number \( e \) was defined as the unique number for which the limit marked with brackets equals one. With this definition \( e^x \) is a fixed point of differentiation, something that will give it a unique standing when we solve differential equations.

Using the basic equality \( a = e^{\ln a} \) we can proceed to derive a whole lotta stuff. First we can learn what the derivative of the logarithm is:

\[
\frac{d}{dx} x = 1 = \frac{d}{dx} e^{\ln x} = e^{\ln x} \frac{d}{dx} \ln x
\]

\[
1 = x \frac{d}{dx} \ln x \quad \rightarrow \quad \frac{d}{dx} \ln x = \frac{1}{x}
\]

We can produce further formulas endlessly by manipulating the exponentials up and down...

\[
\frac{d}{dx} a^x = \frac{d}{dx} (e^{\ln a})^x = \frac{d}{dx} e^{x \ln a} = e^{x \ln a} \ln a = a^x \ln a
\]

which shows that \( e \) is indeed the only number for which the prefactor (\( \ln a \)) is one. We can also do the opposite thingie,

\[
\frac{d}{dx} a^x = \frac{d}{dx} (e^{\ln x})^a = \frac{d}{dx} e^{a \ln x} = e^{a \ln x} \frac{d}{dx} a \ln x = e^{a \ln x} a \ln x = a^x e^a
\]

which is now explicitly valid for any real \( a \). (Discuss: even zero?)

2.9 Practice

Various derivatives shall be(ware?) taken at this point, such as

\[
\frac{d}{dx} x^x = \frac{d}{dx} e^{x \ln x} = e^{x \ln x} (\ln x + \frac{x}{x}) = (1 + \ln x) x^x
\]

2.10 Classes of functions which are closed under derivatives

Many classes of functions are “closed” under the operation of taking derivatives. These include:

- polynomials: these are of the form

\[
P(x) = \sum_{i=0}^{\text{degree}} a_i x^i \quad \rightarrow \quad \frac{d}{dx} P(x) = \sum_{i=0}^{\text{degree}-1} a_i x^i
\]

so their derivative is a polynomial.

- rational functions: these are quotients of polynomials

\[
R(x) = \frac{P(x)}{Q(x)} \quad \rightarrow \quad \frac{d}{dx} R(x) = \frac{P' - PQ'}{Q^2} = \frac{P'Q - Q'P}{Q^2}
\]

and since product, subtraction, and derivatives of polynomials are polynomials it’s still rational.
• sum of exponentials: these are sort-of polynomials in $e^x$.

$$P(x) = \sum_{i=0}^{N} a_i e^{\lambda_i x} \quad \rightarrow \quad \frac{d}{dx} \sum_{i=0}^{N} a_i e^{\lambda_i x} = \sum_{i=0}^{N} a_i \lambda_i e^{\lambda_i x}$$

• and then various other things, like products of polynomials times exponentials etc.

Mostly anything else is *not* closed in form under derivatives, because taking a derivative makes the formula “longer” of sorts.
Chapter 3

Applications of derivation.

3.1 Optimization

A smooth function $f$ can be chopped into segments, each of which is monotonic. Within each segment, the derivative $f'$ will be positive or negative. Since $f'$ changes sign upon changing segment, it must cross zero at the segment boundaries, i.e., the maxima and minima of $f$. ($f'$ may become zero within a segment, simply by glancing zero without ever changing sign, like $f = x^3$). Thus, the equation $f'(x) = 0$ has, as solutions, the extrema of $f$ (in addition to potentially other stuff).

First classic example: find the rectangle of unit area with the smallest perimeter. If we call the width and height $x$ and $y$ respectively, then the area $A = xy = 1$ and the perimeter $P = 2x + 2y = 2x + \frac{2}{x}$.

