Chapter 1

Some Background and Preliminaries

In that it is an attempt to model mathematically a phenomenon that may or may not actually exist, probability theory is a rather peculiar subject. If two people play a game of dice and one of them repeatedly wins, is the explanation that the winner is "a chosen of the gods" or is there something more that can be said? Until the late Middle Ages, most of the Westerners who gave it any thought interpreted *luck* as a manifestation of an individual's standing with whatever divinity or divinities were in vogue. Since other inexplicable phenomena were also assumed to have divine origins, this interpretation was reasonable. On the other hand, like the biblical account of the origins of life, attributing *luck* to divine sources leaves little room for further analysis. Indeed, it is a prerogative of any self-respecting deity to be inscrutable, and therefore one is rude, if not sinful, to subject the motives of one's deity to detailed analysis.

Simply abandoning a divine explanation of *luck* does not solve the problem but only opens it to further inquiry. For instance, if one believes that all experience is a corollary of "the laws of nature," then there is no such thing as *luck*. One person wins more often than the other because "the laws of nature" dictate that outcome. From this hyper-rational perspective, the concept of *luck* is a cop-out: a crutch that need never be used if one is sufficiently diligent in one's application of "the laws of nature." Although its origins may be strictly rational, this reason for denying the existence of *luck* does little to advance one's understanding of many phenomena. Even if one accepts Newton's laws of motion as sacrosanct, it is unlikely that one will ever to able to solve his equations of motion for Avogadro's number of particles, and, if one could, there is considerable doubt that one would be able to extract useful information from the solution. Thus, replacing a divine with a mechanistic explanation of *luck* only substitutes one imponderable by another.

In the 17th century, a few Europeans introduced a wholly new way of thinking about *luck*. Even the idea of thinking about *luck* instead of just accepting it as a phenomenon incapable of analysis requires an entirely new mindset. Spurred by questions posed by Chevalier de Méré (a nobleman with a more than passing interest in gambling), Blaise Pascal (of triangular fame) and Pierre de Fermat (of conjectural fame) began formulating a mathematical model which can be seen as the origins of what we now call *the theory of probability*. Loosely speaking, their thinking was based on the idea that, even if one cannot predict the particular outcome of something like a game of chance, one nonetheless knows all the possible outcomes. Further, one often has reason to believe that one knows with what *probability* each outcome will occur. Hence, one can compute the probability of an *event* (i.e., a collection of the possible outcomes) by adding the probabilities of the individual outcomes making up the event. For instance, if a *fair* (i.e., unbiased) coin is tossed two times, then it is reasonable to suppose that each of the four outcomes (H, H) (i.e., heads on both the first and second tosses), (H, T) (i.e., heads on the first and tails on the second), (T, H), and (T, T) is equally likely to occur. That is, each of these outcomes has probability $\frac{1}{4}$, and the probability of the event that one T and one H occur is therefore $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Alternatively, if one knows that the coin is biased and that heads occur twice as often as tails, then one assigns the preceding list of outcomes probabilities $\frac{4}{9}$, $\frac{2}{9}$, $\frac{2}{9}$, and $\frac{1}{9}$, and therefore the event one H and one T has probability $\frac{4}{9}$.

During the period since their introduction, Pascal's and Fermat's ideas have been refined and applied in venues which their authors could not have anticipated, and the goal of this book is to provide an introduction to some of these developments.

1.1. The Language of Probability Theory

Like any other topic in mathematics, probability has its own language. Because the terminology is chosen to reflect the role of probability theory as a model of random phenomena, it sometimes differs from the choice made elsewhere in mathematics. Thus, although I assume that my readers are familiar with most of the concepts discussed in this section, they may not immediately recognize the terminology that probabilists use to describe them.

1.1.1. Sample Spaces and Events.

I assume that my readers are familiar with the rudiments of naïve set theory and therefore, except for the notation and terminology, will find here very little that is new.

Before one does anything of a set-theoretic nature, one has to specify the universe in which one is working. In probability theory, the role of universe is played by the **sample space**, often denoted by Ω, which is a non-empty set that should be thought of as the space of all possible outcomes of the experiment or game under consideration. An element ω of Ω (abbreviated by $\omega \in \Omega$) is called a **sample point**, and a subset A of Ω (abbreviated by $A \subseteq \Omega$) is called an **event**.

