Lecture Notes: Probability Theory II

Dmitriy (Tim) Kunisky

Spring 2026 (Last Updated: February 21, 2026)

Contents

1 | Introduction

1.1 Main Ideas of Probability Theory I

This course is a continuation of Probability Theory I. In case you have not taken that course, or in any case to summarize, you can think of the following as the main ideas and objects introduced there.

First, **measure** and **integration theory** give us a collection of definitions and associated basic results from analysis that, in the context of probability theory, lead to coherent and useful notions of intuitive ideas like “random”, “uniformly at random”, “expected value”, “convergence in distribution”, “conditional distribution”, and so forth. This level of formalism can usually be avoided when dealing with discrete probability, but it can be dangerous to try to reason about continuous probability without understanding these foundations thoroughly. One simple example yet that appeared quite late in the literature (around 1900) as an illustration of such confusion is *Bertrand’s paradox*, which you can read about online if you are interested. While parts of continuous probability can also be carried out less formally by always using density functions, measure theory also gives us a united language for discrete and continuous probability, and also allows us to reason easily about “hybrid” distributions that have some point masses and some continuous parts.

Next, **independence** is perhaps the main structural assumption on measures that is characteristic of probability theory; in measure-theoretic language this is the study of **product measures**. Very many of the random varaibles we are interested in in probability theory arise, even if they are not themselves independent, as an image of some underlying independent random variables (such as polynomials, matrix functions, or other “simple” functions of i.i.d. random variables).

Third, main business of classical probability theory is to prove **limit theorems** about random variables: results that say that sequences of random variables S_n have some limiting behavior in terms of their full distributions or summary statistics like expectations, moments, or particular tail probabilities as $n \to \infty$. Examples include laws of large numbers, central limit theorems, and large deviations principles. Further, we are usually interested in S_n that somehow involve the “aggregation” of the effect of many random influences on a single random quantity, like the main model we discuss below. This focus stems in part from the motivation for classical probability from statistical mechanics in physics, the study of how the aggregate behavior of many simple particles gives rise to more complicated phenomena. An important parallel pursuit but one that we will only touch upon occasionally in this course is **non-asymptotic** results in probability theory, which also treat large random systems but seek instead to explicitly describe or bound properties of S_n for a given n rather than treating the limit as $n \to \infty$. These two styles of analysis are complementary, and it is useful to be comfortable with both.

Finally, the concrete model you focused on in Probability Theory 1 is the **sum of independent and identically distributed (i.i.d.) random variables**, $S_n = \sum_{i=1}^n X_i$ for X_i drawn i.i.d. from some distribution. Such models are also called **random walks**. This is an very, very widely studied model that is an excellent place to start building probabilistic intuition and to familiarize yourself with the main tools of probability. You may have also taken a few steps beyond this basic model, in particular to the generalization where the X_i have different distributions but are still independent. These were the settings in which you should have seen the law of large numbers, central limit theorem, Poisson limit theorem, and large deviations inequalities, properties of the sequence of random variables $(S_n)_{n \geq 0}$; we will very briskly review these below.

1.2 Our Goals in Probability Theory II

The goal of this class is to take the following two important steps beyond the above topics. First, we will consider more **general discrete sequences** of random variables whose construction deviates in various ways from the sum-of-i.i.d. model. The following are two important examples.

> **Example 1.2.1** (Martingales and functions of i.i.d. variables)**.** The sum-of-i.i.d. model may be reframed as the study of $f(X_1, \dots, X_n)$ for the particular function $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$, the sum function. It is natural to ask how well we can understand other, in particular nonlinear, functions of i.i.d. random variables. In fact we will see that the "partial expectations" given by
>
> $$M_k = \mathop{\mathbb{E}}_{X_{k+1},\dots,X_n} f(X_1, \dots, X_n)$$
>
> form a sequence $(M_k)_{k=0}^n$ that in some regards behaves similarly to the sequence $(S_k)_{k=0}^n$ from the sum-of-i.i.d. model. This object is called a *Doob martingale*; more generally, we will see that *martingales* are an even more general family of random sequences to which a large chunk of the theory of sums-of-i.i.d. can be adapted.

> **Example 1.2.2** (Markov chains)**.** Generalizing in a different direction, suppose that $X_i \in \{\pm 1\}$. Then, the sum-of-i.i.d. or random walk model S_n is, indeed, a random walk on the *graph* formed by the integers $\mathbb{Z}$, where we move to the left or to the right by one step with a given probability. We may dispose of the arithmetic qualities of this setting and ask more generally: what happens if we take random walks on general graphs or other geometric objects, generated by a sequence of independent "steps" to new locations? Again, we will see that some of the theory of sums-of-i.i.d. may be adapted to more general random walks, which are formalized and further generalized in the theory of *Markov chains*. We will see that this leads to interesting connections to graph theory and applications to Monte Carlo methods in machine learning and scientific computing.

Second, we will move away from the discrete setting and take some initial steps in the theory of **continuous families of random variables**. In particular, we will focus on generalizations of random walks $S_n = S(n)$ to continuous variants $(S(t))_{t\geq 0}$, the most important of which is *Brownian motion*. More generally, this will give you a taste of the theory of *random functions* on continuous domains and their various technicalities and intriguing properties; towards the end, we will see a bit of *stochastic calculus*, the interaction of the ordinary calculus and theory of differential equations that you are familiar with with the behavior of functions driven by random processes. Here as in other branches of advanced probability, it is natural to ask the same basic questions as before, the most important being: what limit theorems can we formulate about these new objects, in our case random functions? A notable example we will come back to is the following.

> **Example 1.2.3** (Donsker's invariance principle)**.** Suppose that X_i are i.i.d. centered random variables with unit variance (i.e., $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$). We have discussed above the sum-of-i.i.d. model generated from these, $S_n = \sum_{i=1}^n X_i$. We may view this as $S(n)$, a random function $S : \mathbb{Z}_{\geq 0} \to \mathbb{R}$. Further, interpolating linearly between the values S takes, we may extend that to a random function $S : \mathbb{R}_{\geq 0} \to \mathbb{R}$, the kind of object mentioned above. It turns out that a certain family of renormalizations of S "converges" in a suitable technical sense we will develop. Namely, $S(ct)/\sqrt{c}$ "converges", as $c \to \infty$, to Brownian motion—regardless of the distribution of the steps X_i! You may view this as the "functional" analog of the central limit theorem, thus a comparably foundational result of the theory of random functions.

1.3 Brief Technical Review

Before we start our new topics, let us briefly sketch some proof techniques and applications of them that you should have seen in Probability Theory I. Here we will always be discussing the sum-of-i.i.d. model $S_n = \sum_{i=1}^n X_i$, for $X_1, X_2, \dots \sim \mu$ i.i.d. for some probability measure μ. Let us write $c := \mathbb{E}X_i$.

1.3.1 Laws of Large Numbers

A *weak law of large numbers (WLLN)* is a result of the form

$$\frac{1}{n}S_n \xrightarrow{\mathbb{P}} c,$$

the notation denoting convergence in probability, meaning that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n}S_n - c\right| > \epsilon\right] = 0.$$

The WLLN holds under only the very mild assumption $\mathbb{E}|X_i| < \infty$, which indeed is necessary for the definition of c to make sense in the usual sense. A straightforward way to prove this is to first show that the WLLN holds if $\mathbb{E}X_i^2 < \infty$ by applying Chebyshev's inequality to the above probability. Then, the trick of *truncation*, replacing X_i by $Y_i := X_i \cdot \mathbb{1}\{|X_i| \leq C\}$ for some large constant C, allows the condition to be relaxed as above.

A *strong law of large numbers (SLLN)* is a result of the form

$$\frac{1}{n}S_n \xrightarrow{\text{a.s.}} c,$$

the notation denoting convergence almost surely, meaning that

$$\mathbb{P}\left[\lim_{n\to\infty}\frac{1}{n}S_n = c\right] = 1.$$

Note that this is an extremely different notion on a conceptual level from convergence in probability: convergence almost surely refers to a probability concerning the entire random sequence $(S_n)_{n\geq 0}$, while convergence in probability only refers to a sequence of probabilities concerning each S_n individually. Yet, the two can sometimes be related; in particular, the SLLN holds under the same assumption $\mathbb{E}|X_i| < \infty$, and that can be proved by the methods above along with the *Borel-Cantelli lemma*, an important "bridge" between properties of the S_n individually and the entire sequence.

1.3.2 Weak Convergence and Convergence in Distribution

Before proceeding, let us introduce a common and very handy piece, though not entirely standard, piece of terminology.

Definition 1.3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable taking values in some other measurable space $(\Sigma, \mathcal{G})$, i.e., a measurable function $X : \Omega \to \Sigma$.[a] Then, the *law* of X is the probability measure μ on Σ defined by

$$\mu(A) := \mathbb{P}(X^{-1}(A))$$

for each $A \in \mathcal{G}$. In more conventional probabilistic notation,

$$\mu(A) = \mathbb{P}[X \in A].$$

We denote $\mu = \mathsf{Law}(X)$.

[a] Most often for us Σ is the real numbers and $\mathcal{G}$ their Borel σ-algebra.

It will help greatly to clarify your probabilistic thinking to internalize this notation. Law is just a formal way to talk about "the distribution of a random variable". Thus, if $\mathsf{Law}(X) = \mathsf{Law}(Y)$, then X and Y have the same distribution, i.e., the probabilities with which they take various values are all the same, $\mathbb{P}[X \in A] = \mathbb{P}[Y \in A]$. If $\mathsf{Law}(X)$ is "close" in some sense to $\mathsf{Law}(Y)$, then that should mean that $\mathbb{P}[X \in A] \approx \mathbb{P}[Y \in A]$. The main benefit of but also source of confusion in thinking about laws is that X and Y need not be defined on the same probability space. For instance, central limit theorems can be viewed as saying that the law of a random variable (a normalized sum-of-i.i.d.) is close to a Gaussian, where that Gaussian can and should be viewed as just some other random variable on an unrelated probability space.

With that, you can understand that **convergence in distribution is convergence of laws of random variables**, i.e., a notion of convergence of measures. One way to formalize that notion is as follows; we focus on measures on $\mathbb{R}$ for the sake of simplicity.

Definition 1.3.2 (Weak convergence)**.** Let $\mu_1, \mu_2, \dots, \mu_\infty$ be probability measures on $\mathbb{R}$. We say that μ_n *converges weakly* to μ_∞, denoted $\mu_n \xrightarrow{(w)} \mu_\infty$, if, for any $f : \mathbb{R} \to \mathbb{R}$ bounded and continuous, we have

$$\lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu_\infty.$$

If $X_1, X_2, \dots, X_\infty$ are random variables such that $\mu_n = \mathsf{Law}(X_n)$ for each $n = 1, 2, \dots, \infty$, then the above condition may be written in probabilistic notation as

$$\lim_{n\to\infty} \mathbb{E} f(X_n) = \mathbb{E} f(X_\infty).$$

In this case, weak convergence is also called *convergence in distribution* of X_n to X_∞, and denoted $X_n \Rightarrow X_\infty$.

In my opinion, weak convergence can be a little bit less confusing than convergence in distribution for many limit theorems you come across. The reason is that, as mentioned before, $X_n \Rightarrow X_\infty$ looks like a claim about the random variables X_n, but it actually does not depend on their coupling for different n but rather only on the individual laws of each X_n one at a time. In particular, in statements like central limit theorems, X_∞ needs to be *some* Gaussian random variable, but that random variable does not (and cannot, as your homework will show in a particular sense) have anything in particular to do with the X_n; it is just some other unrelated Gaussian random variable "floating in space". To avoid needing to talk about this unusual object, I think that writing $\mathsf{Law}(X_n) \xrightarrow{(w)} \mathcal{N}(0,1)$ is nicer and clearer, where $\mathcal{N}(0,1)$ (or whatever limiting distribution we are interested in) is a straightforward object—a measure—and the left-hand side of the convergence is just a sequence of other measures.

You may have also seen weak convergence in the following equivalent form.

Proposition 1.3.3 (Part of "Portmanteau Theorem")**.** Let $\mu_1, \mu_2, \dots, \mu_\infty$ be probability measures on $\mathbb{R}$. Then, $\mu_n \xrightarrow{(w)} \mu_\infty$ if and only if, for all $t \in \mathbb{R}$ such that $\mu_\infty(\{t\}) = 0$,[a] we have

$$\lim_{n\to\infty} \mu_n((-\infty, t]) = \mu_\infty((-\infty, t]).$$

If $X_1, X_2, \dots, X_\infty$ are random variables such that $\mu_n = \mathsf{Law}(X_n)$ as above, then the above is equivalent to, for any t such that $\mathbb{P}[X_\infty = t] = 0$, having

$$\lim_{n\to\infty} \mathbb{P}[X_n \le t] = \mathbb{P}[X_\infty \le t].$$

The function $F_X(t) = \mathbb{P}[X \le t]$ is called the *distribution functions* of a random variable X (in other contexts often also the *cumulative distribution function* or *c.d.f.*), and thus this shows that convergence in distribution in the previous sense is equivalent to pointwise convergence of distribution functions at all points where F_{X_∞} is continuous.

[a] A point t with $\mu(\{t\}) > 0$ is called an *atom* or *point mass* of a measure μ.

While the definition in terms of bounded continuous functions turns out to be both more general (since it can be generalized to very broad domains just provided they allow for a notion of continuity, for instance arbitrary metric spaces) and often easier to work with, the above equivalent form is comforting because it tells us that weak convergence indeed implies "convergence of distributions" in the sense of distribution functions, which measure concrete tail probabilities.

1.3.3 Characteristic Functions

How do we establish weak convergence? There are actually many means, but you have probably focused on just one so far, using *characteristic functions*, the functions associated to a probability measure μ or random variable X given by

$$\begin{aligned} \phi_\mu(t) &:= \int \mathsf{exp}(itx)\, d\mu(x), \\ \phi_X(t) &:= \mathbb{E}\,\mathsf{exp}(itX). \end{aligned}$$

Remark 1.3.4. If μ has a density function $\rho(x)$ with respect to the Lebesgue measure, then the characteristic function is merely the *Fourier transform* of ρ:

$$\phi_\mu(t) = \int \mathsf{exp}(itx)\rho(x)\, dx.$$

If you are familiar with a bit of harmonic analysis, this perspective makes the main properties of characteristic functions not so surprising.

The main important properties of the characteristic function are as follows.

Proposition 1.3.5. The characteristic function satisfies the following:

1. $\mathsf{Law}(X) = \mathsf{Law}(Y)$ if and only if $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$.
2. (Lévy continuity theorem) $X_n \Rightarrow X_\infty$ if and only if $\phi_{X_n}(t) \to \phi_{X_\infty}(t)$ for all $t \in \mathbb{R}$.
3. If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.

The point, then, is that the characteristic function on the one hand gives a tool for establishing weak convergence by establishing pointwise convergence of characteristic functions, while also being particularly friendly with sums of independent random variables. Thus, it is an excellent tool in particular for the sum-of-i.i.d. model.

Example 1.3.6 (Law of large numbers). Let us sketch how to use characteristic functions to reprove the WLLN. Let $X_1, X_2, \dots \sim \mu$ be i.i.d. with finite mean and variance and let $S_n := \sum_{i=1}^n X_i$. Then, we compute

$$\begin{aligned}\phi_{\frac{1}{n}S_n}(t) &= \phi_{\frac{1}{n}X_1+\cdots+\frac{1}{n}X_n}(t) \\ &= \phi_{\frac{1}{n}X_1}(t)\cdots\phi_{\frac{1}{n}X_n}(t) \\ &= \left(\phi_{\frac{1}{n}X_1}(t)\right)^n \\ &= \left(\mathbb{E}\exp\left(\frac{itX_1}{n}\right)\right)^n\end{aligned}$$

Now, letting ourselves be a little heuristic (it is not hard to make this step precise), take a Taylor expansion of the exponential function inside, and by our assumption of finite variance we have

$$\begin{aligned}&= \left(1 + \frac{it}{n}\mathbb{E}X_1 + O\left(\frac{1}{n^2}\right)\right)^n \\ &\to \exp(it\mathbb{E}X_1).\end{aligned}$$

This limit is the characteristic function of the constant scalar random variable $c := \mathbb{E}X_1$. Thus, we find that $\frac{1}{n}S_n \Rightarrow c$, and as an exercise you may show that this implies convergence in probability as well. Of course, the WLLN is already easy to prove by other and more direct means, but as we will see below this is an illustrative calculation for the general method around using characteristic functions on sum-of-i.i.d. models.

