

# Lecture Notes: Probability Theory II

Dmitriy (Tim) Kunisky

Spring 2026 (Last Updated: April 9, 2026)

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Main Ideas of Probability Theory I . . . . .	4
1.2	Our Goals in Probability Theory II . . . . .	5
1.3	Brief Technical Review . . . . .	6
1.3.1	Laws of Large Numbers . . . . .	6
1.3.2	Weak Convergence and Convergence in Distribution . . . . .	7
1.3.3	Characteristic Functions . . . . .	9
1.3.4	Central Limit Theorem . . . . .	10
<b>2</b>	<b>Advanced Limit Theorems</b>	<b>12</b>
2.1	Lindeberg Method . . . . .	12
2.1.1	Continuity Theorem . . . . .	12
2.1.2	Proof of Central Limit Theorem . . . . .	13
2.1.3	Universality . . . . .	15
2.1.4	Quantitative Rates of Convergence . . . . .	15
2.2	Moment Method . . . . .	16
2.2.1	Continuity Theorem . . . . .	16
2.2.2	Proof of Central Limit Theorem . . . . .	18
2.2.3	Poisson Limit Theorems for Rare Events . . . . .	20
2.2.4	Proof of Classical Poisson Limit Theorem . . . . .	22
2.2.5	Application: Fixed Points of Random Permutations . . . . .	23
2.2.6	Application: Motifs in Random Graphs . . . . .	23
2.3	Vector Limit Theorems . . . . .	24
2.3.1	Cramér-Wold Device . . . . .	25
2.3.2	Multivariate Central Limit Theorem . . . . .	25
2.3.3	Application: $\chi^2$ Test and Statistic . . . . .	26
<b>3</b>	<b>Conditional Expectation</b>	<b>29</b>
3.1	Motivation . . . . .	29
3.2	Information, $\sigma$ -Algebras, and Restricted Random Variables . . . . .	30
3.3	Characterizing Conditional Expectations . . . . .	32
3.4	Definition, Existence, and Uniqueness . . . . .	33
3.4.1	Radon-Nikodym Theorem . . . . .	34
3.4.2	Existence: Proof of Theorem 3.4.1 . . . . .	35
3.5	Basic Properties . . . . .	36
3.6	Conditional Probability . . . . .	38

3.6.1	Conditional Probabilities as Conditional Expectations . . . . .	38
3.6.2	Incoherence for General Large Measurable Spaces . . . . .	40
3.6.3	Regular Conditional Probability . . . . .	42
3.6.4	Disintegration Theorem . . . . .	43
3.6.5	Warnings . . . . .	44
<b>4</b>	<b>Martingales</b> . . . . .	<b>45</b>
4.1	Basic Definitions . . . . .	45
4.2	Motivating Examples . . . . .	46
4.2.1	Random Walk . . . . .	46
4.2.2	Geometric Random Walk . . . . .	47
4.2.3	Doob Martingale . . . . .	47
4.2.4	Martingale Transform . . . . .	47
4.2.5	Gambling and the Martingale Betting Strategy . . . . .	48
4.3	Stopping Times . . . . .	49
4.3.1	Optional Stopping Theorem . . . . .	51
4.3.2	Application: Simple Random Walk . . . . .	52
4.3.3	Application: Enumeration Problems and Cayley's Theorem . . . . .	55
4.4	Convergence of Martingales . . . . .	56
4.4.1	Almost Sure Convergence . . . . .	56
4.4.2	$L^1$ Convergence and Uniform Integrability . . . . .	60
4.4.3	$L^p$ Convergence and Maximal Inequalities . . . . .	63
4.5	Application: Concentration Inequalities . . . . .	66
4.5.1	Sums of Independent Variables: Hoeffding Inequality . . . . .	67
4.5.2	Martingale Generalization: Azuma Inequality . . . . .	70
4.5.3	Functions of Independent Variables: McDiarmid Inequality . . . . .	71
4.5.4	Interaction with Maximal Inequalities . . . . .	72
4.5.5	Example: Balls, Bins, and Multinomials . . . . .	73
4.6	Application: Random Series . . . . .	74
4.7	Application: Branching Processes . . . . .	75
4.7.1	Extinction . . . . .	75
4.7.2	Associated Martingale and Convergence . . . . .	76
4.7.3	Subcritical Regime ( $m < 1$ ) . . . . .	77
4.7.4	Critical Regime ( $m = 1$ ) . . . . .	77
4.7.5	Supercritical Regime ( $m > 1$ ) . . . . .	78
<b>5</b>	<b>Markov Chains</b> . . . . .	<b>83</b>
5.1	Motivation . . . . .	83
5.2	Measure Theory Toolkit . . . . .	84
5.2.1	Carathéodory Extension Theorem . . . . .	84
5.2.2	Kolmogorov Extension Theorem . . . . .	84
5.2.3	$\pi$ - $\lambda$ Theorem . . . . .	85
5.2.4	Monotone Class Theorem . . . . .	86
5.3	Definition, Existence, and Uniqueness . . . . .	86
5.4	Markov Properties . . . . .	89

5.5	Countable State Space Structure Theory . . . . .	92
5.5.1	Transience and Recurrence . . . . .	92
5.5.2	Application: Lattice Random Walks . . . . .	94
5.5.3	Accessibility Graph . . . . .	95
5.5.4	Fundamental Structure Theorem . . . . .	97
5.6	Stationary Measures, Convergence, and Ergodicity . . . . .	99
5.6.1	Existence and Uniqueness . . . . .	100
5.6.2	Convergence . . . . .	103
	<b>Bibliography</b>	<b>103</b>

# 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 variables 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  $n \rightarrow \infty$ . 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  $n \rightarrow \infty$ . 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 a 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

$$M_k = \mathbb{E}_{X_{k+1}, \dots, X_n} f(X_1, \dots, X_n)$$

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 general-

izations 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  $S : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}$ . Further, interpolating linearly between the values  $S$  takes, we may extend that to a random function  $S : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ , 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  $c \rightarrow \infty$ , 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  $\mu$ . 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  $\epsilon > 0$ ,

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

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

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

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  $X : \Omega \rightarrow \Sigma$ .<sup>a</sup> Then, the *law* of  $X$  is the probability measure  $\mu$  on  $\Sigma$  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  $\mu = \text{Law}(X)$ .

<sup>a</sup>Most often for us  $\Sigma$  is the real numbers and  $\mathcal{G}$  their Borel  $\sigma$ -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  $\text{Law}(X) = \text{Law}(Y)$ , 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  $\text{Law}(X)$  is “close” in some sense to  $\text{Law}(Y)$ , 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  $\mu_n$  *converges weakly* to  $\mu_\infty$ , denoted  $\mu_n \xrightarrow{(w)} \mu_\infty$ , if, for any  $f : \mathbb{R} \rightarrow \mathbb{R}$  bounded and continuous, we have

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

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

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

In this case, weak convergence is also called *convergence in distribution* of  $X_n$  to  $X_\infty$ , 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,  $X_\infty$  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  $\text{Law}(X_n) \xrightarrow{(w)} \mathcal{N}(0, 1)$  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  $\mu_\infty(\{t\}) = 0$ ,<sup>a</sup> we have

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

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

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

The function  $F_X(t) = \mathbb{P}[X \leq t]$  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  $F_{X_\infty}$  is continuous.

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

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  $\mu$  or random variable  $X$  given by

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

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

$$\phi_\mu(t) = \int \exp(itx) \rho(x) dx.$$

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.  $\text{Law}(X) = \text{Law}(Y)$  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  $\phi_{X_n}(t) \rightarrow \phi_{X_\infty}(t)$  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

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

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

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

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

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

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

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

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

$$\int f d\mu_n \rightarrow \int f d\mu_\infty$$

for every bounded continuous  $f : \mathbb{R} \rightarrow \mathbb{R}$ . The “Fourier method” of characteristic functions amounts to showing (in the Lévy continuity theorem) that it suffices to consider the special test functions  $f(x) = \exp(itx)$ . 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 LINDBERG 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  $\int f d\mu_n \rightarrow \int f d\mu_\infty$  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  $g : \mathbb{R} \rightarrow \mathbb{R}$  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  $n \rightarrow \infty$ . 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  $\mu_n$  by something instead involving  $\mu_\infty$ , so that we may obtain a bound as  $n \rightarrow \infty$ . We will also take for granted that there exists a smooth compactly supported  $h : \mathbb{R} \rightarrow \mathbb{R}$  sandwiched between two indicator functions,

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

By assumption,

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

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

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

and thus

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

Putting everything together, we find

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

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

### 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  $\mathbb{E}X_i = c$ ,  $\text{Var} X_i = \sigma^2$ , 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  $f : \mathbb{R} \rightarrow \mathbb{R}$  smooth and compactly supported, we have

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

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

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

Thus, it suffices to control the differences

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

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

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

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

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

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  $\tilde{Z} = \tilde{Z}(Z_k, X_k)$  such that

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

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

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

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

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

Thus, by triangle inequality we have

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

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

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

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

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  $\mathbb{E}|X_i|^3, \mathbb{E}|Y_i|^3 \leq R < \infty$ , 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  $f(x) = \mathbb{1}\{x \leq t\}$ , 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  $\ell$ , 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  $f : \mathbb{R} \rightarrow \mathbb{R}$ , 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.  $\mathbb{E}X_n^k \rightarrow \mathbb{E}X_\infty^k$  for all  $k \geq 0$  (note that this also implies by linearity that  $\mathbb{E}p(X_n) \rightarrow \mathbb{E}p(X_\infty)$  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  $\epsilon : \mathbb{Z}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  such that  $\epsilon(k) \rightarrow 0$  as  $k \rightarrow \infty$ .

Then,  $X_n \Rightarrow X_\infty$ .

**Example 2.2.2.** It is clear that Condition 2 of the Lemma holds for any compactly supported Law( $X_\infty$ ). Further, we will see below results implying that it also holds for  $X_\infty$  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  $\phi_{X_n} \rightarrow \phi_{X_\infty}$ , 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  $X_\infty$ , we have,<sup>1</sup> 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  $X_\infty$  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,

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

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

<sup>1</sup>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  $\mu$  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  $\nu$ , 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.  $\mathbb{E}X_n^k \rightarrow \mathbb{E}X_\infty^k$  for all  $k \geq 0$  (note that this also implies by linearity that  $\mathbb{E}p(X_n) \rightarrow \mathbb{E}p(X_\infty)$  for all polynomials  $p(x)$ ).
2.  $\text{Law}(X_\infty)$  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  $\mathbb{E}X_i = c$ ,  $\text{Var} X_i = \sigma^2$ , 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  $S = S_1 \sqcup \dots \sqcup S_m$ . A *matching* of  $S$  is a partition all of whose parts have size 2. We write  $\text{Part}(S)$  for the set of partitions

of  $S$  and  $\text{Match}(S)$  for the set of matchings of  $S$ . Lastly, we define  $M_k := |\text{Match}([k])|$  and  $P_k := |\text{Part}([k])|$ .

**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  $\mathbb{E}\left(\frac{1}{\sqrt{n}}S_n\right)^k \rightarrow M_k$  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  $\mathbf{i} = (i_1, \dots, i_k) \in [n]^k$  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

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

We make a few preliminary observations about these quantities:

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

Note that the number of terms  $E_{\mathbf{i}}$  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  $E_{\mathbf{i}}$  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

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

Thus, we find

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

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  $X \sim \text{Pois}(\lambda)$  is  $X \in \mathbb{Z}_{\geq 0}$  with probability mass function

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

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

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

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  $\mathbb{E} X_n^k \rightarrow \mathbb{E} X_\infty^k$  for all  $k \geq 0$ , then also  $\mathbb{E} X_n^k \rightarrow \mathbb{E} X_n^\infty$  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  $X \sim \text{Pois}(\lambda)$ , then  $\mathbb{E} X^{\underline{k}} = \lambda^k$ .

From this you may also show together with some combinatorics that  $\text{Pois}(\lambda)$  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  $\mathbb{E}X_n^k \rightarrow \lambda^k$  for all  $k \geq 0$ , for some  $\lambda > 0$ . Then,  $X_n \Rightarrow X_\infty \sim \text{Pois}(\lambda)$ .

Finally, let us describe the quantities  $\mathbb{E}X^k$  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,

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

*Proof.* Since  $X \in \mathbb{Z}_{\geq 0}$ , by (2.2.1) we have that  $X^k$  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.  $\square$

**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,

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

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  $X_n \Rightarrow X_\infty \sim \text{Pois}(\lambda)$  provided that:

1.  $\mathbb{E}X_n = \sum_i \mathbb{P}[E_i^{(n)}] \rightarrow \lambda$ .
2.  $\max_i \mathbb{P}[E_i^{(n)}] \ll 1$ .
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

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

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  $\lambda > 0$ ,  $\text{Bin}(n, \frac{\lambda}{n}) \xrightarrow{(w)} \text{Pois}(\lambda)$  as  $n \rightarrow \infty$ .

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

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

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

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

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

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

$$\rightarrow \lambda^k,$$

completing the proof. □

## 2.2.5 APPLICATION: FIXED POINTS OF RANDOM PERMUTATIONS

Let  $\text{Sym}(n) := \{\sigma : [n] \rightarrow [n] \text{ bijective}\}$ , the set of permutations. By elementary combinatorics,  $|\text{Sym}(n)| = n!$ . We study the random variable  $\sigma = \sigma^{(n)} \sim \text{Unif}(\text{Sym}(n))$ . 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.**  $X_n \Rightarrow X_\infty \sim \text{Pois}(1)$ .

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

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

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

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

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

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

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

## 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

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

Thus, we are not surprise to find:

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

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*  $\mathbf{X}_n \in \mathbb{R}^d$ . 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  $\mathbf{X} \in \mathbb{R}^d$  random and  $\mathbf{t} \in \mathbb{R}^d$ , we define the *characteristic function*

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

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

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

Note here that  $\mathbf{X}_n \Rightarrow \mathbf{X}_\infty$  means that  $\mathbb{E}f(\mathbf{X}_n) \rightarrow \mathbb{E}f(\mathbf{X}_\infty)$  for all bounded continuous  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , 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 appear 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  $\langle \mathbf{t}, \mathbf{X} \rangle$  independently (and similarly for  $\mathbf{Y}$ ,  $\mathbf{X}_n$ , and  $\mathbf{X}_\infty$  in each statement above). Thus, through the intermediate step 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:

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

### 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  $\mathbf{X} \in \mathbb{R}^d$ , we define its *covariance matrix* to be

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

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.**  $\text{Var}[\langle \mathbf{t}, \mathbf{X} \rangle] = \mathbf{t}^\top \text{Cov}[\mathbf{X}] \mathbf{t}$ .

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

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

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

We now state and prove the multivariate central limit theorem.

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

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

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

### 2.3.3 APPLICATION: $\chi^2$ 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  $\text{Mult}(n; p_1, \dots, p_k)$  for the law of  $\mathbf{X} \in \mathbb{Z}_{\geq 0}^k$  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  $\mathbf{X} \sim \text{Mult}(n; p_1, \dots, p_k)$ , then  $\text{Law}(X_i) = \text{Bin}(n, p_i)$ , so in particular  $\mathbb{E}X_i = p_i n$  and  $\text{Var} X_i = p_i(1 - p_i)n$ .

The Pearson  $\chi^2$  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

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

To carry out such a hypothesis test in the style of asymptotic statistics, we must understand the asymptotic distribution of  $S(\mathbf{X})$ , 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,  $S(\mathbf{X}) \Rightarrow Z_1^2 + \dots + Z_{k-1}^2$  as  $n \rightarrow \infty$ .

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

*Proof.* Let  $e_1, \dots, e_k$  be the standard basis of  $\mathbb{R}^k$  and write  $\mu$  for the probability measure drawing  $e_i$  with probability  $p_i$ . Then, if  $v_1, \dots, v_n \sim \mu$  are i.i.d., we have  $\text{Law}(v_1 + \dots + v_n) = \text{Mult}(n; p_1, \dots, p_k)$  ( $v_i$  may be viewed as the indicator vector of the destination bin of the  $i$ th 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  $\mathbf{X}^{(n)} = \sum_{i=1}^n v_i$ . To do this, we calculate the statistics of the  $v_i$ . Let us introduce  $\mathbf{p} = (p_1, \dots, p_k) \in \mathbb{R}^k$ . We have:

$$\begin{aligned} \mathbb{E}v_i &= p_1 e_1 + \dots + p_k e_k \\ &= \mathbf{p}, \\ \text{Cov } v_i &= \mathbb{E}v_i v_i^\top - (\mathbb{E}v_i)(\mathbb{E}v_i)^\top \\ &= p_1 e_1 e_1^\top + \dots + p_k e_k e_k^\top - \mathbf{p}\mathbf{p}^\top \\ &= \text{Diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \\ &=: \Sigma. \end{aligned}$$

By the CLT, we then have

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

Thus, we also have

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

A Gaussian calculation gives:

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

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

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

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

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

## 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  $\mathcal{Y} \rightarrow \mathcal{M}(\mathcal{X})$ . 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  $\mathcal{Y} \rightarrow \mathbb{R}$ .

**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  $\mathcal{Y} = \mathbb{R} \rightarrow \mathcal{M}(\mathcal{X})$ , as above. Likewise, the associated conditional expectation is

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

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

**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

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

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

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

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, $\sigma$ -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  $\mathbb{E}[X | B]$  and  $\mathbb{E}[X | B^c]$  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  $\sigma$ -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- $\sigma$ -algebras*  $\mathcal{G} \subseteq \mathcal{F}$ , constructing the object

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

(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  $\mathbb{E}[X | \mathcal{G}]$  should mean:

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

It remains mysterious what kind of mathematical object  $\mathbb{E}[X | \mathcal{G}]$  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  $Z := \mathbb{E}[X | \mathcal{G}]$  to be a *random variable*, i.e., a function  $Z : \Omega \rightarrow \mathbb{R}$ , 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:

- $\mathcal{G} = \{\emptyset, \Omega\}$ : here,  $Z : \Omega \rightarrow \mathbb{R}$  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,  $Z : \Omega \rightarrow \mathbb{R}$  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  $Y : \Omega \rightarrow \mathbb{R}$  is a random variable and  $Z : \Omega \rightarrow \mathbb{R}$  is  $\sigma(Y)$ -measurable, then there exists an  $f : \mathbb{R} \rightarrow \mathbb{R}$  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  $Z = \mathbb{E}[X | \mathcal{G}]$  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

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

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

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  $\mathbb{E}[X | Y = y] = f(y)$  that we would like our definition of  $Z = \mathbb{E}[X | \mathcal{G}]$  for  $\mathcal{G} = \sigma(Y)$  to recover. In particular, we would like this to be  $Z = f(Y)$ , the function  $f : \mathcal{Y} \rightarrow \mathbb{R}$  evaluated on the random variable  $Y$ . The function  $f(y)$  was given by

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

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

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

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- $\sigma$ -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  $\mathbb{E}[X | \mathcal{G}] := Z$  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  $\mathbb{E}[X | \sigma(Y)]$  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:

- $\mathcal{G} = \{\emptyset, \Omega\}$ : here, on the one hand  $Z = \mathbb{E}[X | \mathcal{G}]$  must be  $\mathcal{G}$ -measurable, i.e., constant. On the other hand, taking  $A = \Omega$ , we must have  $\mathbb{E}[Z] = \mathbb{E}[Z\mathbb{1}_\Omega] = \mathbb{E}[X\mathbb{1}_\Omega] = \mathbb{E}[X]$ , 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  $\mathbb{P}[B] > 0$ ,

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

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  $\mu, \nu$  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  $f : \Omega \rightarrow \mathbb{R}$  such that  $f(\omega) \geq 0$  for all  $\omega$ , and

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

This  $f$  is unique up to  $\mu$ -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  $\nu$  with respect to  $\mu$ .

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  $\mu$  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:

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

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  $g_1, g_2, \dots \in \mathcal{F}$  that achieves the supremum, letting  $f_n := \max\{g_1, \dots, g_n\}$ , and letting  $f := \lim_{n \rightarrow \infty} f_n$ , 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  $\tilde{f} := f + \epsilon\mathbb{1}_P$ . We have  $\tilde{f} \geq 0$  and  $\tilde{f}$  is  $\mathcal{G}$ -measurable by construction, and for  $A \in \mathcal{G}$ ,

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

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

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

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,  $\int_{\Omega} \tilde{f} \, d\mu > K$ , a contradiction to the definition of  $K$ .

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

### 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  $\mu, \nu$  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  $\nu$  and  $\mu = \mathbb{P}$ , but we do so, as indicated above, with respect to the sub-algebra  $\mathcal{G} \subseteq \mathcal{F}$ . This yields  $Z : \Omega \rightarrow \mathbb{R}$  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  $\mathbb{E}[X | \mathcal{G}] := \mathbb{E}[X^+ | \mathcal{G}] - \mathbb{E}[X^- | \mathcal{G}]$  achieves the desired properties, where each of the individual conditional expectations is constructed by the above means.  $\square$

## 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  $Z = \mathbb{E}[X | \mathcal{G}]$  by checking that we indeed have

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

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- $\sigma$ -algebra, and  $a, b \in \mathbb{R}$ . Then,

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

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

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

and using (3.5.1) in each term,

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

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

*Proof.* By Proposition 3.5.1, without loss of generality, we may take  $X = 0$ . Let  $Z := \mathbb{E}[Y | \mathcal{G}]$ . 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.  $\square$

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

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

*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  $Z := X \mathbb{E}[Y | \mathcal{G}] = \mathbb{1}_B \mathbb{E}[Y | \mathcal{G}]$ . We want to show that  $Z = \mathbb{E}[XY | \mathcal{G}]$ , so we verify (3.5.1): let  $A \in \mathcal{G}$ , then we have

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

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.  $\square$

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

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

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

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

*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).  $\square$

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

$$\mathbb{E}\left[\mathbb{E}[X | \mathcal{G}] | \mathcal{H}\right] = \mathbb{E}[X | \mathcal{H}].$$

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

$$\mathbb{E}\left[\mathbb{E}[X | \mathcal{G}]\right] = \mathbb{E}[X].$$

*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,

$$\mathbb{E}\left[\mathbb{E}[X | \mathcal{G}]\mathbb{1}_A\right] = \mathbb{E}[X\mathbb{1}_A] = \mathbb{E}\left[\mathbb{E}[X | \mathcal{H}]\mathbb{1}_A\right].$$

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

**Proposition 3.5.6.** Let  $X \in L^1$  and  $f : \mathbb{R} \rightarrow \mathbb{R}$  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

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

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

## 3.6 CONDITIONAL PROBABILITY

The approach we have taken above might seem a bit unusual, since in our motivating examples conditional *probability* was the more intuitive object from which we derived conditional expectations. However, we will see here that this is a subtler matter that can be treated in different ways and in any case requires more care than conditional expectation. We just give an overview of the main ideas; a good source for details is the paper [CP97] and the textbook [Pol02, Chapter 5].