Plot $P$. Minimize $P$:

$$\frac{dP}{dx} = 2 - \frac{2}{x^2} = 0 \rightarrow x^2 = 1$$

Another classic example. Imagine a lifeguard, who can run on sand with speed $v_l$ and swim with speed $v_w$, standing at coordinates $(0, -1)$. There’s a drowning person at $(a, 1)$, and the water edge lies at $y = 0$. Within land, or within water, the best thing is to beeline; but we can enter the water at an arbitrary point $(x, 0)$. The total time is given by

$$t = \frac{1}{v_l} \sqrt{x^2 + 1} + \frac{1}{v_w} \sqrt{(a-x)^2 + 1}$$

and the best time is given as the solution of $dt/dx = 0$:

$$\frac{dt}{dx} = 0 = \frac{1}{v_l} \frac{2x}{2\sqrt{x^2 + 1}} - \frac{1}{v_w} \frac{2(a-x)}{2\sqrt{(a-x)^2 + 1}}$$

While this equation can then be solved for $x$ yadda yadda, it is a lot more instructive to notice that $x$ is the vertical displacement to the crossing point, and $\sqrt{x^2 + 1}$ the direct distance, so that $x/\sqrt{x^2 + 1}$ is the sine of the angle $\theta_l$ between the normal direction to the beachline and the optimal running line. Simile for $(a-x)$ to get

$$\frac{\sin \theta_l}{v_l} = \frac{\sin \theta_w}{v_w}$$

which is Snell’s law of refraction! (The index of refraction of a medium is inversely proportional to the speed of light in it).

The light paths through every optical system follow the law that the time spent by the light is an extremum, either maximum or minimum. This simple principle (Maupartuis) is a lot more general than the individual laws like Snell or reflection or thin lenses’, since all the individual laws can be derived from it.
3.2 Least squares fit

Given a set of $N$ data, $x_i$ and $y_i$, we would like to “fit” a “best” line of the form $ax + b$. What is “best” is to be defined. We might consider the error in fitting to be the “vertical drop” distance to the line; since distance on the line is $||$, this would lead to a total fit error of the form

$$S_1 = \sum_{i=1}^{N} |ax_i + b - y_i|$$

This error, known as the taxicab error, is pretty much intractable from the analytical point of view, fundamentally because the absolute value function does not have a derivative at zero. It’s a lot better to use

$$S_2 = \sum_{i=1}^{N} (ax_i + b - y_i)^2$$

because the fit can be done analytically. In order to get the optimum, we shall take the derivatives of the error wrt $a$ and $b$ and equate them to zero:

$$\frac{dS}{da} = 0 = \sum_{i=1}^{N} 2(ax_i + b - y_i)x_i$$

and

$$\frac{dS}{db} = 0 = \sum_{i=1}^{N} 2(ax_i + b - y_i)$$

which leads to

$$a \sum_{i=1}^{N} x_i^2 + b \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i$$

$$a \sum_{i=1}^{N} x_i + bN = \sum_{i=1}^{N} y_i$$

which can be solved directly. Define $\sum_{i=1}^{N} x_i y_i = \Sigma_{xy}$ etc. ($\Sigma_1 = N$). The second equation gives us $b = (\Sigma_y - a\Sigma_x)/N$, and so inserting into the first one

$$a\Sigma_{xx} + \frac{(\Sigma_y - a\Sigma_x)}{N} \Sigma_x = \Sigma_{xy}$$

from where

$$a = \frac{N\Sigma_{xy} - \Sigma_x \Sigma_y}{N\Sigma_{xx} - \Sigma_x^2}$$

etc.

3.3 Maximum likelihood

Given a probabilistic model for the generation of some data, then we can try to estimate the model parameters through a maximum-likelyhood estimator: i.e., we assume that the model parameters are what makes the data most likely.

Example: assume independent nucleotides in a DNA sequence. Assume a given GC content $g$, to be estimated. Then if given the sequence ACCG, its probability is $(1-g)g^3$, or in other words $P = (1-g)^{\#AT} \ast g^{\#CG}$. What is the value of $g$ that maximizes $P$?

$$\frac{dP}{dg} = 0 = -\#AT (1-g)^{(\#AT-1)} g^{\#CG} + (1-g)^{\#AT} (\#GC) g^{\#GT-1}$$
from where $(\#AT)g = (\#CG)(1-g)$ or

$$g = \frac{\#CG}{\#ACGT}$$

so for this simple case the maximum likelihood estimator of GC content equals the observed frequency.