1.3.4 Central Limit Theorem

Now let $X_1, X_2, \dots \sim \mu$ be i.i.d. with mean c and variance $\sigma^2 < \infty$, and let $S_n := \sum_{i=1}^n (X_i - c)$. Since we are centering S_n, without loss of generality we may assume $c = 0$. A *central limit theorem (CLT)* is a result of the form

$$\mathsf{Law}\left(\frac{1}{\sqrt{n}}S_n\right) \xrightarrow{\text{(w)}} \mathcal{N}(0, \sigma^2),$$

a weak convergence statement with meaning as discussed above.

The CLT holds just under the above assumptions, but let us review a quick proof using characteristic functions of a version with the extra assumption that $\mathbb{E}|X_i|^3 < \infty$. In this case, we may mimic our proof of the WLLN above: expanding similarly, we find

$$\begin{aligned}\phi_{\frac{1}{\sqrt{n}}S_n}(t) &= \left(\mathbb{E}\exp\left(\frac{itX_1}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{it}{\sqrt{n}}\mathbb{E}X_1 - \frac{t^2}{2n}\mathbb{E}X_1^2 + O\left(\frac{1}{n^{3/2}}\right)\right)^n\end{aligned}$$

and using our assumption on the first two moments of X_1, we have

$$\begin{aligned} &= \left(1 - \frac{\sigma^2 t^2}{2n} \mathbb{E} X_1^2 + O\left(\frac{1}{n^{3/2}}\right)\right)^n \\ &\to \exp\left(-\frac{\sigma^2 t^2}{2}\right), \end{aligned}$$

which an integral calculation shows is precisely $\phi_{\mathcal{N}(0,\sigma^2)}(t)$.

Our next topic will be one of the many other methods of proof of limit theorems in general and CLTs in particular. Before that, it is instructive to reflect on this proof and its advantages and disadvantages. On the one hand, it is very short and simple, and if we take the Lévy continuity theorem for granted, the manipulations with characteristic functions are also quite elementary.

On the other hand, the proof has many disadvantages. First, it depends very rigidly on the structure of the sum-of-i.i.d. model (via the special properties of characteristic functions for independent sums) Second, it is "not probabilistic" in the sense that it operates in the more mysterious Fourier domain without giving us intuition about the probabilistic phenomena that make the CLT hold. For instance, it is hard to use this proof to give an intuitive explanation for why the all sum-of-i.i.d. models up to mild assumptions have the same distributional limit (the *universality* aspect of the CLT) or for why that limit is Gaussian. And lastly, perhaps in part for this reason, it is unclear how (though possible sometimes) to use this proof to give quantitative bounds in the CLT, say on how far certain probabilities or expectations are from their Gaussian limits. Next we will see a very different approach to the CLT that addresses all of these issues, and is a useful technique in probability theory broadly speaking, including some modern applications we will mention.

2 | Advanced Limit Theorems

We will next see some new ways to prove limit theorems, in particular the central and Poisson limit theorems and some variations thereof. To do this, we will develop two additional ways to prove weak convergence. One may view these techniques as associated to different *test functions* and assumptions on them: recall that weak convergence $\mu_n \xrightarrow{(w)} \mu_\infty$ is defined as having

$$\int f \, d\mu_n \to \int f \, d\mu_\infty$$

for every bounded continuous $f : \mathbb{R} \to \mathbb{R}$. The "Fourier method" of characteristic functions amounts to showing (in the Lévy continuity theorem) that it suffices to consider the special test functions $f(x) = \mathsf{exp}(itx)$. However, as a more general principle, any "sufficiently dense" family of test functions that can approximate bounded continuous functions sufficiently well can be used in this way, and leads to a method for proving weak convergence. Thus, an important aspect of weak convergence proofs is to choose to work with test functions that are well-adapted to one's situation.

2.1 Lindeberg Method

The first method we consider uses calculus on test functions, and thus we focus on test functions that are *smooth* (so that we can take derivatives) and *compactly supported* (so that those derivatives are bounded).

2.1.1 Continuity Theorem

First, let us show that there is indeed an associated "continuity theorem" yielding weak convergence.

Lemma 2.1.1 (Weak convergence by smooth functions)**.** Let $\mu_1, \mu_2, \dots, \mu_\infty$ be probability measures. Then, $\mu_n \xrightarrow{(w)} \mu_\infty$ if and only if $\int f \, d\mu_n \to \int f \, d\mu_\infty$ for all f smooth and compactly supported.

Proof. One direction is immediate by definition. For the other, suppose we have the convergence of integrals of smooth and compactly supported functions. Let f be bounded and continuous, such that $|f(x)| \leq K$ for all $x \in \mathbb{R}$. Fix $\epsilon, M > 0$. We will take for granted that there exists a smooth compactly supported $g : \mathbb{R} \to \mathbb{R}$ such that:

1. $|f(x) - g(x)| \leq \epsilon$ for all $x \in [-M, M]$.
2. $|g(x)| \leq 2K$ for all $x \in \mathbb{R}$.

Using this, we may bound by the triangle inequality

$$\left| \int f \, d\mu_n - \int f \, d\mu_\infty \right| \leq \left| \int f \, d\mu_n - \int g \, d\mu_n \right| + \left| \int f \, d\mu_\infty - \int g \, d\mu_\infty \right| + \left| \int g \, d\mu_n - \int g \, d\mu_\infty \right| .$$

Here by assumption the last term goes to zero as $n \to \infty$. For the second term, we have

$$\left| \int f \, d\mu_\infty - \int g \, d\mu_\infty \right| \leq \int_{[-M,M]} |f - g| \, d\mu_\infty + \int_{\mathbb{R} \setminus [-M,M]} |f - g| \, d\mu_\infty \leq \epsilon + 3K\mu_\infty(\mathbb{R} \setminus [-M, M]).$$

Similarly for the first term,

$$\left| \int f \, d\mu_n - \int g \, d\mu_n \right| \leq \epsilon + 3K\mu_n(\mathbb{R} \setminus [-M, M]) = \epsilon + 3K(1 - \mu_n([-M, M])).$$

Here, we would like to control the term involving μ_n by something instead involving μ_∞, so that we may obtain a bound as $n \to \infty$. We will also take for granted that there exists a smooth compactly supported $h : \mathbb{R} \to \mathbb{R}$ sandwiched between two indicator functions,

$$\mathbb{1}_{[-M/2, M/2]}(x) \leq h(x) \leq \mathbb{1}_{[-M,M]}(x).$$

By assumption,

$$\lim_{n \to \infty} \int h \, d\mu_n = \int h \, d\mu_\infty.$$

By bounding the left- and right-hand side using the property of h above, we find

$$\liminf_{n \to \infty} \mu_n([-M, M]) \geq \mu_\infty([-M/2, M/2]),$$

and thus

$$\limsup_{n \to \infty} \left| \int f \, d\mu_n - \int g \, d\mu_n \right| \leq \epsilon + 3K(1 - \mu_\infty([-M/2, M/2])) = \epsilon + 3K(1 - \mu_\infty(\mathbb{R} \setminus [-M/2, M/2])).$$

Putting everything together, we find

$$\limsup_{n \to \infty} \left| \int f \, d\mu_n - \int f \, d\mu_\infty \right| \leq 2\epsilon + 6K(1 - \mu_\infty(\mathbb{R} \setminus [-M/2, M/2])).$$

Since this holds for all $\epsilon, M > 0$, taking $\epsilon \to 0$ and $M \to \infty$ then gives the result. □

2.1.2 Proof of Central Limit Theorem

Now we introduce the Lindeberg method and give a new proof of the central limit theorem. We focus on the following slightly weakened version.

Theorem 2.1.2 (Weak CLT)**.** Let $X_1, X_2, \dots \sim \mu$ be i.i.d. with $\mathbb{E}X_i = c$, $\mathsf{Var}\, X_i = \sigma^2$, and $\mathbb{E}|X_i|^3 < \infty$. Write $S_n := \sum_{i=1}^n (X_i - c)$ and let $N \sim \mathcal{N}(0,1)$. Then, $\frac{1}{\sqrt{n}} S_n \Rightarrow N$.

Proof. First, note that we may assume by translating and scaling the X_i that, without loss of generality, $c = 0$ and $\sigma^2 = 1$. Further, let us define $R := \mathbb{E}|X_i|^3$.

By Lemma 2.1.1, it suffices to show that for all $f : \mathbb{R} \to \mathbb{R}$ smooth and compactly supported, we have

$$\mathbb{E}f\left(\frac{1}{\sqrt{n}} S_n\right) = \mathbb{E}f\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) \to \mathbb{E}f(N).$$

The simple idea of the Lindeberg method is to observe that, if we introduce $Y_1, \dots, Y_n \sim \mathcal{N}(0,1)$ i.i.d., then we have by basic properties of Gaussian distributions that

$$\mathsf{Law}(N) = \mathsf{Law}\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right).$$

Thus, it suffices to control the differences

$$\Delta^{(n)} := \mathbb{E}f\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right).$$

Lindeberg's method is simply to expand this into a telescoping sum where we replace the X_i with the Y_i one by one, and show that each step only incurs a small amount of error. That is, we first expand

$$\begin{aligned}\Delta^{(n)} &= \sum_{k=1}^n \left[\mathbb{E}f\left(\frac{Y_1 + \cdots + Y_{k-1} + X_k + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}f\left(\frac{Y_1 + \cdots + Y_k + X_{k+1} + \cdots + X_n}{\sqrt{n}}\right)\right] \\ &=: \sum_{k=1}^n \Delta_k^{(n)}.\end{aligned}$$

We now try to control the individual $|\Delta_k^{(n)}|$. Fix some k and define

$$Z_k := \frac{Y_1 + \cdots + Y_{k-1} + X_{k+1} + \cdots + X_n}{\sqrt{n}}.$$

Then, we may write

$$\Delta_k^{(n)} = \mathbb{E}f\left(Z_k + \frac{X_k}{\sqrt{n}}\right) - \mathbb{E}f\left(Z_k + \frac{Y_k}{\sqrt{n}}\right).$$

We expect to have $Z_k \sim S_n$, so in particular this quantity should be (typically) of order $\Theta(1)$, while the terms involving X_k and Y_k are of order $O(1/\sqrt{n})$. Thus, it is reasonable to use Taylor expansion on the above expressions, which we may do since we have assumed f is smooth. We find that, using the explicit form of the Taylor remainder, there is some $\widetilde{Z} = \widetilde{Z}(Z_k, X_k)$ such that

$$f\left(Z_k + \frac{X_k}{\sqrt{n}}\right) = f(Z_k) + f'(Z_k)\frac{X_k}{\sqrt{n}} + \frac{1}{2}f''(Z_k)\frac{X_k^2}{n} + \frac{1}{6}f'''(\widetilde{Z})\frac{X_k^3}{n^{3/2}}.$$

We note, importantly, that Z_k is independent of X_k. Therefore, taking expectations, using this independence and that $\mathbb{E}X_k = 0$ and $\mathbb{E}X_k^2 = 1$, and bounding the remainder term crudely, we find

$$\left|\mathbb{E}f\left(Z_k + \frac{X_k}{\sqrt{n}}\right) - \mathbb{E}f(Z_k) - \frac{1}{2n}\mathbb{E}f''(Z_k)\right| \le \frac{1}{6}\|f'''\|_{L^\infty}\frac{R}{n^{3/2}}.$$

But, an identical argument also applies to the version with Y_k instead of X_k, giving:

$$\left|\mathbb{E}f\left(Z_k + \frac{Y_k}{\sqrt{n}}\right) - \mathbb{E}f(Z_k) - \frac{1}{2n}\mathbb{E}f''(Z_k)\right| \le \frac{1}{6}\|f'''\|_{L^\infty}\frac{R}{n^{3/2}}.$$

Thus, by triangle inequality we have

$$\begin{aligned} |\Delta_k^{(n)}| &= \left|\mathbb{E}f\left(Z_k + \frac{X_k}{\sqrt{n}}\right) - \mathbb{E}f\left(Z_k + \frac{Y_k}{\sqrt{n}}\right)\right| \\ &\le \frac{1}{3}\|f'''\|_{L^\infty}\frac{R}{n^{3/2}}. \end{aligned}$$

Finally, since $\Delta^{(n)}$ is a sum of n such terms, we have

$$|\Delta^{(n)}| \le \frac{1}{3}\|f'''\|_{L^\infty}R \cdot \frac{1}{n^{1/2}},$$

and in particular $\lim_{n\to\infty} |\Delta^{(n)}| = 0$, completing the proof. □

Let us make a few additional remarks about this method and its generalizations and applications beyond the CLT.

2.1.3 Universality

The first important quality of the Lindeberg method is that it would still have been very useful even if we did not know that the limiting distribution of $\frac{1}{\sqrt{n}}S_n$ was Gaussian. Even so, we could still show, provided that $\mathbb{E}X_i = \mathbb{E}Y_i$, $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, and $\mathbb{E}|X_i|^3, \mathbb{E}|Y_i|^3 \le R < \infty$, that $|\mathbb{E}f(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i) - \mathbb{E}f(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i)| \lesssim 1/\sqrt{n}$. That is, we could prove the *universality* of a limit without actually knowing what that limit is. This is often useful in more advanced applications when one does not know what the exact limiting distribution of some random variable is, or perhaps does not even expect it to admit a tractable description. One concrete and very much analogous application to what we have seen is in random matrix theory, where many such universality properties are known for, say, matrices with i.i.d. random entries, provided that the first *four* moments of those entries take some fixed values (see, e.g., [TV11]).

2.1.4 Quantitative Rates of Convergence

Also notably, the Lindeberg method provides fully explicit and non-asymptotic bounds on $|\mathbb{E}f(\frac{1}{\sqrt{n}}S_n) - \mathbb{E}f(N)|$ in terms of f, $\mathbb{E}|X_1|^3$, and n. These kinds of bounds can be very useful in applications, and also give us some information about what properties of the setting govern the rate of convergence: in this case, we see that it is the “flatness” of f and the light tails of the X_i that lead to faster convergence. It turns out to be possible to carry out a similar quantitative proof for the non-smooth test function $f(x) = \mathbb{1}\{x \le t\}$, which leads to the following useful bound on the difference of distribution functions:

Theorem 2.1.3 (Berry-Esséen)**.** In the setting of Theorem 2.1.2 with $c = 0$ and $\sigma^2 = 1$, we have for any $t \geq 0$ that

$$\left|\mathbb{P}\left[\frac{1}{\sqrt{n}} S_n \leq t\right] - \mathbb{P}[N \leq t]\right| \leq (1 + \mathbb{E}|X_1|^3)\frac{1}{\sqrt{n}}.$$

The basic idea is to approximate the above indicator function $f(x)$ by a smooth function and handle the resulting error.

Note also that the source of the $1/n^{1/2}$ error rate in n is the Taylor expansion: that we could only match the first two terms of the expansion led to an error of $(1/n^{1/2})^3$, and n such errors led to a bound of $n \cdot (1/n^{1/2})^3 = n^{1-3/2} = n^{-1/2}$. If we have $\mathbb{E}X_i^k = \mathbb{E}N^k = \mathbb{E}Y_i^k$ for all $0 \leq k \leq \ell$ for some $\ell \geq 3$, then we may carry out the same argument with a Taylor expansion of order ℓ, and expect to get an error of $n \cdot (1/n^{1/2})^\ell = n^{-(\ell-1)/2}$. Indeed this can be done, both in the ordinary and Berry-Esséen CLTs:

Theorem 2.1.4 (Moment matching CLTs)**.** Suppose in the setting of Theorem 2.1.2 with $c = 0$ and $\sigma^2 = 1$ that $\mathbb{E}X_i^k = \mathbb{E}N^k$ for all $0 \leq k \leq \ell$ and $R := \mathbb{E}|X_i|^{k+1} < \infty$. Then, for any smooth and compactly supported $f : \mathbb{R} \to \mathbb{R}$, for a constant $C = C(f, R)$, we have

$$\left|\mathbb{E}f\left(\frac{1}{\sqrt{n}}\right) - \mathbb{E}f(N)\right| \leq \frac{C}{n^{(\ell-1)/2}}.$$

Further, for another constant $C' = C'(R)$, for any $t \geq 0$, we also have

$$\left|\mathbb{P}\left[\frac{1}{\sqrt{n}} S_n \leq t\right] - \mathbb{P}[N \leq t]\right| \leq \frac{C'}{n^{(\ell-1)/2}}.$$

Good resources for learning more about these methods include the recent very short paper [Ver26] as well as the more comprehensive Chapter 11 of [O'D14].