### 3.6.1 CONDITIONAL PROBABILITIES AS CONDITIONAL EXPECTATIONS

On the bright side, there is really only one reasonable definition of conditional probability: if  $A \in \mathcal{F}$  and  $\mathcal{G} \subseteq \mathcal{F}$  is a sub- $\sigma$ -algebra, we define

$$\mathbb{P}[A | \mathcal{G}] := \mathbb{E}[\mathbb{1}_A | \mathcal{G}].$$

As with all conditional expectations, this is a  $\mathcal{G}$ -measurable random variable, and comes with the caveat that it is well-defined only up to  $\mathbb{P}$ -null events.

In particular, if  $Y : \Omega \rightarrow \mathcal{Y}$  is a random variable, we write

$$\mathbb{P}[A | Y] := \mathbb{P}[A | \sigma(Y)].$$

We will mostly focus on this case. Let us review our motivating examples and see how this construction compares with those:

EXAMPLE 1: DISCRETE CASE Suppose that  $\Omega = \mathcal{X} \times \mathcal{Y}$  is finite, that  $Y(x, y) = y$ , and that

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

for a probability mass function  $p(x, y) > 0$ . Then for each  $y \in \mathcal{Y}$  the formula

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

defines a probability measure on  $\Omega$  concentrated on the *fiber* of the space of outcomes  $\{(\tilde{x}, \tilde{y}) : \tilde{y} = y\} \subseteq \Omega$ . (Previously we viewed this as a probability measure just on  $\mathcal{X}$ , but we will see that this equivalent perspective of a measure on a single fiber  $\mathcal{X} \times \{y\}$  is easier to generalize.) In particular, this is compatible with the elementary definition of conditional probability,

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

Thus, the above formula may be viewed as specifying a family of probability measures  $\mathbb{P}_y$  for each  $y \in \mathcal{Y}$  a possible value of the random variable  $Y$ .

EXAMPLE 2: CONTINUOUS CASE Suppose that  $\Omega = \mathbb{R}^2$ , that  $Y(x, y) = y$ , and that  $\mathbb{P}$  has a continuous density  $\rho(x, y) > 0$  with respect to Lebesgue measure. The elementary definition of continuous conditional probability is

$$\rho(x \mid y) := \frac{\rho(x, y)}{\int_{\mathbb{R}} \rho(x', y) dx'}.$$

Then for each  $y \in \mathbb{R}$  the formula

$$\mathbb{P}_y[A] := \int_{\mathbb{R}} \mathbb{1}_A(x, y) \rho(x \mid y) dx$$

again defines a probability measure concentrated on the fiber  $\{(\tilde{x}, \tilde{y}) : \tilde{y} = y\} \subset \Omega = \mathbb{R}^2$ . For instance, if  $A = C \times \{y\}$  for some Borel set  $C \subseteq \mathbb{R}$ , then

$$\mathbb{P}_y[A] = \int_C \rho(x \mid y) dx,$$

again recovering the familiar formula from multivariable calculus.

On the other hand, note that this definition allows us to condition on the event  $Y = y$  of measure zero, in a way that seems logical and unique. Below we will clarify when this is possible more generally.

### 3.6.2 INCOHERENCE FOR GENERAL LARGE MEASURABLE SPACES

Based on the above examples, let us summarize what we want a construction of conditional probability to behave like, versus what our construction via conditional expectation achieves.

The examples lead us to look for a family of probability measures  $(\mathbb{P}_y)_{y \in \mathcal{Y}}$ , which satisfy the following properties:

1.  $\mathbb{P}_y$  is indeed a probability measure (which we repeat here since it will become an issue below).
2.  $\mathbb{P}_y$  is supported on the fiber  $\{\omega \in \Omega : Y(\omega) = y\}$ .
3.  $\mathbb{P}_y$  respects the tower-type formula  $\mathbb{E}[\mathbb{P}_y[A]] = \mathbb{P}[A]$ , which may be verified in both of the above cases and should make intuitive sense, allowing us to “distintegrate” probabilities over the conditional probability distributions.

We will see later that, if modified a bit, such a list of conditions specifies an essentially unique family of probability measures, so let us take these natural conditions to be the criteria for a sensible construction of values of  $\mathbb{P}[A | Y = y]$ .

On the other hand, what we have actually built so far is the object  $Z = Z_A = \mathbb{P}[A | Y]$ , which is a random variable  $Z_A(\omega)$  for each  $A \in \mathcal{F}$ . Since  $Z_A$  is  $\sigma(Y)$ -measurable, Lemma 3.2.1 implies that there exists a Borel-measurable function  $f_A : \mathcal{Y} \rightarrow \mathbb{R}$  such that, almost surely,

$$Z_A = f_A(Y).$$

This suggests the provisional definition

$$\tilde{\mathbb{P}}_y[A] := f_A(y),$$

which is really the only sensible definition that the theory we have developed so far leads us to. The question remains: does this definition satisfy the above criteria?

At first glance our construction looks promising. For each fixed set  $A$ , the function  $y \mapsto \tilde{\mathbb{P}}_y[A]$  is measurable, and

$$\tilde{\mathbb{P}}_{Y(\omega)}[A] = \mathbb{P}[A | Y](\omega)$$

holds almost surely.

However, if we look into the details of the above claim we start to see an issue: the claim holds almost surely *for each*  $A$ , meaning that to each  $A \in \mathcal{F}$  there is associated a “bad set”  $N_A$  such that  $\mathbb{P}[N_A] = 0$  and such that we are only guaranteed that

$$\tilde{\mathbb{P}}_{Y(\omega)}[A] = \mathbb{P}[A | Y](\omega) \text{ for } \omega \notin N_A.$$

The problem is that if we want to argue that the  $\tilde{\mathbb{P}}_y$  have various properties for a fixed  $y$ , we must avoid *many* different sets  $N_A$ , which can become impossible if the  $\sigma$ -algebra  $\mathcal{F}$  is sufficiently rich.

As a simple example, consider trying to show that  $\tilde{\mathbb{P}}_\gamma$  is a probability measure, and in particular showing the simple property that it is additive over all disjoint pairs of sets. Fix  $A, B \in \mathcal{F}$  with  $A \cap B = \emptyset$ . We expect to be able to show

$$\tilde{\mathbb{P}}_\gamma[A \cup B] \stackrel{(?)}{=} \tilde{\mathbb{P}}_\gamma[A] + \tilde{\mathbb{P}}_\gamma[B].$$

All three of these expressions are random variables, i.e., functions of  $\omega$ . We should be able to argue as follows: if  $\omega$  is such that  $\gamma = Y(\omega)$ , then

$$\begin{aligned} \tilde{\mathbb{P}}_\gamma[A](\omega) + \tilde{\mathbb{P}}_\gamma[B](\omega) &= \tilde{\mathbb{P}}_{Y(\omega)}[A] + \tilde{\mathbb{P}}_{Y(\omega)}[B] \\ &= \mathbb{P}[A \mid Y](\omega) + \mathbb{P}[B \mid Y](\omega) && \text{(for } \omega \notin N_A \cup N_B) \\ &= \mathbb{E}[\mathbb{1}_A \mid Y](\omega) + \mathbb{E}[\mathbb{1}_B \mid Y](\omega) \\ &= \mathbb{E}[\mathbb{1}_A + \mathbb{1}_B \mid Y](\omega) && \text{(for } \omega \notin N_{A,B}^{(\text{add})}) \\ &= \mathbb{E}[\mathbb{1}_{A \cup B} \mid Y](\omega) \\ &= \mathbb{P}[A \cup B \mid Y](\omega) \\ &= \tilde{\mathbb{P}}_{Y(\omega)}[A \cup B] && \text{(for } \omega \notin N_{A \cup B}) \\ &= \tilde{\mathbb{P}}_\gamma[A \cup B] \end{aligned}$$

But, this only actually holds with the various restrictions on  $\omega$  mentioned above, where  $N_{A,B}^{(\text{add})}$  is another null set away from which additivity of the conditional expectation above holds.

Thus, the entire chain of equations above holds only provided that  $\gamma = Y(\omega)$  for some  $\omega \notin M_{A,B} := N_A \cup N_B \cup N_{A \cup B} \cup N_{A,B}^{(\text{add})}$ . We have  $\mathbb{P}[M_{A,B}] = 0$ , so for any given  $A, B$  the above indeed holds almost surely. Equivalently, we can write  $\gamma = Y(\omega) \notin M_{A,B}$  as  $\gamma \in Y^{-1}(M_{A,B}^c) =: F_{A,B}$ , for some  $F_{A,B} \subseteq \mathcal{Y}$ . Writing  $\mathbb{Q} := \text{Law}(Y)$  a probability measure on  $\mathcal{Y}$ , then we have  $\mathbb{Q}[F_{A,B}] = 1$ . In words, for any fixed  $A, B$ ,  $\tilde{\mathbb{P}}_\gamma$  is additive on  $A$  and  $B$  for  $\mathbb{Q}$ -almost every  $\gamma$ .

But now, when do we actually find that the entire function  $\tilde{\mathbb{P}}_\gamma : \Omega \rightarrow \mathbb{R}$  is additive on disjoint pairs of subsets? Since this requires quantifying over *all* pairs of subsets, it only happens on

$$\gamma \in \bigcap_{\substack{A, B \in \mathcal{F} \\ A \cap B = \emptyset}} F_{A,B} =: F^{(\text{add})}.$$

If  $\mathcal{F}$  is countable, then the above is a countable intersection of sets of full measure, so  $\mathbb{Q}[F^{(\text{add})}] = 1$ . In the same spirit we could check the remaining properties of a probability measure and we would find that, for  $\mathbb{Q}$ -almost every  $\gamma$ ,  $\tilde{\mathbb{P}}_\gamma$  in our definition is indeed a probability measure. However, once  $\mathcal{F}$  is *uncountable* (as happens in all examples with continuous random variables taking values in  $\mathbb{R}$ ), the above intersection is also uncountable, and we are not guaranteed that *any* given  $\tilde{\mathbb{P}}_\gamma$  is a probability measure!

This is the main issue that we will look for other ways to address: at a high level, when we do not have some extra “control” on the nature of the base measurable space  $(\Omega, \mathcal{F})$ , the natural construction of conditional probability via conditional expectation is not guaranteed to give rise to a sensible family of objects. That is precisely the gap that the notion of regular conditional probability is designed to fill.

### 3.6.3 REGULAR CONDITIONAL PROBABILITY

The basic idea is that the above issues can be addressed if we make some additional structural assumptions on  $(\Omega, \mathcal{F})$ . These assumptions will be of a *topological* nature: they will ask for some additional structure beyond just the very “loose” one of being a measurable space. To set up the general statements, let  $(\mathcal{Y}, \mathcal{B})$  be a measurable space, and let  $Y$  be a  $\mathcal{Y}$ -valued random variable as above, i.e., a measurable  $Y : \Omega \rightarrow \mathcal{Y}$ . We write  $\mathbb{Q} := \text{Law}(Y)$ , a probability measure on  $\mathcal{Y}$ . The following encodes a small variation on our criteria listed before:

**Definition 3.6.1** (Strong regular conditional probability). A family  $(\mathbb{P}_y)_{y \in \mathcal{Y}}$  is called a (*strong*) *regular conditional probability* of  $\mathbb{P}$  given  $Y$  if the following hold:

1. For each  $y \in \mathcal{Y}$ ,  $\mathbb{P}_y$  is a probability measure on  $(\Omega, \mathcal{F})$ .
2. For  $\mathbb{Q}$ -almost every  $y \in \mathcal{Y}$ , we have

$$\mathbb{P}_y[Y = y] = 1.$$

3. For every bounded measurable function  $f : \Omega \rightarrow \mathbb{R}$ , the map

$$y \mapsto \int_{\Omega} f(\omega) d\mathbb{P}_y(\omega)$$

is  $\mathcal{B}$ -measurable.

4. For every bounded measurable function  $f : \Omega \rightarrow \mathbb{R}$ ,

$$\int_{\Omega} f(\omega) d\mathbb{P}(\omega) = \int_{\mathcal{Y}} \left( \int_{\Omega} f(\omega) d\mathbb{P}_y(\omega) \right) d\mathbb{Q}(y).$$

When such a family exists (and is “unique enough,” which we will return to below), we are justified in writing

$$\mathbb{P}[A \mid Y = y] := \mathbb{P}_y[A]$$

and, more generally,

$$\mathbb{E}[X \mid Y = y] := \int_{\Omega} X(\omega) d\mathbb{P}_y(\omega).$$

Condition 2 says that  $\mathbb{P}_y$  is concentrated on the fiber  $\{Y = y\}$ , at least for  $\mathbb{Q}$ -almost every  $y$ . Note that some caveat like this is necessary, since if  $Y(\omega)$  never equals  $y$ , then this fiber is empty, while we still want  $\mathbb{P}_y$  have some set on which it is allowed to be supported. Condition 4 says that  $\mathbb{P}$  may be recovered by first drawing  $y$  from  $\mathbb{Q}$  and then drawing  $\omega$  from  $\mathbb{P}_y$ , a kind of reverse Fubini theorem. This is why the above type of construction is also sometimes called *disintegrating*  $\mathbb{P}$  into the measures  $\mathbb{P}_y$ .

The definition has been formulated using bounded measurable test functions because this is the cleanest way to package both measurability and the disintegration formula. Taking  $f(\omega) = \mathbb{1}_A(\omega)$  immediately gives the following.

**Proposition 3.6.2.** If  $(\mathbb{P}_y)_{y \in \mathcal{Y}}$  is a strong regular conditional probability of  $\mathbb{P}$  given  $Y$ , then, for every  $A \in \mathcal{F}$ :

1. The map  $y \mapsto \mathbb{P}_y[A]$  is  $\mathcal{B}$ -measurable.
2. We have

$$\mathbb{P}[A] = \int_{\mathcal{Y}} \mathbb{P}_y[A] d\mathbb{Q}(y).$$

Further, when a strong regular conditional probability exists, it almost surely coincides with our construction of conditional probability from conditional expectation; thus, it is indeed a stronger or “more regular” restricted version of that same construction.

**Proposition 3.6.3.** Let  $(\mathbb{P}_y)_{y \in \mathcal{Y}}$  be a strong regular conditional probability of  $\mathbb{P}$  given  $Y$ . Let  $A \in \mathcal{F}$ . Define

$$\begin{aligned} g_A(y) &:= \mathbb{P}_y[A], \\ G_A(\omega) &:= g_A(Y(\omega)). \end{aligned}$$

Then,  $G_A = \mathbb{P}[A | Y]$  almost surely. More generally, for every bounded measurable  $h : \mathcal{Y} \rightarrow \mathbb{R}$ , almost surely

$$\mathbb{E}[\mathbb{1}_A h(Y)] = \int_{\mathcal{Y}} h(y) \mathbb{P}_y[A] d\mathbb{Q}(y).$$

The proof is a straightforward verification of the defining property of  $\mathbb{P}[A | Y]$ .

In summary, we have at least found the right *definition* of a stronger notion of conditional probability: if we can construct a strong regular conditional probability, then it both almost surely coincides with our first idea for each test set  $A$ , and modifies it within those constraints to behave more sensibly as a family of probability measures.

### 3.6.4 DISINTEGRATION THEOREM

The remaining question is whether such families exist and are unique. The answer is *yes* under mild topological hypotheses on  $(\Omega, \mathcal{F})$ , but *no* in complete generality: there do exist probability spaces  $(\Omega, \mathcal{F}, \mathbb{P})$  and measurable maps  $Y$  for which no strong regular conditional probability exists. So, some extra structure really is necessary. The following is one statement of an existence and uniqueness result.

**Theorem 3.6.4** (Disintegration theorem). Suppose that the following hold.

1.  $\Omega$  is a metric space and  $\mathcal{F}$  is its Borel  $\sigma$ -algebra.
2. There exist finite non-negative measures  $\mu_1, \mu_2, \dots$  on  $(\Omega, \mathcal{F})$ , each with compact

support, such that

$$\mathbb{P} = \sum_{n=1}^{\infty} \mu_n.$$

3. The *graph* set

$$\text{Graph}(Y) := \{(\omega, \mathcal{Y}) \in \Omega \times \mathcal{Y} : \mathcal{Y} = Y(\omega)\}$$

belongs to  $\mathcal{F} \otimes \mathcal{B}$ .

Then, there exists a strong regular conditional probability  $(\mathbb{P}_{\mathcal{Y}})_{\mathcal{Y} \in \mathcal{Y}}$  of  $\mathbb{P}$  given  $Y$ . Moreover, if  $(\mathbb{P}_{\mathcal{Y}})_{\mathcal{Y} \in \mathcal{Y}}$  and  $(\mathbb{P}'_{\mathcal{Y}})_{\mathcal{Y} \in \mathcal{Y}}$  are two such families, then, for  $\mathbb{Q}$ -almost every  $\mathcal{Y}$ , we have  $\mathbb{P}_{\mathcal{Y}} = \mathbb{P}'_{\mathcal{Y}}$ .

We will not prove this theorem here; see [Pol02, Chapter 5] for more detailed discussion. The point is that, once one assumes enough regularity of the underlying measurable spaces, the informal notation  $\mathbb{P}[A \mid Y = \mathcal{Y}]$  can indeed be made rigorous.

You may also verify that this recovers the elementary constructions discussed above both in the discrete and continuous cases. Again, see [Pol02, Chapter 5] for many more examples, including ones in more applied statistical contexts.

### 3.6.5 WARNINGS

There are two basic warnings to keep in mind when venturing into working with regular conditional probabilities and conditioning on events of measure zero:

1. The notation  $\mathbb{P}[A \mid B]$  for an event  $B$  with  $\mathbb{P}[B] = 0$  is *not* meaningful on its own. It only becomes sensible when  $B$  is understood as a fiber  $\{Y = \mathcal{Y}\}$  of some associated random variable  $Y$ , together with a coherent family of probability measures  $(\mathbb{P}_{\mathcal{Y}})$  on the fibers  $\{Y = \mathcal{Y}\}$ . You may look up the famous *Borel-Kolmogorov paradox* for an example of the perils of not being careful about this.
2. Even when a strong regular conditional probability exists, the quantity  $\mathbb{P}[A \mid Y = \mathcal{Y}]$  is, in the theory we have seen, only defined for  $\mathbb{Q}$ -almost every  $\mathcal{Y}$ . At a single exceptional value of  $\mathcal{Y}$ , one may modify  $\mathbb{P}_{\mathcal{Y}}$  arbitrarily without changing any of the properties discussed above. This is why uniqueness in Theorem 3.6.4 is only a  $\mathbb{Q}$ -almost-everywhere statement.

With regard to Warning 2, you might notice that this still slightly violates the nice behavior of our example of continuous conditional probability, where we really could make sense of, say,  $\mathbb{P}[A \mid Y = 0]$  for a real-valued  $Y$  in a coherent way. To formalize this kind of refinement, the right idea is to ask for some kind of continuity of the map  $\mathcal{Y} \mapsto \mathbb{P}_{\mathcal{Y}}$ . For one approach to this in some generality, see [Tju75].

## 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  $\sigma$ -algebras describing the “information” available at time  $n$ .

**Definition 4.1.1** (Filtration). A *filtration*  $(\mathcal{F}_n)_{n \geq 0}$  of a  $\sigma$ -algebra  $\mathcal{F}$  is a nested sequence  $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}_n \subseteq \dots \subseteq \mathcal{F}$  of sub- $\sigma$ -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.  $\mathbb{E}[M_{n+1} | \mathcal{F}_n] = M_n$  almost surely for all  $n \geq 0$ .

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

**Remark 4.1.3.** A sub/super/martingale also automatically has the same property with respect to the filtration  $\tilde{\mathcal{F}}_n := \sigma(M_0, \dots, M_n)$ , 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  $\mathbb{E}[M_{n+k} | \mathcal{F}_n] = M_n$  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.  $(\Delta_n)$  is adapted to  $\mathcal{F}_n$ .
2.  $\Delta_n \in L^1$  for all  $n \geq 1$ .
3.  $\mathbb{E}[\Delta_{n+1} | \mathcal{F}_n] = 0$  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  $\mathbb{E}X_i = 0$ , 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

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

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  $M_0 = 1$ . Then,  $(M_n)$  is a martingale with respect to the same filtration as above. Indeed, we have:

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

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  $X_i \sim \text{Unif}(\{0, 2\})$ . Then, you may check that  $M_n \rightarrow 0$  almost surely. In particular then, we have

$$0 = \mathbb{E} \lim_{n \rightarrow \infty} M_n \neq \lim_{n \rightarrow \infty} \mathbb{E} M_n = 1,$$

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  $M_n := \mathbb{E}[X | \mathcal{F}_n]$  is always a martingale, a simple consequence of the tower property. If further we choose  $\mathcal{F}_0 = \{\emptyset, \Omega\}$  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 random 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.  $\square$

**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  $X_i \sim \text{Unif}(\{\pm 1\})$  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  $h_n : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ . 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  $i$ th 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

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

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

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

The sequence of outcomes is

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

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

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

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:  $\mathbb{E}M_n = 1$ ,  $M_n \rightarrow 0$  almost surely, and so

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

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

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

### 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  $X_i \sim \text{Unif}(\{\pm 1\})$ , 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

$$T := \min\{n : X_n = +1\},$$

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

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).  $T : \Omega \rightarrow \mathbb{Z}_{\geq 0} \cup \{+\infty\}$  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  $X_i \sim \text{Unif}(\{\pm 1\})$  (or any other distribution):

$$T := \min\{n : S_n = a\}, \quad (\text{hitting time})$$

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

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

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

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.  $\square$

**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.  $\square$

**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

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

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

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

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  $1 = \mathbb{E}M_T \neq \mathbb{E}M_0 = 0$ .