### 3.4 Complex numbers

We declare a symbol, spelled $i$, for the imaginary unit. This symbol behaves in every single respect like an ordinary number, only that when it is multiplied by itself, it gives minus one: $i^2 = -1$. Thus $i$ is the square root of minus one, a number that does not otherwise exist within the reals. Since we cannot have polynomials in $i$, because all higher powers immediately become $\pm 1$ or $\pm i$, absolutely any algebraic combination containing $i$ will reduce to the form $a + bi$, where both $a$ and $b$ are real numbers; $a$ is the real part and $b$ is the complex imaginary part of the whole, which we call a complex number. Complex numbers can be added, subtracted, multiplied and divided in the obvious way: we just have to remember to substitute $-1$ every time we get a $i^2$... Thus

- $(a + bi) + (c + di) = (a + c) + (b + d)i$
- $(a + bi)(c + di) = ac + adi + bci + bdi = ac + i(ab + bc) + i^2bd = (ac - bd) + (ad + bc)i$
- $\frac{1}{a + bi} = \frac{a - bi}{a^2 - b^2}$

and from there everything else follows. Try, for instance, $\frac{2+3i}{1-2i}$. Actually, you should try your hand at getting the full formula for division and for squaring a number.

### 3.5 The complex exponential

I will follow the unconventional approach of defining what the value of the complex exponential is:

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

i.e., the exponential of a purely imaginary number has a real and an imaginary part, and we define those to be the sine and cosine. By asserting that the complex exponential is to follow all previous properties of the exponential, like addition of exponents or the derivative formula, we can get directly stuff that we would otherwise have to memorize by rote. Example, the addition formula:

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

$$\cos(\theta + \eta) + i \sin(\theta + \eta) = \cos \theta \cos \eta - \sin \theta \sin \eta + i(\cos \theta \sin \eta + \sin \theta \cos \eta)$$

which you might recognize as the two angle addition formulas together: the real part is the cosine of a sum of angles, and the imaginary part is the sin of a sum of angles. By changing $i$ to $-i$ we get

$$e^{-i\theta} = \cos \theta - i \sin \theta$$

since $-i$ shares all the properties of $i$: which is $i$ and which $-i$ is really a matter of convention. Therefore $\cos(-\theta) = \cos \theta$ and $\sin(-\theta) = -\sin \theta$. Now, an elementary property of sin and cos:

$$e^0 = e^{i(\theta-\theta)} = 1 = \cos^2 \theta + \sin^2 \theta$$
and, finally, the derivative formula:

\[ \frac{d}{dx} e^{ix} = ie^{ix} = -\sin x + i \cos x \]

from where

\[ \frac{d}{dx} \cos x = -\sin x \]
\[ \frac{d}{dx} \sin x = \cos x \]

which shows that there’s a bundle of properties of the trigonometric functions that we can deduce simply by remembering the complex exponential.

One of the most intriguing, symbolic formulas in math is deduced directly from above:

\[ e^{i\pi} = -1 \]

a formula linking \( i \) and \(-1\), the two most important numbers introduced to extend the algebraic operations, with \( e \) and \( \pi \), the fundamental units of exponentiation and trigonometry.

### 3.6 Polar coordinates for complex numbers

From the complex exponential, we can get a very compact form for the polar coordinates expression of a complex number. Remember that any number \((x, y)\) can be converted to polar coordinates \( \rho, \theta \) where \( \rho = \sqrt{x^2 + y^2} \) is the “radius”, or distance to the origin, and \( \theta = \arctan \frac{y}{x} \) is the angle subtended from the \( x \) axis. The reverse conversion reads

\[ x = \rho \cos \theta \]
\[ y = \rho \sin \theta \]

and so, if we write \( x + iy \) as a complex number, we get

\[ x + iy = \rho \cos \theta + i\rho \sin \theta = \rho e^{i\theta} \]

a remarkably compact way of expressing the polar version.

