# Mathematics background for MATH 300

David Dewhurst

September 2017

# Contents

# Preface

Any figures you see can be recreated by running `./demo.py <equation number>` in your terminal, unless the caption says explicitly it's from a paper or something like that. For example, to reproduce the recurrence 1.14, you'd run `./demo.py 1.13` in your terminal. I assume you're using a Linux distribution; if you're not, good luck running the script as I don't know how Windows or Mac work. Python will be the language used throughout and there is an appendix to show you the basics, although the online documentation is excellent (and how I learned python). Any code I write in the text should work perfectly, unless it's designed not to; you'll know if it's designed not to work if I say it's designed not to work.

# Chapter 1

# Discrete and multivariate calculus

The subject of discrete mathematics comprises a large part of the entirety of mathematics; we can't give anything approximating a clear description of it here. You'll have to be content with a brief overview of the aspects most applicable to the study of complex systems. We'll also review a little bit of multivariate calculus, as it'll bring great utility later on in the course.

## 1.1 Sums

Often we wish to add a large number of terms together: $a_1 + a_2 + \cdots + a_N$. We could write it this way, but it is much more convenient to write it using summation notation:

$$a_1 + a_2 + \cdots + a_N \equiv \sum_{i=1}^{N} a_i \tag{1.1}$$

The letter $i$ is called an *index*, and it's a member of an index set $I$: $i \in I$. Here, it's easy to see that $I = \{1, ..., N\}$, a subset of the integers. Of course, we can sum over other index sets too; if we wanted to sum the terms $a_i$ over all even integers, we could write it in a variety of ways:

$$a_0 + a_2 + a_{-2} + \cdots = \sum_{-\infty}^{\infty} a_{2i} = a_0 + \sum_{\substack{n=1 \\ i=2n}}^{\infty} (a_i + a_{-i}) \tag{1.2}$$

You get the picture. By the way, the two sums above are called *infinite series* because their index set, $I = \{i \in \mathbb{Z} : i/2 \in \mathbb{Z}\}$ is an infinite set. Right now we won't worry about what it means to sum over an infinite number of terms; for those of who know some algebra, we're just treating the sums as *formal power series* without notions of convergence. We'll come back to that later.

Sums are the discrete equivalents of integrals; they add stuff up and answer the question "how much of...?" We've glossed over it until now, but the terms of the sum $a_i$ are really functions of $i$; $i \mapsto a_i$. These terms arise because of some system (or algorithm, or theorem, or...) that you're considering. When sums are involved, we're often using them to answer one of a two questions:

a. How much stuff do I have? That is, if I add up all of these terms $a_i$, what do I get?

b. Is there some special property about the terms $a_i$ that I can learn?

Let's do two easy problems to check out how sums operate "in the wild." These are a little cooked up, but you'll find over the course of the semester that problems such as these come up very, very naturally.

**1.** Find a closed form for $\sum_{n=0}^{N} n$.

*Proof.* We note that the functional form for $a_n$ is just $a_n = \mathrm{id}$; $n \mapsto a_n = n$. In doing these kinds of problems, it's often quite helpful to write out the first few *partial sums*:

$$S_2 = \sum_{n=0}^{2} n = 1 + 2 = 3 = 2(2+1)/2$$

$$S_3 = \sum_{n=0}^{3} n = S_2 + 3 = 6 = 3(3+1)/2$$

$$S_4 = \sum_{n=0}^{4} n = S_3 + 4 = 10 = 4(4+1)/2$$

We may have discovered a pattern! Now, we might think that we know $S_N = \frac{N(N+1)}{2}$, but we haven't proved it. (In more complicated examples than this, patterns can hold up for a *long* time, and then suddenly break down!) So we need to prove it. We'll do this by *induction*: show that the formula holds for $N = 1$ (or some other integer; think of it as an anchor for the rest of the proof), assume it holds up to $N$, and then show it works for $N + 1$. Okay, nothing to it: obviously the formula holds for $N = 1$; $1(1+1)/2 = 1$. Let's assume it works for $N$ and show that it works for $N+1$. All right, we'll write it out:

$$\begin{aligned}\sum_{n=1}^{N+1} n &= \frac{N(N+1)}{2} + (N+1) \qquad \text{inductive hypothesis}\\ &= N^2/2 + 3N/2 + 1\end{aligned}$$

Now, note that

$$\frac{(N+1)(N+2)}{2} = \frac{N^2 + 2N + N + 2}{2}$$

Simplify, and they're equal. □

**2.** (This problem's for you!) Find a closed form for $\sum_{n=1}^{N} n^2$. I would use induction if I were you.

Like integration, we don't have to limit ourselves to summation over just one index set. If we have a collection of index sets $I_1, ..., I_K$, we can sum over all of them:

$$\sum_{i_1 \in I_1, \ldots, i_K \in I_K} a_{i_1, \ldots, i_K} \tag{1.3}$$

Note that, in general,

$$\sum_{i_1, \ldots, i_K} a_{i_1, \ldots, i_K} \neq \sum_{i_1} \cdots \sum_{i_K} a_{i_1, \ldots, i_K} \neq \sum_{i_K} \cdots \sum_{i_1} a_{i_1, \ldots, i_K}. \tag{1.4}$$

Blithely changing the order of summation like this can go horribly wrong if, for example, one of the $I_k$s depends on another! Just consider the double sum $\sum_{j=1}^{n} \sum_{i=1}^{j} a_{i,j}$. Changing the order of summation would make the $j$ on the outer sum essentially an unbound variable. If this is hard for you to see, try writing some code to compute a sample sum: for example, let $a_{i,j} = ij$ and compute the sum above.

```
a = lambda i, j: i * j
S = 0
N = 10

for i in range(N):
        for j in range(i):
                S += a(i, j)
```

You'll find that $S = 870$. What happens when we try to compute it the other way?

```
a = lambda i, j: i * j
S = 0
N = 10

for j in range(i):
        for i in range(N):
                S += a(i,j)
```

A very unpleasant `NameError: name 'i' is not defined` results. Other examples abound; see if you can come up with a few! (The more insidious cases come about when you don't have an unbound variable, but changing the order of summation produces wildly different results.)

We can also think of sums as operators going from the space of (partial) sequences to some subset of the real or complex numbers. This is an incredibly useful mindset when dealing with other operators, such as integrals and derivatives, along with sums. There are a few good rules to remember:

a. If the sums are finite (meaning that their index sets are finite), we can just move integrals and derivatives inside the sum:

$$\int dx \sum_{n=1}^{N} a_n(x) = \sum_{n=1}^{N} \int dx \; a_n(x) \tag{1.5}$$

$$\frac{d}{dx}\sum_{n=1}^{N} a_n(x) = \sum_{n=1}^{N} \frac{da_n(x)}{dx} \tag{1.6}$$

Of course, this is obvious; in the finite case, sums are just adding up a bunch of stuff, and from linear algebra we know that integration and differentiation are linear operators. In the infinite case it's a little more tricky; we need to know something about the *convergence* of the sums.

b. If we act on a sum with an operator specific to a particular term of the sum, we must be careful about the rest of the terms. The most common occurrence of this is when we take the derivative with respect to a particular term of a sum:

$$\begin{aligned} \frac{\partial}{\partial a_j}\sum_{i\in I} f(a_i) &= \sum_{i\in I} \frac{\partial f}{\partial a_i}\frac{da_i}{da_j} \\ &= \frac{\partial f}{\partial a_j} \end{aligned} \tag{1.7}$$

since all of the $\frac{da_i}{da_j} = 0$ when $j \neq i$. Of course, this extends to sums over multiple indices as well:

$$\begin{aligned}\frac{\partial}{\partial a_{j_1,...,j_K}} \sum_{i_1,...,i_K} f(a_{i_1,...,i_K}) &= \sum_{i_1,...,i_K} \frac{\partial f}{\partial a_{i_1,...,i_K}} \frac{da_{i_1,...,i_K}}{da_{j_1,...,j_K}} \\ &= \frac{\partial f}{\partial a_{j_1,...,j_K}}\end{aligned} \tag{1.8}$$

But be careful! The entire $K$-fold sum vanishes in this case because we specified differentiation with respect to the entire collection of indices $i_1, ..., i_K$ and, consequently, all of the derivatives $\frac{da_{i_1,...,i_K}}{da_{j_1,...,j_K}}$ vanish except at the point $(j_1, ..., j_K)$. If we don't do this, we'll need to be less cavalier with our differentiation. Consider, for example, the derivative with respect to $x$ of the sum of a sequence of functions.

$$\frac{\partial}{\partial x} \sum_{i,j,k} f_{ijk}(x, y, z) = \sum_{i,j,k} \frac{\partial f_{ijk}(x, y, z)}{\partial x}, \tag{1.9}$$

which is valid as long as the sum isn't infinite. (We'll discuss infinite sums later.) These two types of differentiation are very different–they only look similar! We'll discuss the first type more later on, in Chapter ??.

## 1.2 Products

If we decide to multiply $\{a_i\}$ together instead of adding them, we have a *product.* This is commonly denoted

$$a_1 a_2 \cdots a_N \equiv \prod_{i=1}^{N} a_i \tag{1.10}$$

As with sums, we will (for now) consider ourselves only with finite products. (Infinite products make infinite sums look like a piece of cake!) These often show up when working with certain types of recurrences (see section 1.3), and can also be involved when working with certain differential equations. Often, we'll find ourselves taking the derivative of a product of functions.

It's just repeated application of the chain rule, but it can trip you up:

$$\begin{aligned}\frac{\partial}{\partial x_j}\prod_{i=1}^{N} f_i(x_1,...,x_K) &= \frac{\partial f_1}{\partial x_j}\prod_{\ell=2}^{N} f_\ell(x_1,...,x_K) + \cdots + \\ &\qquad \frac{\partial f_N}{\partial x_j}\prod_{\ell=1}^{N-1} f_\ell(x_1,...,x_K) \\ &= \sum_{k=1}^{N}\frac{\partial f_k}{\partial x_j}\prod_{\substack{\ell=1\\ \ell\neq k}}^{N} f_\ell(x_1,...,x_K)\end{aligned} \tag{1.11}$$

## 1.3 Recurrences

The process of summation naturally arises when considering *recurrence relations*, or equations that relate different terms of the sequence $\{a_i\}_{i\in I}$ to each other. In the course of PoCS, you'll find that recurrences recur quite frequently. Here's an example:

$$x_n = x_{n-1} + n, \quad x_0 = 1, \quad n \geq 1. \tag{1.12}$$

Here's another:

$$f_n = f_{n-1} + 2f_{n-2}, \quad f_0 = 1, \quad f_1 = 1, \quad n \geq 2 \tag{1.13}$$

We'll solve both of these in just a bit.

### 1.3.1 Linear first order (simple!)

Speaking generally, we'd like to figure out an explicit formula for $x_n$ in terms of $n$. There are a few ways to go about doing this:

1. Calculate out the first few terms and see if you can find a pattern. Then, prove it by induction.

2. Rearrange terms to make the expression more amenable to summation (or multiplication), then see if summing (or multiplying) the expression causes many cancellations.

3. Sum both sides and see if you can wrangle out an answer using calculus. Doing this in a principled way is called the method of *ordinary generating functions*, and we'll use it extensively in the next chapter.

4. Multiply both sides by a function and then sum it. These are other generating function methods, some of which we'll discuss.