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

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  $M_{T \wedge n} \rightarrow M_T$  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  $\mathbb{E}M_T = \lim_{n \rightarrow \infty} \mathbb{E}M_{T \wedge n}$ . 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  $(M_{T \wedge n} - M_0) \rightarrow (M_T - M_0)$  almost surely. We also can bound

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

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

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

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.  $\square$

### 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  $X_i \sim \text{Unif}(\{\pm 1\})$  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  $\mathbb{E}T = \infty$ .  $\square$

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  $X_{k+1} = \dots = X_{k+a+b} = +1$ , 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

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

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

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

then the above observation gives that

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

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

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  $a, b > 0$ ,

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

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.  $\square$

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.  $\square$

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

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

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

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

and so  $\tilde{M}_n$  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  $\tilde{M}_n = M_{T \wedge n}$  and the stopping time  $T$ . This gives

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

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.  $\square$

### 4.3.3 APPLICATION: ENUMERATION PROBLEMS AND CAYLEY'S THEOREM

Another charming application of the optional stopping theorem is due to [Wäs20] and gives an unconventional proof of the following classical theorem in combinatorics; see the reference for more applications.

**Theorem 4.3.11** (Cayley). The number of trees on vertex set  $[n]$  (up to isomorphism as labelled graphs) is  $n^{n-2}$ .

*Proof.* Consider the following one-player game, that [Wäs20] calls *padlock solitaire*. We have  $n$  boxes, each with a lock and a key. Key  $i$  opens Box  $i$ . All boxes start out locked. We keep Key 1, and place each Key 2 up to Key  $n$  into a uniformly random box from Box 1 through Box  $n$ .

Thus, at the beginning, we can open Box 1. It might contain some keys, letting us open more boxes. We win the game if we can eventually retrieve all  $n$  keys (or equivalently eventually open every box), and lose otherwise. This outcome is random depending on the distribution of locks in boxes. Note that the total number of possible configurations of locks and boxes is  $n^{n-1}$ , since we place  $n - 1$  locks each into a random one of  $n$  boxes.

Associated to a given configuration of locks and boxes, we can draw a “dependency graph” on  $[n]$ : if Box  $i$  contains Key  $j$ , then we draw a directed edge from  $i$  to  $j$ , representing that opening Box  $i$  allows us to open Box  $j$  later as well. The structure of this graph is that Vertex 1 has no inbound edges, since we have kept Key 1 at the beginning. Every other vertex has exactly one inbound edge. A winning configuration is exactly one where there is a directed path from Vertex 1 to every other vertex. On the other hand, by the above constraints, the connected component of Vertex 1 is a directed tree, rooted at Vertex 1, with all edges directed away from that root. Thus, if  $T(n)$  is the number of labelled trees we want to count, then we have exactly

$$\mathbb{P}[\text{winning}] = \frac{T(n)}{n^{n-1}},$$

and thus

$$T(n) = n^{n-1} \cdot \mathbb{P}[\text{winning}],$$

so calculating the probability of winning will precisely solve our problem.

Let us fix a specific order in which we open the boxes: at each step, if we have any unused keys, we open the box associated to the key of lowest index that we have (e.g., if Box 1 contains Keys 4, 6, and 7, then in the first step we open Box 1, and then open Box 4). Let  $S_j \subseteq [n]$  be the (random) set of indices of keys we have collected after we have opened  $j$  boxes, and  $K_j = |S_j|$  be the number of such keys, with  $S_0 = \emptyset$  and so  $K_0 = 0$ . Then, define

$$M_j := \frac{n - K_j}{n - j} = \frac{1}{n - j} \sum_{i=1}^n \mathbb{1}\{i \notin S_j\}.$$

Since at time  $j$  we have opened  $j$  boxes, this is the number of keys we do not have divided by the number of boxes we have not opened (in which all of the keys we do not have are contained), or the average number of missing keys per unopened box.

Let  $\mathcal{F}_j := \sigma(S_0, \dots, S_j)$ . We then claim that  $(M_j)$  is a martingale with respect to the filtration  $(\mathcal{F}_j)$ . Indeed, we have

$$\begin{aligned}\mathbb{E}[M_j | \mathcal{F}_{j-1}] &= \frac{1}{n-j} \sum_{i=1}^n \mathbb{E}[\mathbb{1}\{i \notin S_j\} | \mathcal{F}_{j-1}] \\ &= \frac{1}{n-j} \sum_{i=1}^n \mathbb{P}[i \notin S_j | \mathcal{F}_{j-1}]\end{aligned}$$

Since  $S_j \supseteq S_{j-1}$ , for  $i \in S_{j-1}$  this probability is zero, and we may restrict to

$$= \frac{1}{n-j} \sum_{i \in [n] \setminus S_{j-1}} \mathbb{P}[i \notin S_j | \mathcal{F}_{j-1}]$$

But now, before opening the  $j$ th box, each key  $i \notin S_{j-1}$  that we do not have is located in a uniformly random one of the  $n - (j - 1)$  unopened boxes, and so

$$\begin{aligned}&= \frac{1}{n-j} \sum_{i \in [n] \setminus S_{j-1}} \left(1 - \frac{1}{n - (j - 1)}\right) \\ &= \frac{1}{n-j} \cdot (n - K_{j-1}) \cdot \left(1 - \frac{1}{n - (j - 1)}\right) \\ &= \frac{1}{n-j} \cdot (n - K_{j-1}) \cdot \left(\frac{n-j}{n - (j - 1)}\right) \\ &= \frac{n - K_{j-1}}{n - (j - 1)} \\ &= M_{j-1}.\end{aligned}$$

Now let us describe the stopping time  $T$  when we are forced to stop playing. This happens either when we have collected all of the keys,  $K_j = n$ , or when the number of boxes we have opened equals the number of keys we have,  $K_j = j$ , so there are no more boxes to open. In the first case we have won and we have  $M_j = 0$ , while in the second we have lost and have  $M_j = 1$ . Thus, since  $M_0 = \frac{n-1}{n}$ , by Theorem 4.3.5 we have

$$\frac{n-1}{n} = \mathbb{E}M_0 = \mathbb{E}M_T = 1 - \mathbb{P}[\text{winning}],$$

thus we find

$$\mathbb{P}[\text{winning}] = 1 - \frac{n-1}{n} = \frac{1}{n},$$

and the result follows. □

## 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 martingale 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  $M_n \rightarrow M_\infty$  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  $M_n \rightarrow M_\infty$  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  $M_\infty$  satisfying the needed condition. Further, by Fatou's Lemma we have, again using the boundedness,

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

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  $\mathbb{E}M_\infty = \mathbb{E} \lim_{n \rightarrow \infty} M_n = \lim_{n \rightarrow \infty} \mathbb{E}M_n$ , 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  $M_n \rightarrow M_\infty$ , 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

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

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  $\mathbb{P}[\lim_{n \rightarrow \infty} M_n \text{ exists}] = 1$ , where we allow for this limit to be  $\pm\infty$ .

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

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

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

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.  $\square$

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  $\lim_{n \rightarrow \infty} M_n =: M_\infty$  exists almost surely, though *a priori* it may be  $\pm\infty$ . But, by Fatou's Lemma we also have

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

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

#### 4.4.2 $L^1$ CONVERGENCE AND UNIFORM INTEGRABILITY

We recall the definition of  $L^p$  norms of random variables, which we will use throughout the next few sections:

**Definition 4.4.6.** For  $X$  a random variable and  $p \geq 1$ , we write

$$\|X\|_{L^p} := (\mathbb{E}|X|^p)^{1/p}$$

and  $L^p = \{X : \|X\|_p < \infty\}$ . We say that  $X_n \xrightarrow{L^p} X_\infty$  if  $\|X_n - X_\infty\|_{L^p} \rightarrow 0$ .

As we saw above, even when  $\lim_{n \rightarrow \infty} M_n$  exists almost surely, Theorem 4.4.1 does not distinguish whether we have the convergence of expectations

$$\mathbb{E} \lim_{n \rightarrow \infty} M_n \stackrel{(?)}{=} \lim_{n \rightarrow \infty} \mathbb{E}M_n.$$

We will probe this question instead from the point of view of modes of convergence, asking when beyond convergence almost surely we have convergence in  $L^p$  for various  $p$ . The following justifies why this is just a slightly stronger version of the above question, and we will see more connections below.

**Proposition 4.4.7.** If  $X_1, X_2, \dots, X_\infty \in L^1$  and  $X_n \xrightarrow{L^1} X_\infty$ , then  $\mathbb{E}X_n \rightarrow \mathbb{E}X_\infty$ .

*Proof.* We have  $|\mathbb{E}X_n - \mathbb{E}X_\infty| \leq \mathbb{E}|X_n - X_\infty| = \|X_n - X_\infty\|_{L^1} \rightarrow 0$ .  $\square$

In fact, for the case of  $L^1$  convergence, we can use tools mostly unrelated to martingales to understand the stronger mode of convergence quite precisely.

**Definition 4.4.8.**  $(X_n)$  are *uniformly integrable (UI)* if

$$\limsup_{t \rightarrow \infty} \sup_n \mathbb{E} \left[ |X_n| \cdot \mathbb{1}_{\{|X_n| \geq t\}} \right] = 0.$$

One way to think of the UI condition is as being stronger than random variables being bounded in  $L^1$ , but weaker than random variables being dominated uniformly by a single  $L^1$  random variable, as the following propositions show.

**Proposition 4.4.9.** If  $(X_n)$  are UI then  $\sup_n \|X_n\|_{L^1} < \infty$ .

*Proof.* We have  $\|X_n\|_{L^1} \leq t + \mathbb{E}[|X_n| \cdot \mathbb{1}_{\{|X_n| \geq t\}}]$  for all  $t > 0$ , and the right-hand side must be finite for some  $t > 0$  by the UI assumption.  $\square$

**Proposition 4.4.10.** If there exists  $Z \in L^1$  such that  $|X_n| \leq Z$  almost surely for all  $n \geq 1$ , then  $(X_n)$  are UI.

*Proof.* Without loss of generality  $Z \geq 0$  almost surely. We have  $\mathbb{E}[|X_n| \cdot \mathbb{1}_{\{|X_n| \geq t\}}] \leq \mathbb{E}[Z \cdot \mathbb{1}_{\{Z \geq t\}}]$ . It remains to show that the right-hand side goes to 0 as  $t \rightarrow \infty$ , a general fact about integrable random variables, proved as follows. Write  $Z^{(t)} := Z \cdot \mathbb{1}_{\{Z \geq t\}}$ . We have  $Z^{(t)} \rightarrow 0$  almost surely as  $t \rightarrow \infty$ , and  $Z^{(t)} \leq Z \in L^1$ . Thus, by the dominated convergence theorem  $\mathbb{E}Z^{(t)} \rightarrow 0$  as required.  $\square$

Thus the following result, which we will not prove here, should not be entirely surprising, saying that in fact the UI condition is the “right” weakest condition under which we can apply the dominated convergence theorem. See [Kle14, Section 6.2] or [Pol02, Section 2.8] for many more details.

**Theorem 4.4.11** (“Strong dominated  $L^1$  convergence”). Let  $X_1, X_2, \dots, X_\infty \in L^1$  and suppose that  $X_n \rightarrow X_\infty$  almost surely. Then,  $X_n \xrightarrow{L^1} X_\infty$  if and only if  $(X_n)$  are UI.

Note that, combined with Proposition 4.4.10, this indeed implies the usual dominated convergence theorem, hence the name.

Applying this to martingales, we find the following.

**Theorem 4.4.12** ( $L^1$  martingale convergence). Suppose  $(M_n)$  is a sub/supermartingale that is UI. Then, there exists a random variable  $M_\infty \in L^1$  such that  $M_n \rightarrow M_\infty$  almost surely and in  $L^1$ .

*Proof.* By Proposition 4.4.9,  $\sup_n \|M_n\|_{L^1} < \infty$  and thus Theorem 4.4.1 applies to show that there exists  $M_\infty \in L^1$  such that  $M_n \rightarrow M_\infty$  almost surely. Then, Theorem 4.4.11 implies convergence in  $L^1$ .  $\square$

It is natural to wonder how useful this result actually is, since it is not clear how to conveniently check UI for concrete sequences of random variables. Below we describe a few criteria that can be used in general or for martingales specifically.

**Lemma 4.4.13.** Suppose  $f : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  has  $\lim_{x \rightarrow \infty} \frac{f(x)}{x} = \infty$ . If  $\mathbb{E}f(|X_n|)$  is bounded, then  $(X_n)$  are UI.

*Proof.* We may bound

$$\begin{aligned} \mathbb{E}[|X_n| \cdot \mathbb{1}\{|X_n| \geq t\}] &= \mathbb{E}\left[\frac{|X_n|}{f(|X_n|)} \cdot \mathbb{1}\{|X_n| \geq t\} \cdot f(|X_n|)\right] \\ &\leq \left(\sup_{x \geq t} \frac{x}{f(x)}\right) \cdot \mathbb{E}[f(|X_n|)], \end{aligned}$$

and the second factor is bounded while the first goes to 0 as  $t \rightarrow \infty$  by assumption.  $\square$

**Corollary 4.4.14.** If  $\sup_n \|X_n\|_{L^p} < \infty$  for any  $p > 1$ , then  $(X_n)$  are UI.

*Proof.* Apply Lemma 4.4.13 with  $f(x) = x^p$ .  $\square$

This is perhaps the most directly useful application, but others come in handy as well, for instance the choice  $f(x) = x \log(1 + x)$ .

For martingales, a special property leads to the following useful condition.

**Corollary 4.4.15.** Suppose  $(M_n)$  is a martingale with  $M_0 \in L^2$  and  $\sum_{i=1}^{\infty} \mathbb{E}(M_i - M_{i-1})^2 = \sum_{i=1}^{\infty} \|M_i - M_{i-1}\|_{L^2}^2 < \infty$ . Then,  $\sup_n \|M_n\|_{L^2} < \infty$ , and in particular  $(M_n)$  is UI.

*Proof.* We have  $M_n = M_0 + \sum_{i=1}^n (M_i - M_{i-1})$ , so from the assumption it follows that  $M_n \in L^2$  for all  $n \geq 0$ . The result follows by expanding  $\mathbb{E}M_n^2 = \|M_n\|_{L^2}^2$  and using that, when  $M_n \in L^2$  for all  $n \geq 0$ , we also have the property of *uncorrelated increments*,

$$\mathbb{E}(M_d - M_c)(M_b - M_a) = 0 \text{ for all } 0 \leq a \leq b < c \leq d,$$

which you may check as a simple exercise with the tower property.  $\square$

### 4.4.3 $L^p$ CONVERGENCE AND MAXIMAL INEQUALITIES

The next result we show addresses the same question as above but for convergence in  $L^p$ . In fact, the main result here is perhaps cleaner, not requiring us to introduce an extra condition like uniform integrability.

**Theorem 4.4.16** ( $L^p$  martingale convergence). Suppose  $(M_n)$  is a martingale,  $p > 1$ , and  $\sup_n \|M_n\|_{L^p} < \infty$ . Then, there exists  $M_\infty \in L^p$  such that  $M_n \rightarrow M_\infty$  almost surely and in  $L^p$ .

As one important detail, note that here it is important that we assume that  $M_n$  is specifically a *martingale*, not just a sub/supermartingale.

As we will see in a moment, this follows essentially immediately from the following also important inequality.

**Theorem 4.4.17** ( $L^p$  martingale maximal inequality). Suppose  $(M_n)$  is a submartingale,  $M_n \geq 0$  almost surely, and  $p > 1$ . Define

$$M_n^* := \max_{0 \leq k \leq n} M_k.$$

Then,

$$\|M_n^*\|_{L^p} \leq \frac{p}{p-1} \|M_n\|_{L^p}.$$

The following will be an important tool in the proof of Theorem 4.4.16 from Theorem 4.4.17, and also will let us clarify part of the statement of the latter.

**Proposition 4.4.18.** Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be convex. Then, the following hold:

1. If  $(M_n)$  is a martingale and  $f(M_n) \in L^1$  for each  $n$ , then  $(f(M_n))$  is a submartingale.
2. If  $(M_n)$  is a submartingale,  $f(M_n) \in L^1$ , for each  $n$ , and  $f$  is non-decreasing, then  $(f(M_n))$  is a submartingale.

*Proof.* The first property is immediate by Jensen's inequality for conditional expectation: almost surely, we have

$$\mathbb{E}[f(M_n) \mid \mathcal{F}_{n-1}] \geq f(\mathbb{E}[M_n \mid \mathcal{F}_{n-1}]) = f(M_{n-1}).$$

For the second property, we use the same argument but at the end must use the monotonicity of  $f$  that we add as an extra assumption, since we only have  $\mathbb{E}[M_n \mid \mathcal{F}_{n-1}] \geq M_{n-1}$ .  $\square$

Then, the following shows that the assumption that  $M_n \geq 0$  in Theorem 4.4.17 is without loss of generality.

**Proposition 4.4.19.** If  $(M_n)$  is a submartingale, then  $(M_n^+) = (\max\{0, M_n\})$  is a submartingale as well.

*Proof.* Use Proposition 4.4.18 with  $f(x) = \max\{0, x\}$ .  $\square$

We now prove the main result, Theorem 4.4.16, using Theorem 4.4.17, whose proof we give afterwards.

*Proof of Theorem 4.4.16.* By Jensen's inequality,  $\sup_n \|M_n\|_{L^1} \leq \sup_n \|M_n\|_{L^p} < \infty$ , so Theorem 4.4.1 applies to give that  $M_n \rightarrow M_\infty \in L^1$  almost surely (Theorem 4.4.12 may also be applied via Corollary 4.4.14 to show convergence in  $L^1$  immediately, though that will not directly help us here). Thus, it remains to upgrade this convergence to convergence in  $L^p$ .

The idea is to use the dominated convergence theorem. Write  $\Delta_n := |M_n - M_\infty|^p$ , then  $\Delta_n \rightarrow 0$  almost surely and it suffices to construct  $Z \in L^1$  such that  $\Delta_n \leq Z$  for each  $n$ . We use

$$\Delta_n \leq (\sup_n |M_n| + |M_\infty|)^p \leq 2^p (\sup_n |M_n|)^p =: Z.$$

Thus it is enough to show  $Z \in L^1$ , for which it is enough in turn to show that  $M^* := \sup_n |M_n| \in L^p$ .

Let us write, as before,

$$M_n^* := \max_{0 \leq k \leq n} |M_k|.$$

Then, we have that  $|M_n^*|^p \geq 0$  and these random variables increase almost surely to  $|M^*|^p$ . So, by the monotone convergence theorem, it suffices to show that  $\sup_n \mathbb{E}|M_n^*|^p < \infty$ . But, by Theorem 4.4.17 together with Proposition 4.4.18 (on the convex function  $f(x) = |x|^p$ ), we have  $\mathbb{E}|M_n^*|^p = \|M_n^*\|_{L^p}^p \leq (\frac{p}{p-1})^p \|M_n\|_{L^p}^p$ , which is bounded over all  $n$  by assumption, completing the proof.  $\square$

It remains to prove the  $L^p$  maximal inequality, Theorem 4.4.17. To do this, we will produce a few preliminary tools.

**Proposition 4.4.20.** Suppose that  $M_n$  is a submartingale and  $T$  is a stopping time. Then, for any  $n \geq 0$ ,

$$\mathbb{E}M_0 \leq \mathbb{E}M_{T \wedge n} \leq \mathbb{E}M_n.$$

*Proof.* The first inequality follows since  $M_{T \wedge n}$  is a submartingale. For the second, define the non-negative predictable process  $H_n := \mathbb{1}\{T < n\}$ . Then,  $H \bullet M$  is a submartingale, and we have  $(H \bullet M)_0 = 0$  while  $(H \bullet M)_n = M_n - M_{T \wedge n}$ . Thus,

$$0 = \mathbb{E}(H \bullet M)_0 \leq \mathbb{E}(H \bullet M)_n = \mathbb{E}[M_n] - \mathbb{E}[M_{T \wedge n}],$$

giving the result.  $\square$

The following is another form of maximal inequality, that can be viewed as a ‘‘maximal Markov inequality’’ to which we will reduce our  $L^p$  maximal inequality. It has uses beyond this, one of which we discuss in Section 4.5.4 below in the context of concentration inequalities.

**Lemma 4.4.21** (Doob maximal inequality). Suppose  $(M_n)$  is a submartingale with  $M_n \geq 0$  almost surely, as before write

$$M_n^* := \max_{0 \leq k \leq n} M_k,$$

and let  $t > 0$ . Then,

$$\mathbb{P}[M_n^* \geq t] \leq \frac{\mathbb{E}[M_n \mathbb{1}\{M_n^* \geq t\}]}{t} \leq \frac{\mathbb{E}M_n}{t}.$$

*Proof.* Define the stopping time  $T := \min\{n : M_n \geq t\}$ . By Proposition 4.4.20,  $\mathbb{E}M_{T \wedge n} \leq \mathbb{E}M_n$ . Let us expand each of these expectations according to whether  $T$  has happened yet. We have

$$\begin{aligned} \mathbb{E}M_n &= \mathbb{E}M_n \mathbb{1}\{M_n^* \geq t\} + \mathbb{E}M_n \mathbb{1}\{M_n^* < t\}, \\ \mathbb{E}M_{T \wedge n} &= \mathbb{E}M_{T \wedge n} \mathbb{1}\{M_n^* \geq t\} + \mathbb{E}M_{T \wedge n} \mathbb{1}\{M_n^* < t\} \end{aligned}$$

and here in the first expectation we have  $M_{T \wedge n} = M_T \geq t$  by definition, while in the second we have  $M_{T \wedge n} = M_n$ . Thus,

$$\begin{aligned} &= \mathbb{E}M_T \mathbb{1}\{M_n^* \geq t\} + \mathbb{E}M_n \mathbb{1}\{M_n^* < t\} \\ &\geq t \mathbb{P}[M_n^* \geq t] + \mathbb{E}M_n \mathbb{1}\{M_n^* < t\}. \end{aligned}$$

Thus  $\mathbb{E}M_{T \wedge n} \leq \mathbb{E}M_n$  implies that

$$t \mathbb{P}[M_n^* \geq t] + \mathbb{E}M_n \mathbb{1}\{M_n^* < t\} \leq \mathbb{E}M_n \mathbb{1}\{M_n^* \geq t\} + \mathbb{E}M_n \mathbb{1}\{M_n^* < t\},$$

and rearranging gives the result.  $\square$

Finally, we use Lemma 4.4.21 to prove the  $L^p$  maximal inequality, Theorem 4.4.17, thereby also completing the proof of Theorem 4.4.16.