Events are usually described in terms of some property that sample points may or may not possess. To wit, if P is a property, then one writes $\{\omega \in \Omega : \omega \text{ has property } P\}$ to denote the event that the observed outcome has property P. To simplify the description of events, it is important to make a judicious choice of the sample space. Above, I took the sample space for two tosses of a coin to be $\{H,T\}^2 = \{(H,H),(H,T),(T,H),(T,T)\}$, the set of ordered pairs whose components are either H or T, and considered the event A determined by the occurrence of one H and one T. For many purposes, a more clever choice would have been the set of ordered pairs $\{0,1\}^2 = \{(1,1),(1,0),(0,1),(0,0)\}$. With the latter choice, the event that one head and one tail occur would have been $\{\omega = (\omega_1,\omega_2) \in \{0,1\}^2 : \omega_1+\omega_2 = 1\}$. More generally, what one looks for is a sample space on which there are functions in terms of which interesting events are easily described. That is, it is often useful to describe an event as $\{\omega : F(\omega) \in \Gamma\}$, where F is a function on Ω and Γ is a subset of its possible values. In the future, I will usually remove the ω from such a description and will abbreviate it by $\{F \in \Gamma\}$.

Very often, one describes an event in terms of other events. Thus, if A and B are events, then their **union** $\{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$, denoted by $A \cup B$, is the event that the outcome has either the property P_A determining A or[1] the property P_B determining B. More generally, if $\{A_i : i \in \mathcal{I}\}$ is a family of events indexed by the index set $\mathcal{I}$, then

$$\bigcup_{i\in\mathcal{I}} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for some } i \in \mathcal{I}\}.$$

The **intersection** $\{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ of A and B, denoted by $A \cap B$, is the event that the outcome has both properties P_A and P_B. Just as in the case of unions, this operation can be applied to a family $\{A_i : i \in \mathcal{I}\}$, and one writes $\bigcap_{i\in\mathcal{I}} A_i$ to denote the event

$$\{\omega \in \Omega : \omega \in A_i \text{ for all } i \in \mathcal{I}\}.$$

[1] The "or" here is non-exclusive. That is, $A \cup B$ includes ω's that are in both A and B.

When the properties P_A and P_B are mutually exclusive and therefore $A \cap B = \emptyset$, where $\emptyset$ denotes the **empty set** (the set having no elements), one says that A and B are **disjoint**.

Writing $\omega \notin A$ to mean that the sample point ω is not an element of the event A, one defines the **complement** of A to be the event $\{\omega \in \Omega : \omega \notin A\}$, denoted by $A\complement$, consisting of those outcomes for which property P_A fails to hold. In this connection, the **difference**, denoted by $A \setminus B$, between events A and B is the event $A \cap B\complement = \{\omega \in A : \omega \notin B\}$. Thus, $A\complement = \Omega \setminus A$.

Exercises 1.1.8 and 1.1.10 are about various more or less obvious relationships between these set-theoretic operations.

1.1.2. Probability Measures. Sample spaces and events have no value unless one has a way to assign probabilities to them. Indeed, the goal of probability theory is to provide a rational way to compute the probability of events. Thus, there is a third, and essential, ingredient. Namely, one wants a function $\mathbb{P}$ that assigns to an event the probability that it occurs. There are three basic requirements that the function $\mathbb{P}$ must satisfy. First, the probability of any event must be non-negative. Second, since Ω contains all the possible outcomes and some outcome must occur, one requires that Ω have probability 1. Third, when events A and B are disjoint, one requires that the probability of their union be the sum of their individual probabilities. Thus,

$$\begin{gathered}\mathbb{P}(A) \geq 0 \text{ for all events } A, \quad \mathbb{P}(\Omega) = 1,\\ \text{and } \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \text{ if } A \cap B = \emptyset.\end{gathered} \tag{1.1.1}$$