5. Use some knowledge of algebra (roots of polynomials, etc.) and a little intuition to convert the analytical problem into a purely algebraic one.

There are many other methods of explicitly solving recurrence relations, but these will be the most useful to us. Right now we should note that the recurrence in 1.12 is linear: all $x_k$ that appear aren't squared, exponentiated, or otherwise mangled. Linear recurrences are far nicer to work with than their cousins, the nonlinear recurrences, because linear recurrences can be solved explicitly. With rare exceptions, nonlinear recurrences can't be solved explicitly and we have to resort to approximation methods.

Enough talk: let's solve 1.12. This one's not too hard; we notice that we can subtract $x_{n-1}$ from both sides to rewrite it as

$$x_n - x_{n-1} = n, \quad x_0 = 1, \quad n \geq 1$$

If we sum both sides, most terms on the left cancel (this is called a *telescoping sum*):

$$\sum_{n=1}^{N}(x_n - x_{n-1}) = x_N - x_0 = \sum_{n=1}^{N} n$$

We recognize that sum from earlier! Thus, we find that

$$x_N = x_0 + \sum_{n=1}^{N} n = 1 + \frac{N(N+1)}{2}$$

Certainly we could have just tried to figure it out and then proved it by induction, but this was a lot easier.

Here's another recurrence that's a little more challenging.

$$\frac{a_{n+1}}{a_n} = 1 + \frac{n}{k}, \quad a_0 = 1, \quad n \geq 0 \tag{1.14}$$

You could try the summing trick, but you'd quickly be in a world of hurt. Instead of summing, why not try "product-ing" (multiplying) both sides?

$$\begin{aligned} \prod_{n=0}^{N-1} \frac{a_{n+1}}{a_n} &= \prod_{n=0}^{N-1}\left(1 + \frac{n}{k}\right) \\ \frac{a_N}{a_{N-1}}\frac{a_{N-1}}{a_{N-2}}\cdots\frac{a_1}{a_0} &= \prod_{n=0}^{N-1}\left(1 + \frac{n}{k}\right) \qquad \text{Cancel everything!!!} \\ a_N &= \prod_{n=0}^{N-1}\left(1 + \frac{n}{k}\right) \end{aligned}$$

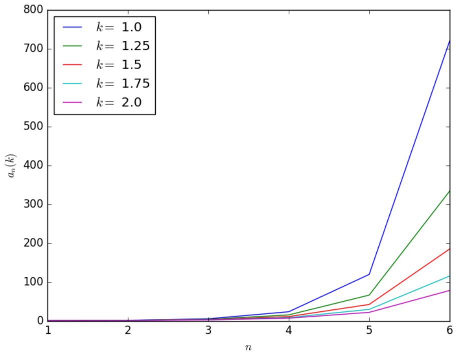

Figure 1.1: Solutions to recurrence 1.14 as functions of the parameter $k$

That's pretty slick; we've reduced the problem of solving a recurrence to finding an explicit expression of evaluating a finite product. (Lest you think this problem is entirely cooked-up, you'll be solving a very similar recurrence when you study rich-get-richer processes...) Let's rewrite the product:

$$\begin{aligned}\prod_{n=0}^{N-1}\left(1+\frac{n}{k}\right) &= \prod_{n=0}^{N-1}\frac{k+n}{k}\\ &= \frac{1}{k^N}(k+N-1)(k+N-2)\cdots(k+0)\end{aligned}$$

This long product on the right-hand side of the equation occurs so frequently it has a name: the *rising factorial*. We typically define it as

$$k^{(N)} \equiv (k+N-1)\cdots(k+1)k, \tag{1.15}$$

so we could write the closed form of the recurrence as

$$a_N(k) = \frac{k^{(N)}}{k^N} \tag{1.16}$$

Figure 1.1 shows solutions to the recurrence for various values of $k$; note that when $k=1$ the solution is just $N!$, as our analytical derivation shows (verify this!)

### 1.3.2 Linear, constant-coefficient, higher order

Okay, what about recurrence 1.13? We might try the sum or product tricks, but they aren't easy, since we have more than two terms present! We'll solve this one two different ways, both algebraic in nature.

**W1**: The *characteristic polynomial.* First, a little diversion. Let's think about the very simple recurrence

$$\frac{a_n}{a_{n-1}} = r, \quad a_0 = 1, \quad n \geq 1$$

To solve this, we just use the product trick to find that $a_N = r^N$. Nothing to it! We'll use this simple guess (also called an *ansatz*, which is the German word for approach or attempt–literally "at sentence") in the solution of multi-term constant-coefficient linear recurrence such as 1.13 by substituting $f_n = r^n$ into the recurrence:

$$f_n = f_{n-1} + 2f_{n-2} \implies r^n - r^{n-1} - 2r^{n-2} = 0 \tag{1.17}$$

Dividing both sides of the equation by $r^{n-2}$ gives us a second order polynomial, $r^2 - r - 2 = 0$. Using the quadratic formula gives $r_\pm = (2, -1)$, so our general solution is

$$f_n = c_1 2^n + c_2(-1)^n \tag{1.18}$$

We still have to solve for the constants $c_1$ and $c_2$. They come from the initial conditions–plug in $n = 0$ and $n = 1$ and write the corresponding equations. Doing this gives the simple linear algebra problem

$$\begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Row reducing gives $c_1 = \frac{1}{3}$ and $c_2 = \frac{2}{3}$, so we find that the specific solution is

$$f_n = \frac{2}{3}2^n + \frac{1}{3}(-1)^n \tag{1.19}$$

This isn't the most powerful way to solve recurrences in general, but for linear, constant-coefficient recurrences it's often the best. Even for recurrences of high order, solving for roots can often be done efficiently using some numerical method (e.g., Newton's method).

**W2** *Ordinary generating functions.* This is complete overkill for a simple recurrence such as this, but it's a good introduction to the generating function method. Just as integration can be thought of as the inverse operation of differentiation, so can summation be thought of as the inverse operation of finite differences, of which these sorts of recurrence relations are examples. The general algorithm is this: rewrite the recurrence in terms of the formal power series $A(x) = \sum_{n=0}^{\infty} a_n x^n$, and solve the corresponding equation for $A(x)$. Then, form $A(x)$'s Taylor series, which will give us the coefficients $a_n = A^{(n)}(0)/n!$ that we so desire.

Starting with recurrence 1.13 (which we'll rewrite as $a_n = a_{n-1} + 2a_{n-2}$ with $a_0 = a_1 = 1$, $n \geq 2$) we're going to "undo" the differencing by summing both sides. We want the indices on the sums to eventually start at $n = 0$, so since our lowest order term is $a_{n-2}$, we'll start the index of each sum at $n = 2$:

$$\sum_{n=2}^{\infty} a_n x^n - \sum_{n=2}^{\infty} a_{n-1} x^n - 2\sum_{n=2}^{\infty} a_{n-2} x^n = 0 \tag{1.20}$$

This is just another way to formally rewrite the recurrence, but now we can use some tricks with the power series. We'll take them one-by-one; the first is relatively easy:

$$\begin{aligned} \sum_{n=2}^{\infty} a_n x^n &= \sum_{n=0}^{\infty} a_n x^n - a_0 - a_1 x \\ &= A(x) - 1 - x \end{aligned} \tag{1.21}$$

Now the second:

$$\begin{aligned} \sum_{n=2}^{\infty} a_{n-1} x^n &= x \sum_{n=2}^{\infty} a_{n-1} x^{n-1} \\ &= x \left( \sum_{n=0}^{\infty} a_n x^n - a_0 \right) \\ &= x(A(x) - 1) \end{aligned} \tag{1.22}$$

And the third:

$$\begin{aligned} 2\sum_{n=2}^{\infty} a_{n-2} x^n &= 2x^2 \sum_{n=0}^{\infty} a_n x^n \\ &= 2x^2 A(x) \end{aligned} \tag{1.23}$$

Cool. Now we can put all of this together into an equation for the generating function $A(x)$. Do some algebra, and you'll find that

$$A(x) = \frac{1}{1 - x - 2x^2} \tag{1.24}$$

From here, we can do one of two things: try to find a pattern for the Taylor series coefficients of $A(x)$ (and thereby find the solution to the recurrence); or just realize that $a_n = A^{(n)}(0)/n!$, and compute whatever terms of the recurrence we need.

This barely scratched the surface of generating functions; they are a *massive* topic of research, even today. The connections between them and areas of pure mathematics such as abstract algebra and complex analysis are incredibly deep; I highly encourage you to purchase the books by GKP [1] and Wilf [2] for further study.

## 1.4 Review of multivariate calculus

This is something you just have to know–and if you don't already, the last two sections were probably rough going!

### 1.4.1 Differentiation

You know about partial derivatives. Remember the formula for a differential of the function $f(x_1, ..., x_N)$:

$$df = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i} dx_i \tag{1.25}$$

Obviously the chain rule works the same way; if we have

$$f(x_1, ..., x_N), \quad x_i(t)$$

then to find $\frac{df}{dt}$ it's the same old process:

$$\frac{df}{dt} = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt} \tag{1.26}$$

One can go on in this way *ad infinitum*. Remember that multivariate functions exist in, well, many variables, so that we don't just ask about derivatives in one direction or another (as in $\partial_x^n f$) but in different directions of different orders. These are the mixed partial derivative operators, e.g., $\frac{\partial^2}{\partial x \partial y} = \frac{\partial}{\partial x}\frac{\partial}{\partial y}$, etc. There are a couple misconceptions about partial derivatives that we should clear up:

a. Just because a function has partial derivatives does not mean it's differentiable!!!

b. Unless the second mixed partial derivatives are both continuous, you can't exchange them! Often I have found that students seem to think Claraiut's theorem holds all the time–it doesn't. This mistake can really get you into trouble in finance; we will discuss this point later.

Something that isn't taught well (or people just don't remember well) is the multivariate Taylor expansion. Recall that, if a function of one variable $f(x)$ satisfies certain conditions it can be expanded in a power series about the point $x = a$ of the form

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

There are subtle questions of convergence of the series and all that; as physicists, economists, and computer scientists, we're often dangerously cavalier with these series and just go around expanding functions willy-nilly. One variable isn't enough sometimes, so we expand multivariate functions. When we do this, we move to partial derivatives, *and we have to remember to include all the mixed partials!* We'll just expand about the point $x_i = 0$ in the $N$-dimensional MacLaurin series:

$$\begin{aligned} f(x_1, ..., x_N) = f(0) + \sum_{i=1}^{N} \partial_{x_i} f(x_1, ..., x_N) x_i + \\ + \frac{1}{2!} \sum_{j,k=1}^{N} \partial_{x_j} \partial_{x_k} f(x_1, ..., x_N) x_j x_k + \cdots \end{aligned} \tag{1.27}$$

The term in the infinite series of order $K$ will thus have $N^K$ terms in its finite sum. In practical work, a lot of times we will truncate the terms at order 2, so that we have at most second partials floating around.

### 1.4.2 Integration

Remember: you can't just interchange integrals! That is to say,

$$\begin{aligned} \int_{\substack{x\in X \\ y \in Y}} f(x,y) d(x,y) &\neq \int_{x\in X} \left( \int_{y \in Y} f(x,y) dy \right) dx \\ &\neq \int_{y\in Y} \left( \int_{x \in X} f(x,y) dx \right) dy \end{aligned} \tag{1.28}$$

in all generality. Now, it is the case that if $\int_{x\in X,\ y\in Y} |f(x,y)| d(x,y) < +\infty$, you can change the order of integration; use this power wisely! (This is called Fubini's theorem.)