2.2 Moment Method

The next method for weak convergence we consider is based instead on *polynomial* test functions. While for technical reasons this is not how we will actually prove the associated continuity theorem below, the idea in the spirit of the above arguments is that results like the Weierstrass approximation theorem imply that polynomials can, on bounded intervals, approximate continuous functions arbitrarily well.

2.2.1 Continuity Theorem

Again, we first prove a general result showing that it suffices to establish convergence of moments along with a certain regularity condition in order to show weak convergence.

Lemma 2.2.1 (Weak convergence by moments)**.** Suppose that $X_1, X_2, \dots, X_\infty \in \mathbb{R}$ are random variables with all moments finite and satisfying the following:

1. $\mathbb{E}X_n^k \to \mathbb{E}X_\infty^k$ for all $k \geq 0$ (note that this also implies by linearity that $\mathbb{E}p(X_n) \to \mathbb{E}p(X_\infty)$ for all polynomials $p(x)$).
2. $\mathbb{E}|X_\infty|^k \leq (\epsilon(k) \cdot k)^k$ for all $k \geq 0$, for some function $\epsilon : \mathbb{Z}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\epsilon(k) \to 0$ as $k \to \infty$.

Then, $X_n \Rightarrow X_\infty$.

Example 2.2.2. It is clear that Condition 2 of the Lemma holds for any compactly supported $\mathsf{Law}(X_\infty)$. Further, we will see below results implying that it also holds for X_∞ that is either Gaussian or Poisson (with any parameters).

Proof of Lemma 2.2.1. We will show that the associated characteristic functions have the convergence $\phi_{X_n} \to \phi_{X_\infty}$, which completes the proof by Lévy's continuity theorem. Fix some $k \geq 1$. We then expand the characteristic functions in Taylor series of order k. For X_∞, we have,[1] using that $\phi_{X_\infty}(t) = \mathbb{E}\exp(itX_\infty)$ and taking a Taylor expansion,

$$\left| \phi_{X_\infty}(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} \mathbb{E}X_\infty^j \right| \leq \frac{|t|^{k+1}}{(k+1)!} \mathbb{E}|X_\infty|^{k+1}$$

and now using our assumption about the moments of X_∞ and a standard bound on the factorial,

$$\begin{aligned} &\leq \frac{|t|^{k+1}}{\left(\frac{k+1}{e}\right)^{k+1}} (\epsilon(k+1) \cdot (k+1))^{(k+1)} \\ &= (e|t| \cdot \epsilon(k+1))^{k+1}. \end{aligned}$$

Also by assumption, for sufficiently large n we have, say, $\mathbb{E}|X_n|^{k+1} \leq 2\mathbb{E}|X_\infty|^{k+1}$. Thus, repeating the same argument, we have for such n that

$$\begin{aligned} \left| \phi_{X_n}(t) - \sum_{j=0}^{k} \frac{(it)^j}{j!} \mathbb{E}X_n^j \right| &\leq \frac{|t|^{k+1}}{(k+1)!} \mathbb{E}|X_n|^{k+1} \\ &\leq 2\frac{|t|^{k+1}}{(k+1)!} \mathbb{E}|X_\infty|^{k+1} \\ &\leq 2\,(e|t| \cdot \epsilon(k+1))^{k+1}. \end{aligned}$$

Thus, again using the convergence of moments,

$$\limsup_{n \to \infty} \left| \phi_{X_n}(t) - \phi_{X_\infty}(t) \right| \leq 3\,(e|t| \cdot \epsilon(k+1))^{k+1},$$

and by taking $k \to \infty$ we find $\phi_{X_n}(t) \to \phi_{X_\infty}(t)$, for all $t \in \mathbb{R}$. The result then follows by Lévy's continuity theorem. □

[1] See [Kle14, Lemma 15.31] for this estimate.

We will not prove the following result, but mention it since it gives the optimal statement of this kind. See [Bil17, Section 30] for details.

Definition 2.2.3. A probability measure μ on $\mathbb{R}$ is *moment-determinate* if all of its moments are finite and, whenever $\int x^k \, d\nu = \int x^k \, d\mu$ for another probability measure ν, then $\nu = \mu$.

Theorem 2.2.4. Suppose that $X_1, X_2, \dots, X_\infty \in \mathbb{R}$ are random variables with all moments finite and satisfying the following:

1. $\mathbb{E}X_n^k \to \mathbb{E}X_\infty^k$ for all $k \geq 0$ (note that this also implies by linearity that $\mathbb{E}p(X_n) \to \mathbb{E}p(X_\infty)$ for all polynomials $p(x)$).
2. $\mathsf{Law}(X_\infty)$ is moment-determinate.

Then, $X_n \Rightarrow X_\infty$.

Note that if X is a random variable and $X_1, X_2, \dots$ are each equal to X, then $X_n \Rightarrow X$. Since a given sequence can only converge in distribution to a random variable with a unique law, any condition we substitute in place of Condition 2 above would have to imply moment-determinacy. In that sense, the above result is optimal.

2.2.2 Proof of Central Limit Theorem

We will prove the following even more weakened version of the CLT here; however, by the standard truncation argument the boundedness condition that appears here is actually without loss of generality (after carrying out some initial bounds).

Theorem 2.2.5 (Very weak CLT)**.** Let $X_1, X_2, \dots \sim \mu$ be i.i.d. with $\mathbb{E}X_i = c$, $\mathsf{Var}\, X_i = \sigma^2$, and $|X_i| < R$ almost surely. Write $S_n := \sum_{i=1}^n (X_i - c)$ and let $N \sim \mathcal{N}(0,1)$. Then, $\frac{1}{\sqrt{n}} S_n \Rightarrow N$.

To prove this by the moment method, we must establish first that the Gaussian distribution satisfies the regularity Condition 2 of Lemma 2.2.1, and second that the moments of $\frac{1}{\sqrt{n}} S_n$ converge to those of N. You may show as an exercise the following characterization of the moments of a standard Gaussian distribution, which we need for both of the above tasks.

Definition 2.2.6 (Partitions and matchings)**.** Let S be a finite set. A *partition* of S is a collection of disjoint, non-empty $S_1, \dots, S_m \subseteq S$ such that $S = S_1 \sqcup \cdots \sqcup S_m$. A *matching* of S is a partition all of whose parts have size 2. We write $\mathsf{Part}(S)$ for the set of partitions of S and $\mathsf{Match}(S)$ for the set of matchings of S. Lastly, we define $M_k := |\mathsf{Match}([k])|$ and $P_k := |\mathsf{Part}([k])|$.

Proposition 2.2.7. Let $N \sim \mathcal{N}(0,1)$. Then, for all $k \geq 1$,

$$\mathbb{E}N^k = M_k = \#\{\text{matchings of } [k]\} = \begin{cases} 0 & \text{if } k \text{ is odd,} \\ (k-1)!! & \text{if } k \text{ is even.} \end{cases}$$

Here we use the notation $(k-1)!! = (k-1)(k-3)\cdots 1$.

In particular, we have $|\mathbb{E}N^k| \leq k^{k/2}$ for all $k \geq 1$, so Lemma 2.2.1 indeed applies to the limiting distribution $\mathcal{N}(0,1)$.

Proof of Theorem 2.2.5. It remains to show that $\mathbb{E}(\frac{1}{\sqrt{n}}S_n)^k \to M_k$ for all $k \geq 0$. We first expand the left-hand side of the convergence, finding

$$\mathbb{E}\left(\frac{1}{\sqrt{n}}S_n\right)^k = \frac{1}{n^{k/2}} \sum_{i_1,\dots,i_k=1}^{n} \mathbb{E}X_{i_1}\cdots X_{i_k}.$$

Let us first work on a single term of this sum. Suppose $\boldsymbol{i} = (i_1,\dots,i_k) \in [n]^k$ is fixed. Define the vector of "frequencies", $f_\alpha = \#\{a \in [k] : i_a = \alpha\}$ for $\alpha \in [n]$. Then, we have by independence of the X_i that

$$E_{\boldsymbol{i}} := \mathbb{E}X_{i_1}\cdots X_{i_k} = \mathbb{E}X_1^{f_1}\cdots X_n^{f_n} = (\mathbb{E}X_1^{f_1})\cdots(\mathbb{E}X_n^{f_n}).$$

We make a few preliminary observations about these quantities:

1. If any $f_\alpha = 1$, then $E_{\boldsymbol{i}} = 0$ since $\mathbb{E}X_i = 0$.
2. If all $f_\alpha \in \{0,2\}$, then $E_{\boldsymbol{i}} = 1$ since $\mathbb{E}X_i^2 = 1$.
3. For all $\boldsymbol{i}$, we have $|E_{\boldsymbol{i}}| \leq R^k$ since $|X_i| \leq R$ almost surely.

Note that the number of terms $E_{\boldsymbol{i}}$ with $f_\alpha \in \{0,2\}$ for all $\alpha \in [n]$ is precisely $M_k \cdot n(n-1)\cdots(n-k/2+1)$, the number of matchings of $[k]$ times the number of ways to label each matched pair with a different index. And, the number of terms $E_{\boldsymbol{i}}$ with all $f_\alpha \neq 1$ and *some* $f_\alpha \geq 3$ is at most $P_k \cdot n^{1+(k-3)/2}$: each such term corresponds to a partition of $[k]$ into parts of size at least 2 and one of which has size at least 3, with parts labelled by indices in $[n]$. The total number of parts is then at most $1+(k-3)/2$, and the number of such partitions is at most P_k, the total number of partitions. Thus, in total we find that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{\sqrt{n}}S_n\right)^k &= \frac{1}{n^{k/2}} \sum_{\boldsymbol{i}\in[n]^K} E_{\boldsymbol{i}} \\ &= M_k \frac{n(n-1)\cdots(n-k/2+1)}{n^{k/2}} + O\left(P_k R^k \frac{n^{1-(k-3)/2}}{n^{k/2}}\right) \\ &= M_k(1+o_k(1)) + O_k\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Thus, we find

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{1}{\sqrt{n}}S_n\right)^k = M_k = \mathbb{E}N^k$$

for all $k \geq 0$, completing the proof. □

2.2.3 Poisson Limit Theorems for Rare Events

As another application of the moment method, we study Poisson limits of integer-valued random variables. Recall that $X \sim \mathsf{Pois}(\lambda)$ is $X \in \mathbb{Z}_{\geq 0}$ with probability mass function

$$\mathbb{P}[X = t] = \exp(-t)\frac{\lambda^t}{t!} \text{ for each } t \in \mathbb{Z}_{\geq 0}.$$

We first develop some general tools for proving convergence to these distributions.

As we will see, in this case it turns out to be useful to use a slightly different polynomial basis than before.

Definition 2.2.8 (Falling factorial polynomials)**.** For any $k \in \mathbb{Z}_{\geq 0}$ and $x \in \mathbb{R}$, define

$$x^{\underline{k}} := x(x-1)\cdots(x-k+1).$$

On the one hand, this is a polynomial in $\mathbb{R}[x]$ of degree k. On the other hand, if $x \in \mathbb{Z}_{\geq 0}$ then we also have the following interpretation.

Proposition 2.2.9. If $x \in \mathbb{Z}_{\geq 0}$, then $x^{\underline{k}}$ is the number of ways to assign labels from $[x]$ to k objects, or equivalently

$$x^{\underline{k}} = \begin{cases} 0 & \text{if } x \leq k-1, \\ x!/(x-k)! & \text{if } x \geq k. \end{cases} \tag{2.2.1}$$

Because these polynomials generate all of the polynomials of $\mathbb{R}[x]$, and in particular the standard monomials x^k are clearly linear combinations of the $1 = x^{\underline{0}}, x^{\underline{1}}, \dots, x^{\underline{k}}$, we have the following.

Proposition 2.2.10. If $\mathbb{E}X_n^{\underline{k}} \to \mathbb{E}X_\infty^{\underline{k}}$ for all $k \geq 0$, then also $\mathbb{E}X_n^k \to \mathbb{E}X_n^\infty$ for all $k \geq 0$.

The following identity, which you may prove as an exercise, shows why the falling factorial polynomials are especially well-suited to working with Poisson random variables.

Proposition 2.2.11. If $X \sim \mathsf{Pois}(\lambda)$, then $\mathbb{E}X^{\underline{k}} = \lambda^k$.

From this you may also show together with some combinatorics that $\mathsf{Pois}(\lambda)$ indeed satisfies Condition 2 of Lemma 2.2.1, and thus that we are justified in using the moment method for convergence to a Poisson distribution. As a result, we find:

Corollary 2.2.12. Suppose that $X_n \in \mathbb{Z}_{\geq 0}$ are random variables such that $\mathbb{E}X_n^{\underline{k}} \to \lambda^k$ for all $k \geq 0$, for some $\lambda > 0$. Then, $X_n \Rightarrow X_\infty \sim \mathsf{Pois}(\lambda)$.

Finally, let us describe the quantities $\mathbb{E}X^{\underline{k}}$ for the kinds of random variables X that one often encounters in such statements. We consider *counting variables* X, ones that count how many of several possible events occurred. Formally, we just view X as a sum of Boolean variables, whose probabilistic interpretation will be as the indicator functions of events. The following is just a deterministic fact about the falling factorial.

Proposition 2.2.13. Suppose that $F_1, \dots, F_m \in \{0, 1\}$, and $X = \sum_{i=1}^m F_i$. Then,

$$X^{\underline{k}} = \sum_{\substack{(i_1,\dots,i_k) \in [m] \\ \text{all } i_a \text{ distinct}}} F_{i_1} \cdots F_{i_k}.$$

Proof. Since $X \in \mathbb{Z}_{\geq 0}$, by (2.2.1) we have that $X^{\underline{k}}$ has the combinatorial interpretation of the number of ways to assign distinct labels from $[X]$ to k objects. The right-hand side above clearly counts the same thing. □

Corollary 2.2.14. Suppose that $E_1, \dots, E_m$ are events, and $X = \sum_{i=1}^m \mathbb{1}_{E_i}$, the (random) number of the E_i that occur. Then,

$$\mathbb{E}X^{\underline{k}} = \sum_{\substack{(i_1,\dots,i_k) \in [m] \\ \text{all } i_a \text{ distinct}}} \mathbb{P}[E_{i_1} \cap \cdots \cap E_{i_k}].$$

Thus we have established a toolkit for proving Poisson limit theorems for random variables counting numbers of rare events that occur that merely asks us to evaluate expressions of the above kind. Let us describe when we expect to be able to do this. Suppose that $X_n = \sum_{i=1}^m \mathbb{1}_{E_i^{(n)}}$ for some events $E_1^{(n)}, \dots, E_m^{(n)}$ for each $n \geq 1$ and some $m = m(n)$. We roughly speaking expect to have $X_n \Rightarrow X_\infty \sim \mathsf{Pois}(\lambda)$ provided that:

1. $\mathbb{E}X_n = \sum_i \mathbb{P}[E_i^{(n)}] \to \lambda$.
2. $\mathsf{max}_i\, \mathbb{P}[E_i^{(n)}] \ll 1$.
3. $E_i^{(n)}$ are only weakly dependent.

That is because, in this case, we should be able to argue, starting from Corollary 2.2.14, that

$$\begin{aligned}
\mathbb{E}X_n^k &= \sum_{\substack{(i_1,\dots,i_k)\in[m] \\ \text{all } i_a \text{ distinct}}} \mathbb{P}[E_{i_1}^{(n)} \cap \cdots \cap E_{i_k}^{(n)}] \\
&\approx \sum_{\substack{(i_1,\dots,i_k)\in[m] \\ \text{all } i_a \text{ distinct}}} \mathbb{P}[E_{i_1}^{(n)}] \cdots \mathbb{P}[E_{i_k}^{(n)}] \\
&\approx \sum_{(i_1,\dots,i_k)\in[m]} \mathbb{P}[E_{i_1}^{(n)}] \cdots \mathbb{P}[E_{i_k}^{(n)}] \\
&= \left(\sum_{i=1}^{m} \mathbb{P}[E_i^{(n)}]\right)^k \\
&\approx \lambda^k,
\end{aligned}$$

as desired.

We will now see several examples where it is indeed possible to carry out such an argument precisely. These also give us a chance, after recalling the basic Poisson limit theorem, to mention two other important models from discrete probability. For a much deeper treatment of these ideas, a great reference is the book [Ald13].

2.2.4 Proof of Classical Poisson Limit Theorem

The simplest case of such Poisson limit theorems is the following result that you are probably familiar with already.

Theorem 2.2.15 (Poisson limit theorem)**.** For any $\lambda > 0$, $\mathsf{Bin}(n, \frac{\lambda}{n}) \xrightarrow{\text{(w)}} \mathsf{Pois}(\lambda)$ as $n \to \infty$.