Several additional properties follow immediately from (1.1.1). For instance, because $\Omega \cup \emptyset = \Omega$ and $\Omega \cap \emptyset = \emptyset$, one has that

$$1 = \mathbb{P}(\Omega) = \mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) = 1 + \mathbb{P}(\emptyset),$$

and therefore $\mathbb{P}(\emptyset) = 0$. Also, if $A \subseteq B$, then,

$$\mathbb{P}(B) = \mathbb{P}\big(A \cup (B \setminus A)\big) = \mathbb{P}(A) + \mathbb{P}(B \setminus A),$$

and so $0 \leq \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$. As an application, since $A \cup B = A \cup \big(B \setminus (A \cap B)\big)$, we have that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}\big(B \setminus (A \cap B)\big) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

for any events A and B. Summarizing, we have shown that

$$\begin{gathered}\mathbb{P}(\emptyset)=0, \quad A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B) \text{ and } \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A),\\ \text{and } \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).\end{gathered} \tag{1.1.2}$$

All these properties should be seen as consistent with our goal of measuring the probability of events. In that some outcome must occur, it is clear that the null event must have probability 0. When event B contains event A, it should be the more probable of the two, and the event that B occurs but not the event A should equal the difference between the probabilities of B and A. Finally, when $A \cap B \neq \emptyset$, then $\mathbb{P}(A)+\mathbb{P}(B)$ fails to take into account that the sample points in both are getting counted twice and therefore must be corrected by subtracting off $\mathbb{P}(A \cap B)$ in order to arrive at the probability of $A \cup B$.

In addition to the preceding, one needs an extension of the additivity property for disjoint events. Namely, by induction on $n \geq 2$, one can easily check that

$$\text{(1.1.3)} \quad \mathbb{P}\big(A_1 \cup \cdots \cup A_n\big) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) \quad \text{if } A_k \cap A_\ell = \emptyset \text{ for } k \neq \ell.$$

However, one wants to know that the same additivity property holds for a countable number of mutually disjoint events. That is, if $\{A_k : k \geq 1\}$ is a sequence of events, then one wants to know that

$$\text{(1.1.4)} \qquad \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k) \quad \text{if } A_k \cap A_\ell = \emptyset \text{ for } k \neq \ell.$$

Equivalently, one is insisting that $\mathbb{P}$ be *continuous under monotone convergence.* To be precise, say that the sequence of events $\{B_n : n \geq 1\}$ **increases** to the event B, denoted by $B_n \nearrow B$, if $B_n \subseteq B_{n+1}$ and $B = \bigcup_{n=1}^{\infty} B_n$. Then (1.1.4) is equivalent to

$$\text{(1.1.5)} \qquad \mathbb{P}(B_n) \nearrow \mathbb{P}(B) \quad \text{if } B_n \nearrow B.$$

To check this equivalence, suppose $B_n \nearrow B$, and set $A_1 = B_1$ and $A_k = B_k \backslash B_{k-1}$ for $k \geq 2$. Then the A_k's are mutually disjoint, $B_n = \bigcup_{k=1}^{n} A_k$, and $B = \bigcup_{k=1}^{\infty} A_k$. Hence, since by (1.1.3), $\mathbb{P}(B_n) = \sum_{k=1}^{n} \mathbb{P}(A_k)$, (1.1.5) holds if and only if (1.1.4) does. Similarly, say that $\{B_n : n \geq 1\}$ **decreases** to B and write $B_n \searrow B$ if $B_n \supseteq B_{n+1}$ and $B = \bigcap_{n=1}^{\infty} B_n$. Then, since $B_n \searrow B \implies B_n\complement \nearrow B\complement$,

$$1 - \mathbb{P}(B_n) = \mathbb{P}(B_n\complement) \nearrow \mathbb{P}(B\complement) = 1 - \mathbb{P}(B),$$

and therefore

$$\text{(1.1.6)} \qquad \mathbb{P}(B_n) \searrow \mathbb{P}(B) \quad \text{if } B_n \searrow B.$$

Finally, observe that, even if the A_k's are not mutually disjoint, nonetheless

$$\text{(1.1.7)} \qquad \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

To see this, take $C_1 = A_1$ and $C_k = A_k \setminus \bigcup_{j=1}^{k-1} A_j$ for $k \geq 2$. Then the C_k's are mutually disjoint, and their union is the same as the union of the A_k's.

Hence, by (1.1.4),

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(C_k) \le \sum_{k=1}^{\infty} \mathbb{P}(A_k),$$

since $C_k \subseteq A_k$ and therefore $\mathbb{P}(C_k) \le \mathbb{P}(A_k)$ for all $k \ge 1$.