It is very helpful, in understanding the geometry in this formula, to think of the following analogy. A real number can be thought of as having an absolute value and a sign as two separate entities, with the sign \( \pm 1 \) and the absolute value being the distance to zero. Similarly, a complex number can be thought of as an absolute value (\( \rho \)) and a sign, \( e^{i\theta} \); the difference being that the sign can now vary continuously. We know from the above formulas that \( e^{i\theta} \) lies on the unit circle.
Chapter 4

A peek of things to come: differential equations and Taylor series.

4.1 Basic Differential equations.

We looked at the problem of the holy bucket. We took a few minutes to discuss granular flow, to discuss the sand clock, and different kinds of flow. We also mumbled how we cannot prove theorems about reality, particularly going over the issue of different kinds of fluid flow, nonnewtonian stuff, starch and cytosol.

In any case we concentrated on the equation

\[ \dot{x} = -\alpha x + I(t) \]

and we looked at two subcases:

4.2 No hole: integrals.

So the equation is \( \dot{x} = I(t) \). “Integrate and fire” neuron was mentioned at this stage. We went though basic integration as an antiderivative.

4.3 No faucet: exponential decay.

So the equation is \( \dot{x} = -\alpha x \), and we solved it. Issues of initial conditions. Got \( x = x_0 e^{-\alpha t} \) as a solution.

4.4 Taylor series.

Given a function \( f(x) \) the straight line that best approximates \( f \) at the origin \( x = 0 \) is \( f(0) + x f'(0) \). Can we make this approximation better? For example is there a parabola which approximates best the function at the origin? The answer is, of course: we want the parabola to have the same intercept as \( f \), the same slope as \( f \), and the same curvature.

Therefore we state: \( p(x) = a + bx + cx^2 \), and setting \( p(0) = f(0) \), \( p'(0) = f'(0) \) and \( p''(0) = f''(0) \) we get

\[
\begin{align*}
    a + 0 + 0 &= f(0) \\
    b + 0 &= f'(0) \\
    2c &= f''(0)
\end{align*}
\]
from where we obtain
\[ p(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} \]
Pause for a second to notice the structure. Because we evaluate \( p \) and its derivatives at 0 we get a single coefficient in each row, in the left hand side: some coefficients were killed because the derivative of a constant is zero, other coefficients are killed because they are multiplied by a nonzero power of \( x \). Therefore in each row there’s a single coefficient, the one in front of the power of \( x \) with the same number as the row (minus one). As a result the coefficient in front of the \( x^1 \) has a single derivative and the one in front of the \( x^2 \) has a second derivative.

This structure suggests that we can keep going. How about finding a polynomial all of whose (nonzero) derivatives agree with those of \( f \) at the origin? We’re going to write that
\[ p(x) = \sum_{i=0}^{N} c_i x^i \]
and then request that the \( n^{th} \) derivative of \( p \) agree with that of \( f \)
\[ \frac{d^n}{dx^n} p = \frac{d^n}{dx^n} f \]
when evaluated at the origin:
\[ \frac{d^n}{dx^n} p \bigg|_{x=0} = \frac{d^n}{dx^n} f \bigg|_{x=0} \]
where the vertical bar on the right side means we’re going to take that function to the left, and evaluate it at the condition specified in the bar. Silliest notation ever but at least it is not ambiguous. To do this we need to compute
\[ \frac{d^n}{dx^n} x^i \bigg|_{x=0} \]
in principle, for any combination of \( i \) and \( n \). The first observation is that if we take too many derivatives we kill it: every derivative ends up with one less power, but as we get to \( x^0 \) we get a constant and the derivative of a constant is zero, and zero is a constant, so if we keep taking derivatives after that we keep getting zero. Therefore if \( n > i \) this number is zero. On the other hand, if we take fewer derivatives, we still have a positive power of \( x \) as we evaluate at \( x = 0 \) and hence we also get a zero. So we’re stuck with computing the fourth derivative of \( x^4 \) and so on.