Proof. For each n, let $F_1^{(n)}, \dots, F_n^{(n)} \sim \mathsf{Ber}(\frac{\lambda}{n})$ be i.i.d. Write $X_n := \sum_{i=1}^n F_i^{(n)}$, then we have $\mathsf{Law}(X_n) = \mathsf{Ber}(n, \frac{\lambda}{n})$.

By Corollary 2.2.12, it suffices to show $\mathbb{E}X_n^k \to \lambda^k$ for each $k \geq 0$. By Proposition 2.2.13, we may expand this expression as

$$\mathbb{E}X_n^{\underline{k}} = \sum_{\substack{(i_1,\dots,i_k)\in[n] \\ \text{all } i_a \text{ distinct}}} \mathbb{E}F_{i_1} \cdots F_{i_k}$$

if $k > n$, then the sum is empty and this value is zero. Otherwise, by independence we may factorize each term of the sum as

$$\begin{aligned}
&= \mathbb{1}\{k \leq n\} \sum_{\substack{(i_1,\dots,i_k)\in[n] \\ \text{all } i_a \text{ distinct}}} (\mathbb{E}F_{i_1}) \cdots (\mathbb{E}F_{i_k}) \\
&= \mathbb{1}\{k \leq n\} \cdot n^{\underline{k}} \cdot \left(\frac{\lambda}{n}\right)^k
\end{aligned}$$

and so as $n \to \infty$ for any fixed $k \geq 0$ we have

$$\to \lambda^k,$$

completing the proof. □

2.2.5 Application: Fixed Points of Random Permutations

Let $\mathsf{Sym}(n) := \{\sigma : [n] \to [n] \text{ bijective}\}$, the set of permutations. By elementary combinatorics, $|\mathsf{Sym}(n)| = n!$. We study the random variable $\sigma = \sigma^{(n)} \sim \mathsf{Unif}(\mathsf{Sym}(n))$. In particular, consider the number of *fixed points*:

$$\begin{aligned} E_i^{(n)} &:= \{\sigma^{(n)}(i) = i\}, \\ X_n &:= \sum_{i=1}^{n} \mathbb{1}_{E_i^{(n)}} \\ &= \#\{i \in [n] : \sigma^{(n)}(i) = i\}. \end{aligned}$$

Note that we have by symmetry

$$\mathbb{E}X_n = \sum_{i=1}^{n} \mathbb{P}[E_i^{(n)}] = n \cdot \frac{1}{n} = 1.$$

Thus, the following is not surprising.

Theorem 2.2.16. $X_n \Rightarrow X_\infty \sim \mathsf{Pois}(1)$.

Proof. As before, we consider the falling factorial moments. We have by Corollary 2.2.14:

$$\mathbb{E}X_n^{\underline{k}} = \sum_{\substack{(i_1,\dots,i_k)\in[n] \\ \text{all } i_a \text{ distinct}}} \mathbb{P}[E_{i_1}^{(n)} \cap \cdots \cap E_{i_k}^{(n)}]$$

where the sum is empty if $k > n$, and otherwise

$$= \mathbb{1}\{k \le n\} \sum_{\substack{(i_1,\dots,i_k)\in[n] \\ \text{all } i_a \text{ distinct}}} \mathbb{P}[\sigma(i_1) = i_1, \dots, \sigma(i_k) = i_k]$$

and we may enumerate the number of such σ as $(n-k)!$, the number of permutations of the points other than those fixed by the above condition. Thus,

$$\begin{aligned} &= \mathbb{1}\{k \le n\} \sum_{\substack{(i_1,\dots,i_k)\in[n] \\ \text{all } i_a \text{ distinct}}} \frac{(n-k)!}{n!} \\ &= \mathbb{1}\{k \le n\} \cdot \frac{n!}{(n-k)!} \cdot \frac{(n-k)!}{n!} \\ &= \mathbb{1}\{k \le n\}. \end{aligned}$$

In particular then, for any fixed $k \ge 0$, as $n \to \infty$ we have $\mathbb{E}X_n^{\underline{k}} \to 1$, and so by Corollary 2.2.12 the proof is complete. □

2.2.6 Application: Motifs in Random Graphs

The following is the foundational model of random graph theory.

Definition 2.2.17 (Erdős-Rényi random graph)**.** $\mathcal{G}(n,p)$ is the law of a random graph G on vertex set $[n]$, where every pair of vertices $1 \leq i < j \leq n$ are connected independently with probability p. We write $i \sim_G j$ for the relation of adjacency in the graph G, or just $i \sim j$ when the graph G is clear from context.

Note that the expected *degree* of any given vertex in such a graph is $p(n-1) \approx pn$. Here we focus on the *sparse* regime of Erdős-Rényi graphs, where this number is of constant order. To achieve this, we take $p = \frac{c}{n}$ for some constant $c > 0$ not changing with n.

One may study in general the number of various *motifs* or subgraph occurrences in $G \sim \mathcal{G}(n,p)$. We take the following simple case:

$$X_n := \#\{\text{triangles in G}\} = \#\{1 \leq i < j < k \leq n : i \sim j, i \sim k, j \sim k\}.$$

We see that this case is a little different from the previous ones, in that the number of events m we are study is not just n. Let $m = \binom{n}{3}$, which we may identify with the number of ways to choose $1 \leq i < j < k \leq n$. Write $\binom{[n]}{3}$ for the set of such subsets $\{i,j,k\}$. Then, writing $E_{\{i,j,k\}} = \{i \sim j, i \sim k, j \sim k\}$, we have

$$X_n = \sum_{\{i,j,k\} \in \binom{[n]}{3}} E_{\{i,j,k\}}.$$

In particular, we have

$$\mathbb{E}X_n = \sum_{\{i,j,k\} \in \binom{[n]}{3}} \left(\frac{c}{n}\right)^3 = \binom{n}{3} \cdot \left(\frac{c}{n}\right)^3 \to \frac{c^3}{6}.$$

Thus, we are not surprise to find:

Theorem 2.2.18. $X_n \Rightarrow X_\infty \sim \mathsf{Pois}(\frac{c^3}{6})$.

We do not give a detailed proof here, but the basic idea is also a little different from the application to random permutations. Here, it is not the case that all of the $E_{\{i,j,k\}}$ are slightly dependent: most collections—for those triangles that do not share any edges—are independent, while some collections—for triangles that do share an edge—are considerably dependent. Thus, the proof amounts to showing that most collections of k triangles are edge-disjoint, a straightforward combinatorial fact.

You may consult the survey [Wor99] for extensive further discussion of applications of such moment methods to random graph theory, including in models with more complicated dependencies than the simple Erdős-Rényi graph we have considered here.

2.3 Vector Limit Theorems

We now consider how one can prove limit theorems for random *vectors* $\boldsymbol{X}_n \in \mathbb{R}^d$. We first develop the following very useful general tool.

2.3.1 Cramér-Wold Device

Let us first recall a few aspects of the theory of characteristic functions for these.

Definition 2.3.1. For $\boldsymbol{X} \in \mathbb{R}^d$ random and $\boldsymbol{t} \in \mathbb{R}^d$, we define the *characteristic function*

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) := \mathbb{E}\exp(i\langle \boldsymbol{t}, \boldsymbol{X}\rangle),$$

for the standard scalar inner product $\langle \boldsymbol{t}, \boldsymbol{X}\rangle = \sum_{i=1}^d t_i X_i$.

These multivariate characteristic functions share many important properties with their univariate versions, the main ones being as follows.

Proposition 2.3.2. The following hold:

- For $\boldsymbol{X}, \boldsymbol{Y}$ random vectors in $\mathbb{R}^d$, $\mathsf{Law}(\boldsymbol{X}) = \mathsf{Law}(\boldsymbol{Y})$ if and only if $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$.
- For $\boldsymbol{X}_n, \boldsymbol{X}_\infty$ random vectors in $\mathbb{R}^d$, $\boldsymbol{X}_n \Rightarrow \boldsymbol{X}_\infty$ if and only if $\phi_{\boldsymbol{X}_n}(\boldsymbol{t}) \to \phi_{\boldsymbol{X}_\infty}(\boldsymbol{t})$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

Note here that $\boldsymbol{X}_n \Rightarrow \boldsymbol{X}_\infty$ means that $\mathbb{E}f(\boldsymbol{X}_n) \to \mathbb{E}f(\boldsymbol{X}_\infty)$ for all bounded continuous $f : \mathbb{R}^d \to \mathbb{R}$, the same exact definition as for real-valued random variables. This generality is an important advantage of defining weak convergence (or convergence in distribution) in terms of test functions rather than objects like distribution functions, though the latter may appears more concrete at first.

Now, we may note that the characterizations in terms of characteristic functions can be viewed as families of properties of the *scalar* random variables $\langle \boldsymbol{t}, \boldsymbol{X}\rangle$ independently (and similarly for $\boldsymbol{Y}$, $\boldsymbol{X}_n$, and $\boldsymbol{X}_\infty$ in each statement above). Thus, through the intermediate stop of characteristic functions we may obtain a more conceptual statement that relates multivariate laws and convergence in distribution to that of these scalar projections:

Corollary 2.3.3 (Cramér-Wold)**.** The following hold in the contexts of Proposition 2.3.2:

- $\mathsf{Law}(\boldsymbol{X}) = \mathsf{Law}(\boldsymbol{Y})$ if and only if $\mathsf{Law}(\langle \boldsymbol{t}, \boldsymbol{X}\rangle) = \mathsf{Law}(\langle \boldsymbol{t}, \boldsymbol{Y}\rangle)$ for all $\boldsymbol{t} \in \mathbb{R}^d$.
- $\boldsymbol{X}_n \Rightarrow \boldsymbol{X}_\infty$ if and only if $\langle \boldsymbol{t}, \boldsymbol{X}_n\rangle \Rightarrow \langle \boldsymbol{t}, \boldsymbol{X}_\infty\rangle$ for all $\boldsymbol{t} \in \mathbb{R}^d$.

2.3.2 Multivariate Central Limit Theorem

As a first application, let us prove a multivariate version of the central limit theorem.

Definition 2.3.4 (Covariance matrix)**.** For a random $\boldsymbol{X} \in \mathbb{R}^d$, we define its *covariance matrix* to be

$$\mathsf{Cov}[\boldsymbol{X}] := \mathbb{E}(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\top = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - (\mathbb{E}\boldsymbol{X})(\mathbb{E}\boldsymbol{X})^\top \in \mathbb{R}^{d\times d}.$$

Here we use linearity of expectation on expressions involving matrix algebra, which you may check as an exercise is valid.

Let us establish a few preliminaries. First, in general, the covariance matrix characterizes the variances of scalar projections of a random vector.

Proposition 2.3.5. $\mathsf{Var}[\langle \boldsymbol{t}, \boldsymbol{X}\rangle] = \boldsymbol{t}^\top \mathsf{Cov}[\boldsymbol{X}]\boldsymbol{t}^\top$.

Proof. Expanding from the definition of the covariance matrix, we immediately have that $\boldsymbol{t}^\top \mathsf{Cov}[\boldsymbol{X}]\boldsymbol{t}^\top = \mathbb{E}(\langle \boldsymbol{t}, \boldsymbol{X}\rangle - \mathbb{E}\langle \boldsymbol{t}, \boldsymbol{X}\rangle)^2$, the definition of the variance. □

Also, we recall the following property of Gaussian random vectors.

Proposition 2.3.6. If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathsf{Law}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) = \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$. In particular, $\mathsf{Law}(\langle \boldsymbol{t}, \boldsymbol{X}\rangle) = \mathcal{N}(\langle \boldsymbol{\mu}, \boldsymbol{t}\rangle, \boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t})$.

We now state and prove the multivariate central limit theorem.

Theorem 2.3.7 (Multivariate CLT)**.** Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots \in \mathbb{R}^d$ be i.i.d. with $\mu = \mathbb{E}\boldsymbol{X}_i$ and $\boldsymbol{\Sigma} = \mathsf{Cov}[\boldsymbol{X}_i]$, with both of these expectations existing. Then,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}) \Rightarrow \boldsymbol{X}_\infty \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Proof. Without loss of generality we may assume $\boldsymbol{\mu} = \boldsymbol{0}$. By Corollary 2.3.3, it suffices to show that, for any $\boldsymbol{t} \in \mathbb{R}^d$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle \boldsymbol{t}, \boldsymbol{X}_k\rangle \Rightarrow \langle \boldsymbol{t}, \boldsymbol{X}_\infty\rangle$. We have that the $\langle \boldsymbol{t}, \boldsymbol{X}_k\rangle$ are i.i.d. with $\mathbb{E}\langle \boldsymbol{t}, \boldsymbol{X}_k\rangle = 0$ and, by Proposition 2.3.5, $\mathsf{Var}[\langle \boldsymbol{t}, \boldsymbol{X}_k\rangle] = \boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}$. On the other hand, by Proposition 2.3.6, we have $\mathsf{Law}(\langle \boldsymbol{t}, \boldsymbol{X}_\infty\rangle) = \mathcal{N}(0, \boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t})$. Thus, this one-dimensional convergence holds by the ordinary central limit theorem, completing the proof. □

2.3.3 Application: χ^2 Test and Statistic

We show as an application how the multivariate central limit theorem can be used to understand a subtle point in statistics that you may have come across. This concerns the following important probability measure.

Definition 2.3.8 (Multinomial distribution)**.** Let $p_1, \dots, p_k > 0$ have $\sum_{i=1}^{k} p_i = 1$. We write $\mathsf{Mult}(n; p_1, \dots, p_k)$ for the law of $\boldsymbol{X} \in \mathbb{Z}_{\geq 0}^k$ where X_i is the number of balls in bin i when n balls are thrown independently at random into bin i with probability p_i.

We note that, if $\boldsymbol{X} \sim \mathsf{Mult}(n; p_1, \dots, p_k)$, then $\mathsf{Law}(X_i) = \mathsf{Bin}(n, p_i)$, so in particular $\mathbb{E}X_i = p_i n$ and $\mathsf{Var}\, X_i = p_i(1 - p_i)n$.

The Pearson χ^2 test of classical statistics involves hypothesis testing whether the outcomes of several trials of an experiment with k possible results arose from the null hypothesis of the multinomial distribution. It proposes to do this by considering the test statistic

$$S(\boldsymbol{X}) := \sum_{i=1}^{k} \left(\frac{X_i - p_i n}{\sqrt{p_i n}} \right)^2 .$$

To carry out such a hypothesis test in the style of asypmtotic statistics, we must understand the asymptotic distribution of $S(\boldsymbol{X})$, which is described by the following foundational result.

Theorem 2.3.9 (Pearson). Let $Z_1, \dots, Z_{k-1} \sim \mathcal{N}(0, 1)$ be independent. Then, in the above setting, $S(\boldsymbol{X}) \Rightarrow Z_1^2 + \cdots + Z_{k-1}^2$ as $n \to \infty$.

The right-hand side has the χ^2 distribution with $k - 1$ degrees of freedom, a rather surprising result since $S(\boldsymbol{X})$ involves a sum of k squares. We will see in the proof how the multivariate CLT elucidates this matter.

Proof. Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_k$ be the standard basis of $\mathbb{R}^k$ and write μ for the probability measure drawing $\boldsymbol{e}_i$ with probability p_i. Then, if $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n \sim \mu$ are i.i.d., we have $\mathsf{Law}(\boldsymbol{v}_1 + \cdots + \boldsymbol{v}_n) = \mathsf{Mult}(n; p_1, \dots, p_k)$ ($\boldsymbol{v}_i$ may be viewed as the indicator vector of the destination bin of the ith ball in our above description of the multinomial distribution). Thus, the multinomial distribution is merely a sum of i.i.d. random vectors with a particular distribution.