*Proof of Theorem 4.4.17.* As before, we write  $M_n^* := \max_{0 \leq k \leq n} M_k \geq 0$ . Further, fix a  $C > 0$  and write  $X_n := M_n^* \wedge C$ . We work with this  $X_n$  in order to avoid integrability issues, and at the end take  $C \rightarrow \infty$  to recover the result. We begin by using the identity  $x^p = \int_0^x p t^{p-1} dt$ :

$$\begin{aligned} \|X_n\|_{L^p}^p &= \mathbb{E}X_n^p \\ &= \mathbb{E} \left[ p \int_0^{X_n} t^{p-1} dt \right] \\ &= p \mathbb{E} \left[ \int_0^\infty t^{p-1} \mathbb{1}\{X_n \geq t\} dt \right] \end{aligned}$$

and by Fubini's theorem on the expectation and the integral (recalling that an expectation is just another integral over a measurable space),

$$\begin{aligned} &= p \int_0^\infty \mathbb{E}[t^{p-1} \mathbb{1}\{X_n \geq t\}] dt \\ &= p \int_0^\infty t^{p-1} \mathbb{P}[X_n \geq t] dt \\ &= p \int_0^C t^{p-1} \mathbb{P}[M_n^* \geq t] dt \end{aligned}$$

and now we are in position to use Lemma 4.4.21, which gives

$$\leq p \int_0^C t^{p-2} \mathbb{E}[M_n \mathbb{1}\{M_n^* \geq t\}] dt$$

after which we reverse all of the previous steps:

$$= p \int_0^\infty t^{p-2} \mathbb{E}[M_n \mathbb{1}\{X_n \geq t\}] dt$$

using Fubini's theorem again,

$$\begin{aligned} &= p \mathbb{E} \left[ \int_0^\infty t^{p-2} M_n \mathbb{1}\{X_n \geq t\} dt \right] \\ &= p \mathbb{E} \left[ M_n \int_0^\infty t^{p-2} \mathbb{1}\{X_n \geq t\} dt \right] \\ &= p \mathbb{E} \left[ M_n \int_0^{X_n} t^{p-2} dt \right] \\ &= p \mathbb{E} \left[ M_n \frac{X_n^{p-1}}{p-1} \right] \\ &= \frac{p}{p-1} \mathbb{E}[M_n X_n^{p-1}] \end{aligned}$$

Lastly, we use Hölder's inequality to make this comparable to the expression we started with,

$$\begin{aligned} &\leq \frac{p}{p-1} (\mathbb{E}|M_n|^p)^{1/p} \mathbb{E}(|X_n|^{(p-1) \cdot \frac{p}{p-1}})^{\frac{p-1}{p}} \\ &= \frac{p}{p-1} \|M_n\|_{L^p} \|X_n\|_{L^p}^{\frac{p-1}{p}}, \end{aligned}$$

and rearranging this gives

$$\|X_n\|_{L^p} \leq \frac{p}{p-1} \|M_n\|_{L^p},$$

and finally taking  $C \rightarrow \infty$  gives the result by using the monotone convergence theorem on the left-hand side.  $\square$

## 4.5 APPLICATION: CONCENTRATION INEQUALITIES

Our first application is to *concentration inequalities*, which say that certain random variables are unlikely to deviate far from their expectations. The general phenomenon of concentration of measure may be summarized informally as follows. Suppose we have random variables  $X_i \in \mathcal{X}_i$  in some domain, and a function  $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \rightarrow \mathbb{R}$ . Under suitable conditions, we expect that

$$\mathbb{P} [ |f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n)| > t ]$$

should be small. It is reasonable to hope for this provided that:

1.  $f$  depends only weakly on any individual one of its inputs, and
2. the random variables  $X_i$  are only weakly dependent.

### 4.5.1 SUMS OF INDEPENDENT VARIABLES: Hoeffding Inequality

The most classical example takes both assumptions to an extreme: it concerns the very simple sum function  $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ , and the random  $X_i$  being exactly independent.

**Theorem 4.5.1** (Hoeffding inequality). Let  $X_1, \dots, X_n$  be independent random variables with  $X_i \in [a_i, b_i]$  almost surely. Define  $S_n := \sum_{i=1}^n X_i$ . Then, for any  $t > 0$ ,

$$\mathbb{P}[|S_n - \mathbb{E}S_n| > t] \leq 2 \exp\left(-\frac{t^2}{2\sigma_{\text{prox}}^2}\right),$$

where

$$\sigma_{\text{prox}}^2 = \frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2.$$

As an example, suppose  $X_i \sim \text{Unif}(\{\pm 1\})$ . Then  $\mathbb{E}S_n = 0$  and  $a_i = -1$ ,  $b_i = 1$ , so  $\sigma_{\text{prox}}^2 = n$ . The above bound becomes

$$\mathbb{P}[|S_n| > t] \leq 2 \exp\left(-\frac{t^2}{2n}\right),$$

or equivalently

$$\mathbb{P}\left[\left|\frac{S_n}{\sqrt{n}}\right| > s\right] \leq 2 \exp\left(-\frac{s^2}{2}\right),$$

which makes it clear how this result is a non-asymptotic tail bound version of the central limit theorem, since the right-hand side is close to the tail behavior of an  $\mathcal{N}(0, 1)$  random variable.

The basic technique behind proving this and many other concentration inequalities might be called the *Chernoff* or *nonlinear Markov method*. This involves applying Markov's inequality after a suitable transformation of a probability, and also appears in Chebyshev-like inequalities and many other settings.

**Definition 4.5.2** (Moment generating function). For a random variable  $X \in \mathbb{R}$ , we denote its *moment generating function*  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  by

$$\phi_X(\lambda) := \mathbb{E} \exp(\lambda(X - \mathbb{E}X)).$$

Centering  $X$  in the definition is unconventional but will be convenient for us. With this, for any random variable  $Y$  and  $\lambda > 0$ ,

$$\mathbb{P}[Y > t] = \mathbb{P}[\exp(\lambda Y) > \exp(\lambda t)] \leq \frac{\mathbb{E} \exp(\lambda Y)}{\exp(\lambda t)} = \frac{\phi_Y(\lambda)}{\exp(\lambda t)}$$

Thus we may bound tail probabilities by optimizing such an expression over  $\lambda$ . Particularly conveniently, the numerator factors over sums of independent random variables, making this a very useful device for sums of independent random variables.

**Proposition 4.5.3.** If  $Y_1$  and  $Y_2$  are independent, then  $\phi_{Y_1+Y_2}(\lambda) = \phi_{Y_1}(\lambda) + \phi_{Y_2}(\lambda)$ .

Building up to the proof of Theorem 4.5.1, we introduce the following class of random variables and some of its basic properties. The following is a simple exercise to verify.

**Proposition 4.5.4.** If  $X \sim \mathcal{N}(0, \sigma^2)$ , then  $\phi_X(\lambda) = \exp(\frac{\sigma^2}{2}\lambda^2)$ .

This justifies the following term:

**Definition 4.5.5** (Subgaussianity). We say that a random variable  $X$  is  $\sigma^2$ -subgaussian if, for all  $\lambda \in \mathbb{R}$ ,  $\phi_X(\lambda) \leq \exp(\frac{\sigma^2}{2}\lambda^2)$ . In this case, we call  $\sigma^2$  the *variance proxy*.

The following is then immediate from Proposition 4.5.3:

**Proposition 4.5.6.** If  $X$  and  $Y$  are independent and  $\sigma^2$ - and  $\tau^2$ -subgaussian respectively, then  $X + Y$  is  $(\sigma^2 + \tau^2)$ -subgaussian.

Finally, subgaussianity gives us the kinds of tail bounds we are trying to establish in Theorem 4.5.1.

**Proposition 4.5.7.** If  $X$  is  $\sigma^2$ -subgaussian, then, for all  $t > 0$ ,

$$\mathbb{P}[|X - \mathbb{E}X| > t] \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

*Proof.* Using the Chernoff method on one tail (the other can be bounded by a symmetric argument), we get

$$\begin{aligned} \mathbb{P}[X - \mathbb{E}X > t] &= \mathbb{P}[\exp(\lambda(X - \mathbb{E}X)) > \exp(\lambda t)] \\ &= \frac{\phi_X(\lambda)}{\exp(\lambda t)} \\ &\leq \exp\left(\frac{\sigma^2}{2}\lambda^2 - t\lambda\right) \end{aligned}$$

and choosing the minimizing value  $\lambda := t/\sigma^2$  gives

$$= \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

as claimed. □

The final ingredient is the following result, which shows that bounded random variables are automatically subgaussian.

**Lemma 4.5.8** (Hoeffding lemma). If  $X \in [a, b]$  almost surely, then  $X$  is  $\frac{1}{4}(b - a)^2$ -subgaussian.

We will prove this in a moment, but let us first see how it implies the main result.

*Proof of Theorem 4.5.1.* By Lemma 4.5.8, each  $X_i$  is  $\frac{1}{4}(b_i - a_i)^2$ -subgaussian. Since they are independent, by Proposition 4.5.6,  $S_n = \sum_{i=1}^n X_i$  is  $\sigma_{\text{prox}}^2$ -subgaussian for  $\sigma_{\text{prox}}^2 := \frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2$ . Finally, the required bound then follows by Proposition 4.5.7.  $\square$

To prove Lemma 4.5.8, we will need a few more tools. The following is a basic fact about variances of bounded random variables that may be viewed as a “baby version” of Hoeffding’s Lemma, bounding the variance by the same variance proxy that appears there.

**Proposition 4.5.9.** Suppose that  $X \in [a, b]$  almost surely. Then,  $\text{Var}[X] \leq \frac{1}{4}(b - a)^2$ .

*Proof.* We have

$$\text{Var}[X] = \text{Var} \left[ X - \frac{b-a}{2} \right] \leq \mathbb{E} \left( X - \frac{b-a}{2} \right)^2 \leq \left( \frac{b-a}{2} \right)^2,$$

giving the result.  $\square$

Next we give a sufficient condition for the property of subgaussianity in terms of a new object.

**Definition 4.5.10** (Cumulant generating function). For a random variable  $X$ , we define  $\psi_X(\lambda) := \log \phi_X(\lambda)$ .

**Proposition 4.5.11.** Suppose that  $\psi_X$  is smooth and  $\psi_X''(\lambda) \leq \sigma^2$  for all  $\lambda \in \mathbb{R}$ . Then,  $X$  is  $\sigma^2$ -subgaussian.

*Proof.* We calculate

$$\begin{aligned} \psi_X(0) &= \log \mathbb{E} \exp(0) = 0, \\ \psi_X'(0) &= \frac{\mathbb{E}(X - \mathbb{E}X) \exp(\lambda(X - \mathbb{E}X))}{\mathbb{E} \exp(\lambda(X - \mathbb{E}X))} \Big|_{\lambda=0} = 0. \end{aligned}$$

By Taylor’s theorem with a remainder estimate we then find  $\psi_X(\lambda) \leq \frac{\sigma^2}{2} \lambda^2$  for all  $\lambda \in \mathbb{R}$ , and thus  $\phi_X(\lambda) = \exp(\psi_X(\lambda)) \leq \exp(\frac{\sigma^2}{2} \lambda^2)$ , as required.  $\square$

*Proof of Lemma 4.5.8.* Without loss of generality we may assume  $\mathbb{E}X = 0$ . Suppose  $X$  is defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . For each  $\lambda \in \mathbb{R}$ , let  $\mathbb{Q}_\lambda$  be another probability measure on  $(\Omega, \mathcal{F})$  such that

$$L_\lambda(\omega) = \frac{d\mathbb{Q}_\lambda}{d\mathbb{P}}(\omega) := \frac{\exp(\lambda X(\omega))}{\mathbb{E} \exp(\lambda X(\omega))}.$$

As an exercise you may check that, since the right-hand side is a non-negative random variable of expectation 1, indeed this specifies a valid probability measure.

Now, note that the first few derivatives of  $\psi_X$  can be written in a way involving  $\mathbb{Q}_\lambda$ :

$$\begin{aligned} \psi_X(\lambda) &= \log \mathbb{E} \exp(\lambda X), \\ \psi'_X(\lambda) &= \frac{\mathbb{E} X \exp(\lambda X)}{\mathbb{E} \exp(\lambda X)} \\ &= \int X(\omega) L_\lambda(\omega) d\mathbb{P}(\omega) \\ &= \int X(\omega) d\mathbb{Q}_\lambda(\omega) \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[X], \end{aligned}$$

where we write  $\mathbb{E}_{\mathbb{Q}}$  for the expectation under a different probability measure  $\mathbb{Q}$  on the same measurable space. In the same way, we find

$$\begin{aligned} \psi''_X(\lambda) &= \frac{\mathbb{E} X^2 \exp(\lambda X)}{\mathbb{E} \exp(\lambda X)} - \left( \frac{\mathbb{E} X \exp(\lambda X)}{\mathbb{E} \exp(\lambda X)} \right)^2 \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[X^2] - (\mathbb{E}_{\mathbb{Q}_\lambda} X)^2 \\ &= \text{Var}_{\mathbb{Q}_\lambda}[X]. \end{aligned}$$

Indeed, we will not need to use it here but if you continue in the same spirit you will find that the higher derivatives of  $\psi_X$  evaluated at  $\lambda$  give the *cumulants* of  $X$  under the probability measure  $\mathbb{Q}_\lambda$ .

In particular, under  $\mathbb{Q}_\lambda$  we still have  $X \in [a, b]$  almost surely. Thus, by Proposition 4.5.9 we have  $\psi''_X(\lambda) \leq \frac{1}{4}(b-a)^2$  for all  $\lambda$ , and so by Proposition 4.5.11  $X$  is  $\frac{1}{4}(b-a)^2$ , as claimed, completing the proof.  $\square$

## 4.5.2 MARTINGALE GENERALIZATION: AZUMA INEQUALITY

Surprisingly, the proof of Hoeffding's inequality can be generalized to martingales, though it seems like it uses the linearity of the sum function in an important way.

**Theorem 4.5.12** (Azuma inequality). Let  $(M_n)$  be a martingale and suppose that its increments satisfy  $a_i \leq M_i - M_{i-1} \leq b_i$  almost surely for all  $i \geq 1$ , for constants  $a_i, b_i \in \mathbb{R}$ . Then,  $M_n - M_0$  is  $\sigma_{\text{prox}}^2$ -subgaussian with

$$\sigma_{\text{prox}}^2 = \frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2.$$

In particular,

$$\mathbb{P}[|M_n - M_0| > t] \leq 2 \exp\left(-\frac{t^2}{2\sigma_{\text{prox}}^2}\right).$$

*Proof.* Without loss of generality we may assume  $M_0 = 0$ . We expand in the usual telescoping series inside of the moment generating function:

$$\begin{aligned} \phi_{M_n}(\lambda) &= \mathbb{E} \exp(\lambda M_n) \\ &= \mathbb{E} \exp\left(\lambda \sum_{i=1}^n (M_i - M_{i-1})\right) \end{aligned}$$

and by the tower and factorization properties of conditional expectation,

$$\begin{aligned} &= \mathbb{E} \left[ \mathbb{E} \left[ \exp\left(\lambda \sum_{i=1}^n (M_i - M_{i-1})\right) \middle| \mathcal{F}_{n-1} \right] \right] \\ &= \mathbb{E} \left[ \exp\left(\lambda \sum_{i=1}^{n-1} (M_i - M_{i-1})\right) \mathbb{E} [\exp(\lambda(M_n - M_{n-1})) \mid \mathcal{F}_{n-1}] \right] \end{aligned}$$

As an exercise you may verify that there is a version of Lemma 4.5.8 for conditional expectations, which gives

$$\begin{aligned} &\leq \mathbb{E} \left[ \exp\left(\lambda \sum_{i=1}^{n-1} (M_i - M_{i-1})\right) \exp\left(\frac{\lambda^2}{2} \cdot \frac{(b_n - a_n)^2}{4}\right) \right] \\ &= \exp\left(\frac{\lambda^2}{2} \cdot \frac{(b_n - a_n)^2}{4}\right) \phi_{M_{n-1}}(\lambda). \end{aligned}$$

Repeating this inductively, we get

$$\phi_{M_n}(\lambda) \leq \exp\left(\frac{\lambda^2}{2} \cdot \frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2\right),$$

giving the result. □

### 4.5.3 FUNCTIONS OF INDEPENDENT VARIABLES: MCDIARMID INEQUALITY

We now describe one of the main applications of Azuma's inequality, to concentration of general functions of independent random variables. Let  $X_1, \dots, X_n$  be independent in arbitrary domains  $X_i \in \mathcal{X}_i$ , let  $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \rightarrow \mathbb{R}$ , and consider the random variable  $f(X_1, \dots, X_n)$ . We will show that this concentrates around its mean, provided a certain version of the principles described earlier hold. Since the  $X_i$  are exactly independent by assumption, it remains to formalize some notion of  $f$  not depending too much on any of its inputs, which we do next.

**Definition 4.5.13.** Associated to  $f$ , define the quantities

$$\delta_i(f) := \sup_{x_1, \dots, x_n, x'_i} |f(x_1, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)|.$$

In words,  $\delta_i(f)$  measures how much  $f$  can possibly change from changing input  $i$  arbitrarily. In this sense, it is a discrete and worst-case notion of “derivative” with respect to the  $i$ th coordinate.

**Corollary 4.5.14** (McDiarmid inequality). If  $X_1, \dots, X_n$  are independent random variables as above, then  $f(X_1, \dots, X_n)$  is  $\sigma_{\text{prox}}^2$ -subgaussian, with

$$\sigma_{\text{prox}}^2 = \sum_{i=1}^n \delta_i(f)^2.$$

*Proof.* Write  $X = (X_1, \dots, X_n) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ . Consider the Doob martingale

$$M_i := \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_i].$$

We have  $M_n = f(X)$  and  $M_0 = \mathbb{E}f(X)$ , so  $M_n - M_0 = f(X) - \mathbb{E}f(X)$ .

We plan to use Azuma’s inequality on  $M_n$ . To control the increments, let  $X'_1, \dots, X'_n$  be independent copies of the  $X_i$ , and write  $X^{(i)} := (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$ , which may be interpreted as  $X$  with coordinate  $i$  *resampled*. We then have

$$\begin{aligned} M_i - M_{i-1} &= \mathbb{E}[f(X) \mid X_1, \dots, X_i] - \mathbb{E}[f(X) \mid X_1, \dots, X_{i-1}] \\ &= \mathbb{E}[f(X) \mid X_1, \dots, X_i] - \mathbb{E}[f(X^{(i)}) \mid X_1, \dots, X_{i-1}] \end{aligned}$$

and since the second expression does not include  $X_i$  at all anymore, we may condition on  $X_i$  without changing the value,

$$\begin{aligned} &= \mathbb{E}[f(X) \mid X_1, \dots, X_i] - \mathbb{E}[f(X^{(i)}) \mid X_1, \dots, X_{i-1}, X_i] \\ &= \mathbb{E}[f(X) - f(X^{(i)}) \mid X_1, \dots, X_i], \end{aligned}$$

and thus  $|M_i - M_{i-1}| \leq \delta_i(f)$  almost surely. The result then follows from Theorem 4.5.12.  $\square$

#### 4.5.4 INTERACTION WITH MAXIMAL INEQUALITIES

We mention one “upgrade” of such inequalities that is lesser known but sometimes useful. The idea is to use Lemma 4.4.21 in the middle of a proof using the Chernoff method. In particular, suppose that  $M_n$  is a martingale, and as before define

$$M_n^* := \max_{0 \leq k \leq n} M_k.$$

We have

$$\begin{aligned} \mathbb{P}[M_n^* \geq t] &= \mathbb{P}[\max_{0 \leq k \leq n} M_k \geq t] \\ &= \mathbb{P}[\max_{0 \leq k \leq n} \exp(\lambda M_k) \geq \exp(\lambda t)] \end{aligned}$$

But now, since  $f(t) = \exp(\lambda t)$  is a convex function, by Proposition 4.4.18 we have that  $(\exp(\lambda M_n))$  is a non-negative submartingale, and we may use Lemma 4.4.21 to bound

$$\leq \frac{\mathbb{E} \exp(\lambda M_n)}{\exp(\lambda t)}.$$

Thus, both Hoeffding's inequality (Theorem 4.5.1) and Azuma's inequality (Theorem 4.5.12) can be upgraded to pertain not just to the value at time  $n$  of a random walk or martingale, but to the maximum value taken up to time  $n$ . (This also applies to McDiarmid's inequality, but there does not seem to be any immediate application of applying this idea to a Doob martingale.)

#### 4.5.5 EXAMPLE: BALLS, BINS, AND MULTINOMIALS

As an example, consider an experiment of throwing  $m$  balls independently and uniformly at random into  $n$  bins. (This is actually a special case of the multinomial distribution discussed in Section 2.3.3.) Let

$$Z := \#\{\text{empty bins}\}.$$

A computation by linearity of expectation gives

$$\mathbb{E}Z = n \cdot \mathbb{P}[\text{bin 1 empty}] = n \cdot \left(1 - \frac{1}{n}\right)^m.$$

We will aim to show that  $Z$  concentrates around this expectation.