What do we get? The first derivative of \( x^4 \) gets us \( 4x^3 \), the second \( 4 \times 3 x^2 \) the third \( 4 \times 3 \times 2 x \) and the last one \( 4 \times 3 \times 2 \times 1 \), and then we better stop or we’ll kill it. Not hard to see where this is going is it?
\[ \frac{d^n}{dx^n} x^n = n! \]
(\text{where it is immaterial now whether you evaluate at zero or not, since it is now a constant}).
Thus, coming back to
\[ \frac{d^n}{dx^n} p \bigg|_{x=0} = \frac{d^n}{dx^n} f \bigg|_{x=0} \]
we get
\[ \frac{d^n}{dx^n} p \bigg|_{x=0} = c_n n! = \frac{d^n}{dx^n} f \bigg|_{x=0} \]
or in other words,
\[ c_n = \frac{1}{n!} \frac{d^n}{dx^n} f \bigg|_{x=0} \]
or in a format that may be easier to remember

\[ p(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(0)\frac{x^3}{3!} + f''''(0)\frac{x^4}{4!} + f'''''(0)\frac{x^5}{5!} + \cdots \]

We say that \( p(x) \) is the Taylor expansion of \( f \) in the limit in which we make \( N \) arbitrarily large.

### 4.5 Properties of the Taylor expansion

So is this like the hottest thing since sliced bread or what? Let’s take a look at what we get for various functions before getting all crazy.

Obviously the easiest function to try is one in which we can compute all the derivatives explicitly.

**Taylor expansion of a polynomial**: If we take two different polynomials and we request that all of their derivatives match, something interesting happens. First of all, if the two polynomials have different orders this cannot be done: if we take more derivatives than the smallest of the two orders, we get coefficients (on the longest polynomial) that can only be matched by setting the coefficient to zero, thereby lowering the order of the higher-order polynomial. Then if they have the same order, we will just make the coefficients equal. Therefore, the **Taylor expansion of a polynomial is the same polynomial**.

The next obvious thing to try is the exponential. All the derivatives of the exponential are the same exponential, and \( e^0 = 1 \). So calling \( e_T(x) \) the Taylor expansion of the exponential we get

\[ e_T(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \frac{x^7}{5040} + \cdots \]

Let us see if this works. Let’s write a Python program to evaluate this Taylor series.

```python
x=1
e=1
f=1.0
for i in range(1,10):
    f=f*i  # what does this mean?
    e=e+x**i/f
    print i,e
```

which outputs

<table>
<thead>
<tr>
<th>i</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>2.66666666667</td>
</tr>
<tr>
<td>4</td>
<td>2.70833333333</td>
</tr>
<tr>
<td>5</td>
<td>2.71666666667</td>
</tr>
<tr>
<td>6</td>
<td>2.71805555556</td>
</tr>
<tr>
<td>7</td>
<td>2.71825396825</td>
</tr>
<tr>
<td>8</td>
<td>2.71827876984</td>
</tr>
<tr>
<td>9</td>
<td>2.71828152557</td>
</tr>
</tbody>
</table>

where we see the variable \( e \) converges to the mathematical value of \( e \).

For those who are quick and care about efficiency: notice we’re not evaluating the factorial from scratch every iteration up there. Why? On the other hand we’re using \( x**i \) in the loop, is there a way to make this more efficient?

Take the above code and “instrument it”, namely add facilities in the code to monitor whether the variable \( e \) converges to the correct value. Then explore whether this in fact happens, for various values of \( x \), so as to survey what are the values of \( x \) for which the Taylor expansion of the exponential converges to the right value?
Chapter 5

Linear Algebra

5.1 Vector spaces

We are going to work with sets called vector spaces. They are defined so as to have two elementary operations: you can add two vectors, and the result is within the vector space; and you can change the size of a vector by multiplying it with a “scalar”, namely the number that sets the scale or size.