Here is another useful thing to remember. If you are able to factor a function $f(x_1, ..., x_N) = \prod_1^N f_i(x_i)$ somehow, *and* if each domain $X_i$ does not depend on the values of any of the other variables, we can factor out an integral as follows:

$$\begin{aligned}
&\int_{x_1 \in X_1, ..., x_N \in X_N} d(x_1, ..., x_N) f(x_1, ..., x_N) \\
&\quad = \int_{x_1 \in X_1, ..., x_N \in X_N} d(x_1, ..., x_N) \prod_1^N f_i(x_i) \\
&\quad = \prod_1^N \int_{x_i \in X_i} dx_i \; f_i(x_i)
\end{aligned} \tag{1.29}$$

This can be handy when dealing with statistics of large systems.

# Chapter 2

# Probability

The theory of probability is wide and deep, and one of the most difficult subjects in mathematics. Here we will cover discrete probability in depth, and touch on the applied aspects of continuous probability theory; to enter into a discussion of the continuous theory requires rather advanced mathematics that won't really add to your understanding of probability concepts.

## 2.1 Axioms and fundamentals

For right now we will concern ourselves only with spaces of finitely many or countably infinitely many events that can occur; the index set $I$ is either finite or can be put into one-to-one correspondence with the integers. If we denote an event by $X_i$, we can write the sample space $S$ as

$$S = \bigcup_i X_i \tag{2.1}$$

We will denote a discrete *probability measure* by $p(\cdot)$. It is best, for now, to not worry about what the fundamental meaning of this is too much. You can think about the frequentist interpretation (if I roll a fair die many, many times, about one-sixth of the rolls will land on five; I have a one-fifth probability of rolling five) for now. We will later discuss other interpretations.

### 2.1.1 Axioms

There are three axioms of probability.

**PA1.** $0 \leq p(X_i) \leq 1$ for all $i$.

**PA2.** $p(S) = 1$

**PA3.** Let $J \subseteq I$. Then $p\left(\bigcup_{j\in J} X_j\right) \leq \sum_{j\in J} p(X_j)$, with equality holding if and only if the events $X_j$ are mutually exclusive.

Let us briefly unpack the third axiom. We are considering the set of events $\bigcup_{j\in J} X_j$ and would like to know the probability that we observe this set. This, of course, requires that we know how to interpret $\bigcup$ in the context of probability. A union can be interpreted as an *or*:

$$\bigcup_{j\in J} X_j = X_1 \text{ or } X_2 \text{ or ... or } X_{|J|} \tag{2.2}$$

Armed with this, we now know that we're really seeking to find the probability that *at least one of* the events $X_j$ happens. If the events are mutually exclusive–that is, if $X_i \cap X_j = \emptyset$ for all $i, j \in J$, then it's pretty obvious that we just add the probabilities to find the answer. But if they aren't mutually exclusive, then adding the probabilities would be overcounting, so the relationship becomes a strict inequality.

### 2.1.2 Important subsets

The third axiom tells us (almost) how to compute the probability that at least one of something happens. But we discussed that it overcounts; we need to remove what it's overcounting. To that end, suppose $I = \{A, B\}$ and we wished to find $p(A \cup B)$:

$$p(A \cup B) = \underbrace{p(A) + p(B)}_{\text{Count both}} - \underbrace{p(A \cap B)}_{\text{prob that both occur}} \tag{2.3}$$

If we instead tried to find $p(A \cup B \cup C)$, it'd be almost as easy:

$$\begin{aligned} p(A \cup B \cup C) = p(A) + p(B) + p(C) - p(A \cap B) \\ - p(A \cap C) - p(B \cap C) + p(A \cap B \cap C) \end{aligned} \tag{2.4}$$

Again, we first overcounted, then undercounted by taking *too much* away, then put back some of the stuff we took away. This is, in general, called the *inclusion-exclusion* principle. In all generality we can write

$$p\left(\bigcup_{i=1}^{N} X_i\right) = \sum_{j=1}^{N} (-1)^{j-1} \sum_{\substack{\mathcal{L}\subset I \\ |\mathcal{L}|=j}} p\left(\bigcap_{\ell\in\mathcal{L}} X_\ell\right) \tag{2.5}$$

This looks intimidating, but it's actually a big softie: the first sum simply says whether to over- or under-count; the second sum says to sum over all $j$-sized subsets of $I = \{1, ..., N\}$, and the intersection is as above.

Once we know how to deal with unions, intersections are actually really easy to deal with. We can interpret $\bigcap_{j\in J} X_j$ as "$X_1$ and $X_2$ and ... and $X_{|J|}$". How do we calculate this statement? Well, the *opposite* of it is "at least one of the events $X_j$ does not occur", which we can express as $\bigcup_{j\in J} X_j^{(c)}$, where the superscript $c$ stands for compliment; $X_j^{(c)} \equiv S \backslash X_j$. Thus, we have that

$$p\left(\bigcap_{j\in J} X_j\right) = 1 - p\left(\bigcup_{j\in J} X_j^{(c)}\right), \tag{2.6}$$

and we know how to calculate the probability of that union thanks to Eq. 2.5.

### 2.1.3 Independent probabilities

One of the most important special cases is where the events $X_j$ are all independent; that is $S = \bigcup_{i\in I} X_i$ and $X_i \cap X_j$ for all $i, j \in I$. This restriction makes the above formulae a lot simpler:

$$p\left(\bigcup_{j\in J} X_j\right) = \sum_{j\in J} p(X_j) \tag{2.7}$$

$$p\left(\bigcap_{j\in J} X_j\right) = \prod_{j\in J} p(X_j) \tag{2.8}$$

Independent probabilities occur very often in physics, economics, and computer science. For example, in classical mechanics a physical system can only be in one state at a time. We may wonder what the probability is that a system is in a particular subset of states, say $J \subset I$. These probabilities must be independent, and so we have $p(\text{state is in } J) = \sum_{j\in J} p(\text{state is } j)$, for a simple example. Much of the following material in this chapter is based on the idea of independence.

Here's a little toy example: suppose there's a forest full of animals of various types.[1] The probability of observing one animal is unchanged by the observation of any other animal. What's the probability of reaching into the forest, taking out two animals at the same time, and observing that the two animals are of the same species? (Assume that each species is equinumerous.) We

[1] This comes from Peter Dodds.

can just write this out:

$$\begin{aligned} p\left(\text{both type 1} \cup \text{both type 2} \cup ...\right) &= p\left(\bigcup_{i=1}^{N} \text{type } i \text{ and type } i\right) \\ &= \sum_{i=1}^{N} p(\text{type } i \text{ and type } i) \\ &= \sum_{i=1}^{N} p_i^2 \end{aligned}$$

The above result uses independence twice: once in rewriting the probability of the union as a sum of probabilities, and once in rewriting the union as a product of probabilities. For another example of independence in action, see [3].

### 2.1.4 Conditional probability

The concept of conditioning is fundamental to both science and the human condition. It is very rare that we actually ask the question, "what is the probability that...?" Rather, we tend to say, "given that such-and-such thing happened, what is the probability that...?" Formally, we *define* conditional probability as follows: given two events $A$ and $B$, the probability of $A$ given $B$ is

$$p(A|B) = p(A \cap B)/p(B) \tag{2.9}$$

You may also see this written $p(A|B) = p(A, B)/p(B)$; the distribution $p(A, B)$ is called the joint distribution and is another way of denoting "the probability of $A$ and $B$". (The concept of a joint distribution is more convenient when considering probability in many dimensions, and especially useful in the continuum.) Equation 2.9 is a fundamental equation of science, and is often introduced as an axiom of probability!

You should be aware of a few things related to conditional probability. First of all, since $p(A \cap B) = p(B \cap A)$–who cares what order the events are in, we want both of them to occur!–we could just as easily write $p(A \cap B) = p(B|A)p(A)$. Substituting this into Eq. 2.9 and rearranging terms gives us *Bayes's theorem*:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} \tag{2.10}$$

I cannot stress enough how important this equation is. Along with the definition (axiom) of conditional probability, this equation is **the key** to understanding probability, statistics, and much of logic. We will return repeatedly to Bayes's theorem in the coming sections. One other note about Eq. 2.10: you may often see this written as

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \iff p(\text{evidence}|\text{data}) = \frac{p(\text{data}|\text{evidence})p(\text{evidence})}{p(\text{data})} \tag{2.11}$$

This is an excellent interpretation of the theorem, and I encourage you to think of it thus. We call $p(\theta) \equiv p(\text{evidence})$ the *prior*, $p(X|\theta) \equiv p(\text{data}|\text{evidence})$ the *likelihood*, $p(X) \equiv p(\text{data})$ the *marginal likelihood*, and $p(\theta|X) \equiv p(\text{evidence}|\text{data})$ the *posterior*.

## 2.2 Discrete probability

We will confine ourselves to index sets $I$ as outlined in Section 2.1. We will be considering probability distributions $p_i \equiv \Pr(\text{event } X_i \text{ occurs})$, and we will consider the events $X_i$ to be independent unless otherwise stated.

### 2.2.1 Normalization

There are a few terms to know. A discrete probability function–the probability, say, that the random variable $N = n$, will be written as $p(N = n) \equiv p_n$. Often I will say something like "the random variable $n$," but I really mean "the random variable $N$ observed to be $n$". The function $p_n$ is known as a *probability mass function* (PMF). The function $p(n \leq N) \equiv \sum_{n=0}^{N} p_n$ is known as the *cumulative distribution function* (PDF). We will also occasionally talk about the *complimentary cumulative distribution function* (CCDF), which is just $p(n \geq N) = \sum_{n=N}^{\infty} p_n$, if, for example, $n \in [0, \infty)$.

Let's deal with some actual probability distributions now. For example, we might have the *uniform* distribution over the integers $\{0, ..., N\}$, denoted $\mathcal{U}[0, N]$. Since the probability distribution is uniform over all possible states, we thus have $p_n \propto \frac{1}{N}$. (The symbol $\propto$ means "proportional to".) We don't yet write = because we need to make sure that **PA2** is satisfied; the probabilities need to sum to one. Let's do this. We know that $p_n = c\frac{1}{N}$, and we'd like to figure out the value of $c$.

$$1 = \sum_{n=1}^{N} \frac{c}{N} = c\sum_{n=1}^{N} \frac{1}{N} = c,$$

so in this (admittedly silly!) case, we have $c = 1$ and $p_n = \frac{1}{N}$. Other distributions are, of course, less trivial. For example, suppose we have $p_n \propto$

$e^{-\lambda n}$, where $n \in [0, \infty)$. (This is called the *Boltzmann distribution* and arises frequently in statistical mechanics.) We can go through the same process:

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} c e^{-\lambda n} \\ &= c \sum_{n=0}^{\infty} e^{-\lambda n} = c \sum_{n=0}^{\infty} \left( \frac{1}{e^{\lambda}} \right)^n \\ &= c \frac{1}{1 - \frac{1}{e^{\lambda}}} = \frac{e^{\lambda}}{e^{\lambda} - 1}, \end{aligned}$$

so that $c = \frac{e^{\lambda}-1}{e^{\lambda}}$. Thus, the probability of observing the number $n$ is given by $p_n = \left(1 - e^{-\lambda}\right) e^{-\lambda n}$.