Let us apply the multivariate CLT to $\boldsymbol{X}^{(n)} = \sum_{i=1}^{n} \boldsymbol{v}_i$. To do this, we calculate the statistics of the v_i. Let us introduce $\boldsymbol{p} = (p_1, \dots, p_k) \in \mathbb{R}^k$. We have:

$$\begin{aligned}
\mathbb{E}\boldsymbol{v}_i &= p_1 \boldsymbol{e}_1 + \cdots + p_k \boldsymbol{e}_k \\
&= \boldsymbol{p}, \\
\mathsf{Cov}\, \boldsymbol{v}_i &= \mathbb{E}\boldsymbol{v}_i \boldsymbol{v}_i^\top - (\mathbb{E}\boldsymbol{v}_i)(\mathbb{E}\boldsymbol{v}_i)^\top \\
&= p_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + \cdots + p_k \boldsymbol{e}_k \boldsymbol{e}_k^\top - \boldsymbol{p}\boldsymbol{p}^\top \\
&= \mathsf{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top \\
&=: \boldsymbol{\Sigma}.
\end{aligned}$$

By the CLT, we then have

$$\frac{1}{\sqrt{n}} (\boldsymbol{X}^{(n)} - n\boldsymbol{p}) \Rightarrow \boldsymbol{X}^{(\infty)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Thus, we also have

$$S(\boldsymbol{X}_n) = \sum_{i=1}^{k} \left(\frac{X_i^{(n)} - p_i n}{\sqrt{p_i n}} \right)^2 \Rightarrow \sum_{i=1}^{k} \left(\frac{X_i^{(\infty)}}{\sqrt{p_i}} \right)^2 = \|\mathsf{Diag}(\boldsymbol{p})^{-1/2} \boldsymbol{X}^{(\infty)}\|_2^2.$$

A Gaussian calculation gives:

$$\begin{aligned}\mathsf{Law}(\mathsf{Diag}(\boldsymbol{p})^{-1/2}\boldsymbol{X}^{(\infty)}) &= \mathcal{N}(\boldsymbol{0}, \mathsf{Diag}(\boldsymbol{p})^{-1/2}\Sigma\mathsf{Diag}(\boldsymbol{p})^{-1/2}) \\ &= \mathcal{N}(\boldsymbol{0}, \mathsf{Diag}(\boldsymbol{p})^{-1/2}(\mathsf{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\mathsf{Diag}(\boldsymbol{p})^{-1/2}) \\ &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k - \boldsymbol{q}\boldsymbol{q}^\top),\end{aligned}$$

where $q_i = \sqrt{p_i}$. Note that this vector has $\|\boldsymbol{q}\|_2 = \sum_{i=1}^k p_i = 1$, a unit vector. In particular then, $\boldsymbol{I}_k - \boldsymbol{q}\boldsymbol{q}^\top$ is a projection matrix to a $(k-1)$-dimensional subspace of $\mathbb{R}^k$.

Let us write $\boldsymbol{g} := \mathsf{Diag}(\boldsymbol{p})^{-1/2}\boldsymbol{X}^{(\infty)}$, which we have established has law $\mathsf{Law}(\boldsymbol{g}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k - \boldsymbol{q}\boldsymbol{q}^\top)$. Now, for any orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(k)$, we have $\|\boldsymbol{g}\|_2^2 = \|\boldsymbol{Q}\boldsymbol{g}\|_2^2$. On the other hand,

$$\mathsf{Law}(\boldsymbol{Q}\boldsymbol{g}) = \mathcal{N}(0, \boldsymbol{Q}(\boldsymbol{I}_k - \boldsymbol{q}\boldsymbol{q}^\top)\boldsymbol{Q}^\top) = \mathcal{N}(0, \boldsymbol{I}_k - (\boldsymbol{Q}\boldsymbol{q})(\boldsymbol{Q}\boldsymbol{q})^\top).$$

Choosing $\boldsymbol{Q}$ appropriately, we may in particular arrange to have $\boldsymbol{Q}\boldsymbol{q} = \boldsymbol{e}_k$. Thus, we have that $\mathsf{Law}(\|\boldsymbol{g}\|_2^2) = \mathsf{Law}(\|\boldsymbol{h}\|_2^2)$ for $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k - \boldsymbol{e}_k\boldsymbol{e}_k^\top)$. In particular, $\|\boldsymbol{h}\|_2^2 = \sum_{i=1}^{k-1} h_i^2$, where $h_1, \dots, h_{k-1} \sim \mathcal{N}(0,1)$ are i.i.d. The above establishes that $S(\boldsymbol{X}_n) \Rightarrow \|\boldsymbol{h}\|_2^2$, completing the proof. □

3 | Conditional Expectation

3.1 Motivation

Our next goal will be to develop a general measure-theoretic framework for *conditional* probability and expectation. The intuitive idea behind these constructions is to perform probabilistic reasoning where certain random variables are "fixed" to given values while others still fluctuate randomly, or where certain events are "fixed" to occur. Let us first review a few elementary constructions of conditional probability and expectation that you have probably seen before, which we will try to unify and generalize in our abstract formulation.

Example 1: Discrete Random Variables Suppose that $(X, Y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2$ only take on finitely many values (we could allow countably many just as well). Their joint distribution is then specified by the probability mass function,

$$p(x, y) := \mathbb{P}[X = x, Y = y].$$

Let us assume for the sake of simplicity that $p(x, y) > 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In this case, the elementary notion of conditional probability is given by the definition

$$\mathbb{P}[X = x \mid Y = y] := \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} = \frac{p(x, y)}{\sum_{x' \in \mathcal{X}} p(x', y)}.$$

For any given $y \in \mathcal{Y}$, the above is a probability measure over $\mathcal{X}$. Thus, the conditional probability is a *parametric family* of probability measures, or a function $\mathcal{Y} \to \mathcal{M}(\mathcal{X})$. The associated notion of conditional expectation is given in terms of these conditional probabilities by

$$\mathbb{E}[X \mid Y = y] := \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}[X = x \mid Y = y].$$

This yields a number for each $y \in \mathcal{Y}$; i.e., the "type" of this notion of conditional expectation is a function $\mathcal{Y} \to \mathbb{R}$.

Example 2: Continuous Random Variables Suppose that $(X, Y) \in \mathbb{R}^2$ instead have a continuous joint density $\rho(x, y) > 0$. The conditional density on a given value of Y is then

$$\rho(x \mid Y = y) = \rho(x \mid y) := \frac{\rho(x, y)}{\int \rho(x', y)\, dx'}.$$

This again is a probability density for each $y \in \mathcal{Y} = \mathbb{R}$, and thus may be viewed as the same kind of object, a function $\mathcal{Y} = \mathbb{R} \to \mathcal{M}(\mathcal{X})$, as above. Likewise, the associated conditional expectation is

$$\mathbb{E}[X \mid Y = y] := \int x\rho(x \mid y)\,dx,$$

which is again a function $\mathcal{Y} = \mathbb{R} \to \mathbb{R}$.

Example 3: Events For a different example of an elementary notion of conditional expectation, suppose that A and B are events and $\mathbb{P}[B] > 0$. Then, the conventional definition of conditional probability of events is

$$\mathbb{P}[A \mid B] := \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Similarly, if X is a random variable, then we may define the conditional expectation on an event as

$$\mathbb{E}[X \mid B] := \frac{\mathbb{E}X\mathbb{1}_B}{\mathbb{P}[B]}.$$

This example, while clearly similar in spirit to the above two, is suspiciously different: now the "type" both of the conditional probability and the conditional expectation is merely that of a single scalar, not a function. We will next concern ourselves with how to reconcile and generalize all of these notions into one unifying setup.

3.2 Information, σ-Algebras, and Restricted Random Variables

The way we will unite the above examples is by formalizing how to condition on a general "collection of information". We think of the first two examples above as our being told the value of Y, and taking probabilities or expectations over the "remaining randomness". Likewise, we think of the last example as being told whether event B did or did not happen (the formalism we recalled only talks about assuming that B *did* happen, but we will soon see that it is more sensible to consider $\mathbb{E}[X \mid B]$ and $\mathbb{E}[X \mid B^c]$ as bundled together into one object).

What will be the "collections of information" that we work with? In Kolmogorov's foundations of probability that we have been working in, this is precisely the role that σ-algebras play, describing coherent collections of information over which we can do probability (and the underlying measure theory).

Suppose that we are working over an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the objects we will consider conditioning on are *sub-σ-algebras* $\mathcal{G} \subseteq \mathcal{F}$, constructing the object

$$\text{“ } \mathbb{E}[X \mid \mathcal{G}] \text{ ”}$$

(We will see that focusing on conditional expectations and then defining conditional probabilities in terms of those is a clearer path to take.) These represent a "partial outcome" that we have access to, and our conditional expectations will be over the "remaining randomness"; the larger $\mathcal{G}$, the more of the outcome we have access to, and the less our conditional expectation should average. We will work from the following guiding examples of what $\mathbb{E}[X \mid \mathcal{G}]$ should mean:

- $\mathcal{G} = \{\varnothing, \Omega\}$: this, the smallest possible sub-σ-algebra, should represent taking expectations without conditioning at all, so we expect to have $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$.
- $\mathcal{G} = \sigma(B) = \{\varnothing, B, B^c, \Omega\}$: this algebra represents conditioning on whether event B happened or not, so $\mathbb{E}[X \mid \mathcal{G}]$ should in some sense contain just the two numbers $\mathbb{E}[X \mid B]$ and $\mathbb{E}[X \mid B^c]$.
- $\mathcal{G} = \sigma(Y)$ for a random variable Y: this algebra represents conditioning on the value of Y, so $\mathbb{E}[X \mid \mathcal{G}]$ should be a function from values of Y to $\mathbb{R}$.
- $\mathcal{G} = \mathcal{F}$: this, the largest possible sub-σ-algebra, should represent not taking an expectation at all, so we expect to have $\mathbb{E}[X \mid \mathcal{G}]$ containing all of the information of the value of the random variable X itself.

It remains mysterious what kind of mathematical object $\mathbb{E}[X \mid \mathcal{G}]$ should be, since the various "types" it needs to have for different choices of $\mathcal{G}$ seem quite different. However, we will see next that there is an elegant choice that indeed captures all of these examples, which is to take $Z := \mathbb{E}[X \mid \mathcal{G}]$ to be a *random variable*, i.e., a function $Z : \Omega \to \mathbb{R}$, which is $\mathcal{G}$-*measurable*. Let us understand why this is compatible with the above examples by reviewing what $\mathcal{G}$-measurability means in each case. Three of the examples are simple:

- $\mathcal{G} = \{\varnothing, \Omega\}$: here, $Z : \Omega \to \mathbb{R}$ is $\mathcal{G}$-measurable if and only if it is constant. Thus, such Z indeed "contains" or "encodes" a single scalar value, as we expect.
- $\mathcal{G} = \sigma(B)$: here, $Z : \Omega \to \mathbb{R}$ is $\mathcal{G}$-measurable if and only if $Z(\omega) = b$ when $\omega \in B$ and $Z(\omega) = c$ when $\omega \in B^c$, for some $b, c \in \mathbb{R}$. Thus again this definition is compatible with the above, containing exactly two scalar values.
- $\mathcal{G} = \mathcal{F}$: here, any random variable Z is $\mathcal{G} = \mathcal{F}$-measurable by definition. Indeed, we will see that here it is sensible to take $Z = X$, so that the expectation "conditional on everything" has no effect on a random variable.

The last case $\mathcal{G} = \sigma(Y)$ is trickier: recall that here we hope that an $\mathcal{G}$-measurable random variable will be a function of values of Y. This is true by an important but seldom mentioned result of measure theory:

> **Lemma 3.2.1** (Doob-Dynkin)**.** If $Y : \Omega \to \mathbb{R}$ is a random variable and $Z : \Omega \to \mathbb{R}$ is $\sigma(Y)$-measurable, then there exists an $f : \mathbb{R} \to \mathbb{R}$ Borel-measurable such that $Z(\omega) = f(Y(\omega))$ for all $\omega \in \Omega$.

Thus we are in luck and indeed a $\mathcal{G} = \sigma(Y)$-measurable Z contains precisely the information of a function of values of Y.

We have arrived at a proposal for what type of object $Z = \mathbb{E}[X \mid \mathcal{G}]$ should be: a $\mathcal{G}$-measurable random variable on the same space that X is defined on. Now let us see how we can specify what such a Z should be.

3.3 Characterizing Conditional Expectations

First, we show the following description of a collection of data that specifies a $\mathcal{G}$-measurable random variable.

Proposition 3.3.1. Suppose that Z, Z' are $\mathcal{G}$-measurable random variables such that $\mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[Z'\mathbb{1}_A]$ for all $A \in \mathcal{G}$. Then, $Z = Z'$ $\mathbb{P}$-almost surely.

In other words, the collection of numbers $(\mathbb{E}[Z\mathbb{1}_A])_{A\in\mathcal{G}}$ determines Z up to differences on null sets.

Proof. We can calculate

$$\begin{aligned}\mathbb{E}|Z - Z'| &= \mathbb{E}[(Z - Z')\mathbb{1}_{\{Z-Z'>0\}}] + \mathbb{E}[(Z' - Z)\mathbb{1}_{\{Z-Z'\le 0\}}] \\ &= \mathbb{E}[Z\mathbb{1}_A] - \mathbb{E}[Z'\mathbb{1}_A] + \mathbb{E}[Z'\mathbb{1}_{A^c}] - \mathbb{E}[Z\mathbb{1}_{A^c}] \\ &= 0\end{aligned}$$

for $A = \{Z - Z' > 0\} \in \mathcal{G}$. At the end above we apply our assumption, and the result follows. □

With this, let us revisit the first of our examples and see how this collection of data looks in that case. Recall that in this case we have $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ for $\mathcal{X}, \mathcal{Y}$ finite sets, and we gave an explicit definition of $\mathbb{E}[X \mid Y = y] = f(y)$ that we would like our definition of $Z = \mathbb{E}[X \mid \mathcal{G}]$ for $\mathcal{G} = \sigma(Y)$ to recover. In particular, we would like this to be $Z = f(Y)$, the function $f : \mathcal{Y} \to \mathbb{R}$ evaluated on the random variable Y. The function $f(y)$ was given by

$$f(y) = \sum_{x\in X} x \cdot \mathbb{P}[X = x \mid Y = y] = \sum_{x\in X} x \cdot \frac{p(x, y)}{\mathbb{P}[Y = y]}.$$

Suppose we have $A \in \mathcal{G} = \sigma(Y)$. Such sets are of the form $A = \{Y \in B\}$ for some $B \subseteq \mathcal{Y}$. Thus, we may calculate

$$\begin{aligned}\mathbb{E}[Z\mathbb{1}_A] &= \mathbb{E}f(Y)\mathbb{1}\{Y \in B\} \\ &= \sum_{y\in B} \mathbb{P}[Y = y] \cdot f(y) \\ &= \sum_{y\in B} \mathbb{P}[Y = y] \cdot \sum_{x\in X} x \cdot \frac{p(x, y)}{\mathbb{P}[Y = y]} \\ &= \sum_{y\in B}\sum_{x\in X} x \cdot p(x, y) \\ &= \mathbb{E}[X\mathbb{1}\{Y \in B\}] \\ &= \mathbb{E}[X\mathbb{1}_A].\end{aligned}$$

Thus, omitting the intermediate details, we find the relation

$$\mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A] \text{ for all } A \in \mathcal{G}.$$

You may check as an exercise that in fact the same relation holds in the second example of continuous random variables as well, essentially by doing the same calculation with sums replaced with integrals as appropriate. By Proposition 3.3.1, this criterion characterizes the random variable Z up to $\mathbb{P}$-null events.

3.4 Definition, Existence, and Uniqueness

Our next step is to take the above as the *definition* of the conditional expectation. We will show that such a random variable Z always exists, and will denote by it the conditional expectation:

Theorem 3.4.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X \in L^1(\mathbb{P})$ a random variable (i.e., one having $\mathbb{E}|X| < \infty$). Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then, there exists a random variable Z satisfying the following:

1. Z is $\mathcal{G}$-measurable.
2. $\mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A]$ for all $A \in \mathcal{G}$.
3. $Z \in L^1(\mathbb{P})$.

Further, if Z and Z' both satisfy the above conditions, then $Z = Z'$ $\mathbb{P}$-almost surely.

Definition 3.4.2. We denote $\mathbb{E}[X \mid \mathcal{G}] := Z$ for Z as in Theorem 3.4.1. Note that Z is only well-defined up to $\mathbb{P}$-null events.

We have already seen that $\mathbb{E}[X \mid \sigma(Y)]$ defined in this more abstract way behaves in a way compatible with our elementary definitions, since we calculated above that the elementary definition yields a Z that indeed satisfies the above conditions (in particular the main Condition 2), which must therefore be unique up to $\mathbb{P}$-null events by the uniqueness clause of the Theorem. Let us see what happens in the other examples we looked at:

- $\mathcal{G} = \{\varnothing, \Omega\}$: here, on the one hand $Z = \mathbb{E}[X \mid \mathcal{G}]$ must be $\mathcal{G}$-measurable, i.e., constant. On the other hand, taking $A = \Omega$, we must have $\mathbb{E}[Z] = \mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[X]$, so we must have $Z = \mathbb{E}[X]$, as we expected above.
- $\mathcal{G} = \sigma(B)$: here, we must have $Z(\omega) = b$ if $\omega \in B$ and $Z(\omega) = c$ if $\omega \in B^c$ for some $b, c \in \mathbb{R}$. Let us derive these values. Taking $A = B$,

$$b\mathbb{P}[B] = \mathbb{E}[Z\mathbb{1}_B] = \mathbb{E}[X\mathbb{1}_B],$$

and thus, if $\mathbb{P}[B] > 0$,

$$b = \frac{\mathbb{E}[X\mathbb{1}_B]}{\mathbb{P}[B]} = \mathbb{E}[X \mid B],$$

the right-hand side referring to the elementary notion of expectation conditional on an event. Similarly, you may derive that $c = \mathbb{E}[X \mid B^c]$. Thus, as we hoped above, $\mathbb{E}[X \mid \sigma(B)]$ contains precisely these two elementary conditional expectation values.