Thus all vector spaces are defined “over” a space of scalars $K$, which for us shall be the real numbers or the complex numbers. (Watch out: since each complex number packs two reals, a 3-dimensional complex space is 6-dimensional w.r.t. the reals). We express this symbolically as

\[
\forall v, w \in V \rightarrow v + w \in V
\]

(which is read alouds as “forall v and w in V then v plus w is in V”) and

\[
\forall v \in V \forall a \in K \rightarrow av \in V
\]

(which is read aloud as “forall v in V and forall a in K, a times v is in V”).

If we were free to make up any stuff we wanted we would soon run into a contradiction. Thus the vector space needs to satisfy some “axioms”, or conditions telling us it is a well-behaved and tame vector space that does everything a nice vector space is supposed to. There are four axioms related to vector addition, and four axioms related to scalar multiplication. In the following, $u, v, w \in V$, the vector space, and $a, b \in K$, the scalar field (usually the reals). The gist of the axioms is to guarantee that the vector space operations work together correctly with the operations on the scalars proper.

- $u + v = v + u$ (conmutativity)
- $(u + v) + w = u + (v + w)$ (associativity: addition does not depend on parenthesis grouping)
- $\exists 0 \in V \forall v + 0 = v \forall v \in V$ (there is a zero element, and adding it does not move anyone)
- $\exists -v \forall v + (-v) = 0 \forall v \in V$ (every element has an additive inverse)
- $(a+b)v = av+bv$ (Addition in $K$ maps to addition in $V$: please note that the first + is addition in $K$ while the second one is in $V$. We could write this as $(a+Kb)v = av+Vbv$ but such a thing is strictly for children)
- $a(v + w) = av + aw$ Scalar multiplication distributes over a sum in $V$.
- $(ab)v = a(bv)$ (Multiplication in $K$ maps to scalar multiplication in $V$. Note, as above, on the left hand side first multiplication is $K$ while second is $V$, while on the RHS both multiplications are in $V$).
• $1v = v$ Multiplication by the unit of the scalars does not move the scalars. Might be thought of redundant with the previous one since $1 \cdot 1 = 1$ for any field $K$ and thus $(1 \cdot 1)v = 1(1v) = (1v)$, but this would require that every vector $w$ should be writable as $1v$ for some $v$...

Please note that the four addition axioms only tell us that addition of vectors looks roughly like the normal addition we know and love and forbid strange things from happening. The four multiplication axioms, on the other hand, are telling us that the arithmetic properties of the field are compatible with scalar multiplication.

5.2 Examples of vector spaces.

A large number of things are vector spaces, or can potentially be endowed with vector space structure. Let’s give a pseudobiological example of exactly what is needed to become a vector space. Let’s imagine we have a collection of test tubes with stuff. Now, the contents of test tubes can be mixed together, so we could say that we have a space of test tubes, and dumping tubes together is test tube addition. Furthermore, we have the empty tube, and we can dump only half (or any other fraction) of a test tube, providing multiplication by scalars. However, we’re missing several requisite properties: which are they?

(Answer: scalar multiplication by factors $> 1$ or $< 0$, associativity of addition (if the contents of the tubes react with one another, such reactions would be history dependent), and additive inverses. They might even not commute: when making sulfonitric mixture, sulfuric acid should be poured into the nitric acid and not vice versa because the heat generated can cause splashing!)

So, as to the examples.

• $\mathbb{R}^n$, the set of ordered $n$-tuples of real numbers. This is the “canonical” vector space we usually think of.

• $\mathbb{C}^n$ ordered $n$-tuples of complex numbers. Notice that since a complex number packs two reals, this is twice as large a space as the previous one.

• $P^n(x)$, the set of all polynomials of order $n$ over a variable $x$.

• The set of all functions. Functions can be regarded as vectors with “continuous” indices: $x_i$ is an object which returns a number for each value of the index $i$; similarly, $f(x)$ returns a number for each value of the “index” $x$. You do want to verify that this set satisfies all of the definitions above.

etc.

5.3 Subspaces

The word subspace in math is usually meant to denote a subset of the full space which retains some characteristic of the full; thus, not every old subset, but something that is special, retaining and inheriting properties. In the context of linear algebra, a linear subspace is a subset of a linear space which is itself a linear space. The trick is that it inherits the addition and multiplication from the big guy—you cannot define these independently.