A function $\mathbb{P}$ on events that possesses the properties in (1.1.1) and (1.1.4) is called a **probability measure** on Ω. It turns out that, when Ω is uncountable, there are logical obstructions[2] to constructing probability measures which are defined for *all* events, and, to overcome these obstructions, one usually has to settle for a probability measure that is defined only on a carefully specified class of events. However, in this chapter we avoid these problems by restricting our attention to countable sample space, where these technicalities do not arise.

Exercises for § 1.1

Exercise 1.1.8. Here are a couple of elementary exercises in set theory.

(**i**) One of Norbert Wiener's least renowned contributions to mathematics was his set-theoretic formulation of an ordered pair. Namely, given a and b, define the ordered pair $(a, b) = \big\{\{a\}, \{a, b\}\big\}$, and show that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.

(**ii**) Let $\mathcal{I} \neq \emptyset$ be an index set, and show that $\big(\bigcup_{i\in\mathcal{I}} A_i\big)\complement = \bigcap_{i\in\mathcal{I}} A_i\complement$. This equality is sometimes called **De Morgan's Law**.

Exercise 1.1.9. Given a probability measure $\mathbb{P}$ on a sample space Ω, $N \ge 2$, and events $A_1, \dots, A_N$, use (1.1.2) and induction on N to show that

$$\mathbb{P}\big(A_1 \cup \cdots \cup A_N\big) = -\sum_{F} (-1)^{\mathrm{card}(F)} \mathbb{P}(A_F),$$

where F runs over non-empty subsets of $\{1, \dots, N\}$ and $A_F = \bigcap_{j\in F} A_j$.[3]

Exercise 1.1.10. Given a sequence $\{A_n : n \ge 1\}$ of events, define

$$\varlimsup_{n\to\infty} A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \quad \text{and} \quad \varliminf_{n\to\infty} A_n = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n.$$

Show that $\varlimsup_{n\to\infty} A_n$ is the set of points that are in infinitely many A_n's and that $\varliminf_{n\to\infty} A_n$ is the set of points that are in all but a finite number of A_n's and therefore that $\varliminf_{n\to\infty} A_n \subseteq \varlimsup_{n\to\infty} A_n$. One says that $\lim_{n\to\infty} A_n$ exists if $\varlimsup_{n\to\infty} A_n = \varliminf_{n\to\infty} A_n$.

[2] See, for example, Theorem 2.2.18 in [**10**].

[3] card(F) is the number of elements in the set F.

Exercise 1.1.11. Notice that the preceding are natural, set-theoretic versions of the corresponding notions for real numbers, and show that

$$\mathbb{P}\left(\underline{\lim}_{n\to\infty} A_n\right) \le \underline{\lim}_{n\to\infty} \mathbb{P}(A_n) \quad \text{and} \quad \overline{\lim}_{n\to\infty} \mathbb{P}(A_n) \le \mathbb{P}\left(\overline{\lim}_{n\to\infty} A_n\right), \tag{1.1.12}$$

and therefore

$$\lim_{n\to\infty} A_n \text{ exists} \implies \mathbb{P}\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Finally, show that, as a consequence of (1.1.7) and (1.1.6),

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}\left(\overline{\lim}_{n\to\infty} A_n\right) = 0. \tag{1.1.13}$$

This last statement, which is attributed to E. Borel, has many applications.

1.2. Finite and Countable Sample Spaces

The technicalities alluded to at the end of § 1.1.1 do not arise when the sample space Ω is finite or countable, and therefore we will begin by considering examples of such sample spaces.

1.2.1. Probability Theory on a Countable Space. To get started, I need to specify what we will mean when we sum over a countable index set $\mathcal{I}$. The definition that we will adopt looks a little cumbersome at first, but it is the one which corresponds to Lebesgue's theory of integration (cf. § 2.4) and is therefore the natural one for our purposes.