3.4.1 Radon-Nikodym Theorem

We now move towards the proof of Theorem 3.4.1. The main difficulty will be in the existence proof, for which we allude to the following important and useful result of general measure theory. We sketch the main ideas of its proof here as well.

Theorem 3.4.3 (Radon-Nikodym)**.** Let $(\Omega, \mathcal{G})$ be a measurable space and μ, ν be finite, non-negative measures on this space. Suppose that $\nu \ll \mu$, meaning that $\nu(A) = 0$ whenever $\mu(A) = 0$. Then, there exists a $\mathcal{G}$-measurable $f : \Omega \to \mathbb{R}$ such that $f(\omega) \geq 0$ for all ω, and

$$\nu(A) = \int_A f \, d\mu \text{ for all } A \in \mathcal{G}. \tag{3.4.1}$$

This f is unique up to μ-null events (i.e., if f and g both satisfy the above, then $\mu(\{f - g \neq 0\}) = 0$).

The function f as in the Theorem is usually denoted

$$\frac{d\nu}{d\mu} := f$$

and called the *relative density* of ν with respect to μ.

The proof will rely on the following standard measure-theoretic result, whose proof we omit (see [Bil17, Theorem 32.1] for details).

Theorem 3.4.4 (Hahn decomposition)**.** Let $(\Omega, \mathcal{G})$ be a measurable space and μ be a signed measure on it. Then, there exist $P, N \subseteq \Omega$ disjoint and such that $P \sqcup N = \Omega$ such that $\mu(A) \geq 0$ for all $A \in \mathcal{G}$ with $A \subseteq P$, and $\mu(A) \leq 0$ for all $A \in \mathcal{G}$ with $A \subseteq N$.

Proof Sketch of Theorem 3.4.3. We outline the proof of the existence part of the statement. We begin by constructing a candidate function f, and then show that it works. Consider the set of functions that "underestimate" the equality that f is supposed to have:

$$\mathcal{F} := \left\{ f : \Omega \to \mathbb{R} : f \geq 0, \, \mathcal{G}\text{-measurable}, \, \int_A f \, d\mu \leq \nu(A) \text{ for all } A \in \mathcal{G} \right\}.$$

Note that $\mathcal{F} \neq \emptyset$ since the zero function belongs to $\mathcal{F}$. Also, you may check that $\mathcal{F}$ is closed under taking the maximum of functions and monotone limits of functions. Define

$$K := \sup_{f \in \mathcal{F}} \int_\Omega f \, d\mu.$$

Since $\mathcal{F}$ is closed under maxima and monotone limits, by taking a sequence $g_1, g_2, \cdots \in \mathcal{F}$ that achieves the supremum, letting $f_n := \max\{g_1, \dots, g_n\}$, and letting $f := \lim_{n\to\infty} f_n$, we see that there exists an $f \in \mathcal{F}$ that achieves the supremum, i.e., that has

$$K = \int_\Omega f \, d\mu.$$

Next, we argue that this f we have constructed in fact satisfies the conditions of being a relative density. Define the "remainder measure" associated to this f,

$$\delta(A) := \nu(A) - \int_A f \, d\mu.$$

This is a non-negative finite measure on $(\Omega, \mathcal{G})$. We will be done if we can show that $\delta = 0$. For the sake of contradiction, suppose otherwise. Then, $\delta(\Omega) > 0$, and so there exists some $\epsilon > 0$ such that $\delta(\Omega) - \epsilon\mu(\Omega) > 0$. Consider the signed measure $\gamma := \delta - \epsilon\mu$. By Theorem 3.4.4, there exists some $P \in \mathcal{G}$ such that $\gamma(A) \geq 0$ for all $A \subseteq P$ and such that $\gamma(\Omega) \leq \gamma(P)$. In particular, $\gamma(P) > 0$.

Define $\widetilde{f} := f + \epsilon\mathbb{1}_P$. We have $\widetilde{f} \geq 0$ and $\widetilde{f}$ is $\mathcal{G}$-measurable by construction, and for $A \in \mathcal{G}$,

$$\begin{aligned}
\int_A \widetilde{f} \, d\mu &= \int_A f \, d\mu + \epsilon\mu(P \cap A) \\
&= \nu(A) - \delta(A) + \epsilon\mu(P \cap A) \\
&\leq \nu(A) - \delta(P \cap A) + \epsilon\mu(P \cap A) \\
&= \nu(A) - \gamma(P \cap A) \\
&\leq \nu(A),
\end{aligned}$$

so $\widetilde{f} \in \mathcal{F}$. And, we have

$$\int_\Omega \widetilde{f} \, d\mu = \int_\Omega f \, d\mu + \epsilon\mu(P) = K + \epsilon\mu(P).$$

Also,

$$0 < \gamma(P) = \delta(P) - \epsilon\mu(P) \leq \delta(P) \leq \nu(P).$$

Using our assumption of absolute continuity, since $\nu(P) > 0$ we must have $\mu(P) > 0$ as well. Thus, $\int_\Omega \widetilde{f} \, d\mu > K$, a contradiction to the definition of K.

Thus, we must have $\delta = 0$, and thus f satisfies the required condition. □

3.4.2 Existence: Proof of Theorem 3.4.1

We are now ready for the proof we skipped earlier. Before continuing, we note that we will use Theorem 3.4.3 in a way that might seem unusual: given a fixed pair of measures μ, ν on $(\Omega, \mathcal{F})$, we may apply the Theorem with respect to any $\mathcal{G} \subseteq \mathcal{F}$. This gives *different* relative densities $f = f_\mathcal{G}$: they must be $\mathcal{G}$-measurable, which is more restrictive the smaller $\mathcal{G}$ is, but also they must satisfy the family of conditions (3.4.1), which are fewer the smaller $\mathcal{G}$ is.

Proof of Theorem 3.4.1. Consider first the case $X \geq 0$. From X, we define a measure

$$\nu(A) := \mathbb{E}[X \mathbb{1}_A] = \int_A X(\omega)\, d\mathbb{P}(\omega).$$

Note that, in the notation of Theorem 3.4.3, this makes it so that $X = \frac{d\nu}{d\mathbb{P}}$, using a random variable as a relative density. Since $X \geq 0$ this measure is non-negative, and since $X \in L^1$ it is finite. Thus, we may apply Theorem 3.4.3 to ν and $\mu = \mathbb{P}$, but we do so, as indicated above, with respect to the sub-algebra $\mathcal{G} \subseteq \mathcal{F}$. This yields $Z : \Omega \to \mathbb{R}$ a non-negative $\mathcal{G}$-measurable function, which satisfies

$$\int_A Z(\omega)\, d\mathbb{P}(\omega) = \nu(A)$$

for all $A \in \mathcal{G}$. But, rewriting either side and viewing Z as a random variable, this equivalently says

$$\mathbb{E} Z \mathbb{1}_A = \mathbb{E} X \mathbb{1}_A,$$

and so this Z satisfies precisely the desired condition.

For general X, write $X^+ := X \vee 0$ and $X^- := X \wedge 0$, so that $X^\pm \geq 0$ while $X = X^+ - X^-$. Then, you may check that setting $\mathbb{E}[X \mid \mathcal{G}] := \mathbb{E}[X^+ \mid \mathcal{G}] - \mathbb{E}[X^- \mid \mathcal{G}]$ achieves the desired properties, where each of the individual conditional expectations is constructed by the above means. □

3.5 Basic Properties

We now establish some basic "building block" properties that make it easier to work with conditional expectations. To prove these, we work directly from the defining property in Theorem 3.4.1, constructing $Z = \mathbb{E}[X \mid \mathcal{G}]$ by checking that we indeed have

$$\mathbb{E}[Z \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A] \text{ for all } A \in \mathcal{G}. \tag{3.5.1}$$

We will be a little sloppy with the other conditions, but you may check that they hold in all cases below. We also note that the conditional expectation is only defined up to $\mathbb{P}$-null events, so all equalities of random variables below involving conditional expectations are also to be interpreted in this sense.

Proposition 3.5.1 (Linearity)**.** Let $X, Y \in L^1$, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and $a, b \in \mathbb{R}$. Then,

$$\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}].$$

Proof. We check that the right-hand side satisfies (3.5.1): letting $A \in \mathcal{G}$,

$$\mathbb{E}\Big[(a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}])\mathbb{1}_A\Big] = a\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{G}]\mathbb{1}_A + b\mathbb{E}\Big[\mathbb{E}[Y \mid \mathcal{G}])\mathbb{1}_A\Big]$$

and using (3.5.1) in each term,

$$\begin{aligned} &= a\mathbb{E}[X\mathbb{1}_A] + b\mathbb{E}[Y\mathbb{1}_A] \\ &= \mathbb{E}[(aX + bY)\mathbb{1}_A]. \end{aligned}$$

□

Proposition 3.5.2 (Monotonicity)**.** Let $X, Y \in L^1$, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and suppose that $X \leq Y$ almost surely. Then, $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}]$ almost surely.

Proof. By Proposition 3.5.1, without loss of generality, we may take $X = 0$. Let $Z := \mathbb{E}[Y \mid \mathcal{G}]$. Then, using (3.5.1) and then that $Y \geq 0$ almost surely, we have

$$\mathbb{E}[Z\mathbb{1}\{Z < 0\}] = \mathbb{E}[Y\mathbb{1}\{Z < 0\}] \geq 0,$$

so $Z \geq 0$ almost surely as claimed. □

Proposition 3.5.3 (Factorization)**.** Let $X, Y, XY \in L^1$, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and suppose that X is $\mathcal{G}$-measurable. Then,

$$\mathbb{E}[XY \mid \mathcal{G}] = X\mathbb{E}[Y \mid \mathcal{G}].$$

Proof. The proof follows the standard "simple function ladder" method of establishing the result for $X = \mathbb{1}_A$ for $A \in \mathcal{G}$, then for linear combinations of such indicator random variables, then for general non-negative random variables, i.e., $\mathcal{G}$-measurable functions, and finally for all random variables. We only describe the first part and leave the rest as an exercise. Suppose $X = \mathbb{1}_B$ for $B \in \mathcal{G}$. Let $Z := X\mathbb{E}[Y \mid \mathcal{G}] = \mathbb{1}_B\mathbb{E}[Y \mid \mathcal{G}]$. We want to show that $Z = \mathbb{E}[XY \mid \mathcal{G}]$, so we verify (3.5.1): let $A \in \mathcal{G}$, then we have

$$\mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[\mathbb{1}_{A\cap B}\mathbb{E}[Y \mid \mathcal{G}]]$$

and since $A \cap B \in \mathcal{G}$, by the property (3.5.1) we have

$$\begin{aligned} &= \mathbb{E}[\mathbb{1}_{A\cap B}Y] \\ &= \mathbb{E}[XY\mathbb{1}_A], \end{aligned}$$

as required. □

Proposition 3.5.4 (Independence)**.** Let $X \in L^1$, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and suppose X is independent of $\mathcal{G}$. Then,

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X].$$

In particular, if X and Y are independent random variables, then

$$\mathbb{E}[X \mid Y] = \mathbb{E}[X].$$

Proof. Let $A \in \mathcal{G}$ and write $z := \mathbb{E}[X]$, using a lowercase letter since this is just a deterministic number. We have

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[X] \cdot \mathbb{E}[\mathbb{1}_A] = z \cdot \mathbb{E}[\mathbb{1}_A] = \mathbb{E}[z\mathbb{1}_A],$$

thus the constant random variable z satisfies (3.5.1). □

Proposition 3.5.5 (Tower property)**.** Let $X \in L^1$ and $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ nested σ-algebras. Then,

$$\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}\Big] = \mathbb{E}[X \mid \mathcal{H}].$$

In particular, taking $\mathcal{H} = \{\varnothing, \Omega\}$, we have

$$\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{G}]\Big] = \mathbb{E}[X].$$

Proof. Let $A \in \mathcal{H}$; note that then $A \in \mathcal{G}$ as well by assumption. Then, we have using (3.5.1) twice,

$$\mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{G}]\mathbb{1}_A\Big] = \mathbb{E}[X\mathbb{1}_A] = \mathbb{E}\Big[\mathbb{E}[X \mid \mathcal{H}]\mathbb{1}_A\Big].$$

But, this also verifies (3.5.1) for the stated equality of conditional expectations. □

Proposition 3.5.6. Let $X \in L^1$ and $f : \mathbb{R} \to \mathbb{R}$ convex such that $f(X) \in L^1$ as well. Then, almost surely (noting that either side of the below is a random variable!), we have

$$\mathbb{E}[f(X) \mid \mathcal{G}] \geq f(\mathbb{E}[X \mid \mathcal{G}]).$$

Proof Sketch. Use that any convex f is a supremum of linear functions, along with Propositions 3.5.1 and 3.5.2. □

4 | Martingales

4.1 Basic Definitions

We first give the main definitions surrounding martingales, and then give some motivating examples in the following Section to show the wide range random sequences martingales can describe. Martingales are *discrete-time stochastic processes*, sequences $(M_n)_{n\geq 0}$ of random variables with a discrete time index $n \in \mathbb{Z}_{\geq 0}$. We define the following structure of σ-algebras describing the "information" available at time n.

Definition 4.1.1 (Filtration)**.** A *filtration* $(\mathcal{F}_n)_{n\geq 0}$ of a σ-algebra $\mathcal{F}$ is a nested sequence $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \cdots \subseteq \mathcal{F}$ of sub-σ-algebras. We say that a sequence $(X_n)_{n\geq 0}$ of random variables is *adapted* to $(\mathcal{F}_n)$ if X_n is $\mathcal{F}_n$-measurable for each $n \geq 0$, and that a sequence $(X_n)_{n\geq 1}$ is *predictable* with respect to $(\mathcal{F}_n)$ if X_n is $\mathcal{F}_{n-1}$-measurable for each $n \geq 1$.

Definition 4.1.2 (Martingale)**.** A sequence $(M_n)_{n\geq 0}$ of random variables is a *martingale* with respect to a filtration $(\mathcal{F}_n)$ if the following hold:

1. (M_n) is adapted to $(\mathcal{F}_n)$.

2. $M_n \in L^1$ for all $n \geq 0$.

3. $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$ almost surely for all $n \geq 0$.

Such a sequence is a *submartingale* if instead $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] \geq M_n$, and a *supermartingale* if instead $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] \leq M_n$.

Remark 4.1.3. A sub/super/martingale also automatically has the same property with respect to the filtration $\widetilde{\mathcal{F}}_n := \sigma(M_0, \dots, M_n)$, the minimal filtration to which the sequence is adapted.

The following is an important basic property to keep in mind—in expectation, martingales stay constant, submartingales go up, and supermartingales go down.

Proposition 4.1.4. If (M_n) is a martingale, then $\mathbb{E}M_n = \mathbb{E}M_0$ and $\mathbb{E}[M_{n+k} \mid \mathcal{F}_n] = M_n$ almost surely for all $n, k \geq 1$. If (M_n) is a sub/supermartingale, then the same holds with $=$ replaced by $\geq$ or $\leq$, respectively.

Proof. By the tower property and induction. □

There is also another way to view the definition of martingales in terms of the differences between consecutive values.

Definition 4.1.5 (Martingale increments)**.** A sequence $(\Delta_n)_{n\geq 1}$ of random variables is a sequence of *martingale increments* with respect to a filtration $(\mathcal{F}_n)$ if the following hold:

1. (Δ_n) is adapted to $\mathcal{F}_n$.
2. $\Delta_n \in L^1$ for all $n \geq 1$.
3. $\mathbb{E}[\Delta_{n+1} \mid \mathcal{F}_n] = 0$ for all $n \geq 0$.

Definition 4.1.6 (Finite differences)**.** Given a sequence $(M_n)_{n\geq 0}$, we write $\Delta(M)$ for the sequence $\Delta(M)_n = M_n - M_{n-1}$ for $n \geq 1$.

Proposition 4.1.7. Let (M_n) be an adapted sequence of L^1 random variables. Then, M_n is a martingale if and only if $\Delta(M)$ is a sequence of martingale increments.