The fundamental property of a subspace is to be closed under addition and multiplication by scalars. All other properties are satisfied by being member of the bigger space. The first and most trivial test is that 0 should belong to the subspace. If it don’t have zero it ain’t a subspace. Let’s try a few things:

• The $x$ axis in the plane: all vectors of the form $(x, 0)$. Similarly, the $y$ axis, and any line going through the origin.
The $xy$ plane in 3-D space is a subspace. (And any plane going through the origin)

The set of all continuous functions within the space of all functions. (Continuity is not broken by addition or multiplication by scalars!)

The set of all functions with $d$ continuous derivatives.

Things that are NOT linear subspaces (why!!??):

- The real axis in the complex plane: not invariant under multiplication by $i$.
- The line $y = ax + b$ in the plane (if $b \neq 0$ then the origin is not in the subspace thus violating multiplication by zero)
- The curve $y = ax^2$ in the plane (though it goes through zero, it’s not closed under multiplication or addition).
5.4 Operations with representations

Consider three vector spaces $V$, $W$, and $Z$ with basis $b_i$, $d_i$, and $q_i$ respectively, and linear functions $f: V \to W$ and $g: W \to Z$.

Any vector $v \in V$ can be represented in the basis of $V$ through coefficients $v_i$ by the equation

$$v = \sum_i v_i b_i$$

We call the $v_i$ the representation of $v$ on the basis $b$. Similarly, the representation of $f$ on the basis $b$ and $d$ is given by a matrix $A$ defined through the equation

$$f(b_i) = \sum_j A_{ji} d_j$$

where care has to be taken to note that the left hand side $f(b_i)$ is an actual vector in $W$, and that the sum is a sum of vectors and the multiplication inside is the multiplication of vectors by scalars. The conventional order of the indices of $A$ is row, column.

By using the first representation we get

$$f(v) = \sum_i v_i f(b_i) = \sum_i v_i \sum_j A_{ji} d_j = \sum_j \sum_i v_i A_{ji} d_j$$

from where

$$f(v) = \sum_j \left( \sum_i A_{ji} v_i \right) d_j$$

which defines that the representation of $f$ applied to $v$ in the target basis of $W$ is a sum of products; remembering that $f(v) = w \in W$ and that by analogy $w$ is represented in the basis $d$ of $W$ by $w = f(v) = \sum_j w_j d_j$, we see we have no choice other than to say that the sum within the parenthesis above must be equal to the $w_j$; we finally can write the operation exclusively on the representations:

$$w_j = \sum_i A_{ji} v_i$$

which operation is the common definition of multiplication of a matrix times a vector.

Let us now see how to compose functions and what happens to their representations. The linear function $g$ has a representation $B$ in the basis $d$ and $q$ which is entirely analogous to the one of $f$:

$$g(d_j) = \sum_k B_{kj} q_k$$

and then if we consider $h = g \circ f$

$$h(b_i) = \sum_k C_{ki} q_k$$

we get

$$h(b_i) = g(f(b_i)) = \sum_j A_{ji} g(d_j) = \sum_j A_{ji} \sum_k B_{kj} q_k = \sum_k \left( \sum_j A_{ji} B_{kj} \right) q_k$$

from where matrix multiplication is functional composition.
which states that the representation $C$ of the functional composition of linear functions is obtained by multiplying the matrices $A$ and $B$ representing the functions being composed—functional composition $h = g \circ f$ becomes matrix multiplication $C = B \times A$. Please note that the order of the multiplication matters, because first $f$ is applied (hence the matrix $A$ is on the right) and then $B$ is applied, and the index being summed over is the one in the middle.

Slight aside: please please note that in the last equation, the “free indices” $k$ and $i$ appear in the same order on both sides and the “mute index” $j$ (mute because it is being summed over, and hence its name could be changed at will) is in the middle. This defines the appropriate order.

5.5 Examples