Here are two normalization exercises for you to do:

1. Normalize $p_n \propto e^{-\lambda n}$ when $n \in [0, N]$ (not, as above, in $[0, \infty)$). This is the most commonly-used form of the Boltzmann distribution.

2. Normalize $p_n \propto n^{-\gamma}$ when $\gamma > 1$. Why is $\gamma > 1$ required to normalize this distribution?

In general, normalization isn't that hard. Since we know that $p_n = cq(n)$, where $q(n)$ is the function such that $p_n \propto q(n)$, we know that we just need to find $\sum_{n \in I} q(n)$, and then do some algebra. The only tricky piece may be actually computing that sum–the chapter on special functions in this book may occasionally be of some use in this task!

### 2.2.2 Moments

We often use probability as a tool when considering large systems of which deterministic simulation would be completely infeasible. Often we would like to understand the *mean behavior* of such systems–what they do in the average case–as well as the *standard deviation* from this average. You have surely heard of the mean and standard deviation before, and likely know how to calculate them. We will again introduce them here in a more formal sense.

We define the $k$-th moment of a discrete random variable $X \sim p(x)$ as

$$\mathbb{E}_p[X^k] \equiv \langle X^k \rangle = \sum_{x \in \mathbb{X}} x^k p(x) \tag{2.12}$$

(Probabilists and statisticians usually use the notation $\mathbb{E}_p[X^k]$, while physicists often use $\langle X^k \rangle$. I will use the physicists notation unless the meaning becomes unclear, which can happen.) It's pretty clear what's happening: we are seeing what the value of $x^k$ is at every $x$, and multiplying that by the probability of actually seeing $x$. Thinking about this another way, if $x^k$ is very large, but the probability of actually observing $x$–given by $p(x)$–is vanishingly small, the contribution to the sum of the term $x^k p(x)$ is not going to be large! We also aren't restricted to ordinary moments of the form $X^k$; we can find the value of the functional moment $\langle f(X) \rangle = \sum_{x \in \mathbb{X}} f(x)p(x)$ just as easily. The ordinary moments are clearly a subset of this more general case.

The most commonly-used moment statistic is the mean, with which you're familiar. It is defined as the first ($k = 1$) expectation:

$$\langle X \rangle = \sum_{x \in \mathbb{X}} x \; p(x) \tag{2.13}$$

You have probably calculated the mean from observed data $(x_1, ..., x_N)$ as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.14}$$

It makes sense that this is the empirical approximation to Eq. 2.13; if $p(x)$ is large relative to $p(y)$, it is more likely that we observe $x$ than $y$, and thus more $x_i$ will be equal to $x$ than $y$. Expressed a little more formally, we can rewrite Eq. 2.14 as

$$\begin{aligned} \mu &= \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \frac{x_i}{N} \\ &= \sum_{j=1}^{M} x_{\text{unique}}^{(j)} \frac{n_j}{N} \end{aligned} \tag{2.15}$$

where each $x_{\text{unique}}^{(j)}$ is a unique observed value and $n_j$ counts the number of times that value was observed. With some mild conditions on the true probability distribution $p(x)$, we know that as we observe many draws from the distribution– as $N \to \infty$–we will have $\frac{n_x}{N} \to p(x)$.

Calculation of the mean is an important exercise, and one that you will do quite frequently in PoCS! There are some practice problems below, but before you do them, we should note something about the mean (and moments in general): they don't have to exist! There are many probability distributions that are perfectly well-defined–that is, they sum (or integrate, in the continuous case) to one–and yet have infinite mean and other moments. These distributions are usually very interesting; you will encounter many of them this semester.

1. Calculate $\langle X \rangle$ if $p(x)$ is the uniform distribution on $[0, N]$

2. Calculate $\langle X \rangle$ if $p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ with $x \in [0, \infty)$. This is the all-important Poisson distribution, and describes the number of independently and identically distributed random variables arriving at some location in a given amount of time, among other things. (Hint: Taylor series.)

There is an incredibly useful property of the expectation called the *tower property.* Succinctly, it says that the expectation of an expectation is equal to the original expectation;

$$\mathbb{E}[\mathbb{E}[f(X)]] = \mathbb{E}[f(X)], \tag{2.16}$$

It is pretty easy to see why this is so:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[f(X)]] &= \sum_{y \in \mathbb{X}} \left( \sum_{x \in \mathbb{X}} f(x) p(x) \right) p(y) \\
&= \sum_{y \in \mathbb{X}} \langle f(X) \rangle p(y) \\
&= \langle f(X) \rangle \sum_{y \in \mathbb{X}} p(y) \\
&= \mathbb{E}[f(X)]
\end{aligned}$$

The key thing to understand is that the mean (or any moment!) of a function of a random variable $f(X)$ is *just a number*, at least, in relationship to its own probability distribution $p(x)$, and so we can pull it out of the second sum–which, of course, sums to one.

The other moment statistic with which you're likely most familiar is the *standard deviation*, often denoted $\sigma$. In fact, the more fundamental statistic is the *variance*, denoted $\sigma^2$, since the standard deviation is equal the square root of the variance. It is the expectation of the squared deviation of an observation $X$ from $p(x)$ and the mean of $p(x)$, $\langle X \rangle$:

$$\begin{aligned}
\mathrm{Var}(X) &= \langle (X - \langle X \rangle)^2 \rangle \\
&= \langle X^2 - 2X\langle X \rangle + \langle X \rangle^2 \rangle \\
&= \langle X^2 \rangle - 2\langle X \rangle \langle X \rangle + \langle X \rangle^2 \\
&= \langle X^2 \rangle - \langle X \rangle^2
\end{aligned} \tag{2.17}$$

Make sure you understand the properties of the expectation used to derive this statistic! (In moving between the second and third lines we used both the linearity of the expectation and the tower property.) This derivation makes the empirical estimator of the variance rather obvious:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i^2}{N} - x_i\right) \tag{2.18}$$

Depending on your choice of programming language, this may not be a stable method to compute the variance; division of large numbers (such as $x_i^s 2$) by other large numbers (like $N^2$) might not be a great idea. Consult a textbook on numerical analysis to fix this problem if you need to. You can see a demonstration of the convergence of these estimators to the true mean and variance in Figure 2.2.2. There are many other moment statistics, some of which are even important. However, these two should be the dearest to your heart.

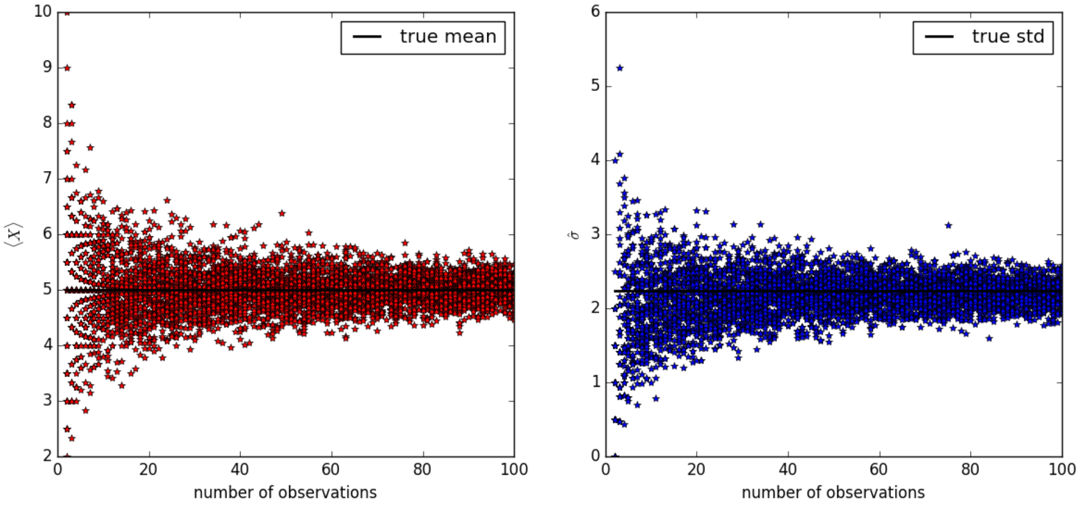

Figure 2.1: Estimated mean and standard deviation of random variates $X \sim \text{Poisson}(\lambda = 5)$ converging to the true mean and standard deviation. Call `./demo.py moment-demo` to generate this image yourself.

## 2.2.3 Generating functions

You have noticed that we do a lot of work with sums in order to find moment statistics. Sometimes actually manipulating the sums to get an answer is less than trivial–you may have seen this already–and we will resort to using generating functions. (We introduced these back in Chapter 1.) There are a few types that are used commonly in probability:

1. (*Probability generating function (PGF)*) This is the original generating function we introduced when solving linear recurrences. Given some sequence of numbers–in this case, the numbers are probabilities $p_n$–we will write the PGF as
$$P(x) = \sum_{n=0}^{\infty} p_n x^n \tag{2.19}$$

2. (*Moment generating function (MGF)*) Instead of attacking $p_n$ with the monomial $x^n$, we'll instead use the exponential $e^{tn}$:
$$M_p(t) = \sum_{n=0}^{\infty} p_n e^{tn} \tag{2.20}$$
Can you guess why they're called moment-generating functions?

3. (*Characteristic function* (CF)) This is the most important generating function, though we will not use it too much in this short book. We will use $e^{itn}$, where $i \equiv \sqrt{-1}$:
$$\Psi_p(t) = \sum_{n=0}^{\infty} p_n e^{itn} \tag{2.21}$$

Generating functions are sort of like magic. Suppose, for sake of argument, that we know $P(x)$ for some distribution $p_n$. Here are some very handy things to remember about the PGF. First, consider the fact that $P(1) = \sum_{n=0}^{\infty} p_n 1^n = \sum_{n=0}^{\infty} p_n = 1$. So finding the constant of normalization becomes trivial. Calculating the mean isn't hard either:

$$\begin{aligned} \langle n \rangle &= \sum_{n=0}^{\infty} n p_n \\ &= \sum_{n=0}^{\infty} \frac{d}{dx} \left(p_n x^n\right) |_{x=1} \\ &= \frac{d}{dx} \left( \sum_{n=0}^{\infty} p_n x^n \right) \Bigg|_{x=1} \\ &= \frac{dP(x)}{dx} \Bigg|_{x=1} \end{aligned} \tag{2.22}$$

We can take higher order moments using the PGF too. For example, here is the calculation of the second moment, $\langle n^2 \rangle$:

$$\begin{aligned}
\langle n^2 \rangle &= \sum_{n=0}^{\infty} n^2 p_n \\
&= \frac{d}{dx} \left( \sum_{n=0}^{\infty} n p_n x^n \right) \Bigg|_{x=1} \\
&= \left( \frac{d}{dx} \circ x \right) \left( \sum_{n=0}^{\infty} n p_n x^{n-1} \right) \Bigg|_{x=1} \\
&= \left( \frac{d}{dx} \circ x \circ \frac{d}{dx} \right) \left( \sum_{n=0}^{\infty} p_n x^n \right) \Bigg|_{x=1} \\
&= \left( \frac{d}{dx} \circ x \circ \frac{d}{dx} \right) (P(x)) \Big|_{x=1}
\end{aligned} \tag{2.23}$$

In general, moments can be computed with the PGF with the formula $\langle n^k \rangle = \theta_k P(x)|_{x=1}$, with $\theta_k$ defined recursively[2] as

$$\theta_k = \frac{d}{dx} \circ x \circ \theta_{k-1}, \quad \theta_1 = \frac{d}{dx} \tag{2.24}$$

The *falling factorial moments*, defined

$$\langle n_{(k)} \rangle \equiv \left\langle \prod_{\ell=1}^{k} (n - \ell + 1) \right\rangle \tag{2.25}$$

are much easier to compute with the PGF:

$$\left\langle \prod_{\ell=1}^{k} (n - \ell + 1) \right\rangle = \frac{d^k}{dx^k} \left( P(x) \right) |_{x=1} \tag{2.26}$$

[2] This is a recurrence relation similar to the ones we studied in chapter 1! However, this is an *operator* recurrence relation, one of the more sublime structures in mathematics. I recommend you seek them out and study them in depth.