If  $X_j \in [n]$  denotes the index of the bin where ball  $j$  lands, we have

$$Z = f(X_1, \dots, X_m)$$

and the  $X_i$  are independent. As is intuitively obvious, changing the location of a single ball can only change the number of empty bins by at most 1, so  $\delta_i(f) \leq 1$  for all  $i \in [m]$ . Thus,  $\sigma_{\text{prox}}^2 = m$  in McDiarmid's inequality, and we find

$$\mathbb{P}\left(\left|Z - n \left(1 - \frac{1}{n}\right)^m\right| > t\right) \leq 2 \exp\left(-\frac{t^2}{2m}\right).$$

Roughly speaking, we may summarize this by saying that, with high probability,

$$Z = n \left(1 - \frac{1}{n}\right)^m \pm O(\sqrt{m})$$

with high probability. In some regimes this gives quite strong notions of concentration. For instance, if  $m = n$ , then  $\mathbb{E}Z \approx n/e$ , and we find

$$Z = \frac{n}{e} \pm O(\sqrt{n}).$$

## 4.6 APPLICATION: RANDOM SERIES

One class of martingales for which the  $L^p$  convergence theorems are useful is sums of independent but not identically distributed random variables that converge to the values of random *infinite series*. That is, let  $X_1, X_2, \dots$  be independent, and consider

$$S_n := \sum_{i=1}^n X_i,$$

$$S_\infty := \sum_{i=1}^{\infty} X_i$$

When does  $S_\infty$  exist and how does it behave?

The following is one of the main theorems on random series, which we may prove easily using martingale convergence theorems.

**Theorem 4.6.1** (Kolmogorov two-series theorem). Suppose that, in the above setting,  $X_i \in L^2$  and  $\sum_{i=1}^{\infty} |\mathbb{E}[X_i]| < \infty$  and  $\sum_{i=1}^{\infty} \text{Var}[X_i] < \infty$ . Then,  $S_n \rightarrow S_\infty \in L^2$  almost surely and in  $L^2$ .

*Proof.* Without loss of generality we may suppose  $\mathbb{E}X_i = 0$ . The result is then immediate from the  $p = 2$  case of Theorem 4.4.17, using that  $\|S_n\|_{L^2}^2 = \mathbb{E}S_n^2 = \sum_{i=1}^n \text{Var}[X_i]$ .  $\square$

An important special case is the construction and study of certain *Gaussian processes*. For instance, we may build the *Gaussian analytic function*,

$$f(z) = \sum_{k=0}^{\infty} \frac{g_k}{\sqrt{k!}} z^k$$

where  $g_k \sim \mathcal{N}(0, 1)$  are i.i.d. It turns out to be possible to define this for all  $z \in \mathbb{C}$ , yielding a random holomorphic/analytic function (hence the name), but for the sake of simplicity let us focus on  $z \in \mathbb{R}$ . Theorem 4.6.1 implies that the limit defining  $f(z)$  exists almost surely and is an  $L^2$  random variable.

But further, the decay of the  $1/\sqrt{k!}$  coefficients is so fast that, as you may show as an exercise, the series converges in *any*  $L^p$ . As a consequence of convergence in all  $L^p$ , as you may again show as an exercise, we have for all  $\ell \geq 0$  that the moments of  $f(z)$  are given by

$$\begin{aligned} \mathbb{E}f(z)^\ell &= \lim_{n \rightarrow \infty} \mathbb{E} \left( \sum_{k=0}^n \frac{g_k}{\sqrt{k!}} z^k \right)^\ell \\ &= \lim_{n \rightarrow \infty} \mathbb{E}_{g \sim \mathcal{N}(0, \sum_{k=0}^n \frac{1}{k!} z^{2k})} g^\ell \\ &= \mathbb{E}_{g \sim \mathcal{N}(0, \sum_{k=0}^{\infty} \frac{1}{k!} z^{2k})} g^\ell \\ &= \mathbb{E}_{g \sim \mathcal{N}(0, \exp(z^2))} g^\ell. \end{aligned}$$

Thus, by Lemma 2.2.1 on determining distributions by moments that we saw earlier, in fact the infinite sum of Gaussian random variables  $f(z)$  itself has a Gaussian law,

$$\text{Law}(f(z)) = \mathcal{N}(0, \exp(z^2)).$$

By similar methods you may probe the properties of vectors  $(f(z_1), \dots, f(z_k))$ , and so forth. The martingale convergence theorems give a convenient tool to avoid many technicalities: by calculations like the above, you may effectively work as though the result for manipulating finite sums of Gaussian random variables also apply to infinite Gaussian series.

## 4.7 APPLICATION: BRANCHING PROCESSES

We now show how martingales can be applied to understand a central object in discrete probability, the *Galton-Watson branching process*. This is the following simple model of population growth. Let  $\mu$  be a probability measure on  $\mathbb{Z}_{\geq 0}$  representing the number of offspring each individual in the population has. Let  $X_{n,i} \sim \mu$  for  $i, n \geq 1$  be i.i.d. We define the process  $(Z_n)$  by

$$\begin{aligned} Z_0 &:= 1, \\ Z_n &:= \sum_{i=1}^{Z_{n-1}} X_{n,i} \text{ for } n \geq 1. \end{aligned}$$

That is, the population at time  $n$  is determined by each individual in the population at time  $n - 1$  having a number of offspring distributed as  $\mu$ .

We will study the question of how the long-term behavior of this process depends on the distribution  $\mu$ . To that end, we define the parameters

$$\begin{aligned} p(k) &:= \mathbb{P}_{X \sim \mu} [X = k], \\ m &:= \mathbb{E}_{X \sim \mu} [X] = \sum_{k=0}^{\infty} k \cdot p(k). \end{aligned}$$

We assume that  $X_{n,i} \in L^1$ , i.e., that  $m < \infty$ .

Defining the filtration

$$\mathcal{F}_n := \sigma(X_{k,i} : k \leq n),$$

we see that  $(Z_n)$  is adapted to  $(\mathcal{F}_n)$ .

### 4.7.1 EXTINCTION

We say that the process  $(Z_n)$  *goes extinct* if  $Z_n = 0$  for some  $n \geq 1$ , in which case by definition also  $Z_N = 0$  for all  $N \geq n$ . Also, since  $Z_n$  is integer-valued, this is equivalent to having  $Z_n \rightarrow 0$ . We emphasize that these events are the same:

$$E := \{Z_n = 0 \text{ for some } n \geq 1\} = \{Z_n = 0 \text{ for all } n \text{ sufficiently large}\} = \{Z_n \rightarrow 0\}.$$

We define the probability of extinction as

$$\rho := \mathbb{P}[E].$$

We also set

$$\begin{aligned} E_n &:= \{Z_n = 0\}, \\ \rho_n &:= \mathbb{P}[E_n], \end{aligned}$$

the event and probability of extinction by time  $n$ .

**Proposition 4.7.1.** The  $\rho_n$  are non-decreasing in  $n$ , and  $\rho = \lim_{n \rightarrow \infty} \rho_n$ .

*Proof.* We have

$$E_1 \subseteq E_2 \subseteq \cdots \subseteq E,$$

and further by the above

$$E = \bigcup_{n=1}^{\infty} E_n,$$

and the result follows by the continuity property of probability measures.  $\square$

#### 4.7.2 ASSOCIATED MARTINGALE AND CONVERGENCE

We will see in a moment that the following normalization of  $Z_n$  is a natural one:

$$M_n := \frac{Z_n}{m^n}.$$

**Proposition 4.7.2.**  $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] = mZ_n$ , and thus  $\mathbb{E}[Z_n] = m^n$  and  $M_n$  is a martingale.

*Proof.* Using similar tricks to what we have seen before in dealing with sums and integrals with random quantities in the limits, we compute

$$\begin{aligned} \mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}\left[\sum_{i=1}^{Z_n} X_{n+1,i} \mid \mathcal{F}_n\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{\infty} X_{n+1,i} \mathbb{1}\{i \leq Z_n\} \mid \mathcal{F}_n\right] \\ &= \sum_{i=1}^{\infty} \mathbb{E}[X_{n+1,i} \mathbb{1}\{i \leq Z_n\} \mid \mathcal{F}_n] \\ &= \sum_{i=1}^{\infty} \mathbb{1}\{i \leq Z_n\} \mathbb{E}[X_{n+1,i} \mid \mathcal{F}_n] \\ &= \sum_{i=1}^{Z_n} \mathbb{E}[X_{n+1,i}] \\ &= mZ_n, \end{aligned}$$

as claimed. The only not quite rigorous step above is exchanging the *infinite* sum with the conditional expectation, which you may verify using linearity of conditional expectation and the monotone convergence theorem. That  $\mathbb{E}Z_n = m^n$  then follows by the tower rule and induction, and that  $M_n$  is a martingale is immediate.  $\square$

We then get the following result, which we will see is a powerful tool for reasoning about the branching process, for free from our general theory.

**Corollary 4.7.3.** There exists a  $M_\infty \in L^1$  such that  $M_n \rightarrow M_\infty$  almost surely.

*Proof.* We have  $M_n \geq 0$  and  $\mathbb{E}M_n = \|M_n\|_{L^1} = 1$  by construction, so the result follows by Theorem 4.4.1.  $\square$

### 4.7.3 SUBCRITICAL REGIME ( $m < 1$ )

We divide our subsequent analysis of the behavior of  $(Z_n)$  and  $(M_n)$  into three cases according to the value of  $m$ . The following describes the situation in the simplest case:

**Theorem 4.7.4.** If  $m < 1$ , then  $\rho = 1$ , i.e.,  $Z_n \rightarrow 0$  almost surely. Equivalently, almost surely  $Z_n = 0$  for all sufficiently large  $n$ , and thus  $M_\infty = \lim_{n \rightarrow \infty} M_n = 0$  almost surely as well.

*Proof.* We have by Markov's inequality

$$\mathbb{P}[Z_n > 0] = \mathbb{P}[Z_n \geq 1] \leq \mathbb{E}Z_n = m^n.$$

Thus  $\sum_{n=1}^{\infty} \mathbb{P}[Z_n > 0] < \infty$ , so by the Borel-Cantelli lemma almost surely we have  $Z_n > 0$  for only finitely many  $n$ .  $\square$

### 4.7.4 CRITICAL REGIME ( $m = 1$ )

The next case is slightly but not that much more complicated:

**Theorem 4.7.5.** Suppose that  $m = 1$ . Then:

1. If  $p(1) = 1$ , i.e.,  $X_{n,i} = 1$  almost surely, then  $Z_n = 1$  almost surely for all  $n$  and  $\rho = 0$ .
2. If  $p(1) < 1$ , then  $\rho = 1$ ,  $Z_n \rightarrow 0$  almost surely, and  $M_\infty = 0$  almost surely (indeed, in this case  $M_n = Z_n$ ).

*Proof.* The first case is immediate from the definition of  $Z_n$ . Suppose we are in the second case. Since  $m = 1$ ,  $M_n = Z_n$  is a martingale, which by Corollary 4.7.3 converges almost surely. On the other hand, since  $M_n$  is integer-valued, it can only converge if it is eventually constant, so almost surely we have  $M_n = M_\infty$  for all sufficiently large  $n$ .

We consider the probability that this limit is some  $k \geq 1$ : introducing parameters  $N, \ell$  to be chosen later,

$$\begin{aligned}
\mathbb{P}[M_\infty = k] &= \mathbb{P}[M_n = Z_n = k \text{ for all } n \text{ sufficiently large}] \\
&= \lim_{N \rightarrow \infty} \mathbb{P}[M_n = Z_n = k \text{ for all } n \geq N] \\
&\leq \lim_{N \rightarrow \infty} \mathbb{P}[Z_N = \cdots = Z_{N+\ell} = k] \\
&\leq \lim_{N \rightarrow \infty} \mathbb{P}\left[\sum_{i=1}^k X_{N+1,i} = \cdots = \sum_{i=1}^k X_{N+\ell,i} = k\right] \\
&= \lim_{N \rightarrow \infty} \left(\mathbb{P}\left[\sum_{i=1}^k X_{1,i} = k\right]\right)^\ell \\
&\leq (1 - p(0)^k)^\ell.
\end{aligned}$$

Now, note that this holds for any  $\ell \geq 1$ , and since  $m = \mathbb{E}_{X \sim \mu} X = 1$  while  $\mathbb{P}[X = 1] < 1$ , we must have  $p(0) = \mathbb{P}[X = 0] > 0$ , and thus taking  $\ell \rightarrow \infty$  shows that  $\mathbb{P}[M_\infty = k] = 0$ . Thus, indeed almost surely  $M_\infty = 0$ .  $\square$

#### 4.7.5 SUPERCRITICAL REGIME ( $m > 1$ )

This case is more complicated, because very different behaviors are possible in the limit. It is easy to see that, whenever  $p(0) > 0$ , extinction happens with some positive probability  $\rho > 1$ ; in particular, we already have  $Z_1 = 0$  with probability  $p(0)$ . On the other hand, at least some of the time we expect  $Z_n \approx \mathbb{E}Z_n = m^n$ , which when  $m > 1$  is diverging exponentially. While more detailed analysis is possible, here we will just try to understand these two extreme behaviors.

First, we give a description of the probability of extinction. It turns out that there is an elegant description in terms of a particular generating function. Here let us define  $X := X_{1,1} \sim \mu$ , a random variable having the offspring distribution.

**Definition 4.7.6.** Let  $Y \in \mathbb{Z}_{\geq 0}$  be a random variable. Then, we define for  $t \in [-1, 1]$

$$\zeta_Y(t) := \mathbb{E}[t^Y] = \sum_{k \geq 0} \mathbb{P}[Y = k] \cdot t^k.$$

We call  $\zeta_Y$  the *probability generating function (pgf)* of  $Y$ .

**Remark 4.7.7** (Generating functions and moment problems). We have previously discussed the moment generating function (mgf), which looks quite similar,

$$\phi_Y(t) = \mathbb{E}[\exp(tY)].$$

Indeed, for  $t > 0$  these are merely related by a change of variables,

$$\sum_{k \geq 0} \mathbb{P}[Y = k] \cdot t^k = \zeta_Y(t) = \phi_Y(\log t) = \sum_{k \geq 0} \frac{\mathbb{E}Y^k}{k!} (\log t)^k.$$

If the series defining  $\phi_Y(t)$  converges absolutely in an open interval containing  $t = 0$ , we may expand this relation in power series in  $t$  (the assumption allowing us to expand  $\log(t) = \log(\epsilon + (t - \epsilon))$ ) and match up the coefficients. This gives equations, at first surprising, that express the  $\mathbb{P}[Y = k]$  explicitly in terms of the moments  $\mathbb{E}Y^k$ , giving a direct solution of the moment problem for such discrete random variables (of the kind discussed in Section 2.2). This formula is most conveniently expressed in terms of the falling factorial moments (Definition 2.2.8), in which case it takes the form

$$\mathbb{P}[Y = k] = \frac{1}{k!} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \mathbb{E}[Y^{n+k}].$$

For a simple example, if  $Y \sim \text{Pois}(\lambda)$ , we may check that this is compatible with our Proposition 2.2.11: we have

$$\frac{1}{k!} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \mathbb{E}[Y^{n+k}] = \frac{1}{k!} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \lambda^{n+k} = \frac{\lambda^k}{k!} \sum_{n=0}^{\infty} \frac{(-\lambda)^n}{n!} = \frac{\lambda^k}{k!} e^{-\lambda} = \mathbb{P}[Y = k].$$

See [Tak65] for more details.

One use of such generating functions is that they interact especially nicely with sums of random numbers of random variables.

**Proposition 4.7.8.** Let  $N \in \mathbb{Z}_{\geq 0}$  be random and  $X_1, X_2, \dots \sim \mu$  be i.i.d. and independent of  $N$ . Define  $S := \sum_{i=1}^N X_i$ . Then,

$$\zeta_S(t) = \zeta_N(\zeta_{X_1}(t)) = (\zeta_N \circ \zeta_{X_1})(t).$$

*Proof.* We have

$$\begin{aligned} \zeta_S(t) &= \mathbb{E}[t^{\sum_{i=1}^N X_i}] \\ &= \mathbb{E} \left[ \sum_{n=1}^{\infty} \mathbb{1}\{N = n\} t^{\sum_{i=1}^n X_i} \right] \\ &= \sum_{n=1}^{\infty} \mathbb{E} \left[ \mathbb{1}\{N = n\} \prod_{i=1}^n t^{X_i} \right] \\ &= \sum_{n=1}^{\infty} \mathbb{P}[N = n] \cdot (\mathbb{E}[t^{X_1}])^n \\ &= \sum_{n=1}^{\infty} \mathbb{P}[N = n] \cdot \zeta_{X_1}(t)^n \\ &= \zeta_N(\zeta_{X_1}(t)), \end{aligned}$$

as claimed. □

Applying this to branching processes gives an especially nice structure from the point of view of such generating functions.

**Corollary 4.7.9.** We have  $\zeta_{Z_{n+1}}(t) = \zeta_{Z_n}(\zeta_X(t))$  and  $\zeta_{Z_0}(t) = t$ , so  $\zeta_{Z_n}(t) = \zeta_X^{(n)}(t)$ , with the exponent  $(n)$  denoting repeated composition.

For the  $\rho_n$  whose limit is the extinction probability, we find:

**Corollary 4.7.10.** We have  $\rho_n = \mathbb{P}[Z_n = 0] = \phi_{Z_n}(0) = \phi_X(\rho_{n-1}) = \phi_X^{(n)}(0)$ .

Recalling that the extinction probability is  $\rho = \lim_{n \rightarrow \infty} \rho_n$ , we then see that  $\rho$  is the limit of repeated applications of the same function  $\phi_X$ . We should expect such repeated applications to converge to a *fixed point* of  $\phi_X$ , and indeed this is what happens, and further this property exactly characterizes the extinction probability:

**Theorem 4.7.11.** Suppose that  $m > 1$ . Then, there is a unique  $\rho \in [0, 1)$  such that  $\phi_X(\rho) = \rho$ , and this  $\rho$  is exactly the extinction probability of the associated branching process.

We note that the first part of the result is merely a fact about random variables  $X \in \mathbb{Z}_{\geq 0}$  with  $\mathbb{E}[X] > 1$ .

*Proof.* Consider the value and first two derivatives of  $\zeta_X$ :

$$\begin{aligned}\zeta_X(t) &= \sum_{k=0}^{\infty} p(k)t^k, \\ \zeta'_X(t) &= \sum_{k=1}^{\infty} kp(k)t^{k-1}, \\ \zeta''_X(t) &= \sum_{k=2}^{\infty} k(k-1)p(k)t^{k-2}.\end{aligned}$$

Since  $m = \mathbb{E}[X] > 1$ , there must be some  $k \geq 2$  such that  $p(k) = \mathbb{P}[X = k] > 0$ . In particular,  $\zeta'_X$  and  $\zeta''_X$  are both strictly positive on  $t > 0$ , so  $\zeta_X$  is strictly increasing and strictly convex on  $t \in [0, 1]$ . Further, we have

$$\begin{aligned}\zeta_X(0) &= p(0) \in [0, 1), \\ \zeta_X(1) &= \sum_{k=0}^{\infty} p(k) \\ &= 1, \\ \zeta'_X(1) &= \sum_{k=0}^{\infty} kp(k) \\ &= m > 1.\end{aligned}$$

Such a function has a unique fixed point on  $[0, 1)$  (note that we must omit the upper endpoint 1 because  $\zeta_X(1) = 1$  is always another fixed point), and the iterates  $\rho_n = \zeta_X^{(n)}(1)$  are a strictly increasing sequence converging to this fixed point, whereby this  $\rho$  is the extinction probability.  $\square$

**Corollary 4.7.12.** For any value of  $m$ , we have that  $p(0) = 0$  if and only if  $\rho = 0$  (i.e., almost surely the population does not go extinct).

*Proof.* We consider the three cases of  $m$  separately:

- If  $m > 1$ , the equivalence follows from Theorem 4.7.11 since the unique fixed point is  $\rho = 0$  if and only if  $p(0) = \phi_X(0) = 0$ .
- If  $m = 1$ , the equivalence follows from Theorem 4.7.5 since in that case  $p(0) = 0$  if and only if  $p(1) = 1$ .
- If  $m < 1$ , then we never have  $\rho = 0$  by Theorem 4.7.4 (since we always have  $\rho = 1$ ), and correspondingly we also never have  $p(0) = 0$  since in that case  $X \geq 1$  almost surely and  $m = \mathbb{E}[X] \geq 1$ .

So, the equivalence indeed holds in all cases. □

**Corollary 4.7.13.** If  $m > 1$ , then we always have  $\rho < 1$ , i.e., the probability that the population does not go extinct is  $1 - \rho > 0$ , a strictly positive probability.

Above we have given an explicit characterization of the probability  $\rho$  of extinction when  $m > 1$ , which together with the simpler characterizations in Theorems 4.7.4 and 4.7.5 characterizes  $\rho$  in all cases. Also, the last Corollary above says that when  $m > 1$  there is an event of positive probability when the population does not go extinct. We now turn our attention to the asymptotic behavior of  $Z_n$  on this latter event.

By Corollary 4.7.3, we have that  $M_n = Z_n/m^n$  converges almost surely to some  $M_\infty$ . However, it is still possible that  $M_\infty = 0$  almost surely, in which case we will have that  $Z_n$  grows asymptotically slower than  $m^n$ . We will conclude our discussion by giving sufficient conditions under which this does not happen, so that there is another event of positive probability under which  $Z_n$  grows comparably to its expectation  $\mathbb{E}Z_n = m^n$ .