Let $\mathcal{I}$ be a finite or countable index set and let $\{a_i : i \in \mathcal{I}\} \subseteq (-\infty, \infty]$ be given. If $\mathcal{I} = \emptyset$, we will take $\sum_{i\in\mathcal{I}} a_i = 0$, and if it is non-empty but finite and $\{a_i : i \in \mathcal{I}\} \subseteq \mathbb{R}$, the meaning of $\sum_{i\in\mathcal{I}} a_i$ is unambiguously defined by ordinary arithmetic. When $\mathcal{I}$ is finite and $a_i = \infty$ for some $i \in \mathcal{I}$, we will say that $\sum_{i\in\mathcal{I}} a_i = \infty$. Finally, assume that $\mathcal{I}$ is infinite, and write $F \subset\subset \mathcal{I}$ when F is a finite subset of $\mathcal{I}$. We will say that $\sum_{i\in\mathcal{I}} a_i$ **converges** to $s \in \mathbb{R}$ if and only if for each $\epsilon > 0$ there is a finite $F_\epsilon \subset\subset \mathcal{I}$ such that $\left|s - \sum_{i\in F} a_i\right| < \epsilon$ whenever $F_\epsilon \subseteq F \subset\subset \mathcal{I}$, in which case I will use $\sum_{i\in\mathcal{I}} a_i$ to denote s. Finally, say that $\sum_{i\in\mathcal{I}} a_i$ converges to ∞ and write $\sum_{i\in\mathcal{I}} a_i = \infty$ if for all $R \in (0, \infty)$ there exists an $F_R \subset\subset \mathcal{I}$ such that $\sum_{i\in F} a_i \ge R$ whenever $F_R \subseteq F \subset\subset \mathcal{I}$.

In the following, and elsewhere, for $a \in (-\infty, \infty]$ I will use a^+ to denote the **positive part** $a \vee 0$ and a^- to denote the **negative part** $-(a \wedge 0) = (-a)^+$ of a. Obviously, $|a| = a^+ + a^-$ and $a = a^+ - a^-$.

Lemma 1.2.1. *Assume that $\mathcal{I}$ is infinite and that $\{a_i : i \in \mathcal{I}\} \subseteq (-\infty, \infty]$. If $a_i \geq 0$ for all $i \in \mathcal{I}$, then*

$$\sum_{i\in\mathcal{I}} a_i \text{ converges to } \sup\left\{\sum_{i\in F} a_i : F \subset\subset \mathcal{I}\right\}.$$

More generally, if $\sum_{i\in\mathcal{I}} a_i^- < \infty$, then

$$\sum_{i\in\mathcal{I}} a_i \text{ converges to } \sum_{i\in\mathcal{I}} a_i^+ - \sum_{i\in\mathcal{I}} a_i^-.$$

In fact, if $a_i \in \mathbb{R}$ for all $i \in \mathcal{I}$, then $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in (-\infty, \infty]$ if and only if $\sum_{i\in\mathcal{I}} a_i^- < \infty$, and $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in \mathbb{R}$ if and only if $\sum_{i\in\mathcal{I}} |a_i| < \infty$. Finally, if $\{i_k : k \geq 1\}$ is an enumeration of $\mathcal{I}$ (i.e., $k \in \mathbb{Z}^+ \longmapsto i_k \in \mathcal{I}$ is one-to-one and onto) and $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in (-\infty, \infty]$, then the sequence $\{\sum_{k=1}^n a_i : n \geq 1\}$ converges to s.

Proof. To prove the first assertion, set $s = \sup\left\{\sum_{i\in F} a_i : F \subset\subset \mathcal{I}\right\}$. Since, $s \geq \sum_{i\in F} a_i$ for all $F \subset\subset \mathcal{I}$, what we must show is that for each $s' < s$ there is an $F_{s'} \subset\subset \mathcal{I}$ such that $\sum_{i\in F} a_i \geq s'$ for all $\mathcal{I} \supset\supset F \supseteq F_{s'}$. But, by definition, there exists an $F_{s'} \subset\subset \mathcal{I}$ such that $\sum_{i\in F_{s'}} a_i \geq s'$, and so, because the a_i's are non-negative, $\sum_{i\in F} a_i \geq s'$ for all $\mathcal{I} \supset\supset F \supseteq F_{s'}$.