Moments are (surprise!) much easier to calculate using the moment-generating function, or MGF. For example,

$$\begin{aligned}\langle n\rangle &= \sum_{n=0}^{\infty} np_n \\ &= \sum_{n=0}^{\infty} np_n e^{tn}|_{t=0} \\ &= \frac{d}{dt}\left(\sum_{n=0}^{\infty} p_n e^{tn}\right)\Bigg|_{t=0} \\ &= \frac{dM_p(t)}{dt}\bigg|_{t=0}\end{aligned} \tag{2.27}$$

You can see that, in general,

$$\langle n^k\rangle = \frac{d^kM_p(t)}{dt^k}\bigg|_{t=0} \tag{2.28}$$

Now, moments are all well and good when they exist. But, as we mentioned before, some distributions just don't have them! For an example, think of $p_n \propto n^{-1}$. You should immediately recognize the (divergent) harmonic series. But generating functions can still be helpful! It is a theorem, for example, that a probability distribution is completely characterized by its *characteristic function* (not a surprise). The characteristic function can be used to calculate moments–for example, consider the calculation of the mean:

$$\begin{aligned}\langle n\rangle &= \sum_{n=0}^{\infty} np_n \\ &= -i\sum_{n=0}^{\infty} inp_n e^{itn}|_{t=0} \\ &= -i\frac{d}{dt}\left(\sum_{n=0}^{\infty} p_n e^{itn}\right)\Bigg|_{t=0} \\ &= -i\frac{d\Psi_p(t)}{dt}\bigg|_{t=0}\end{aligned} \tag{2.29}$$

but they can also be used for other purposes (along with the PGF and the MGF) and *they always exist, even when the PGF and MGF don't.* So, if you're trying to prove something using generating functions but the others don't exist, use the characteristic function.

What else can we do with generating functions? Well, how about this: suppose that we have $X_1, ..., X_N$ independent random variables. They don't need to be identically distributed. We would like to determine the distribution of the sum of these random variables. Since a distribution is uniquely determined by its characteristic function, let's just compute that instead:

$$\begin{aligned}\Psi_{\sum X_i}(t) &= \left\langle \exp\left(it \sum X_i\right)\right\rangle \\ &= \left\langle \prod_{i=1}^{N} \exp(itX_i) \right\rangle \\ &= \prod_{i=1}^{N} \langle \exp(itX_i)\rangle \\ &= \prod_{i=1}^{N} \Psi_{X_i}(t)\end{aligned} \tag{2.30}$$

This is what I mean when I say generating functions are like magic. You can think of many more examples like this–and I won't spoil them for you, since I think you'll work through some of them in your PoCS homework![3] For more information on generating functions as applied to probability, an excellent reference is [4]; of course [2] is the master resource.

## 2.3 Continuous probability

Now that you know (something) of discrete probability, continuous probability won't be too hard to figure out. I hasten to say that the actual mathematics behind continuous probability is quite a bit more difficult than that (essentially finite combinatorics) behind the discrete material. But, in terms of application, you will see that we mostly replace sums with integrals, and everything just works out.

### 2.3.1 From discrete to continuous

The continuum is very different from even countably infinite sets. Instead of a function $p_n$ defined over, say, $n \in [0, \infty)$, we now have a *probability*

[3] For those of you who are well-versed in higher mathematics, you will have noticed by now that these generating functions seem awfully familiar. In fact this is because they are exactly the frequency-space transforms that you have encountered in your differential equations courses; the PGF is the $z$-transform; the MGF is (essentially, map $t \mapsto -t'$) the Laplace transform; and the CF is the Fourier transform. All of the theorems and properties of those transforms work just as well here. Use this power wisely!

*density function*, usually denoted $p(x)$ or $f_X(x)$, for $x \in \Omega \subseteq \mathbb{R}$. If you took multivariate calculus, you'll remember mass density functions–this is exactly like that. Since we're going to be integrating instead of summing, *the probability that $X = x$ is always zero. There is no probability whatsoever that the random variable $X$ will equal $x$.* This is due to nuances of the real numbers.

Aside from this, most things stay the same. For example, the CDF is just

$$F_X(x \in S) = \int_{S \subseteq \mathbb{R}} dx\ f_X(x), \tag{2.31}$$

which is essentially the same as the discrete definition–if $X \in \mathbb{R}$, then this is just $F_X(X \leq x) = \int_{-\infty}^{x} dx\ f_X(x)$. For example, if $X \in \mathbb{R}$ and we wanted to know the probability that we observed $X \in [4, 12]$, we'd just integrate:

$$F_X(4 \leq x \leq 12) = \int_4^{12} dx\ f_X(x).$$

Nothing to it. The CCDF is essentially the same too. If $X \in \mathbb{R}$, then

$$\begin{aligned} F_{X,\geq}(x) &= 1 - F_X(x) \\ &= \int_x^{\infty} dx\ f_X(x) \end{aligned} \tag{2.32}$$

Does anything else change? Not really. For example, the moments are defined in the identical manner:

$$\langle g(X) \rangle = \int_{x \in \Omega} dx\ g(x) f_X(x). \tag{2.33}$$

And the three generating functions are also essentially identical:

$$P(z) = \int_{-\infty}^{\infty} dx\ f_X(x) z^x \quad \text{(this is not used much)} \tag{2.34}$$

$$M_f(t) = \int_{-\infty}^{\infty} dx\ f_X(x)\ e^{xt} \tag{2.35}$$

$$\Psi_f(t) = \int_{-\infty}^{\infty} dx\ f_X(x)\ e^{ixt} \tag{2.36}$$

Of these three, by far the most used is the characteristic function $\Psi_f(t)$. This is because it is identical to the *Fourier transform*, which is absolutely ubiquitous throughout science and engineering and there are many theorems about it that provide us with useful results. You can verify for yourself that the properties we defined above all still hold for each of these transforms.

### 2.3.2 Limit theorems

I won't give too much away here–meaning that I won't prove what follows, since you'll be doing a lot of that in class–but there are two incredibly important theorems that you should know. We will first introduce the *normal distribution*, also known as the Gaussian distribution. This is probably the most important probability distribution and is defined

$$f_X(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.37}$$

If a random variable $X$ is distributed according to the normal distribution we write $X \sim \mathcal{N}(\mu,\sigma)$; $\mu$ is the mean of the distribution and $\sigma^2$ is the variance. (You should definitely attempt to prove this; do not take my word for it.)

The first theorem is reassuring:

**Theorem 1.** *(Weak law of large numbers) Suppose that $X_1,...,X_N$ are i.i.d. random variables $\langle X_n \rangle$ is defined and is finite with mean $\mu$. Then, for any $\varepsilon > 0$,*

$$\lim_{N\to\infty} \Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| > \varepsilon\right) = 0$$

The second theorem relies on the normal distribution:

**Theorem 2.** *(Lindeberg-Lévy central limit theorem) Suppose that $X_1,...,X_N$ are i.i.d. random variables such that $\langle X_n \rangle$ and Var$(X_n)$ are defined and finite, with mean $\mu$ and variance $\sigma^2$. Then*

$$\sqrt{n}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right) \overset{d}{\to} \mathcal{N}(0,\sigma^2)$$

The symbol $\overset{d}{\to}$ means "converges in distribution"; it is a technical point, but there are several types of convergence and this is a relatively strong one. This theorem is of utmost importance; you should seek to understand it as best you can.

There are other central limit theorems–lots of them–but Theorem 2 is the most widely-used of them. You should know, however, that the *generalized central limit theorem* states that any i.i.d. random variables with finite mean and infinite variance also converge in distribution to a limit distribution, known as the *Lévy stable distribution*. Look it up. There is a lot more to probability than what is presented here–as I stated at the beginning of this chapter, it is one of the most well-studied and difficult areas of mathematics, and there are still many open questions–so you should read some references on the subject if you want to know more. Feller's texts [4, 5] are phenomenal guides.

# Chapter 3

# Special functions

There are some functions that just appear over and over again when studying physics, economics, computer science, and complex systems more generally. I will cover some of them here in an attempt to ease the pain when you're confronted by them in a dark alley (e.g., in PoCS).

## 3.1 The gamma function

We all know about the factorial:

$$n! = n(n-1)!, \quad 0! = 1$$

You can also think about it in product notation: $n! = \prod_{k=1}^{n} k$, where we take the empty product to be equal to one. Now, the factorial shows up a lot: probably its most common appearance is in the Taylor series expansion of a function, although it also describes the number of permutations of $n$ elements of a set, etc. It also appears in the binomial coefficient,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{3.1}$$

which (among other things) describes the number of ways to choose $k$ things from $n$ things and appears in the binomial formula:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \tag{3.2}$$

Well, that's just fine. But suppose, for example, we wanted to expand $f(x,y) = (x+y)^{\alpha}$, where $\alpha \in \mathbb{R}\backslash\mathbb{Z}$; that is, where $\alpha$ was not an integer.

How are we supposed to interpret the binomial coefficient? We don't know what $\alpha!$ means when $\alpha$ isn't an integer!

Enter the gamma function. Its easiest interpretation is the continuation of the factorial to non-integer values: we want some function $\Gamma(x)$ that satisfies the all-important recursive property of the factorial. It turns out that a function that does this is

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt \tag{3.3}$$

To see this, let us compute the integral via integration by parts:

$$\begin{aligned}\int_0^\infty t^x e^{-t}\,dt &= -t^x e^{-t}|_0^\infty + x\int_0^\infty t^{x-1}e^{-t}\,dt\\ &= x\int_0^\infty t^{x-1}e^{-t}\,dt,\end{aligned}$$

so we see indeed that $\Gamma(x+1) = x\Gamma(x)$. And we have got the right function for interpolation of the factorial too; it is the case that $\Gamma(n+1) = n!$:

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n)\\ &= n(n-1)\Gamma(n-1)\\ &\ \ \vdots\\ &= n(n-1)\cdots 1 = n!\end{aligned} \tag{3.4}$$

There are so many beautiful identities related to the gamma function that it is difficult to know what to write and what to leave to you. I will mention a few; you should look up others in [6] or a similar function reference.[1]

[1] Doing this is incredibly enjoyable. I have spent many hours just reading special function identities in this text and others, such as Abramowitz [7]. At times they are overwhelming in their depth; imagining the work that has gone into each identity is inspirational, to say the least.