**Theorem 4.7.14.** If  $X_{n,i} \in L^2$ , then  $M_n \rightarrow M_\infty$  almost surely and in  $L^2$ , and therefore also in  $L^1$ . In particular,  $\mathbb{E}M_n \rightarrow \mathbb{E}M_\infty$ , so  $\mathbb{E}M_\infty = 1$ , and  $M_\infty$  is not almost surely equal to zero. Therefore, there exist  $\delta, \epsilon > 0$  and  $N \geq 1$  such that  $\epsilon < \mathbb{P}[M_n \geq \delta] = \mathbb{P}[Z_n \geq \delta m^n]$  for all  $n \geq N$ .

*Proof.* We argue that we may apply the  $L^2$  case of Theorem 4.4.16. This will imply the first convergence statement provided we show that  $\|M_n\|_{L^2}$  is uniformly bounded. We compute:

$$\begin{aligned}
 \|M_n\|_{L^2}^2 &= \mathbb{E}M_n^2 \\
 &= \frac{1}{m^{2n}} \mathbb{E}Z_n^2 \\
 &= \frac{1}{m^{2n}} \mathbb{E}[\mathbb{E}[Z_n^2 \mid \mathcal{F}_{n-1}]] \\
 &= \frac{1}{m^{2n}} \mathbb{E} \left[ \mathbb{E} \left[ \left( \sum_{i=1}^{Z_{n-1}} X_{n,i} \right)^2 \mid \mathcal{F}_{n-1} \right] \right]
 \end{aligned}$$

We omit the details of the next step, which you may verify as an exercise:

$$\begin{aligned}
 &= \frac{1}{m^{2n}} \mathbb{E} \left[ Z_{n-1} \mathbb{E}[X^2] + Z_{n-1}(Z_{n-1} - 1)m^2 \right] \\
 &= \frac{1}{m^{2n}} \mathbb{E} \left[ Z_{n-1} \text{Var}[X] + Z_{n-1}^2 m^2 \right] \\
 &= \frac{1}{m^{2n}} \cdot m^{n-1} \cdot \text{Var}[X] + \mathbb{E}[M_{n-1}^2] \\
 &= \frac{\text{Var}[X]}{m^{n+1}} + \|M_{n-1}\|_{L^2}^2
 \end{aligned}$$

and, unfolding the recursion

$$\leq \text{Var}[X] \sum_{n=0}^{\infty} \frac{1}{m^n},$$

a finite upper bound. Thus, Theorem 4.4.16 applies to give that  $M_n \rightarrow M_\infty$  in  $L^2$ , and the other results follow immediately.  $\square$

In fact, it is possible to give a “sharp” version of such a result, which is important and often cited in more advanced literature on branching processes.

**Theorem 4.7.15** ([KS66]). If  $m > 1$ , then  $M_\infty$  is not almost surely equal to zero if and only if  $\mathbb{E}[X \log(1 + X)] < \infty$ .

Both this result and Theorem 4.7.14 require some care in their interpretation. They may seem puzzling at first because they say that provided that the offspring random variables  $X$  are “small” then  $Z_n$  grows somewhat quickly. However, the point is that if  $\mathbb{E}[X \log(1 + X)] = \infty$ , then  $X$  is very heavy-tailed, and the size of the mean  $m$  is driven by very rare outcomes of  $X$ . Thus,  $m$  is quite large while  $Z_n$  usually grows more slowly, and thus  $M_n = Z_n/m^n \rightarrow 0$ .

## 5 | MARKOV CHAINS

### 5.1 MOTIVATION

A Markov chain can be viewed as a general type of stochastic process that obeys the basic logic of time we observe in physics: if you know the present configuration of a system, then learning about its past does not help you in determining the future.

You have probably seen the definition for a finite or countable state space  $S$ : a Markov chain is a sequence of random variables  $(Y_n)_{n \geq 0} \in S$  satisfying the *Markov property* that

$$\mathbb{P}[Y_{n+1} = j \mid Y_n = i, Y_{n-1} = i_{n-1}, \dots, Y_1 = i_1, Y_0 = i_0] = \mathbb{P}[Y_{n+1} = j \mid Y_n = i] =: p_{n+1}(i, j). \quad (5.1.1)$$

In this case the functions  $p_n : S \times S \rightarrow [0, 1]$  are called *transition kernels*. We focus here on the *time-homogeneous* case, where these  $p_n = p$  are all the same.

The following are some examples we have seen already in the class, as well as one important one that we have not (yet).

**Example 5.1.1** (Random walk). Consider the simple random walk  $Y_n = \sum_{i=1}^n X_i$  for  $X_i \sim \text{Unif}(\{\pm 1\})$  drawn i.i.d. We may view this as a Markov chain on the state space  $S = \mathbb{Z}$ , with transition kernel

$$p(i, j) = \begin{cases} \frac{1}{2} & \text{if } j = i + 1, \\ \frac{1}{2} & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases}.$$

**Example 5.1.2** (Branching process). The branching process we saw earlier,  $Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}$  for  $X_{n,i} \sim \mu$  i.i.d. for  $\mu$  a probability measure on  $\mathbb{Z}_{\geq 0}$ , may be viewed as a Markov chain on the state space  $S = \mathbb{Z}_{\geq 0}$ . The transition kernel now depends on  $\mu$  and may be summarized conveniently in terms of the convolution operation:

$$p(i, j) = \mathbb{P} \left[ \sum_{a=1}^i X_{1,a} = j \right] = \underbrace{\mu * \mu * \dots * \mu}_{i \text{ times}}(j).$$

**Example 5.1.3** (Random walk on graph). A kind of example we have not yet seen takes values that are not numbers. Suppose  $G = (V, E)$  is a finite simple graph. The *simple random walk* is the stochastic process of random variables  $(Y_n)_{n \geq 0}$  taking values  $Y_n \in V$  and starting from some  $Y_0 = v_0$  where, to form  $Y_n$  for each  $n \geq 1$ , we choose a random

neighbor of  $Y_{n-1}$  in  $G$ . This is a Markov chain on state space  $S = V$ , whose transition kernel may be written

$$p(i, j) = \begin{cases} \frac{1}{\deg_G(i)} & \text{if } \{i, j\} \in E, \\ 0 & \text{otherwise} \end{cases}.$$

As we will return to later, in this case the transition kernel can be usefully viewed as a matrix  $\mathbf{P} \in \mathbb{R}^{V \times V}$ , which is then  $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$  for  $\mathbf{A}$  the adjacency matrix of  $G$  and  $\mathbf{D} = \text{diag}(\deg_G(i))_{i \in V}$ .

## 5.2 MEASURE THEORY TOOLKIT

### 5.2.1 CARATHÉODORY EXTENSION THEOREM

We first review the basic ideas of constructing probability measures on product spaces. Let  $(\Omega, \mathcal{F})$  be a measurable space and  $\mathcal{I}$  be some other set. We denote by  $\text{Pow}(\Omega)$  the *power set* of  $\Omega$ , the set of all subsets.

**Definition 5.2.1** (Cylinder set). Let  $i_1, \dots, i_m \in \mathcal{I}$  and  $B_1, \dots, B_m \in \mathcal{F}$ . A set of the form

$$\{\omega \in \Omega^{\mathcal{I}} : \omega_{i_a} \in B_a \text{ for each } a \in [m]\}$$

is called a *cylinder set*. We denote the set of cylinder sets by  $C(\Omega, \mathcal{I}) \subseteq \text{Pow}(\Omega)$ .

Recall that we *define* the product  $\sigma$ -algebra  $\mathcal{F}^{\mathcal{I}}$  to be the one generated by (i.e., the smallest one containing) all cylinder sets in  $\Omega^{\mathcal{I}}$ . We denote this  $\mathcal{F}^{\mathcal{I}} = \sigma(C(\Omega, \mathcal{I}))$ . Further, the following important result says that any consistent assignment of measure values to  $C(\Omega, \mathcal{I})$  can be extended to a probability measure.

**Theorem 5.2.2** (Carathéodory). Let  $\mu : C(\Omega, \mathcal{I}) \rightarrow [0, 1]$  be countably additive on disjoint cylinder sets whose union is another cylinder set, and have  $\mu(\Omega) = 1$ . Then, there exists a unique extension of  $\mu$  to a probability measure on  $\mathcal{F}^{\mathcal{I}}$ .

We will see one way to prove the uniqueness part below in Corollary 5.2.9.

### 5.2.2 KOLMOGOROV EXTENSION THEOREM

The following is a close relative of the Carathéodory extension theorem, that lets us instead extend sequences of finite-dimensional marginal probability distributions to infinite-dimensional ones.

**Theorem 5.2.3** (Discrete Kolmogorov extension). Suppose  $(\Omega, \mathcal{F})$  is a measurable space and, for each  $n \geq 1$ ,  $\mu_n$  is a probability measure on  $(\Omega^n, \mathcal{F}^n)$ . Suppose that these satisfy

the *compatibility* property that, for each  $1 \leq m < n$  and  $A_1, \dots, A_m \in \mathcal{F}$ ,

$$\mu_n(A_1 \times \cdots \times A_m \times \underbrace{\Omega \times \cdots \times \Omega}_{n-m \text{ times}}) = \mu_m(A_1 \times \cdots \times A_m).$$

Then, there exists a unique probability measure  $\mathbb{P}$  on  $(\Omega^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}})$  such that, for all  $m \geq 1$  and all  $A_1, \dots, A_m \in \mathcal{F}$ ,

$$\mathbb{P}[A_1 \times \cdots \times A_m \times \Omega \times \Omega \times \cdots] = \mu_m(A_1 \times \cdots \times A_m).$$

### 5.2.3 $\pi$ - $\lambda$ THEOREM

Let  $\Omega$  be a set. We write  $\text{Pow}(\Omega)$ , sometimes denoted  $2^\Omega$ , for the *power set* of all subsets of  $\Omega$ . Recall that a  $\sigma$ -algebra  $\mathcal{F} \subseteq \text{Pow}(\Omega)$  may be defined as any set of subsets that contains  $\Omega$ , is closed under complements, and is closed under countable intersections or unions. The following are two different weakenings of this notion:

**Definition 5.2.4** ( $\pi$ -system). We say  $\Pi \subseteq \text{Pow}(\Omega)$  is a  $\pi$ -system if it is closed under finite intersections.

**Definition 5.2.5** ( $\lambda$ -system). We say  $\Lambda \subseteq \text{Pow}(\Omega)$  is a  $\lambda$ -system if  $\Omega \in \Lambda$  and  $\Lambda$  is closed under complement and countable union of disjoint sets.

The following is the first tool that goes along with these notions.

**Theorem 5.2.6** (Dynkin  $\pi$ - $\lambda$  theorem). Suppose that  $\Pi$  is a  $\pi$ -system,  $\Lambda$  is a  $\lambda$ -system, and  $\Pi \subseteq \Lambda$ . Then,  $\sigma(\Pi) \subseteq \Lambda$ .

We will use this with the following particular choice of  $\Pi$ , called the  $\pi$ -system of *cylinder sets*.

**Proposition 5.2.7.** Let  $(\Omega, \mathcal{F})$  be a measurable space and  $\mathcal{I}$  be another set. Define the set

$$\Pi := \left\{ \{\omega \in \Omega^{\mathcal{I}} : \omega_{i_a} \in B_a \text{ for all } a = 1, \dots, m\} : m \geq 1, i_1, \dots, i_m \in \mathcal{I}, B_1, \dots, B_m \in \mathcal{F} \right\} \\ \subseteq \text{Pow}(\Omega^{\mathcal{I}}).$$

Then,  $\Pi$  is a  $\pi$ -system.

The following is then immediate from Theorem 5.2.6.

**Corollary 5.2.8.** In the setting of Proposition 5.2.7, suppose that  $\Lambda$  is a  $\lambda$ -system on  $\Omega^I$ . Let  $\Pi$  be the  $\pi$ -system of cylinder sets. If  $\Pi \subseteq \Lambda$ , then  $\mathcal{F}^I \subseteq \Lambda$  (here  $\mathcal{F}^I$  is the product  $\sigma$ -algebra, which equals  $\sigma(\Pi)$  by definition).

The pattern by which we use this result is that, if we want to show that some property holds for all of  $\mathcal{F}^I$ , then it suffices to show that it holds for cylinder sets by hand, and then show the closure properties (under complement and disjoint countable union) that show that the set of sets for which the property holds is a  $\lambda$ -system.

One important application is the following, essentially the same as the uniqueness clause of both the Kolmogorov and Carathéodory extension theorems, stating that a probability measure on a product space is uniquely determined by its values on cylinder sets.

**Corollary 5.2.9.** Let  $\mu, \nu$  be two probability measures on  $(\Omega^I, \mathcal{F}^I)$  and let  $\Pi$  be the  $\pi$ -system of cylinder sets. If  $\mu(A) = \nu(A)$  for all  $A \in \Pi$ , then  $\mu = \nu$ .

*Proof.* Let  $\Lambda := \{A \in \mathcal{F}^I : \mu(A) = \nu(A)\}$ . By the properties of probability measures  $\Lambda$  is a  $\lambda$ -system, and  $\Pi \subseteq \Lambda$  by assumption. The result then follows by Corollary 5.2.8.  $\square$

#### 5.2.4 MONOTONE CLASS THEOREM

The following can be viewed as an analog of the  $\pi$ - $\lambda$  theorem for functions.

**Theorem 5.2.10** (Monotone class theorem). Let  $\Pi$  be a  $\pi$ -system on  $\Omega$  such that  $\Omega \in \Pi$ . Suppose that  $\mathcal{H} \subseteq \{h : \Omega \rightarrow \mathbb{R}\}$  is a set of functions satisfying the following:

1. If  $A \in \Pi$ , then  $\mathbb{1}_A \in \mathcal{H}$ .
2. If  $h_1, h_2 \in \mathcal{H}$  and  $c_1, c_2 \in \mathbb{R}$  then  $c_1 h_1 + c_2 h_2 \in \mathcal{H}$  (i.e.,  $\mathcal{H}$  is closed under linear combinations).
3. If  $h_n, h : \Omega \rightarrow \mathbb{R}_{\geq 0}$  are bounded and non-negative, and  $h_n$  increase pointwise to  $h$ , then  $h \in \mathcal{H}$  (i.e.,  $\mathcal{H}$  is closed under bounded and monotone limits).

Then,  $\mathcal{H}$  contains all bounded measurable functions  $h : \Omega \rightarrow \mathbb{R}$ .

### 5.3 DEFINITION, EXISTENCE, AND UNIQUENESS

We will define and construct discrete-time Markov chains on a general measurable state space  $(S, \mathcal{G})$ . The simple Markov property for countable state spaces from (5.1.1) is sensible to generalize as follows. First, we choose a filtration  $(\mathcal{F}_n)_{n \geq 0}$  of  $\mathcal{G}$  and suppose  $(Y_n)_{n \geq 0}$  is adapted to this filtration. Then, we ask that, for each  $A \in \mathcal{G}$ ,

$$\mathbb{P}[Y_{n+1} \in A \mid \mathcal{F}_n] = \mathbb{P}[Y_{n+1} \in A \mid Y_n]$$

By Lemma 3.2.1, we expect such a random variable to be a function of  $Y_n$ , and thus for some  $p : S \times \mathcal{G} \rightarrow [0, 1]$  we will have

$$= p(Y_n, A). \quad (5.3.1)$$

However, we now run into the same issues discussed in Section 3.6, namely that it is not necessarily always possible to parametrize conditional probabilities such that this function  $p$  has some very basic reasonable properties. On the other hand, in this case we will usually think of ourselves as being *given* a function  $p$  that defines a Markov chain, so we may simply define a “valid”  $p$  to have these basic properties. They are as follows:

**Definition 5.3.1.** A function  $p : S \times \mathcal{G} \rightarrow [0, 1]$  is a (*Markov*) *transition kernel* if:

1. For all  $y \in S$ , the function  $A \mapsto p(y, A)$  is a probability measure on  $(S, \mathcal{G})$ .
2. For all  $A \in \mathcal{G}$ , the function  $y \mapsto p(y, A)$  is a measurable function  $S \rightarrow [0, 1]$ .

We also denote  $p_y(A) := p(y, A)$ , so that each  $p_y$  is a probability measure.

This is a slightly milder version of Definition 3.6.1 that we encountered earlier.

With this, we may also define a Markov chain. Intuitively, a Markov chain is supposed to be the law of a process where we first draw  $Y_0 \sim \mu_0$  for some probability measure  $\mu_0$  on  $(S, \mathcal{G})$ , and then repeatedly draw  $Y_n \sim p_{Y_{n-1}}$ . This is not an entirely rigorous formulation when we want to describe sampling an *infinite* string  $(Y_n)_{n \in \mathbb{N}}$ , so instead we ask for (5.3.1) to hold, which is a more formal way of asking for the same structure.

**Definition 5.3.2.**  $(Y_n)$  is a *Markov chain* with respect to the filtration  $(\mathcal{F}_n)$  having transition kernel  $p$  and initial distribution  $\mu_0$  if:

1.  $(Y_n)$  is adapted to  $(\mathcal{F}_n)$ .
2.  $\text{Law}(Y_0) = \mu_0$ .
3.  $\mathbb{P}[Y_{n+1} \in A \mid \mathcal{F}_n] = p(Y_n, A)$  almost surely for each  $n \geq 0$  and  $A \in \mathcal{G}$ . We call this the *weak Markov property*.

We call  $(\mu_0, p)$  the *data* of the Markov chain.

The main result we will show is the following, stating that we may construct an essentially unique Markov chain with any given initial distribution and transition kernel.

**Theorem 5.3.3.** Let  $\mu_0$  be a probability measure on  $(S, \mathcal{G})$  and  $p : S \times \mathcal{G} \rightarrow [0, 1]$  a transition kernel. Then, there exists a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and a family of random variables  $Y_n : \Omega \rightarrow S$  such that  $(Y_n)_{n \in \mathbb{N}}$  is a Markov chain with data  $(\mu_0, p)$ . Further, if  $(Y'_n)_{n \in \mathbb{N}}$  is another such Markov chain, then  $\text{Law}((Y_n)_{n \in \mathbb{N}}) = \text{Law}((Y'_n)_{n \in \mathbb{N}})$ .

*Proof.* We construct  $\mu_n$  that we hope to be a suitable law for  $(Y_0, \dots, Y_n) \in S^{n+1}$  and then invoke the Kolmogorov extension theorem, Theorem 5.2.3. We do this by induction. First, we are forced to choose  $Y_0 = \mu_0$ . Now, if the weak Markov property is to hold, on cylinder sets we must have

$$\begin{aligned}
& \mathbb{P}[(Y_0, \dots, Y_n) \in A_0 \times \dots \times A_n] \\
&= \mathbb{P}[Y_0 \in A_0, \dots, Y_n \in A_n] \\
&= \mathbb{E}[\mathbb{1}\{Y_0 \in A_0\} \cdots \mathbb{1}\{Y_n \in A_n\}] \\
&= \mathbb{E}[\mathbb{E}[\mathbb{1}\{Y_0 \in A_0\} \cdots \mathbb{1}\{Y_n \in A_n\} \mid \mathcal{F}_{n-1}]] \\
&= \mathbb{E}[\mathbb{1}\{Y_0 \in A_0\} \cdots \mathbb{1}\{Y_{n-1} \in A_{n-1}\} \mathbb{P}[Y_n \in A_n \mid \mathcal{F}_{n-1}]] \\
&= \mathbb{E}[\mathbb{1}\{Y_0 \in A_0\} \cdots \mathbb{1}\{Y_{n-1} \in A_{n-1}\} p(Y_{n-1}, A_n)]. \quad (\text{weak Markov property})
\end{aligned}$$

In integral notation and rewriting in terms of the marginal distributions  $\mu_n$ , we then expect the  $\mu_n$  to satisfy

$$\begin{aligned}
\mu_n(A_0 \times \dots \times A_n) &= \int_{(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \in A_0 \times \dots \times A_{n-1}} p(Y_{n-1}, A_n) d\mu_{n-1}(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \\
&= \int_{(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \in A_0 \times \dots \times A_{n-1}} d\mu_{n-1}(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \int_{\mathcal{Y}_n \in A_n} dp_{\mathcal{Y}_{n-1}}(\mathcal{Y}_n).
\end{aligned}$$

We take the above to be the recursive definition of the  $\mu_n$  on cylinder sets, which, by the Carathéodory extension theorem, Theorem 5.2.2, determines unique probability measures  $\mu_n$  (you may check the additivity condition as an exercise). This definition is also sometimes written fully expanded, as

$$\mu_n(A_0 \times \dots \times A_n) = \int_{\mathcal{Y}_0 \in A_0} d\mu_0(\mathcal{Y}_0) \int_{\mathcal{Y}_1 \in A_1} dp_{\mathcal{Y}_0}(\mathcal{Y}_1) \cdots \int_{\mathcal{Y}_n \in A_n} dp_{\mathcal{Y}_{n-1}}(\mathcal{Y}_n),$$

which is equivalent.

Next, we plan to use the Kolmogorov extension theorem, Theorem 5.2.3, to build  $\mathbb{P}$  from these  $\mu_n$ . To check the compatibility condition, note that, by definition,

$$\begin{aligned}
& \mu_n(A_0 \times \dots \times A_{n-1} \times S) \\
&= \int_{(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \in A_0 \times \dots \times A_{n-1}} p(Y_{n-1}, S) d\mu_{n-1}(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \\
&= \int_{(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) \in A_0 \times \dots \times A_{n-1}} d\mu_{n-1}(\mathcal{Y}_0, \dots, \mathcal{Y}_{n-1}) = \mu_{n-1}(A_0 \times \dots \times A_{n-1}),
\end{aligned}$$

and the more general form follows inductively.

Thus Theorem 5.2.3 applies and gives that there is a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and random variables  $Y_n$  such that  $\text{Law}((Y_0, \dots, Y_n)) = \mu_n$ . By our construction of  $\mu_n$  above it then follows that these satisfy  $\text{Law}(Y_0) = \mu_0$ . And, Theorem 5.2.3 actually gives such a realization with  $\Omega = S^{\mathbb{N}}$  where the random variable  $Y_n$  is just the  $n$ th coordinate of  $\omega \in S^{\mathbb{N}}$ , i.e., a realization  $\omega \in S^{\mathbb{N}}$  is itself the string  $(Y_n)_{n \in \mathbb{N}}$ . We have seen above that the weak Markov property uniquely specifies the marginal distributions  $\mu_n$  of  $\text{Law}((Y_n)_{n \in \mathbb{N}})$ , and thus the uniqueness of this law follows from the uniqueness clause of Theorem 5.2.3, provided  $\mathbb{P}$  satisfies the weak Markov property.