4.2 Motivating Examples

4.2.1 Random Walk

The simplest example of a martingale is the random walk or sum of independent random variables that you are already familiar with. Let X_i be independent with $\mathbb{E}X_i$, and define $S_n := \sum_{i=1}^n X_i$, with $S_0 = 0$. Then, with respect to the filtration $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$, the sequence (S_n) is a martingale. To check this, we compute

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[S_n + X_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[S_n \mid \mathcal{F}_n] + \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = S_n + \mathbb{E}[X_{n+1}] = S_n$$

using the linearity, factorization, and independence properties of conditional expectation.

To compare with other examples, we note some aspects of the limiting behavior of this sequence: S_n almost surely does not converge, but $\frac{1}{n}S_n$ does by the law of large numbers, while $\frac{1}{\sqrt{n}}S_n$ converges in distribution (provided $X_i \in L^2$ are i.i.d.) by the central limit theorem, and we can prove concentration estimates on $\mathbb{P}[S_n \geq tn]$ in large deviations principles. We will see a little later how some of these properties enjoy generalizations to martingales.

4.2.2 Geometric Random Walk

Suppose now that X_i are independent with $X_i \geq 0$ almost surely and $\mathbb{E}X_i = 1$. Define $M_n := \prod_{i=1}^n X_i$, with $M_n = 1$. Then, (M_n) is a martingale with respect to the same filtration as above. Indeed, we have:

$$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[M_n X_{n+1} \mid \mathcal{F}_n] = M_n \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = M_n \mathbb{E}[X_{n+1}] = M_n.$$

This kind of object is called a *geometric random walk*, and for example has many applications in mathematical finance.

These examples can have quite different convergence behavior than random walks. For instance, suppose $X_i \sim \mathsf{Unif}(\{0,2\})$. Then, you may check that $M_n \to 0$ almost surely. In particular then, we have

$$0 = \mathbb{E} \lim_{n\to\infty} M_n \lneq \lim_{n\to\infty} \mathbb{E}M_n = 1,$$

or in other words, the M_n converge almost surely but not in L^1. We will later study under what conditions martingales converge, and when we can avoid the above situation and have convergence both almost surely and in L^1.

4.2.3 Doob Martingale

Let $X \in L^1$ be *any* random variable, and $(\mathcal{F}_n)$ any filtration. Then, you may check that $M_n := \mathbb{E}[X \mid \mathcal{F}_n]$ is always a martingale, a simple consequence of the tower property. If further we choose $\mathcal{F}_0 = \{\varnothing, \Omega\}$ while $\mathcal{F}_N = \mathcal{F}$ for some N, then we will have $M_0 = \mathbb{E}X$ while $M_N = X$. In between, in some way according to the filtration, the martingale gradually "reveals information" about the outcome of the random variable X.

One important practical application of this construction is to $X = f(Z_1, \dots, Z_N)$ for Z_i independent random variables and $\mathcal{F}_n := \sigma(Z_1, \dots, Z_n)$. We will see soon that this can be used to prove concentration inequalities for quite general nonlinear functions of independent ranodm variables.

4.2.4 Martingale Transform

Another important general construction of martingales is the following, which can be viewed as a kind of discrete-time stochastic integral.

Definition 4.2.1 (Martingale transform)**.** Let $(H_n)_{n\geq 1}$ and $(M_n)_{n\geq 0}$ be sequences of random variables. Then, we define

$$(H \bullet M)_n := \sum_{i=1}^n H_i \Delta(M)_i = \sum_{i=1}^n H_i (M_i - M_{i-1})$$

for each $n \geq 1$, and $(H \bullet M)_0 := 0$.

Proposition 4.2.2. Suppose $H = (H_n)_{n\geq 1}$ is predictable and $M = (M_n)_{n\geq 0}$ is a martingale. Then, $H \bullet M$ is a martingale provided each of its elements is L^1. If $H_n \geq 0$ almost surely for each $n \geq 1$ as well, and M is a sub/supermartingale, then $H \bullet M$ is a sub/supermartingale as well.

Proof. We check that the increments satisfy the required property: we have $\Delta(H \bullet M)_n = H_n \Delta(M)_n$, and so

$$\mathbb{E}[\Delta(H \bullet M)_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[H_{n+1}\Delta(M)_{n+1} \mid \mathcal{F}_n] = H_{n+1}\mathbb{E}[\Delta(M)_{n+1} \mid \mathcal{F}_n]$$

by the predictability of H. Then, the result follows by Proposition 4.1.7. □

Remark 4.2.3. In general, $H \bullet M = \sum H_i \Delta(M)_i$ should be viewed as a discrete analog of the Stieltjes integral, and we will later develop some theory around Stieltjes integrals, in which martingales and a version of the above result play a crucial role.

4.2.5 Gambling and the Martingale Betting Strategy

For example, consider martingale transforms of the simple random walk: let $X_i \sim \mathsf{Unif}(\{\pm 1\})$ and $S_n := \sum_{i=1}^n X_i$, a martingale with respect to $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$. Consider a predictable sequence $(H_n)_{n\geq 1}$ with respect to this filtration. This means H_n is $\sigma(X_1, \dots, X_{n-1})$-measurable, so by Lemma 3.2.1 we have $H_n = h_n(X_1, \dots, X_{n-1})$ almost surely for some measurable $h_n : \mathbb{R}^{n-1} \to \mathbb{R}$. Then, the martingale transform takes the form

$$(H \bullet S)_n = \sum_{i=1}^n H_i X_i = \sum_{i=1}^n h_i(X_1, \dots, X_{i-1}) X_i.$$

We may interpret this in terms of gambling strategies: X_i is the outcome of a simple fair game where we bet some amount and either lose our bet or win an amount equal to our bet. $H_i = h_i(X_1, \dots, X_{i-1})$ the size of our bet on the ith round of the game, which may depend on the previous outcomes of the game. The martingale transform $H \bullet S$ then gives the sequence of our profit over the course of the game.

There is a special case of this that is commonly cited as a paradox of probability theory, confusingly also called "the martingale" sometimes. Formally, we set

$$\begin{aligned} H_1 &:= 1, \\ H_n &:= \left\{ \begin{array}{ll} 2H_{n-1} & \text{if } X_{n-1} = -1, \\ 0 & \text{if } X_{n-1} = +1 \end{array} \right\}. \end{aligned}$$

In words, we double our bet until the first time we win, at which point we stop playing.

To analyze the behavior of this strategy, note that almost surely there will be some $X_i = +1$, so let $n := \min\{i : X_i = +1\}$. Then, we have the sequence of bets

$$H_1 = 1 = 2^0, \quad H_2 = 2^1 \quad \cdots \quad H_n = 2^{n-1}, \quad H_{n+1} = H_{n+2} = \cdots = 0.$$

The sequence of outcomes is

$$X_1 = \cdots = X_{n-1} = -1, \quad X_n = +1.$$

Thus, for all $N \geq n$, we have

$$(H \bullet S)_N = \sum_{i=1}^{N} H_i X_i = -2^0 - 2^1 - \cdots - 2^{n-2} + 2^{n-1} = +1.$$

That is, we appear to always systematically win money!

Of course, the sensible resolution of the paradox is that there is a risk we have not modelled: if we have a finite budget, then we will be "stopped out" of the game and forced to stop playing if we lose all of our money. We will see below that, once we take this into account, we can rigorously show that it is impossible to systematically win money in this way.

For now, let us notice three diverse possible behaviors of martingales that the above examples show us:

1. Random walk: $\mathbb{E}M_n = 0$, M_n almost surely does not converge.

2. Geometric random walk: $\mathbb{E}M_n = 1$, $M_n \to 0$ almost surely, and so

$$\lim_{n\to\infty} \mathbb{E}M_n = 1 > 0 = \mathbb{E} \lim_{n\to\infty} M_n.$$

3. Martingale betting strategy: $\mathbb{E}M_n = 0$, $M_n \to 1$ almost surely, and so

$$\lim_{n\to\infty} \mathbb{E}M_n = 0 < 1 = \mathbb{E} \lim_{n\to\infty} M_n.$$

4.3 Stopping Times

One alternative way in which we can view the example of the martingale betting strategy is as an ongoing random walk that we decide to "stop" at a particular random time. In particular, consider the martingale, for $X_i \sim \mathsf{Unif}(\{\pm 1\})$, given by

$$M_n := \sum_{i=1}^{n} X_i \cdot 2^{i-1}.$$

This is just a weighted random walk, and so is a martingale. Then, if we define

$$\begin{aligned} T &:= \min\{n : X_n = +1\}, \\ \widetilde{M}_n &:= M_{T \wedge n}, \end{aligned}$$

then $\widetilde{M}_n$ has the same law as the martingale associated to the martingale betting strategy formed as a martingale transform above.

We will see that it is useful for the general theory of martingales as well as other applications to develop some general tools around such random times. The following is a reasonable notion of random time with respect to a filtration.

Definition 4.3.1 (Stopping time)**.** $T : \Omega \to \mathbb{Z}_{\geq 0} \cup \{+\infty\}$ is a *stopping time* with respect to the filtration $(\mathcal{F}_n)$ if $\{T = n\} \in \mathcal{F}_n$ for all $n \geq 0$, or equivalently if $\{T \leq n\} \in \mathcal{F}_n$ for all $n \geq 0$.

Two natural examples can be formulated in terms of the simple random walk $S_n = \sum_{i=1}^{n} X_i$, with $X_i \sim \mathsf{Unif}(\{\pm 1\})$ (or any other distribution):

$$T := \mathsf{min}\{n : S_n = a\}, \qquad \text{(hitting time)}$$
$$T := \mathsf{min}\{n : S_n \notin [-a, b]\}. \qquad \text{(exit time)}$$

One natural example of a random time that is *not* a stopping time is

$$T := \mathsf{sup}\{n : S_n = a\}. \qquad \text{(last visit time)}$$

We focus on the behavior of M_T for T a stopping time, a single random variable, or of the sequence $(M_{T \wedge n})$, which follows the trajectory of (M_n) until T happens (if ever), and then "stops" at its current value and remains there forever.

The following is a useful property to keep in mind for building predictable processes out of stopping times.

Proposition 4.3.2. If T is a stopping time, then, for any $n \geq 1$, $\{n > T\}, \{n \leq T\} \in \mathcal{F}_{n-1}$.

Proof. The two events are complementary, so it suffices to consider either one, and we have $\{n > T\} = \{T \leq n - 1\} \in \mathcal{F}_{n-1}$ by definition. □

Proposition 4.3.3. If T is a stopping time and (M_n) is a sub/supermartingale, then $(M_{T \wedge n})$ is also a sub/supermartingale.

Proof. Let $H_n := \mathbb{1}\{n \leq T\}$, then (H_n) is predictable by Proposition 4.3.2. Also, $H_n \geq 0$, so the martingale transform $H \bullet M$ retains the properties of being a sub/supermartingale. And, we have

$$(H \bullet M)_n = \sum_{i=1}^{n} H_i(M_i - M_{i-1}) = \sum_{i=1}^{T \wedge n} (M_i - M_{i-1}) = M_{T \wedge n} - M_0,$$

and shifting by M_0 an $\mathcal{F}_0$-measurable random variable also does not affect the properties of being a sub/supermartingale. □

Corollary 4.3.4. Suppose that T is a stopping time that is almost surely bounded, i.e. there exists some $\tau \in \mathbb{Z}_{\geq 0}$ such that $T \leq \tau$ almost surely. If (M_n) is a martingale, then $\mathbb{E}M_T = \mathbb{E}M_0$ (and similarly with inequalities of (M_n) is a sub/supermartingale).

Proof. Write $\widetilde{M}_n := M_{T\wedge n}$, which is a martingale by Proposition 4.3.3. Then, we have

$$\mathbb{E}M_T = \mathbb{E}M_{T\wedge\tau} = \mathbb{E}\widetilde{M}_\tau = \mathbb{E}\widetilde{M}_0 = \mathbb{E}M_{T\wedge 0} = \mathbb{E}M_0. \qquad \square$$

We will next pursue the question of, for (M_n) a martingale and T a stopping time, when we have

$$\mathbb{E}M_T \overset{(?)}{=} \mathbb{E}M_0$$

Note that this can be viewed as asking when the property that $\mathbb{E}M_n = \mathbb{E}M_0$ can be extended to random times n. So far, we know that the answer is "sometimes": the above Corollary says this holds provided T is bounded, while the example of the martingale betting strategy gives a construction where $1 = \mathbb{E}M_T \gneq \mathbb{E}M_0 = 0$.

4.3.1 Optional Stopping Theorem

The following gives several useful conditions under which the above kind of result holds.

Theorem 4.3.5 (Doob optional stopping)**.** Suppose that (M_n) is a martingale and T is a stopping time. Suppose also that *any* one of the following conditions holds:

1. ("Boundedness in time") $T \le \tau$ almost surely for some $\tau \in \mathbb{Z}_{\ge 0}$.
2. ("Boundedness in space") $T < \infty$ almost surely and, for some $C \in \mathbb{R}_{\ge 0}$, $|M_n| \le C$ almost surely for all $n \ge 0$.
3. ("Boundedness in increments") $\mathbb{E}T < \infty$ (i.e., $T \in L^1$) and, for some $C \in \mathbb{R}_{\ge 0}$, $|M_n - M_{n-1}| \le C$ almost surely for all $n \ge 1$.

Then, $M_T \in L^1$ and $\mathbb{E}M_T = \mathbb{E}M_0$.

Proof. The sufficiency of Condition 1 is just the statement of Corollary 4.3.4 above.

For Condition 2, note that if $T < \infty$ almost surely then $M_{T\wedge n} \to M_T$ almost surely, and this sequence of random variables is uniformly bounded by C. Therefore, by the bounded convergence theorem, $M_T \in L^1$ and $\mathbb{E}M_T = \lim_{n\to\infty} \mathbb{E}M_{T\wedge n}$. But, since $(M_{T\wedge n})$ is a martingale, for any n we have $\mathbb{E}M_{T\wedge n} = \mathbb{E}M_{T\wedge 0} = \mathbb{E}M_0$, and thus $\mathbb{E}M_T = \mathbb{E}M_0$.

For Condition 3, we use a slightly subtler version of the above argument. Since $\mathbb{E}T < \infty$, we also have $T < \infty$ almost surely, and so, as above, we have $(M_{T\wedge n} - M_0) \to (M_T - M_0)$ almost surely. We also can bound

$$\begin{aligned} |M_{T\wedge n} - M_0| &= \left| \sum_{i=1}^{T\wedge n} (M_i - M_{i-1}) \right| \\ &\le C(T \wedge n) \\ &\le CT. \end{aligned}$$

Since $T \in L^1$, this means that $M_{T\wedge n} - M_0$ are dominated by the integrable random variable CT. So, we can apply the dominated convergence theorem to this sequence, which gives $(M_T - M_0) \in L^1$ whereby $M_T \in L^1$, and

$$\mathbb{E}M_T - \mathbb{E}M_0 = \mathbb{E}[M_T - M_0] = \lim_{n\to\infty} \mathbb{E}[M_{T\wedge n} - M_0].$$

But again since $M_{T\wedge n}$ is a martingale every term in the limit is 0, so we find $\mathbb{E}M_T = \mathbb{E}M_0$ again. □

4.3.2 Application: Simple Random Walk

We now show how the seemingly abstract optional stopping theorem can actually give very concrete insights about the behavior of the fundamental simple random walk model, $S_n = \sum_{i=1}^n X_i$ with $X_i \sim \mathsf{Unif}(\{\pm 1\})$ i.i.d. Recall that this S_n is a martingale. We consider the two stopping times discussed above.

Hitting Time Let $a \neq 0$ and

$$T = T_a := \min\{n : S_n = a\}.$$

We will show the following result, which should be quite surprising on its face if you try picturing it for a small value like $a = 1$.

Theorem 4.3.6. For any $a \neq 0$, $\mathbb{E}T_a = \infty$.

Proof. If we do not have $T = T_a < \infty$ almost surely, then the result follows immediately. So, suppose $T < \infty$ almost surely (you may show separately that this is in fact the case). Then, $S_T = a$ by definition, so $\mathbb{E}S_T = a \neq 0 = \mathbb{E}S_0$. Thus, Theorem 4.3.5 must not apply to this example, so Conditions 1, 2, and 3 must all fail. Condition 1 fails because T is not bounded almost surely, and Condition 2 because $|S_n|$ is not bounded almost surely uniformly in n. But, S_n satisfies the bounded increments part of Condition 3. So, the only way Condition 3 can fail is if $\mathbb{E}T = \infty$. □

Note that, unlike treatments of this by direct combinatorics that you might have seen in more elementary probability theory, this generalizes immediately to random walks with arbitrary bounded step sizes by exactly the same proof.