**G1.** There is another convenient integral form of the gamma function:

$$\begin{aligned}
\Gamma(x) &= \int_0^1 \left(\ln\frac{1}{t}\right)^{x-1} dt \\
&= \int_0^1 (-\ln t)^{x-1}\, dt \qquad \text{let } t \mapsto e^{-s},\ e^{-s} \in (0,1) \text{ so } s \in (\infty, 0) \\
&= \int_\infty^0 \left(-\ln e^{-s}\right)^{x-1} d\left(e^{-s}\right) \\
&= -\int_\infty^0 s^{x-1} e^{-s} ds \\
&= \int_0^\infty s^{x-1} e^{-s} ds
\end{aligned} \tag{3.5}$$

which of course we recognize as Eq. 3.3.

**G2.** We have $\Gamma(1-x) = -x\Gamma(-x)$. For, after all,

$$\begin{aligned}
\Gamma(1-x) &= \int_0^\infty t^{(1-x)-1} e^{-t} dt \\
&= \int_0^\infty t^{-x} e^{-t} dt \\
&= -e^{-t}t^{-x}\Big|_0^\infty - \int_0^\infty (-x)t^{-x-1}(-e^{-t})dt \\
&= (-x)\int_0^\infty t^{(-x)-1} e^{-t} dt \\
&= -x\Gamma(-x)
\end{aligned} \tag{3.6}$$

**G3.** Here is a beautiful identity that proves surprisingly useful: $\Gamma(1/2) = \sqrt{\pi}$. I will prove this so that you can get used to some of the integral manipulations that you'll use frequently in PoCS. First, note that we can rewrite the gamma function with the transformation $t \mapsto s^2$:

$$\begin{aligned}
\Gamma(x) &= \int_0^\infty t^{x-1} e^{-t} dt \\
&= \int_0^\infty (s^2)^{x-1} e^{-s^2} d(s^2) \\
&= 2\int_0^\infty s^{2(x-1)} e^{-s^2} s\ ds \\
&= 2\int_0^\infty s^{2x-1} e^{-s^2}\ ds
\end{aligned}$$

Now, let us take $x = \frac{1}{2}$ to write the gamma function as[2] $\Gamma(1/2) = 2\int_0^\infty e^{-s^2}\ ds$. We're looking for something square-rooted; let's square something else! (Don't question this questionable logic.)

$$\begin{aligned}\Gamma(1/2)\Gamma(1/2) &= \left(2\int_0^\infty e^{-s^2}\ ds\right)\left(2\int_0^\infty e^{-u^2}\ du\right)\\ &= 4 \iint\limits_{(0,\infty)\times(0,\infty)} \exp\left(-(s^2+u^2)\right) d(s,u)\end{aligned} \tag{3.7}$$

We would like to simplify this integral. Noting that we are integrating strictly over the first quadrant and that the term $s^2 + u^2$ looks suspiciously like $r^2$–the radius squared in polar coordinates–we establish the coordinate change $(s,u) \mapsto (r\cos\theta, r\sin\theta)$. The new differential is

$$\begin{aligned}\left|\frac{\partial(s,u)}{\partial(r,\theta)}\right| drd\theta &= \begin{vmatrix}\cos\theta & -r\sin\theta\\ \sin\theta & r\cos\theta\end{vmatrix} drd\theta\\ &= rdrd\theta,\end{aligned}$$

so we can convert Eq. 3.7 to

$$\begin{aligned}4 \iint\limits_{(0,\infty)\times(0,\infty)} \exp\left(-(s^2+u^2)\right) d(s,u) &= 4\int\limits_0^{\pi/2}\int\limits_0^{\infty} e^{-r^2} r\ dr\ d\theta\\ &= 4\int_0^{\pi/2} d\theta \int_0^\infty dr\ re^{-r^2}\\ &= 2\pi\int_0^\infty dr\ re^{-r^2}\\ \text{change to } \rho = -r^2,\ \rho\in(0,-\infty)\quad &= -\pi\int_0^{-\infty} d\rho\ e^\rho\\ &= \pi\end{aligned}$$

Take roots and you're finished.

Now you can actually decide what $\binom{1/2}{k}$ means in some principled way. We can just define the real-valued binomial coefficient as

$$\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)} \tag{3.8}$$

[2] You should note a similarity between this integral and the CDF of the normal distribution given by the integral of Eq. 2.37. This derivation is intimately related to the normal distribution; in fact, the gamma function is entwined with much of continuous probability theory in rather remarkable ways.

## 3.2 The beta function

The beta function is defined as

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{3.9}$$

From this, we will find an expression for it in terms of an integral. (It is often that this is done in the opposite direction; one defines the beta function as an integral and then derives this property. But it is my opinion that it is this property that is the fundamental characteristic of the beta function, not the other way around.) Without further ado, write the product of two gamma functions as

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \left(\int_0^\infty t^{\alpha-1}e^{-t}\,dt\right)\left(\int_0^\infty s^{\alpha-1}e^{-s}\,ds\right)\\ &= \iint\limits_{(0,\infty)\times(0,\infty)} t^{\alpha-1}s^{\beta-1}e^{-(s+t)}d(t,s)\end{aligned}$$

Make the variable change $(t,s)\mapsto(xy,x(1-y))$ and calculate $|\frac{\partial(t,s)}{\partial(x,y)}|$. You will find that $dt\ ds=-x\ dx\ dy$. Since $s,t\in(0,\infty)$ and $s+t=x$, we must have $x\in(0,\infty)$ and consequently $y\in(0,1)$. Then rewrite the integral again:

$$\begin{aligned}\iint\limits_{(0,\infty)\times(0,\infty)} t^{\alpha-1}s^{\beta-1}e^{-(s+t)}d(t,s) &= \int\limits_0^\infty\int\limits_0^1 (xy)^{\alpha-1}\left(x(1-y)\right)^{\beta-1}e^{-(xy+x(1-y))}x\ dx\ dy\\ &= \int\limits_0^\infty\int\limits_0^1 x^{\alpha+\beta-1}y^{\alpha-1}(1-y)^{\beta-1}e^{-x}dx\ dy\\ &= \int_0^\infty x^{\alpha+\beta-1}e^{-x}dx\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\ dy\\ &= \Gamma(\alpha+\beta)\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\ dy\end{aligned}$$

So we have found that

$$B(\alpha,\beta)=\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\ dy \tag{3.10}$$

as its integral definition. You will find that Eq. 3.9 is the most important beta function identity to know; it comes up a lot, especially in probability. Note that, if $n,m\in\mathbb{Z}_+$, then we can actually write

$$B(n,m)=\frac{(n-1)!(m-1)!}{(n+m-1)!} \tag{3.11}$$

by the definition of the gamma function.

There are many more phenomenal identities involving the gamma and beta function together. Here are two that you might find exceptionally useful:

**B1.** $B(\alpha, \beta) = B(\beta, \alpha)$. Pretty obvious.

**B2.** We can also rewrite the generalized binomial coefficient in terms of the beta function.

$$\begin{aligned}\binom{x}{y} &= \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}\\ &= \frac{\Gamma(x+2)}{(x+1)\Gamma(y+1)\Gamma(x-y+1)}\\ &= \frac{1}{(x+1)B(x-y+1, y+1)}\end{aligned}$$

We will derive more identities as exercises in the chapter on tricks and asymptotics.

## 3.3 The zeta function

This is probably the most second-most famous special function among mathematicians in general, and most revered among number theorists. It is usually defined as a power series:

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \tag{3.12}$$

It arises in many applications in physics and optimization, but is probably most famous for an unsolved conjecture due to Riemann, known as the *Riemann hypothesis*:

**Theorem 3.** ***Riemann hypothesis*** *(B. Riemann 1859): the real part of all nontrivial zeros of $\zeta(s)$ lie on the line $x = \frac{1}{2}$ in the complex plane.*

The trivial zeros of $\zeta(s)$ are the negative even integers. (We will not prove this here; it requires some complex analysis to see.) You should know that this function also has a representation as an infinite product over all prime numbers! Although this sounds exotic, it actually makes a lot of sense. In deriving this we will use the concept of a *number sieve*, which algorithmically factors out primes from numbers. Consider the following:

$$\begin{aligned}\zeta(s) &= 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots\\ \left(1 - \frac{1}{2^s}\right)\zeta(s) &= 1 + \cancel{\frac{1}{2^s}} + \frac{1}{3^s} + \cancel{\frac{1}{4^s}} + \frac{1}{5^s} + \cdot\end{aligned}$$

We can just get rid of all factors of any prime number $p$ on the RHS that we want by "sieving" them out; multiply the LHS by $(1 - \frac{1}{p^s})$. Doing this for every prime gives

$$\prod_p \left(1 - \frac{1}{p^s}\right) \zeta(s) = 1$$

which can just be rewritten as[3]

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^{-s}}\right). \tag{3.13}$$

This form of the function actually shows up quite often in physics; you may also find yourself factoring it out of some other infinite product you're considering–say, in the analysis of the steady state of a recurrence relation!

[3] If you're an analysis nerd (like me), you can make this completely rigorous by noting that $\sum_{n\geq 1} n^{-s}$ converges for all $s > 1$; thus, this infinite product converges since it is equal to the sum.

# Chapter 4

# Optimization

You're already familiar with univariate optimization from basic calculus. You *should* have learned basic multivariate continuous optimization in multivariate calculus, but I have found that this skill isn't always as well-developed as it should be; we'll go over this. We'll also dig into functional optimization, as that will play an important part of some problems in PoCS (as well as the successor course, complex networks). I won't go too much into other types, such as discrete and combinatorial optimization, since those are (perhaps surprisingly) much, much harder problems.

## 4.1 Multivariate continuous optimization

### 4.1.1 Unconstrained optimization

Suppose we have a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$ of which we'd like to find the maximum (minimum) value. No problem; essentially in parallel with the univariate case, we will

a. take its "derivative" and set it equal to zero

b. make sure its "second derivative" at the optimal point is negative (positive)

I used quotes because the actual action of differentiation is the only part of this algorithm that really changes. Instead of setting the univariate derivative to zero, we will set the gradient to zero; instead of checking that the second derivative is negative (positive), we will check that the eigenvalues of the Hessian are negative (positive). Recall that the gradient of $f$ is defined

$$\nabla f = \sum_{i=1}^{n} \partial_{x_i} f \; \mathbf{e}_i \tag{4.1}$$

where $\mathbf{e}_i$ are the standard Euclidean basis vectors in $\mathbb{R}^n$.[1] The Hessian operator is just $H_{ij} = \partial_{x_i x_j}$, so that the Hessian matrix of $f$ is

$$H(f) = \begin{pmatrix} \partial^2_{x_1} f & \partial_{x_1 x_2} f & \cdots & \partial_{x_1 x_n} f \\ \partial_{x_2 x_1} f & \partial^2_{x_2} f & \cdots & \partial_{x_2 x_n} f \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{x_n x_1} f & \partial_{x_n x_2} f & \cdots & \partial^2_{x_n} f \end{pmatrix} \tag{4.2}$$

We remember how to find eigenvalues, of course: solve the equation

$$\det\left(H(f) - \lambda I\right) = 0$$

for the eigenvalues $\lambda$. In $R^n$ there will be $n$ eigenvalues, although they are not guaranteed to be unique. The multidimensional equivalent to the sign of the second derivative is analysis of the signs of the eigenvalues. If all eigenvalues are negative, the solution of $\nabla f(x) = 0$ is a maximum; if they're all positive, the solution is a minimum. Of course, it may be the case that some eigenvalues are positive and some are negative; this is called a *saddle point.*

### 4.1.2 Constrained optimization

Of course, we often have constraints on our optimization problem. A typical problem you'll encounter (in PoCS and, more generally, in any sort of applied and industrial mathematics) is

$$\begin{aligned} \min_{x_1,\ldots,x_n} & f(x_1, ..., x_n) \\ \text{s.t. } & g_1(x_1, ..., x_n) = k_1 \\ & \vdots \\ & g_m(x_1, ..., x_n) = k_m \end{aligned} \tag{4.3}$$

[1] The gradient is an interesting thing. It is the only possible choice for the derivative of $f : \mathbb{R}^n \to \mathbb{R}$, but it is not necessarily the derivative of such a function; $f$ may have a perfectly well-defined gradient and be non-differentiable. Technically, a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable if and only if there exists a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(c + v) - f(c) = T_c(v) + ||v||E_c(v)$$

where $E_c(v)$ is a nonlinear error function that goes to zero as $||v|| \to 0$. In the $m = 1$ dimensional case that we are considering, it is obvious that $T$ must be the gradient. In the general $m$-dimensional case, it is not hard to see that actually $T = \frac{\partial(f_1,\ldots,f_m)}{\partial(x_1,\ldots,x_n)}$, the Jacobian matrix!