Next, suppose that $u \equiv \sum_{i\in\mathcal{I}} a_i^- < \infty$. If $\sum_{i\in\mathcal{I}} a_i^+ = \infty$ and $R < \infty$, then there exists an $F_R \subset\subset \mathcal{I}$ such that $\sum_{i\in F_R} a_i^+ \geq R + u$, and therefore, for any $F_R \subseteq F \subset\subset \mathcal{I}$, $\sum_{i\in F} a_i \geq R + u - \sum_{i\in F} a_i^- \geq R$. Thus, in this case, $\sum_{i\in\mathcal{I}} a_i$ converges to $\infty = \sum_{i\in\mathcal{I}} a_i^+ - \sum_{i\in\mathcal{I}} a_i^-$. If $v \equiv \sum_{i\in\mathcal{I}} a_i^+ < \infty$ and $\epsilon > 0$, then there exists an $F_\epsilon \subset\subset \mathcal{I}$ such that $\left|v - \sum_{i\in F} a_i^+\right| + \left|u - \sum_{i\in F} a_i^-\right| < \epsilon$ and therefore $\left|v - u - \sum_{i\in F} a_i\right| < \epsilon$ for all $F_\epsilon \subseteq F \subset\subset \mathcal{I}$. Thus, in this case also, $\sum_{i\in\mathcal{I}} a_i$ converges to $\sum_{i\in\mathcal{I}} a_i^+ - \sum_{i\in\mathcal{I}} a_i^-$. Hence, $\sum_{i\in\mathcal{I}} a_i^- < \infty$ always implies that $\sum_{i\in\mathcal{I}} a_i$ converges to $\sum_{i\in\mathcal{I}} a_i^+ - \sum_{i\in\mathcal{I}} a_i^-$.

Now suppose that $\{a_i : i \in \mathcal{I}\} \subseteq \mathbb{R}$ and that $\sum_{i\in\mathcal{I}} a_i$ converges to $s \in (-\infty, \infty]$. To see that $\sum_{i\in\mathcal{I}} a_i^- < \infty$, choose some $t \in (-\infty, s)$ and note that there exists an $F \subset\subset \mathcal{I}$ such that $\sum_{i\in H} a_i \geq t$ whenever $F \subseteq H \subset\subset \mathcal{I}$. Thus, if $G \subset\subset \{i \in \mathcal{I} \setminus F : a_i < 0\}$, then

$$t \leq \sum_{i\in F\cup G} a_i = \sum_{i\in F} a_i - \sum_{i\in G} a_i^-,$$

and so $\sum_{i\in G} a_i^- \leq \sum_{i\in F} a_i - t$, which means that, for any $H \subset\subset \mathcal{I}$, $\sum_{i\in H} a_i^- \leq 2\sum_{i\in F} |a_i| - t$ and therefore $\sum_{i\in\mathcal{I}} a_i^- < \infty$. Hence, we now know that when $\{a_i : i \in \mathcal{I}\} \subseteq \mathbb{R}$, $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in (-\infty, \infty]$ if and only if $\sum_{i\in\mathcal{I}} a_i^- < \infty$.

If $\sum_{i\in\mathcal{I}} |a_i| < \infty$, then we already know that $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in \mathbb{R}$. Conversely, suppose that $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in \mathbb{R}$. Then

$a_i < \infty$ for all $i \in \mathcal{I}$. Indeed, if $a_{i_0} = \infty$ for some $i_0 \in \mathcal{I}$, then $\sum_{i\in F} a_i = \infty$ for any $F \subset\subset \mathcal{I}$ with $i_0 \in F$. Thus, by the preceding, we know that $\sum_{i\in\mathcal{I}} a_i^- < \infty$. Applying the same reasoning to $\{-a_i : i \in \mathcal{I}\}$, we see that $\sum_{i\in\mathcal{I}} a_i^+ < \infty$ and therefore $\sum_{i\in\mathcal{I}} |a_i| < \infty$.