It remains to check that the weak Markov property actually holds. Recall that this asks that, for all  $n \geq 1$  and all  $A \in \mathcal{G}$ ,

$$\mathbb{P}[Y_n \in A \mid \mathcal{F}_{n-1}] \stackrel{(?)}{=} p(Y_{n-1}, A) \text{ almost surely.}$$

By the definition of conditional expectation, this asks that, for all  $B \in \mathcal{F}_{n-1}$ , we have

$$\mathbb{E}[\mathbb{1}\{B\}\mathbb{1}\{Y_n \in A\}] \stackrel{(?)}{=} \mathbb{E}[\mathbb{1}\{B\}p(Y_{n-1}, A)]. \quad (5.3.2)$$

Our definition of the  $\mu_n$  above shows that this is satisfied for cylinder sets  $B = A_0 \times \cdots \times A_{n-1}$ . To extend it to general  $B \in \mathcal{F}_{n-1}$ , we use the  $\pi$ - $\lambda$  theorem, Theorem 5.2.6. In particular, the result follows by Corollary 5.2.8 using the above observation, once we check that the sets  $B \in \mathcal{F}_{n-1}$  that satisfy (5.3.2) are a  $\lambda$ -system, which you may do as an exercise.  $\square$

**Definition 5.3.4.** We write  $\mathbb{P}_{\mu_0}$  for the unique law of  $(Y_n)_{n \in \mathbb{N}}$  a Markov chain with data  $(\mu_0, p)$ , keeping the dependence on the transition kernel  $p$  implicit. We write  $\mathbb{E}_{\mu_0}$  for the expectation with respect to  $\mathbb{P}_{\mu_0}$ . And, we define  $\mathbb{P}_x := \mathbb{P}_{\delta_x}$  and  $\mathbb{E}_x := \mathbb{E}_{\delta_x}$ .

The following intuitive result says that a Markov chain started from a random point has the law of a mixture of Markov chains started from deterministic points with suitable weights. It is another nice exercise in using the  $\pi$ - $\lambda$  theorem.

**Proposition 5.3.5.** Suppose  $A \in \mathcal{G}^{\mathbb{N}}$  (so  $A \subseteq S^{\mathbb{N}}$ ). Then, for any  $\mu_0$  a probability measure on  $S$ ,

$$\mathbb{P}_{\mu_0}[A] = \int \mathbb{P}_y[A] d\mu_0(y). \quad (5.3.3)$$

*Proof.* Let  $\Lambda \subseteq \mathcal{G}^{\mathbb{N}}$  be the set of all  $A$  for which (5.3.3) holds.  $\Lambda$  contains every cylinder set  $A$  by the definition of the  $\mu_n$  in our construction, and is a  $\lambda$ -system by standard properties of probability measures. The result then follows by Corollary 5.2.8.  $\square$

## 5.4 MARKOV PROPERTIES

Let us write  $\mathbf{Y} = (Y_0, Y_1, \dots) \in S^{\mathbb{N}}$  for the entire string of outcomes of a Markov chain. The weak Markov property may be written as the statement that, for all  $A \in \mathcal{G}$ ,

$$\mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} [\mathbb{1}\{Y_{n+1} \in A\} \mid \mathcal{F}_n] = p(Y_n, A) = \mathbb{E}_{Z \sim \mathbb{P}_{Y_n}} [\mathbb{1}\{Z_1 \in A\}].$$

Here the second expectation involves drawing from the *random measure*  $\mathbb{P}_{Y_n}$  indexed by the random variable  $Y_n$ . It requires a bit of care to make sure that an expectation with respect to such a random measure actually satisfies the measurability property required of a random variable, but you may check that this follows from the measurability condition we impose on Markov transition kernels.

Another way to write this property is to introduce the operator

$$\text{shift}_m(\mathbf{Y}) := (Y_m, Y_{m+1}, \dots).$$

Recall that, in our construction of Markov chains, we may as well view  $(S^{\mathbb{N}}, \mathcal{G}^{\mathbb{N}}) = (\Omega, \mathcal{F})$  to be the measurable space we are working over. Then, any given measurable function  $F : S^{\mathbb{N}} \rightarrow \mathbb{R}$ , which we may think of as a function  $F(\mathbf{Y})$  of the outcome of the Markov chain, is actually itself a random variable. Likewise,  $F \circ \text{shift}_m$  is also a random variable. Thus, if we define  $F(\mathbf{y}) := \mathbb{1}\{y_1 \in A\}$ , then the above is equivalently the more compact equality

$$\mathbb{E}[F \circ \text{shift}_n \mid \mathcal{F}_n] = \mathbb{E}_{Y_n}[F].$$

We now explore some additional, strengthened Markov properties, which give stronger senses in which “the future and the past are independent conditional on the present” in a Markov chain.

**Theorem 5.4.1** (Intermediate Markov property). Let  $F : S^{\mathbb{N}} \rightarrow \mathbb{R}$  be bounded and measurable and  $n \geq 0$ . Then, almost surely,

$$\mathbb{E}[F \circ \text{shift}_n \mid \mathcal{F}_n] = \mathbb{E}_{Y_n}[F]. \quad (5.4.1)$$

Equivalently in the previous notation, almost surely

$$\mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} [F(Y_n, Y_{n+1}, \dots) \mid \mathcal{F}_n] = \mathbb{E}_{Z \sim \mathbb{P}_{Y_n}} [F(Z_0, Z_1, \dots)].$$

*Proof Sketch.* We apply the monotone class theorem, Theorem 5.2.10, with the  $\pi$ -system  $\Pi$  of cylinder sets of  $S^{\mathbb{N}}$ . You may check as an exercise using the properties of conditional expectation that the collection of bounded measurable  $F$  for which (5.4.1) holds indeed satisfies Conditions 2 and 3 of the Theorem. It then suffices to check that (5.4.1) holds for  $F = \mathbb{1}_A$  the indicator of a cylinder set  $A = A_0 \times \dots \times A_m \times S \times S \times \dots$  for  $A_i \in \mathcal{G}$ . In this case, (5.4.1) reduces to the statement

$$\mathbb{P}[Y_n \in A_0, Y_{n+1} \in A_1, \dots, Y_{n+m} \in A_m \mid \mathcal{F}_n] = \mathbb{1}\{Y_n \in A_0\} \Pr_{Z \sim \mathbb{P}_{Y_n}} [Z_1 \in A_1, \dots, Z_m \in A_m],$$

which amounts to checking the defining property of conditional expectation again, using a slightly generalized version of the argument above for the weak Markov property using the  $\pi$ - $\lambda$  theorem.  $\square$

The following is an important application that lets us “factorize”

**Corollary 5.4.2** (Chapman-Kolmogorov equations). For any  $m, n \geq 0$ ,  $y \in S$ , and  $A \in \mathcal{G}$ ,

$$\mathbb{P}_y[Y_{m+n} \in A] = \Pr_{Y \sim \mathbb{P}_y} [Y_{m+n} \in A] = \mathbb{E}_{Y \sim \mathbb{P}_y} \left[ \Pr_{Z \sim \mathbb{P}_{Y_m}} [Z_n \in A] \right].$$

*Proof.* Using the tower property and intermediate Markov property, we have

$$\begin{aligned} \Pr_{Y \sim \mathbb{P}_y} [Y_{m+n} \in A] &= \mathbb{E}_{Y \sim \mathbb{P}_y} [\mathbb{1}\{Y_{m+n} \in A\}] \\ &= \mathbb{E}_{Y \sim \mathbb{P}_y} \left[ \mathbb{E}_{Y \sim \mathbb{P}_y} [\mathbb{1}\{Y_{m+n} \in A\} \mid \mathcal{F}_n] \right] \\ &= \mathbb{E}_{Y \sim \mathbb{P}_y} \left[ \mathbb{E}_{Z \sim \mathbb{P}_{Y_m}} [\mathbb{1}\{Z_n \in A\}] \right], \end{aligned}$$

as claimed.  $\square$

The next result is the strongest kind of Markov property we consider, which will be very useful for studying the times at which we reach various states (if at all) in Markov chains.

**Theorem 5.4.3** (Strong Markov property). Let  $N$  be a stopping time. Recall that we define the associated  $\sigma$ -algebra  $\mathcal{F}_N := \{A : A \cap \{N \leq n\} \in \mathcal{F}_n \text{ for all } n \geq 0\}$ . For  $F : S^{\mathbb{N}} \rightarrow \mathbb{R}$  bounded and measurable, viewed as before as a random variable on the probability space associated to a Markov chain, we have

$$\mathbb{E}_{\mu_0} [F \circ \text{shift}_N \mid \mathcal{F}_N] \cdot \mathbb{1}\{N < \infty\} = \mathbb{E}_{Y_N} [F] \cdot \mathbb{1}\{N < \infty\}.$$

We note that the restriction to the event  $\{N < \infty\}$  is crucial, or else neither side is well-defined.

*Proof.* It suffices to check that, for each  $A \in \mathcal{F}_N$  with  $A \subseteq \{N < \infty\}$ , we have

$$\mathbb{E}_{\mu_0} [(F \circ \text{shift}_N) \mathbb{1}\{A\}] \stackrel{(?)}{=} \mathbb{E}_{\mu_0} \left[ \mathbb{E}_{Y_N} [F] \mathbb{1}\{A\} \right],$$

or equivalently

$$\mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} [F(Y_N, Y_{N+1}, \dots) \mathbb{1}\{A\}] \stackrel{(?)}{=} \mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} \left[ \mathbb{E}_{Z \sim Y_N} [F(Z_0, Z_1, \dots)] \mathbb{1}\{A\} \right].$$

Starting from the left-hand side, we expand by the possible values of  $N$ :

$$\begin{aligned} \mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} [F(Y_N, Y_{N+1}, \dots) \mathbb{1}\{A\}] &= \mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} \left[ \sum_{n=0}^{\infty} F(Y_n, Y_{n+1}, \dots) \mathbb{1}\{A, N = n\} \right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} [F(Y_n, Y_{n+1}, \dots) \mathbb{1}\{A, N = n\}] \end{aligned}$$

but now, by the definition of  $\mathcal{F}_N$ , we have  $\{A, N = n\} \in \mathcal{F}_n$ , and so we may use the intermediate Markov property on every inner expectation, obtaining

$$= \sum_{n=0}^{\infty} \mathbb{E}_{Y \sim \mathbb{P}_{\mu_0}} \left[ \mathbb{E}_{Z \sim Y_n} [F(Z_0, Z_1, \dots)] \mathbb{1}\{A, N = n\} \right],$$

and undoing the first manipulations then gives the result.  $\square$

## 5.5 COUNTABLE STATE SPACE STRUCTURE THEORY

We now focus on the special case where the state space  $S$  is countable. In this case, we may consider the transition kernel between points,

$$p^{(1)}(x, y) = p(x, y) := \mathbb{P}_x[Y_1 = y] = p(x, \{y\}).$$

We also define the  $k$ -step transition kernels

$$p^{(k)}(x, y) := \mathbb{P}_x[Y_k = y]$$

and by the Chapman-Kolmogorov equation from Corollary 5.4.2 expanded in the countable case, we have

$$= \sum_{x_1, \dots, x_{k-1} \in S} p(x, x_1)p(x_1, x_2) \cdots p(x_{k-1}, y). \quad (5.5.1)$$

We may also associate a countable matrix  $\mathbf{P} \in [0, 1]^{S \times S}$  to the transition kernel, having entries  $P_{xy} = p(x, y)$ . In that case, by the above equation, the transition kernels with other numbers of steps are given by the matrix powers:

$$p^{(k)}(x, y) = (\mathbf{P}^k)_{xy}.$$

### 5.5.1 TRANSIENCE AND RECURRENCE

We now consider one of the fundamental questions about the structure of countable state space Markov chains: if we start from a given  $x \in S$ , which  $y \in S$  will we ever visit? (In some cases one can literally ask the same question for continuous state spaces, but often it makes more sense to ask about some quantitative version of approaching arbitrarily close to another point if, say,  $S = \mathbb{R}^d$ .) To this end, we define the stopping times at which we visit a point:

$$\begin{aligned} T_y^{(0)} &:= 0, \\ T_y^{(k+1)} &:= \min\{n > T_y^{(k)} : Y_n = y\} \text{ for each } k \geq 0. \end{aligned}$$

In words,  $T_y^{(k)}$  is the time of the  $k$ th visit to  $y$ , where even if we start at  $y$  we do not count that as a visit. We then consider

$$\rho_{xy} := \mathbb{P}_x[T_y^{(1)} < \infty] = \mathbb{P}[\text{ever visit } y \text{ starting from } x].$$

**Definition 5.5.1.** We call  $x \in S$  *recurrent* if  $\rho_{xx} = 1$  and *transient* if  $\rho_{xx} < 1$ .

The following intuitive calculation captures the idea that, to visit  $y$  some number  $k$  times, one must first visit  $y$  once, and then “circle back”  $k - 1$  more times, where each time  $y$  is visited one effectively “starts over” by the Markov property.

**Theorem 5.5.2.**  $\mathbb{P}_x[T_y^{(k)} < \infty] = \rho_{xy}\rho_{yy}^{k-1}$ .

*Proof.* The case  $k = 1$  is by definition of  $\rho_{xy}$ . For  $k \geq 2$ , we will apply the strong Markov property. Define  $F(\mathbf{y}) := \mathbb{1}\{\mathbf{y}_i = y \text{ for some } y_i \geq 1\}$ . The strong Markov property with respect to the stopping time  $T_y^{(k-1)}$  then gives

$$\begin{aligned} \mathbb{P}_x[T_y^{(k)} < \infty] &= \mathbb{E}[\mathbb{1}\{T_y^{(k)} < \infty\}] \\ &= \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} [F(Y_{T_y^{(k-1)}}), Y_{T_y^{(k-1)}+1}, \dots) \mathbb{1}\{T_y^{(k-1)} < \infty\}] \\ &= \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} \left[ \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} [F(Y_{T_y^{(k-1)}}), Y_{T_y^{(k-1)}+1}, \dots) \mid \mathcal{F}_{T_y^{(k-1)}}] \mathbb{1}\{T_y^{(k-1)} < \infty\} \right] \\ &= \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} \left[ \mathbb{E}_{\mathbf{Z} \sim \mathbb{P}_{Y_{T_y^{(k-1)}}}} [F(Z_0, Z_1, \dots)] \mathbb{1}\{T_y^{(k-1)} < \infty\} \right] \end{aligned}$$

but now, since by definition  $Y_{T_y^{(k-1)}} = y$ , we have

$$\begin{aligned} &= \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} \left[ \mathbb{E}_{\mathbf{Z} \sim \mathbb{P}_y} [F(Z_0, Z_1, \dots)] \mathbb{1}\{T_y^{(k-1)} < \infty\} \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim \mathbb{P}_y} [F(Z_0, Z_1, \dots)] \cdot \mathbb{E}_{\mathbf{Y} \sim \mathbb{P}_x} [\mathbb{1}\{T_y^{(k-1)} < \infty\}] \\ &= \rho_{yy} \cdot \mathbb{P}[T_y^{(k-1)} < \infty], \end{aligned}$$

and the result follows by induction.  $\square$

This result has some very useful consequences. First, we can derive the following basic more qualitative properties about visiting states infinitely many times.

**Corollary 5.5.3.** The following hold for any  $y \in S$ :

1. If  $y$  is recurrent, then  $\mathbb{P}_y[\text{return infinitely many times to } y] = 1$ .
2. If  $y$  is recurrent, then  $\mathbb{P}_x[\text{visit } y \text{ infinitely many times to}] = \rho_{xy}$  for any  $x \in S$ .
3. If  $y$  is transient, then  $\mathbb{P}_x[\text{visit } y \text{ infinitely many times to}] = 0$  for any  $x \in S$ .

Second, we obtain the following important characterization of recurrence. We define the random variable

$$N(y) := \#\{\text{visits to } y\} = \sum_{n=1}^{\infty} \mathbb{1}\{Y_n = y\}.$$

**Theorem 5.5.4.**  $y \in S$  is recurrent if and only if  $\mathbb{E}_y[N(y)] = \sum_{n=1}^{\infty} \mathbb{P}_y[Y_n = y] = \infty$ . If

$\mathcal{Y}$  is transient, then for any  $x \in S$ ,

$$\mathbb{E}_x[N(\mathcal{Y})] = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty.$$

One interpretation of the first part is as a statement that the Borel-Cantelli lemma is tight when applied to the events  $\{Y_n = \mathcal{Y}\}$  for  $n \geq 1$ .

*Proof.* We expand

$$\begin{aligned} \mathbb{E}_y N(\mathcal{Y}) &= \sum_{k=1}^{\infty} \mathbb{P}_y[N(\mathcal{Y}) \geq k] \\ &= \sum_{k=1}^{\infty} \mathbb{P}_y[T_y^{(k)} < \infty] \\ &= \sum_{k=1}^{\infty} \rho_{yy}^k, \end{aligned}$$

and the stated equivalence follows. The second claim follows from the same calculation from initial state  $x$ .  $\square$

### 5.5.2 APPLICATION: LATTICE RANDOM WALKS

As an application, we consider the simple random walk on  $\mathbb{Z}^d$ : this may be viewed as the lattice or “grid” graph where two integer vectors are adjacent if and only if they differ by exactly 1 in exactly one coordinate.

**Theorem 5.5.5 (Pólya).**  $\mathbf{0} \in \mathbb{Z}^d$  is recurrent for the simple random walk if and only if  $d \in \{1, 2\}$ .

*Proof Sketch.* It suffices to estimate  $\mathbb{P}[Y_n = \mathbf{0}]$  and apply Theorem 5.5.4. We give a heuristic way to do this. Note that we may view

$$Y_n = \sum_{i=1}^n \mathbf{v}_i$$

for  $\mathbf{v}_i \sim \text{Unif}(\{e_1, -e_1, \dots, e_d, -e_d\})$  i.i.d. random vectors. We calculate  $\mathbb{E}\mathbf{v}_i = \mathbf{0}$  and  $\mathbb{E}\mathbf{v}_i \mathbf{v}_i^\top = \frac{1}{d} \mathbf{I}_d$ . By the multivariate central limit theorem, we then have

$$\text{Law}\left(\frac{1}{\sqrt{n}} Y_n\right) \approx \mathcal{N}\left(\mathbf{0}, \frac{1}{d} \mathbf{I}_d\right).$$

More heuristically, for  $Y_n$  itself we have

$$\text{Law}(Y_n) \approx \mathcal{N}\left(\mathbf{0}, \frac{n}{d} \mathbf{I}_d\right).$$

Even more heuristically, let us estimate the probability that  $Y_n$  is exactly  $\mathbf{0}$  by the probability of falling in a small box around  $\mathbf{0}$ :

$$\begin{aligned}\mathbb{P}[Y_n = \mathbf{0}] &= \mathbb{P}\left[Y_n \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d\right] \\ &\approx \int_{\mathbb{R}^d} \mathbb{P}_{g \sim \mathcal{N}(\mathbf{0}, \frac{n}{d} \mathbf{I}_d)} \left[ \mathbf{g} \in \left[-\frac{1}{2}, \frac{1}{2}\right]^d \right] \\ &= \int_{\left[-\frac{1}{2}, \frac{1}{2}\right]^d} \frac{1}{\sqrt{\det(2\pi \frac{n}{d} \mathbf{I}_d)}} \exp\left(-\frac{1}{2} \mathbf{x}^\top \left(\frac{n}{d} \mathbf{I}_d\right)^{-1} \mathbf{x}\right) d\mathbf{x} \\ &= \left(\frac{2\pi}{d} n\right)^{-d/2} \int_{\left[-\frac{1}{2}, \frac{1}{2}\right]^d} \exp\left(-\frac{d}{2n} \|\mathbf{x}\|^2\right) d\mathbf{x}\end{aligned}$$

for large  $n$ , the integral is close to 1 (the volume of the box), and the leading term is up to a constant depending only on  $d$ ,

$$\asymp n^{-d/2}.$$

Thus, we have  $\mathbb{E}_0 N(\mathbf{0}) \asymp \sum_{n \geq 1} n^{-d/2}$ , which indeed diverges if and only if  $d \leq 2$ .  $\square$

### 5.5.3 ACCESSIBILITY GRAPH

We now return to the general structure theory for countable state spaces. In particular, we focus on the following notion.

**Definition 5.5.6.** We say that  $y \in S$  is *accessible* from  $x \in S$  if  $\rho_{xy} > 0$ . We also always let  $x \in S$  be accessible from itself. We denote this relation  $x \rightarrow y$ , and write  $x \leftrightarrow y$  if both  $x \rightarrow y$  and  $y \rightarrow x$ . This defines a directed graph on  $S$ , which we call the *accessibility graph*.

We next establish some of the basic properties of the accessibility graph.

**Proposition 5.5.7.** If  $\rho_{xy} > 0$ , then there exists some  $k \geq 1$  and  $x_1, \dots, x_{k-1} \in S$  such that  $p^{(k)}(x, y) \geq p(x, x_1)p(x_1, x_2) \cdots p(x_{k-1}, y) > 0$ .

*Proof.* We have by union bound

$$\rho_{xy} = \mathbb{P}_x[Y_k = y \text{ for some } k \geq 1] \leq \sum_{k \geq 1} \mathbb{P}_x[Y_k = y] = \sum_{k \geq 1} p^{(k)}(x, y),$$

so if  $\rho_{xy}$  then there must exist  $k$  such that  $p^{(k)}(x, y) > 0$ . The existence of a suitable path  $x_1, \dots, x_{k-1}$  then follows by the expansion in the Chapman-Kolmogorov equation (5.5.1).  $\square$

**Proposition 5.5.8.** If  $x \rightarrow y$  and  $y \rightarrow z$ , then  $x \rightarrow z$ .

*Proof.* If either  $x = y$  or  $y = z$ , the result is immediate. Otherwise, we have  $\rho_{xy}, \rho_{yz} > 0$ . By Proposition 5.5.7, there exist  $k, \ell$  such that  $p^{(k)}(x, y), p^{(\ell)}(y, z) > 0$ . By the Chapman-Kolmogorov equation, we then have  $p^{(k+\ell)}(x, z) > 0$ .  $\square$

**Theorem 5.5.9.** If  $x$  is recurrent and  $x \rightarrow y$ , then  $y$  is recurrent and  $\rho_{xy} = \rho_{yx} = 1$ . In particular, we also have  $x \leftrightarrow y$ .