We will assume that $f$ and $g_j$ are all continuous and at least twice-differentiable. To find the optimum of $f$, we just extend the method of *Lagrange multipliers* a little bit. Instead of having only one multiplier, we have $m$ of them–one for each constraint. The method works like this: form the objective function $\mathcal{L}$, defined

$$\mathcal{L}(x_1, ..., x_n) = f(x_1, ..., x_n) + \sum_{j=1}^{m} \lambda_j \left(k_j - g_j(x_1, ..., x_n)\right) \tag{4.4}$$

Then, solve $\nabla\mathcal{L}(x) = 0$. All the solutions of this equation are possible optima; certainly they are critical points of $f$, and you know how to figure out (using the Hessian) if they are local maxima, minima, or saddle points. But here there are $m$ more things we need to do: figure out if each $x_{\text{opt}}$ is compatible with the $m$ equalities. We do this by taking partial derivatives with respect to each multiplier: $\partial_{\lambda_j}\mathcal{L} = 0$. Now we must find those solutions that solve *all of the equations* that we have.[2] Do yourself a favor and solve a few practice problems.

1. Minimize $f(x, y) = x + y$ subject to $x^2 + y^2 = r^2$.

2. Maximize $H(x) = -\sum_{i=1}^{n} x_i \ln x_i$ subject to $\sum_{i=1}^{n} x_i = 1$. The function $H$ is called the *entropy* of $x$; you will learn about it and related functions in PoCS.

## 4.2 Functional optimization

Many of the functions we care about in physics, economics, and computer science are actually functions from some space of functions to the real numbers. For example, consider the problem of minimizing total energy of a physical system over time:

$$\min_{x(t)} \int_{t_0}^{T} dt\ E(x(t), \dot{x}(t), t) \tag{4.5}$$

[2] The simpler, though more abstract, way to think of this is simply to take the $n+m$-dimensional gradient of $\mathcal{L}$, since it really is now a function in $\mathbb{R}^{n+m}$–$n$ of these coming from the variables $x_1, ..., x_n$, and $m$ of these coming from the $m$ multipliers. We can thus summarize the process by writing $\nabla\mathcal{L}(x;\lambda) = 0$, where here

$$\nabla = \sum_{j=1}^{n} \partial_{x_j} \mathbf{e}_j + \sum_{\ell=j+1}^{m} \partial_{\lambda_\ell} \mathbf{e}_\ell$$

Physical systems do this; this is called the principle of stationary energy or *Hamilton's principle.*

We can formulate these problems in a manner similar to problem 4.3. We want to maximize or minimize some quantity, in this case a functional represented by an integral (or sum), and we must do this subject to $m$ constraints, also functionals represented by integrals or sums. Thus, problems of this form will generally look like

$$\begin{aligned} \min_{r(x)} \int_\Omega dx\ F(r(x), r'(x), x) \\ \text{s.t.} \int_\Omega dx\ G_i(r(x), r'(x), x) = K_i, \quad i = 1, ..., m \end{aligned} \tag{4.6}$$

Solving these problems is a little bit trickier than above. We will first consider the case where $F$ and each $G_i$ are functions only of $r(x)$; they don't contain $r'(x)$ terms or terms with $x$ alone. In this case, we first form what's called the action integral:

$$J = \int_\Omega \left[ F(r(x)) - \sum_{i=1}^{m} \lambda_i \left(K_i - G_i(r(x))\right) \right] dx \tag{4.7}$$

Heuristically-speaking, we want to take the derivative of this integral with respect to $r(x)$. If we want to be formal about it, this is a pretty tall order; this is called functional differentiation, defined by the *Fréchet derivative*, and is denoted $\frac{\delta}{\delta r(x)}$. However, since we're (as I've stated before) cool and cavalier, we'll just do what comes naturally:

$$\begin{aligned} \frac{\delta J}{\delta r(x)} &= \frac{\delta}{\delta r(x)} \int_\Omega \left[ F(r(x)) - \sum_{i=1}^{m} \lambda_i \left(K_i - G_i(r(x))\right) \right] dx \\ &= \int_\Omega \frac{\partial}{\partial r(x)} \left[ F(r(x)) - \sum_{i=1}^{m} \lambda_i \left(K_i - G_i(r(x))\right) \right] dx \\ &= \int_\Omega \left[ \partial_{r(x)} F + \sum_{i=1}^{m} \lambda_i \partial_{r(x)} G_i \right] dx \end{aligned} \tag{4.8}$$

Okay, now what? Well, in problems like those described by Eq. 4.4 we take the derivative and set it equal to zero; we'll do that here too:

$$\int_\Omega \left[ \partial_{r(x)} F + \sum_{i=1}^{m} \lambda_i \partial_{r(x)} G_i \right] dx = 0 \implies \partial_{r(x)} F + \sum_{i=1}^{m} \lambda_i \partial_{r(x)} G_i = 0$$

This implication is by something called the *fundamental lemma of the calculus of variations*, into which I will not go here.[3] It's pretty obvious now that we just solve $\partial_{r(x)}F + \sum_{i=1}^{m} \lambda_i \partial_{r(x)} G_i = 0$ for $r(x)$; this will give us the optimal *function* $r(x)$ that solves the problem associated with the action in Eq. 4.7.

The harder case is when $F$ or $G_i$ are functions of $r'(x)$ as well. The action is, in this case, given by a similar integral to that in Eq. 4.7:

$$\begin{aligned} J &= \int_\Omega \left[ F(r(x), r'(x), x) - \sum_{i=1}^{m} \lambda_i \left( K_i - G_i(r(x), r'(x), x) \right) \right] dx \\ &= \int_\Omega L(r(x), r'(x), x)\ dx, \end{aligned} \tag{4.9}$$

where we have simply defined $L = F - \sum_1^m \lambda_i (K_i - G_i)$. By a theorem (proved in Section 4.2.1 for the truly brave) the fundamental equation here is not $\partial_{r(x)} L = 0$, as above, but rather the celebrated *Euler-Lagrange equation*, ubiquitous throughout classical mechanics:

$$\frac{\partial L}{\partial r(x)} = \frac{d}{dx} \frac{\partial L}{\partial r'(x)} \tag{4.10}$$

This, then, is the equation to be solved for $r(x)$!

### 4.2.1 Derivation of Euler-Lagrange equation

**Optional!!!** You need to be *very* familiar with vector calculus to begin to follow this derivation.

Let $x \in \mathbb{R}^d$, $\psi : \Omega \to \mathbb{R}$ where $\Omega \subseteq \mathbb{R}^d$ closed, and consider the problem of finding an extremum of the functional

$$J(\psi) = \int_\Omega L(x, \psi(x), \nabla \psi(x)) dx$$

where we assume that $L \in C^1$ and that the integral actually exists and has a finite value. We will impose the boundary condition $\psi(x) = f(x)$ for all $x \in \partial\Omega$. Let $h$ be some test function–any arbitrary function in $C^1$ such

[3] FLCV: Suppose $f$ and $g$ are functions $(a, b) \to \mathbb{R}$ with $g$ compactly supported. If

$$\int_a^b f(x) g(x) dx = 0$$

for *any* such $g$, then $f(x) = 0$.

that $h(x) = 0$ for all $x \in \partial\Omega$ will do–and consider the functional derivative $\delta J \equiv \lim_{\varepsilon \to 0} \frac{J(\psi+\varepsilon h)-J(\psi)}{\varepsilon} = \frac{d}{d\varepsilon} J(\psi + \varepsilon h)$. (I will not prove here that the usual rules of differential calculus hold here–it is a remarkable fact that most of them carry over!) Let us define the vector field $F = \sum_{i=1}^{d} \frac{\partial L}{\partial \psi_{x_i}} e_i$, where $\psi_{x_i} \equiv \frac{\partial \psi}{\partial x_k}$. Then we can calculate the functional derivative:

$$\begin{aligned}
\delta J &= \frac{d}{d\varepsilon} \int_\Omega L(x, \psi(x+\varepsilon h), \nabla\psi(x+\varepsilon h))dx \\
&= \int_\Omega \Big( \frac{\partial L}{\partial \psi} \frac{\partial \psi}{\partial \varepsilon} + \sum_1^d \frac{\partial L}{\partial \psi_{x_i}} \frac{\partial \psi_{x_i}}{\partial \varepsilon} \Big) dx \\
&= \int_\Omega \Big( \frac{\partial L}{\partial \psi} h + \sum_1^d \frac{\partial L}{\partial \psi_{x_i}} h_{x_i} \Big) dx
\end{aligned}$$

Substituting in the definition of $F$, we can rewrite this as

$$\begin{aligned}
\int_\Omega \Big( \frac{\partial L}{\partial \psi} h + F \cdot \nabla h \Big) dx &= \int_\Omega \Big( \frac{\partial L}{\partial \psi} - \nabla \cdot F \Big)\ h\ dx + \\
&\qquad \underbrace{\oint_{\partial\Omega} (hF) \cdot n\ dS}_{\text{zero, since } J \text{ is linear in } \frac{\partial L}{\partial \psi}} \qquad \text{by the divergence theorem} \\
&= \int_\Omega \Big( \frac{\partial L}{\partial \psi} - \nabla \cdot F \Big) h dx
\end{aligned}$$

We want $\delta J = 0$ for any test function $h$, and thus, by the fundamental theorem of calculus of variations, we have

$$\frac{\partial L}{\partial \psi} - \nabla \cdot F = \frac{\partial L}{\partial \psi} - \sum_1^d \frac{\partial}{\partial x_i} \frac{\partial L}{\partial \psi_{x_i}} = 0$$

This is called the Euler-Lagrange equation.

# Chapter 5

# Asymptotics and tricks

Physical processes often generate mathematical phenomena that are prohibitively difficult to express analytically. One may be able to reduce the complexity of some expressions using the tools of asymptotic theory, which I will describe in some detail below. I will also outline some handy tricks that will guaranteed be useful at some point in PoCS.