Exit Time Let $a, b > 0$ and consider the slightly modified version of the exit time

$$T = T_{a,b} := \min\{n : S_n \in \{-a, b\}\}.$$

That is, this is the first time that S_n hits the boundary of the interval $[-a, b]$.

As a preliminary, we establish the following, which shows that the situation for hitting times will not occur here.

Proposition 4.3.7. $\mathbb{E}T_{a,b} \leq (a+b)2^{a+b} < \infty$.

Proof. Note that, if $X_{k+1} = \cdots = X_{k+a+b} = +1$, then $T \leq k + (a+b)$, i.e., if a run of $a+b$ steps in the same direction has just occurred, then T must have occurred by the end of such a run. So, define the random variables

$$Y_\ell := \mathbb{1}\{X_{(\ell-1)(a+b)+1} = \cdots = X_{\ell(a+b)} = +1\}$$

for $\ell \geq 1$. These are independent and $\mathsf{Law}(Y_\ell) = \mathsf{Ber}(2^{-(a+b)})$. If we set

$$G := \mathsf{min}\{\ell : Y_\ell = 1\},$$

then the above observation gives that

$$T \leq (a+b)G.$$

On the other hand, G is a geometric random variable, $\mathsf{Law}(G) = \mathsf{Geom}(2^{-(a+b)})$, and you may compute directly that $\mathbb{E}[G] = 2^{a+b}$, giving the result. □

We will then obtain the following by applying the optional stopping theorem to the exit time, describing the probability with which S_n exits $[-a, b]$ on either side of the interval.

Theorem 4.3.8. For any $a, b > 0$,

$$\mathbb{P}[S_{T_{a,b}} = -a] = \frac{b}{a+b},$$
$$\mathbb{P}[S_{T_{a,b}} = b] = \frac{a}{a+b}$$

Note that the dependence on a and b looks inverted but is intuitively correct on reflection: as b gets larger, it becomes less likely to exit $[-a, b]$ on the upper side, and vice-versa.

Proof. Write $T = T_{a,b}$. Note that $S_T \in \{-a, b\}$ almost surely, and write $p := \mathbb{P}[S_T = -a]$. Condition 3 of Theorem 4.3.5 holds, so we have

$$0 = \mathbb{E}S_0 = \mathbb{E}S_T = p \cdot (-a) + (1-p) \cdot b = b - p(a+b),$$

and solving for p gives the result. □

It turns out that we may derive much more information about the joint distribution of the random variables (T, S_T) by building other martingales out of S_n. Let us demonstrate the first natural such extension.

Proposition 4.3.9. Let (S_n) be the simple random walk as above. Then, $M_n := S_n^2 - n$ is a martingale.

Proof. We may calculate directly

$$\begin{aligned}
\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] &= \mathbb{E}[(S_{n-1} + X_n)^2 - n \mid \mathcal{F}_{n-1}] \\
&= \mathbb{E}[S_{n-1}^2 \mid \mathcal{F}_{n-1}] + \mathbb{E}[2X_n S_{n-1} \mid \mathcal{F}_{n-1}] + \mathbb{E}[X_n^2 \mid \mathcal{F}_{n-1}] - n \\
&= S_{n-1}^2 + 2S_{n-1}\mathbb{E}[X_n] + \mathbb{E}[X_n^2] - n \\
&= S_{n-1}^2 - (n-1) \\
&= M_{n-1},
\end{aligned}$$

as required. □

We would like to use Theorem 4.3.5 on this new martingale (M_n) together with the exit stopping time T. As before, Conditions 1 and 2 of Theorem 4.3.5 do not apply, so we can only hope to use Condition 3. However, in this case Condition 3 does not apply either! That is because the increments of M_n are unbounded. These are:

$$\begin{aligned}
M_n - M_{n-1} &= S_n^2 - n - (S_{n-1}^2 - (n-1)) \\
&= S_n^2 - S_{n-1}^2 - 1 \\
&= 2X_n S_{n-1} + X_n^2 - 1,
\end{aligned}$$

which is unbounded since $|X_n| = 1$ while S_{n-1} can be as large as $n-1$.

However, we can get around this issue by a sneaky trick: consider the stopping time T we are interested in, and define the stopped martingale

$$\widetilde{M}_n := M_{T \wedge n}.$$

Consider the increments of this martingale: if $T \le n-1$, then we have $\widetilde{M}_n - \widetilde{M}_{n-1} = M_T - M_T = 0$. If $T \ge n$, then we have $|S_n|, |S_{n-1}| \le a \vee b$, since by definition $T \ge n$ means that S_n has not exited the interval $[-a, b]$ by time n. Thus, in this case we have

$$|\widetilde{M}_n - \widetilde{M}_{n-1}| = |M_n - M_{n-1}| = |2X_n S_{n-1} + 1| \le 2(a \vee b) + 1,$$

and so $\widetilde{M}_n$ *does* have bounded increments and can be used with Theorem 4.3.5. Doing this, we get the following new piece of information about T:

Theorem 4.3.10. For any $a, b > 0$, $\mathbb{E}[T_{a,b}] = ab$.

Proof. By the above remarks, we may apply Theorem 4.3.5 to the martingale $\widetilde{M}_n = M_{T \wedge n}$ and the stopping time T. This gives

$$\mathbb{E}M_T = \mathbb{E}\widetilde{M}_T = \mathbb{E}\widetilde{M}_0 = \mathbb{E}M_0 = 0.$$

Expanding the definition of M_T here, we find

$$\begin{aligned}
0 &= \mathbb{E}[S_T^2 - T] \\
&= \mathbb{E}[S_T^2] - \mathbb{E}[T] \\
&= a^2\mathbb{P}[S_T = a] + b^2\mathbb{P}[S_T = b] - \mathbb{E}[T] \\
&= a^2 \cdot \frac{b}{a+b} + b^2 \cdot \frac{a}{a+b} - \mathbb{E}[T] \\
&= ab - \mathbb{E}[T],
\end{aligned}$$

and rearranging gives the result. □

4.4 Convergence of Martingales

4.4.1 Almost Sure Convergence

We now prove the following fundamental result about the convergence of submartingales and supermartingales under a relatively mild condition.

> **Theorem 4.4.1** (Doob almost sure convergence)**.** Suppose (M_n) is a sub/supermartingale that is *bounded in* L^1, i.e., such that $\sup_n \mathbb{E}|M_n| < \infty$. Then, there exists a random variable $M_\infty \in L^1$ such that $M_n \to M_\infty$ almost surely.

Before giving the proof, let us see what this says about our discussion of gambling strategies from Section 4.2.5.

> **Corollary 4.4.2.** Suppose (M_n) is a supermartingale that is *bounded below*, i.e., such that, for some $C > 0$, $M_n \geq -C$ almost surely for all $n \geq 0$. Then, there exists a random variable $M_\infty \in L^1$ such that $M_n \to M_\infty$ almost surely, and further $\mathbb{E}M_\infty \leq \inf_n \mathbb{E}M_n \leq \mathbb{E}M_0$.

Proof. Under the boundedness assumption, we have

$$|M_n| \leq 2C + M_n,$$

and therefore, together with the monotonicity of expectations for supermartingales,

$$\mathbb{E}|M_n| \leq 2C + \mathbb{E}M_n \leq 2C + \mathbb{E}M_0.$$

Thus, Theorem 4.4.1 applies and gives the existence of M_∞ satisfying the needed condition. Further, by Fatou's Lemma we have, again using the boundedness,

$$\begin{aligned} \mathbb{E}M_\infty &= \mathbb{E}\liminf_{n\to\infty} M_n \\ &\leq \liminf_{n\to\infty} \mathbb{E}M_n \end{aligned}$$

and, since for a supermartingale this is a non-increasing sequence, we have

$$\begin{aligned} &= \inf_n \mathbb{E}M_n \\ &\leq \mathbb{E}M_0, \end{aligned}$$

as claimed. □

Since any martingale is a supermartingale, this applies to any martingale transform $H \bullet S$ of the simple random walk process $S = (S_n)$. In particular, you can read this result as saying that, if your betting strategy H is such that you make sure (by any mechanism you like) that you never lose an amount of money more than $C > 0$, then your sequence of profits must eventually converge (i.e., you must gradually diminish the size of your bets over time), and you cannot win money on average.

Relatedly, note that it is *not* necessarily the case that $\mathbb{E}M_\infty = \mathbb{E}\lim_{n\to\infty} M_n = \lim_{n\to\infty}\mathbb{E}M_n$, as for instance the geometric random walk example we saw earlier shows. Later we will address separately when this happens, which is a consequence of a stronger mode of convergence for $M_n \to M_\infty$, namely convergence in L^1.

Towards the proof of Theorem 4.4.1, let us describe some objects that allow us to talk about the convergence of a sequence. Our general idea is control convergence by controlling the *oscillations* of a sequence of numbers.

To that end, we introduce the following times associated to a general sequence (M_n): for a given $a < b \in \mathbb{R}$,

$$\begin{aligned} N_0 &:= -1, \\ N_{2k-1} &:= \min\{n \geq N_{2k-2} : M_n \leq a\} \text{ for each } k \geq 1, \\ N_{2k} &:= \min\{n \geq N_{2k-1} : M_n \geq b\} \text{ for each } k \geq 1. \end{aligned}$$

Said in words, each odd-indexed time $N_1, N_3, N_5, \dots$ is the first time after the previous even-indexed time that our sequence drops below level a, and each even-indexed time $N_2, N_4, N_6, \dots$ is the first time after the previous odd-indexed time that our sequence rises above level b.

We call each interval $[N_{2k-1}, N_{2k}]$ for $k \geq 1$ an *upcrossing* of the interval $[a, b]$ by our sequence, for obvious reasons. Further, we define

$$\begin{aligned} U_n = U_n(a,b) &:= \max\{k : N_{2k} \leq n\}, \\ U_\infty = U_\infty(a,b) &:= \lim_{n\to\infty} U_n(a,b). \end{aligned}$$

These are respectively the number of complete upcrossings that have occurred by time n, and the total number of upcrossings (possibly infinite) over all time.

Our first Lemma towards the main proof relates the $U_\infty(a,b)$ to the event that M_n converges:

Lemma 4.4.3. Suppose that, for all $a < b \in \mathbb{R}$, we have $\mathbb{P}[U_\infty(a,b) < \infty] = 1$. Then, we also have $\mathbb{P}[\lim_{n\to\infty} M_n \text{ exists}] = 1$, where we allow for this limit to be $\pm\infty$.

Proof. We relate the event that $\lim_{n\to\infty} M_n$ does not exist to a *countable* union of events related to $U_\infty(a,b)$, by restricting our attention to $a < b \in \mathbb{Q}$:

$$\begin{aligned} &\mathbb{P}[\lim_{n\to\infty} M_n \text{ does not exist}] \\ &= \mathbb{P}[\liminf_{n\to\infty} M_n \lneq \limsup_{n\to\infty} M_n] \\ &= \mathbb{P}[\text{there exist } a < b \in \mathbb{Q} \text{ such that } \liminf_{n\to\infty} M_n < a < b < \limsup_{n\to\infty} M_n] \\ &\leq \mathbb{P}[\text{there exist } a < b \in \mathbb{Q} \text{ such that } U_\infty(a,b) = \infty] \\ &= 0 \end{aligned}$$

by assumption, since the last event is a countable union of events of probability zero. □

Next, we develop a quantitative tool to control the number of upcrossings, which will let us verify the condition of Lemma 4.4.3.

Lemma 4.4.4 (Doob upcrossing lemma)**.** Suppose M_n is a supermartingale. Then, for all $n \geq 1$ and $a < b \in \mathbb{R}$,

$$\mathbb{E}U_n(a,b) \leq \frac{\mathbb{E}(M_n - a)^-}{b-a} \leq \frac{|a| + \mathbb{E}|M_n|}{b-a}.$$

We will use the following fact that is easy to verify from the definition of the times N_i:

Proposition 4.4.5. The N_i for each $i \geq 1$ are all stopping times.

Proof of Lemma 4.4.4. The argument is a lovely "financial proof", arguing that on the one hand, a supermartingale is a "losing game", in the sense that we do not expect to be able to profit by betting on it. On the other hand, when a process has many upcrossings, we *can* profit from it by "buying low and selling high."

To implement this idea, consider the sequence

$$H_n := \mathbb{1}\{n \in (N_{2k-1}, N_{2k}] \text{ for some } k \geq 1\}.$$

Roughly speaking, up to careful treatment of the boundary conditions, this is the indicator that the time n occurs during an upcrossing. By Proposition 4.3.2, the treatment of the boundary conditions is such that H_n is predictable. Also, $H_n \geq 0$. So, by Proposition 4.2.2, $H \bullet M$ is a supermartingale, and in particular

$$\mathbb{E}(H \bullet M)_n \leq \mathbb{E}(H \bullet M)_0 = 0.$$

This formalizes the first part of our intuitive argument above, that we cannot make money by betting (in this case, according to the strategy H) on the supermartingale M.

We now argue that if $U_n(a,b)$ is large then so is $(H \bullet M)_n$. Indeed, we have, letting $u = U_n(a,b)$,

$$\begin{aligned}(H \bullet M)_n &= \sum_{i=1}^{n} H_i(M_i - M_{i-1}) \\ &= \sum_{i=N_1+1}^{N_2} (M_i - M_{i-1}) + \cdots + \sum_{i=N_{2u-1}+1}^{N_{2u}} (M_i - M_{i-1}) + \sum_{i=N_{2u+1}+1}^{n} (M_i - M_{i-1})\end{aligned}$$

Here, the last sum might be empty, in which case we view it as being zero. Continuing by telescoping these sums,

$$= (M_{N_2} - M_{N_1}) + \cdots + (M_{N_{2u}} - M_{N_{2u-1}}) + \mathbb{1}\{n > N_{2u+1}\}(M_n - M_{N_{2u+1}})$$

Each of the first u differences is at least $(b-a)$ by definition of the times N_i, so we have

$$\geq u \cdot (b-a) - (M_n - M_{N_{2u+1}})^-$$

and since $M_{N_{2u+1}} \leq a$ and $u = U_n(a,b)$, we have

$$\geq U_n(a,b) \cdot (b-a) - (M_n - a)^-.$$

Taking expectations and combining our two observations,

$$0 \geq \mathbb{E}(H \bullet M)_n \geq (b-a)\mathbb{E}U_n(a,b) - \mathbb{E}(M_n - a)^-,$$

and rearranging gives the result. □

Finally, by putting together our two Lemmas, we are ready for the proof of the main result, which is quick using these tools.

Proof of Theorem 4.4.1. If (M_n) is a martingale, then it is also a supermartingale, and if (M_n) is a submartingale, then $(-M_n)$ is a supermartingale, so without loss of generality we may assume that (M_n) is a supermartingale. Suppose that $K := \sup_n \mathbb{E}|M_n|$, which is finite by assumption.

By Lemma 4.4.4, for all $a < b \in \mathbb{R}$ and $n \geq 1$, we have

$$\mathbb{E}U_n(a,b) \leq \frac{|a| + \mathbb{E}|M_n|}{b-a} \leq \frac{|a| + K}{b-a}.$$

In particular, this bound is independent of n. Also, $U_n(a,b)$ are non-negative random variables that increase to $U_\infty(a,b)$, so by the monotone convergence theorem we have

$$\mathbb{E}U_\infty(a,b) \leq \frac{|a| + K}{b-a} < \infty.$$

In particular then, $U_\infty(a,b) < \infty$ almost surely.

Thus, the condition of Lemma 4.4.3 is satisfied. So, we obtain that $\lim_{n\to\infty} M_n =: M_\infty$ exists almost surely, though *a priori* it may be $\pm\infty$. But, by Fatou's Lemma we also have

$$\mathbb{E}|M_\infty| = \mathbb{E}\lim_{n\to\infty} |M_n| \leq \liminf_{n\to\infty} \mathbb{E}|M_n| \leq K < \infty,$$

thus $M_\infty \in L^1$ and is finite almost surely. □

BIBLIOGRAPHY

[Ald13] David Aldous. *Probability approximations via the Poisson clumping heuristic*, volume 77. Springer Science & Business Media, 2013.

[Bil17] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

[Kle14] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, second edition, 2014.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[TV11] Terence Tao and Van Vu. Random matrices: universality of local eigenvalue statistics. 2011.

[Ver26] Roman Vershynin. A friendly proof of the berry-esseen theorem. *arXiv preprint arXiv:2602.06234*, 2026.

[Wor99] Nicholas C Wormald. Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298, 1999.