You can think of the first claim as saying that *recurrence is contagious under accessibility*.

*Proof.* The case  $x = y$  is immediate, so suppose  $x \neq y$ .

We first show  $\rho_{yx} = 1$ . Suppose otherwise. Since  $\rho_{xy} > 0$ , by Proposition 5.5.7 there exist  $x_1, \dots, x_{k-1} \in S$  such that  $p(x, x_1) \cdots p(x_{k-1}, y) > 0$ . Choose the smallest possible such  $k$ , then we must have that  $x_1, \dots, x_{k-1} \neq x$  (or else a suffix of these sequence would be a shorter one with the same property). Then, as you may check by the intermediate Markov property, we have

$$\begin{aligned} 0 &= \mathbb{P}_x[\text{never return to } x] \\ &\geq p(x, x_1) \cdots p(x_{k-1}, y) \mathbb{P}_y[\text{never visit } x] \\ &= p(x, x_1) \cdots p(x_{k-1}, y) \cdot (1 - \rho_{yx}). \end{aligned}$$

The right-hand side must equal 0, but the first factor is positive, so we must have  $\rho_{yx} = 1$ .

We next show that  $y$  is recurrent. We verify the condition of Theorem 5.5.4 for this. Since  $\rho_{yx} = 1 > 0$ , again by Proposition 5.5.7 there must exist an  $\ell$  such that  $p^{(\ell)}(y, x) > 0$ . Then, we have

$$\begin{aligned} \mathbb{E}_y N(y) &= \sum_{n \geq 1} \mathbb{P}_y[Y_n = y] \\ &= \sum_{n \geq 1} p^{(n)}(y, y) && \geq \sum_{n \geq 1} p^{(\ell+n+k)}(y, y) \end{aligned}$$

and by the Chapman-Kolmogorov equation

$$\begin{aligned} &\geq \sum_{n \geq 1} p^{(\ell)}(y, x) p^{(n)}(x, x) p^{(k)}(x, y) \\ &= p^{(\ell)}(y, x) p^{(k)}(x, y) \cdot \sum_{n \geq 1} p^{(n)}(x, x) \\ &= \infty, \end{aligned}$$

the conclusion following since the first factor is positive and the second diverges by the recurrence of  $x$ .

Finally, to show  $\rho_{xy} = 1$  we may simply reverse the roles of  $x$  and  $y$  and repeat the argument for  $\rho_{yx} = 1$ .  $\square$

From this, we learn some structure theory about the subgraph of the accessibility graph induced on the recurrent states. Let us define

$$R := \{x \in S : x \text{ recurrent}\},$$

$$R_x := \{y \in S : x \rightarrow y\}.$$

**Corollary 5.5.10.** The following hold:

1.  $\leftrightarrow$  is an equivalence relation on  $R$ , and  $R_x$  is the equivalence class of  $x$  under this relation.
2. There exist  $\{x_i\}_{i \in \mathcal{I}} \subseteq S$  for some countable (finite or countably infinite) set  $\mathcal{I}$  such that  $R_{x_i}$  are disjoint and  $R = \bigsqcup_{i \in \mathcal{I}} R_{x_i}$ .
3. The induced subgraph of the accessibility graph on any  $R_{x_i}$  for  $i \in \mathcal{I}$  contains every edge oriented in both directions.

#### 5.5.4 FUNDAMENTAL STRUCTURE THEOREM

Finally we arrive at our main decomposition of the accessibility graph.

**Definition 5.5.11.** We say that a set of states  $A \subseteq S$  is:

- *closed* if whenever  $x \in A$  and  $\rho_{xy} > 0$  then  $y \in A$ , and
- *irreducible* if  $\rho_{xy} > 0$  for all  $x, y \in A$ .

The following are some basic properties of these notions that you can verify as an exercise.

**Proposition 5.5.12.** The following hold:

1. If  $A$  is closed and  $x \in A$ , then  $\mathbb{P}_x[Y_n \in A] = 1$  for any  $n \geq 1$ .
2. Each  $R_x$  is both closed and irreducible.

The following gives a partial converse to the second point of the previous Proposition.

**Proposition 5.5.13.** Let  $A \subseteq S$  be finite. Then, the following hold:

1. If  $A$  is closed, then there exists  $x \in A$  that is recurrent.
2. If  $A$  is closed and irreducible, then all  $x \in A$  are recurrent and  $A = R_x$  for any

$x \in A$ . Thus, if  $A$  is finite, then  $A$  is closed and irreducible if and only if  $A = R_x$  for some recurrent  $x$ .

*Proof.* For the first part, suppose that  $A$  is closed but all  $x \in A$  are transient. Pick some  $x \in A$ . Then, using Theorem 5.5.4 and the first part of Proposition 5.5.12, we have

$$\infty < \sum_{y \in A} \frac{\rho_{xy}}{1 - \rho_{yy}} = \sum_{y \in A} \mathbb{E}_x[N(y)] = \mathbb{E}_x \sum_{y \in A} \sum_{n=1}^{\infty} \mathbb{1}\{Y_n = y\} = \mathbb{E}_x \sum_{n=1}^{\infty} 1 = \infty,$$

a contradiction. For the second part, by the first part there exists  $x \in A$  recurrent, so by the irreducibility of  $A$  and Theorem 5.5.9, all  $x \in A$  are recurrent.  $\square$

**Example 5.5.14.** The case of the simple random walk on  $S = \mathbb{Z}^3$  and  $A = S$  gives an example of an infinite  $A$  that is closed but consists only of transient states.

Thus the partition  $R = \bigsqcup_{i \in \mathcal{I}} R_{x_i}$  achieved above is into disjoint subsets that are both closed and irreducible. Our main theorem generalizes this to a partition of all of  $S$ .

**Theorem 5.5.15** (Structure theorem for countable state space Markov chains). The relation  $\leftrightarrow$  is an equivalence relation on  $S$  whose equivalence classes are irreducible sets of states. Let  $S = \bigsqcup_{i \in \mathcal{I}} C_i$  be the partition into these classes for some countable index set  $\mathcal{I}$ . Define a directed graph  $H$  on vertex set  $\{C_i\}_{i \in \mathcal{I}}$ , where we include an edge  $C_i \rightarrow C_j$  if and only if there exists  $x \in C_i$  and  $y \in C_j$  such that  $x \rightarrow y$  (or equivalently  $\rho_{xy} > 0$ ). This may be viewed as a suitable kind of quotient of the full accessibility graph  $G$  on vertex set  $S$ . We call a vertex a *leaf* of a directed graph if it has no outgoing edges. This decomposition and graph structure have the following properties:

1. For each  $i \in \mathcal{I}$ , the induced subgraph of  $G$  on vertex set  $C_i$  includes every edge in both orientations.
2. The directed graph  $H$  is acyclic (containing no directed cycles).
3.  $C_i$  is a leaf of  $H$  if and only if it is closed (recall that all  $C_i$  are by construction also irreducible).
4. Every  $R_x$  is some leaf of  $H$ ; in particular, all recurrent states lie in the leaves of  $H$ .
5. Every finite leaf of  $H$  is  $R_x$  for some recurrent  $x$ .

*Proof.* All results are immediate from our previous ones except that  $H$  is acyclic. To see this, note that if there were a cycle among some  $C_{i_1}, \dots, C_{i_k}$ , then since all pairs of vertices in a given  $C_i$  are mutually accessible, there would be  $x_{i_a} \in C_{i_a}$  such that  $x_{i_a} \leftrightarrow x_{i_{a+1}}$  for each  $a \in [k]$ , with index arithmetic modulo  $k$ . But then, any  $x \in C_{i_a}$  and  $y \in C_{i_b}$  would be mutually accessible, and so the  $C_{i_a}$  would all belong to the same equivalence class of the relation  $\leftrightarrow$ .  $\square$

**Corollary 5.5.16.** If  $S$  is finite, then the leaves of  $H$  are all among the  $R_x$  and their union is exactly  $R$ .

**Corollary 5.5.17.** If all of  $S$  is irreducible and any  $x \in S$  is recurrent, then all  $x \in S$  are recurrent and  $H$  has a single vertex that is all of  $S$ . Further, if  $S$  is finite and irreducible then all  $x \in S$  are recurrent.

**Example 5.5.18.** Any leaf of  $H$  that contains transient states must be infinite. A simple example is a random walk on  $\{v\} \sqcup \mathbb{Z}^3$  that from  $v$  stays at  $v$  with probability  $1/2$  and goes to  $0 \in \mathbb{Z}^3$  with probability  $1/2$ , and on  $\mathbb{Z}^3$  evolves according to the simple random walk. Then, the equivalence classes of  $\leftrightarrow$  are  $\{v\}$  and  $\mathbb{Z}^3$ , the graph  $H$  is  $\{v\} \rightarrow \mathbb{Z}^3$ , and  $\mathbb{Z}^3$  is an infinite leaf all of whose states are transient.

## 5.6 STATIONARY MEASURES, CONVERGENCE, AND ERGODICITY

In the above context, we say that  $S$  is *recurrent* if all  $x \in S$  are recurrent. We now focus mostly on the case where  $S$  is irreducible and recurrent and consider “how much time” a Markov chain spends in each state. To make this more precise, consider the probability measures on  $S$

$$\mu_{x,n}(\mathcal{Y}) := \mathbb{P}_x[Y_n = \mathcal{Y}].$$

We ask whether the  $\mu_{x,n}$  converge to some limiting  $\mu$ . Note that we have, by the Chapman-Kolmogorov equation,

$$\begin{aligned} \mu_{x,n}(\mathcal{Y}) &= \mathbb{P}_x[Y_n = \mathcal{Y}] \\ &= \mathbf{p}^{(n)}(x, \mathcal{Y}) \\ &= \sum_{z \in S} \mathbf{p}^{(n-1)}(x, z) \mathbf{p}(z, \mathcal{Y}) \\ &= \sum_{z \in S} \mathbf{p}^{(n-1)}(x, z) \mathbf{p}(z, \mathcal{Y}) \\ &= \sum_{z \in S} \mu_{x,n-1}(z) \mathbf{p}(z, \mathcal{Y}). \end{aligned}$$

Thus, since the hypothetical limiting  $\mu$  should be a fixed point of this relation, we make the following definition.

**Definition 5.6.1** (Stationary measure). A measure  $\mu$  (not necessarily a probability mea-

sure) is *stationary* if, for all  $x \in S$ ,

$$\mu(x) = \sum_{y \in S} \mu(y) p(y, x) = \sum_{y \in S} \mu(y) p^{(k)}(y, x)$$

for any  $k \geq 1$ .

We note that the case  $k \geq 2$  follows from  $k = 1$  by the Chapman-Kolmogorov equation.

The following examples show that we will need to be careful in understanding when there exists a stationary probability measure, and when that is unique.

**Example 5.6.2.** Consider the simple random walk on  $\mathbb{Z}$ .  $\mu$  is stationary if and only if, for all  $x \in \mathbb{Z}$ ,  $\mu(x) = \frac{1}{2}(\mu(x-1) + \mu(x+1))$ . You may check that this implies that there is no stationary probability measure, and all stationary measures are of the form  $\mu(x) = c$  for all  $x \in S$ , for some  $c \geq 0$ .

**Example 5.6.3.** Consider the Markov chain on a set  $S$  where  $p(x, x) = 1$  for all  $x \in S$ . Then, *any* measure  $\mu$  (probability or otherwise) is stationary.

### 5.6.1 EXISTENCE AND UNIQUENESS

Define the time

$$T_x := \min\{n \geq 1 : Y_n = x\}.$$

When we initialize from  $x$ , then this is the first time we return to  $x$ .

**Theorem 5.6.4 (Existence).** Suppose  $x$  is recurrent. Define

$$\mu_x(y) := \mathbb{E}_x \sum_{n=0}^{T_x-1} \mathbb{1}\{Y_n = y\} = \sum_{n=0}^{\infty} \mathbb{P}_x[Y_n = y, T_x > n].$$

Then,  $\mu_x$  is a stationary measure.

**Remark 5.6.5.** We make the following notes:

- It is an exercise in your homework to show that  $\mu_x(y) < \infty$  for all  $y$ , so we leave this out of the proof below.
- $\mu_x$  is a finite measure, and therefore renormalizable to a probability measure, if and only if  $\mathbb{E}_x T_x < \infty$ . Thus for instance from the above reasoning in Example 5.6.2 we find that for the simple random walk on  $\mathbb{Z}$  we have  $\mathbb{E}_0 T_0 = \infty$ .
- We have  $\mu_x(y) > 0$  if and only if  $\rho_{xy} > 0$  if and only if  $y \in R_x$ .
- We have  $\mu_x(x) = 1$ .

*Proof.* Let us define

$$q_n(x, y) := \mathbb{P}_x[Y_n = y, T_x > n].$$

After expanding the definition of stationarity and  $\mu_x$ , we need to show that, for each  $z \in S$ , we have

$$\sum_{n=0}^{\infty} q_n(x, z) = \sum_{n=0}^{\infty} \sum_{y \in S} q_n(x, y) p(y, z).$$

We consider two cases separately.

First, suppose  $z \neq x$ . Then, we have

$$\begin{aligned} \sum_{n=0}^{\infty} \sum_{y \in S} q_n(x, y) p(y, z) &= \sum_{n=0}^{\infty} \sum_{y \in S} \mathbb{P}_x[Y_n = y, T_x > n] p(y, z) \\ &= \sum_{n=0}^{\infty} \sum_{y \in S} \mathbb{P}_x[Y_n = y, T_x > n, Y_{n+1} = z] \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x[Y_{n+1} = z, T_x > n + 1] \\ &= \sum_{n=0}^{\infty} q_{n+1}(x, z), \end{aligned}$$

and this gives the result since  $q_0(x, z) = 0$  whenever  $x = z$ .

Next, suppose  $z = x$ . We have by a similar calculation

$$\begin{aligned} \sum_{n=0}^{\infty} \sum_{y \in S} q_n(x, y) p(y, x) &= \sum_{n=0}^{\infty} \sum_{y \in S} \mathbb{P}_x[Y_n = y, T_x > n] p(y, x) \\ &= \sum_{n=0}^{\infty} \sum_{y \in S} \mathbb{P}_x[Y_n = y, T_x > n, Y_{n+1} = x] \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x[T_x = n + 1] \\ &= \sum_{n=0}^{\infty} \mathbb{P}_x[T_x = n] \\ &= 1 \\ &= \mu_x(x), \end{aligned}$$

where we have used that  $\mathbb{P}_x[T_x = 0] = 0$  by construction. □

**Theorem 5.6.6 (Uniqueness).** If  $S$  is irreducible and recurrent, and  $\mu, \nu \neq 0$  are stationary, then  $\mu = c \cdot \nu$  for some  $c > 0$ .

**Example 5.6.7.** Combining with the reasoning in Example 5.6.2, we find that indeed all of the stationary measures are the constant measures. Thus,  $\mu_0$  must also be a constant measure,  $\mu_0(y) = c$  for all  $y \in \mathbb{Z}$ . Since  $\mu_0(0) = 1$ , we must have  $c = 1$ . Interpreting this in words, the expected number of visits by the simple random walk to *any*  $y \in \mathbb{Z}$  before returning to 0, no matter how large or small, equals 1.

**Proposition 5.6.8.** Suppose  $S$  is irreducible and recurrent and  $\mu \neq 0$  is a stationary measure. Then,  $\mu(x) > 0$  for all  $x \in S$ .

*Proof.* Since  $\mu \neq 0$ , we may pick some  $x \in S$  for which  $\mu(x) > 0$ . Let  $y \in S$ . Since  $S$  is irreducible,  $\rho_{xy} > 0$ , and thus by Proposition 5.5.7 there exists some  $k \geq 1$  such that  $p^{(k)}(x, y) > 0$ . Then, by stationarity,

$$\mu(y) = \sum_z \mu(z) p^{(k)}(z, y) \geq \mu(x) p^{(k)}(x, y) > 0,$$

as claimed. □

We note that this also admits a partial converse, as follows.

**Proposition 5.6.9.** Suppose  $\mu$  is a stationary probability measure and  $\mu(x) > 0$ . Then,  $x$  is recurrent.

*Proof.* We have, using Theorem 5.5.4, since  $x$  is recurrent

$$\begin{aligned} \infty &= \sum_{n=1}^{\infty} \mu(x) \\ &= \sum_{n=1}^{\infty} \sum_{y \in S} \mu(y) p^{(n)}(y, x) \\ &= \sum_{y \in S} \mu(y) \sum_{n=1}^{\infty} \mathbb{P}[Y_n = x] \\ &= \sum_{y \in S} \mu(y) \mathbb{E}_y[N(x)] \\ &= \sum_{y \in S} \mu(y) \frac{\rho_{yx}}{1 - \rho_{xx}} \end{aligned}$$

and using that  $\mu(y)$  is a probability measure and  $\rho_{yx} \leq 1$ , we have

$$\leq \frac{1}{1 - \rho_{xx}}.$$

Thus we must have  $\rho_{xx} = 1$  and  $x$  is recurrent. □

*Proof of Theorem 5.6.6.* Let  $\mu^{(0)} \neq 0$  be stationary and pick some  $x \in S$ . By Proposition 5.6.8,  $\mu^{(0)}(x) > 0$ . Let  $\mu(y) := \mu^{(0)}(y)/\mu^{(0)}(x)$ . This is also a stationary measure that is proportional to  $\mu^{(0)}$  and satisfies  $\mu(x) = 1$ . It then suffices to show that *any* stationary measure  $\mu$  with  $\mu(x) = 1$  has  $\mu = \mu_x$ .

It further suffices to show that  $\mu \geq \mu_x$ . If we have this, then let  $\mu'(y) := \mu(y) - \mu_x(y)$ . This is again another stationary measure, with  $\mu'(x) = 0$ . But then, by Proposition 5.6.8 we must have  $\mu' = 0$ , and so  $\mu = \mu_x$ .

To show  $\mu = \mu_x$ , we consider the following kind of expansion, using stationarity repeatedly:

$$\begin{aligned} \mu(y) &= \sum_{z_0} \mu(z_0)p(z_0, y) \\ &= p(x, y) + \sum_{z_0 \neq x} \mu(z_0)p(z_0, y) \\ &= p(x, y) + \sum_{z_0 \neq x} \sum_{z_1} \mu(z_1)p(z_1, z_0)p(z_0, y) \\ &= p(x, y) + \sum_{z_0 \neq x} p(x, z_0)p(z_0, y) + \sum_{z_0, z_1 \neq x} \mu(z_1)p(z_1, z_0)p(z_0, y). \end{aligned}$$

Note we may interpret the first two terms here as  $\mathbb{P}_x[Y_1 = y, T_x > 1]$  and  $\mathbb{P}_x[Y_2 = y, T_x > 2]$ . Repeating this procedure and throwing out the last sum, we get that, for each  $N \geq 1$ ,

$$\mu(y) \geq \sum_{n=1}^N \mathbb{P}_x^x[Y_n = y, T_x > n] = \sum_{n=1}^N q_n(x, y).$$

Taking  $N \rightarrow \infty$  then gives  $\mu(y) \geq \mu_x(y)$ . □

## 5.6.2 CONVERGENCE

We now focus on the case where a Markov chain has  $S$  irreducible and recurrent and has a stationary probability measure  $\mu$ , which then by Theorem 5.6.6 is unique. We ask, when do we have that  $\mu_{x,n} \xrightarrow{(w)} \mu$ ?

One simple obstruction is *periodicity*: consider for example the simple random walk on a cycle graph of even length. It is easy to check that the unique stationary probability measure is the uniform measure on the vertices. On the other hand,  $\mu_{x,n}$  is supported on disjoint subsets of vertices depending on whether  $n$  is even or odd. We make a definition to exclude this possibility.

**Definition 5.6.10.** A Markov chain is *aperiodic* if, for all  $x \in S$  recurrent, there exists some  $n_0 = n_0(x) \in \mathbb{N}$  such that  $p^{(n)}(x, x) > 0$  for all  $n \geq n_0$ . Equivalently, the greatest common divisor of the integers  $\{n \in \mathbb{N} : p^{(n)}(x, x) > 0\}$  equals 1 (you may verify this equivalence as a simple number-theoretic exercise).

The following expresses that this is in fact the *only* obstruction.

**Theorem 5.6.11.** Suppose a Markov chain on  $S$  is aperiodic and has  $S$  irreducible and recurrent. Suppose also this chain has a stationary probability measure  $\mu$ . Then, for all  $x \in S$ ,  $\mu_{x,n} \xrightarrow{(w)} \mu$ .

## 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.
- [CP97] Joseph T Chang and David Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.
- [Kle14] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, second edition, 2014.
- [KS66] Harry Kesten and Bernt P Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966.
- [O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.
- [Pol02] David Pollard. *A user's guide to measure theoretic probability*. Number 8. Cambridge University Press, 2002.
- [Tak65] Lajos Takács. A moment problem. *Journal of the Australian Mathematical Society*, 5(4):487–490, 1965.
- [Tju75] Tue Tjur. A constructive definition of conditional distributions. Technical report, Institute of Mathematical Statistics, University of Copenhagen, 1975.
- [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.
- [Wäs20] Johan Wästlund. Padlock solitaire: A martingale trick for combinatorial enumeration. *arXiv preprint arXiv:2008.13017*, 2020.
- [Wor99] Nicholas C Wormald. Models of random regular graphs. *London Mathematical Society Lecture Note Series*, pages 239–298, 1999.