## 5.1 Asymptotics: outline and notation

Let $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. We will make no assumption on their continuity, differentiability, etc., but the reader will note in what follows that certain notions of asymptotic behavior require these constraints. There are two common uses of the sentence "$f$ is asymptotic to $g$":

a. For all $\varepsilon > 0$ there exists $x' \in \mathbb{R}$ such that, for all $x > x'$, $|f(x) - g(x)| < \varepsilon$. This is the infinite-limit, or large-$x$, definition of asymptotic, and is generally what is meant when no other clarifications are present.

b. For all $\varepsilon > 0$ there exists $x' \in \mathbb{R}$ such that, for all $x > x'$, $|f(\frac{1}{x}) - g(\frac{1}{x})| < \varepsilon$. This is the zero-limit definition of asymptotic. The reader will note that this definition is practically identical to the infinite-limit definition.

Here is a good exercise: using infinite sequences, propose a rigorous "$x = a$-limit" definition of asymptotic functions, and show that the two cases above are just special cases of this one.[1]

Asymptotics are very useful to us. Here is an obvious example: suppose, in the course of solving some problem, we come across the hyperbolic cosine

[1] You should start with the following: let $\{a_n\}_1^\infty$ be a sequence of real numbers such that $a_n \to a$ as $n \to +\infty$. Then...

function $\cosh x = \frac{1}{2}(e^x + e^{-x})$. We could carry this function along with us for the rest of the problem. However, we could also note that $|\cosh x - \frac{1}{2}e^x| = |\frac{1}{2}(e^x + e^{-x}) - \frac{1}{2}e^x| = \frac{1}{2}e^{-x}$ decays quickly to zero as $x$ gets large, and thus, for large $x$, the difference between $\cosh x$ and $\frac{1}{2}e^x$ is almost nothing at all.

The reader familiar with computer science will note a similarity between the definition of asymptotic functions and so-called "big theta notation" in the analysis of algorithms, although the two are not quite identical. See [8] for more details; we will not need this notation for our course.

We should also note the mathematician's use of the so-called "big oh" notation. In the coming sections you will see me write, for example, $f(x) = 1 + x + \mathcal{O}(x^2)$. This means that $f(x)$ looks like $1 + x$, and then the remaining error described by approximating the function $f(x)$ by $1 + x$ is no more than some constant times $|x^2|$ *when* $0 < x \ll 1$.. The connections between this notation and asymptoticity can be formally developed; I recommend that you do so.

## 5.2 Asymptotic formulas

### 5.2.1 Power / MacLaurin series

We may be interested in representing the function $f(x)$ by a power series of the form $\sum_0^\infty \frac{a_n}{n!}x^n$. We should note that **not all functions $f$ have this representation**, but that, as physicists, we're–say it with me–cool and cavalier, and so we'll often just assume that such a representation is possible.[2] It is very common to truncate Taylor series after either the first or second order. Truncation after the first order is principled, since if a function is differentiable at any point it is well-approximated by a linear function. Truncation after the second order is also principled when one wishes to take into account nonlinear effects that may be present; this is particularly true in the case of a multivariate function $f(x_1, ..., x_n)$, whose second order Taylor expansion (if it exists!) is

$$f(x_1, ..., x_n) = f(0) + \sum_{i=1}^{n}(\partial_{x_k} f(0))x_k + \frac{1}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}(\partial_{x_k x_j} f(0))x_k x_j + \mathcal{O}(x^3).$$

Without the second order expansion in this case, we lose all information about interactions between $f$'s arguments!

[2] Here is a good problem due to James Wilson: let $f(x) = e^{1/x^2}$. Show that $f(x)$ has derivatives of all orders and that $f^{(n)}(0) = 0$, meaning that $f$ cannot be expanded in a Taylor series. Be very careful of functions like this!

Here are some good expansions to know:

a. $e^u = \sum_0^\infty \frac{u^n}{n!}$, so that, for example, $e^u - 1 \approx u$ for small $u$.

b. $\ln(1-u) = -\sum_1^\infty u^n/n$, so that $\ln(1-u) \approx -u$ for small $u$.

c. $\sin x = \sum_0^\infty \frac{(-1)^n}{(2n+1)!} x^{2n+1}$ and $\cos x = \sum_0^\infty \frac{(-1)^n}{(2n)!} x^{2n}$. This, combined with (a.), leads to one of the most beautiful identities in mathematics, which you should now prove: $e^{i\theta} = \cos\theta + i\sin\theta$.[3] Common expansions here are $\sin x \approx x$, the famed "small-angle approximation" in the theory of classical oscillators, and $\cos x \approx 1 - x$, both valid for small $x$.

d. The geometric series, which is of paramount importance: $\frac{1}{1-u} = \sum_0^\infty u^n$. This series converges for $u \in (-1, 1)$ and quite clearly diverges elsewhere.

You should prove all of these from the definition of a Taylor series; by no means are these the only expansions you will need, and you must be able to derive them for yourself.

### 5.2.2 Sterling's formula and Laplace's method

One of the best-known and most impressive asymptotic approximations is one derived by James Sterling for the factorial. I will derive this below, and in doing so, show one of the many uses for the Gaussian integral $\int_{-\infty}^{\infty} dx\ e^{-x^2}$. (We have seen a modified version of this integral before, in 3.1.)

Write the factorial as the gamma function and make a change of variables:

$$\begin{aligned} n! &= \int_0^\infty dt\ t^n e^{-t} \\ &= \int_0^\infty dt\ e^{n\ln t - t} \\ &= \int_0^\infty d(ny)\ e^{n\ln ny - ny} \\ &= \int_0^\infty dy\ n\ e^{n(\ln n + \ln y) - ny} \\ &= n\ e^{n\ln n} \int_0^\infty dy\ e^{n(\ln y - y)}, \end{aligned}$$

[3] This is an incredibly deep result and has all sorts of ramifications in the theory of ordinary and partial differential equations, classical and quantum mechanics, and complex analysis.

so that the question becomes one of calculating the integral $\int_0^\infty dy\ e^{n(\ln y - y)}$. To do this we need a theorem due to Laplace.

**Theorem 4.** *(Laplace's method) If $f(x)$ is a function on $[a,b]$ with global maximum $x_0$ and $f(x)$ is not close to $f(x_0)$ unless $x$ is close to $x_0$, and if $f''(x_0) < 0$ and $x_0$ is not $a$ or $b$, then*

$$\int_a^b dx\ e^{Mf(x)} \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}} e^{Mf(x_0)}$$

*for large $M$.*

This is not an entirely rigorous statement of the theorem (although it is good enough for our purposes) and the proof of the theorem requires significant real analysis; see me for pointers if interested. However, we can still get a very good understanding of it in general by considering an asymptotic expansion of $f(x)$ about the point $x_0$:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \mathcal{O}((x - x_0)^3)$$

Since $x_0$ is not $a$ or $b$ and it is a global maximum, it is also a stationary point and thus $f'(x_0) = 0$; the function thus has expansion $f(x) \approx f(x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$. We can therefore write the integral as

$$\begin{aligned}\int_a^b dx\ e^{Mf(x)} &\approx \int_a^b dx\ e^{M(f(x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2)} \\ &= e^{Mf(x_0)} \int_a^b dx\ e^{-\frac{M}{2}f''(x_0)(x-x_0)^2}.\end{aligned}$$

Now, the remarkable thing is that this integral we've found is actually proportional to the CDF of the normal distribution with variance $\frac{1}{Mf''(x_0)}$ and mean $x_0$. Just as in our derivation of gamma function identity **G3**, we will now calculate this integral with the assumption that $a$ and $b$ are far enough apart that we can approximate $I(x) = \int_a^b dx\ e^{-\frac{M}{2}f''(x_0)(x-x_0)^2}$ by the identical integral over all of $\mathbb{R}$.[4] Write the square of the integral:

$$\begin{aligned}I(x)I(y) &= \left(\int_{-\infty}^{\infty} dx\ e^{-\frac{M}{2}f''(x_0)(x-x_0)^2}\right)\left(\int_{-\infty}^{\infty} dy\ e^{-\frac{M}{2}f''(x_0)(y-x_0)^2}\right) \\ &= \iint\limits_{\mathbb{R}^2} d(x,y)\ e^{-\frac{M}{2}f''(x_0)((x-x_0)^2+(y-x_0)^2)}\end{aligned}$$

[4] This is called the Gaussian integral. It is ubiquitous throughout science, engineering, and mathematics.

Make the variable change to polar coordinates with $r^2 = (x-x_0)^2+(y-x_0)^2$. The integral becomes

$$\iint\limits_{\mathbb{R}^2} d(x,y)\ e^{-\frac{M}{2}f''(x_0)((x-x_0)^2+(y-x_0)^2)} = \int_0^{2\pi} d\theta \int_0^\infty dr\ r\ e^{-\frac{M}{2}f''(x_0)r^2}$$
$$= 2\pi \int_0^\infty dr\ r\ e^{-\frac{M}{2}f''(x_0)r^2}$$

Make another change of coordinates, this time to $\rho = --\frac{M}{2}f''(x_0)r^2$, so that the integral is now just

$$2\pi \int_0^\infty dr\ r\ e^{-\frac{M}{2}f''(x_0)r^2} = -\frac{2\pi}{Mf''(x_0)} \int_0^{-\infty} d\rho\ e^\rho$$
$$= \frac{2\pi}{Mf''(x_0)}.$$

Thus we have $I(x) = \sqrt{\frac{2\pi}{Mf''(x_0)}}$, and so

$$e^{Mf(x_0)} \int_a^b dx\ e^{-\frac{M}{2}f''(x_0)(x-x_0)^2} \approx e^{Mf(x_0)} \sqrt{\frac{2\pi}{Mf''(x_0)}}. \tag{5.1}$$

What about in our particular case, that of the integral $\int_0^\infty dy\ e^{n(\ln y - y)}$? Well, noting that $f'(y) = y^{-1} - 1$ and $f''(y) = -y^{-2}$, we conclude that the maximum of $f$ occurs at $y = 1$ and is given by $y_0 = -1$. Using Laplace's method, we conclude that

$$n! = n\ e^{n\ln n} \int_0^\infty dy\ e^{n(\ln y - y)}$$
$$\approx ne^{n\ln n}\left(e^{-n}\sqrt{\frac{2\pi}{n}}\right)$$
$$= \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

You should take a minute to understand how beautiful this formula is.

Here is a useful exercise for you: derive a (slightly) weaker form of Sterling's approximation by taking the logarithm of the factorial:

$$\ln n! = \sum_1^n \ln n$$

Replace the sum by an integral and do what comes naturally. For more on this type of brutality involving sums and integrals, see the next section:

## 5.3 Replacing sums with integrals

Coming soon...

# Bibliography

[1] Ronald L Graham, Donald E Knuth, and Oren Patashnik. Concrete mathematics (1989). *Massachusetts: Addison-Wesley.*

[2] H Wilf. Generatingfunctionology,(1990). *ISBN: 0-12-751956-4.*

[3] David Dewhurst. Random walks on networks. `https://daviddewhurst.github.io/random-walks-networks/`, 2017. [Online; accessed 22-August-2017].

[4] William Feller. *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons New York, 1968.

[5] Willliam Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

[6] Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products.* Academic press, 2014.

[7] Milton Abramowitz, Irene A Stegun, et al. Handbook of mathematical functions. *Applied mathematics series*, 55(62):39, 1966.

[8] Big O Notation. `https://en.wikipedia.org/wiki/Big_O_notation`.