Finally, suppose $\sum_{i\in\mathcal{I}} a_i$ converges to some $s \in (-\infty, \infty]$ and that $\{i_k : k \geq 1\}$ is an enumeration of $\mathcal{I}$. Set $s_n = \sum_{k=1}^n a_{i_k}$. Assuming that $s < \infty$, for a given $\epsilon > 0$, choose $F_\epsilon \subset\subset \mathcal{I}$ accordingly, and choose $N_\epsilon \geq 1$ so that $F \supseteq \{i_1, \dots, i_{N_\epsilon}\}$. Then $|s_n - s| < \epsilon$ for all $n \geq N_\epsilon$. The argument when $s = \infty$ is essentially the same. □

Remark 1.2.2. Given $\{a_k : k \in \mathbb{Z}^+\} \subseteq \mathbb{R}$, in calculus one says that $\sum_{k=1}^\infty a_k$ converges to $s \in (-\infty, \infty]$ if the sequence of partial sums $s_n = \sum_{k=1}^n a_k$ converge to s. The final part of Lemma 1.2.1 says that if $\sum_{k\in\mathbb{Z}^+} a_k$ converges, then $\sum_{k=1}^\infty a_k$ does as well. However, the converse is not true. Indeed, $\sum_{k=1}^\infty \frac{(-1)^k}{k}$ converges in the calculus sense to $-\log 2$, whereas $\sum_{k\in\mathbb{Z}^+} \frac{(-1)^k}{k}$ does not converge. The point is that the convergence or divergence of $\sum_{k=1}^\infty a_k$ depends on the order in which the a_k's are added, whereas the convergence or divergence of $\sum_{k\in\mathbb{Z}^+} a_k$ does not depend on how one orders the summands. In the future it will be important to distinguish between these two notions of summation, and for that reason I will continue to reserve $\sum_{k=1}^\infty a_k$ for the standard calculus notion.

Suppose that $\{a_i : i \in \mathcal{I}\} \cup \{b_i : i \in \mathcal{I}\} \subseteq (-\infty, \infty]$. Then it should be clear that

$$\begin{aligned} &\sum_{i\in\mathcal{I}} a_i^- < \infty \text{ and } a_i \leq b_i \text{ for all } i \in \mathcal{I} \\ &\implies \sum_{i\in\mathcal{I}} b_i^- < \infty \text{ and } \sum_{i\in\mathcal{I}} a_i \leq \sum_{i\in\mathcal{I}} b_i. \end{aligned} \tag{1.2.3}$$

Equally clear should be the fact that

$$\begin{aligned} &\sum_{i\in\mathcal{J}\cup\mathcal{K}} a_i = \sum_{i\in\mathcal{J}} a_i + \sum_{i\in\mathcal{K}} a_i \\ &\text{if } \sum_{i\in\mathcal{J}\cup\mathcal{K}} a_i^- < \infty \text{ and } \mathcal{J} \text{ and } \mathcal{K} \text{ are disjoint subsets of } \mathcal{I}. \end{aligned} \tag{1.2.4}$$

With the preceding at hand, we can now discuss probability measures on a finite or countable sample space Ω. Indeed, if (1.1.4) is going to hold, then a probability measure $\mathbb{P}$ is completely determined by the probability it assigns to events consisting of precisely one sample point. That is, once one knows $p(\omega) \equiv \mathbb{P}(\{\omega\})$ for each $\omega \in \Omega$, one knows that

$$\mathbb{P}(A) = \sum_{\omega\in A} p(\omega) \tag{1.2.5}$$

for every $A \subseteq \Omega$. Obviously, $p(\omega) \geq 0$ and $1 = \mathbb{P}(\Omega) = \sum_{\omega\in\Omega} p(\omega)$. Conversely, if $p : \Omega \longrightarrow [0,1]$ and $1 = \sum_{\omega\in\Omega} p(\omega)$, then there is a unique probability measure on Ω such that $p(\omega) = \mathbb{P}(\{\omega\})$ for all $\omega \in \Omega$. Namely, one simply defines $\mathbb{P}(A)$ by (1.2.5). Thus, there is a one-to-one correspondence between functions $p : \Omega \longrightarrow [0,1]$ satisfying $1 = \sum_{\omega\in\Omega} p(\omega)$ and probability measures $\mathbb{P}$ on Ω. A non-negative function p on Ω satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$ is sometimes called a **probability function**, and the associated probability measure $\mathbb{P}$ given by (1.2.5) is called the **probability measure determined by** p.

1.2.2. Uniform Probabilities and Coin Tossing. Perhaps the most easily understood examples of the preceding are those in which Ω is finite and each of its elements occurs with the same probability. That is, there is a $p \geq 0$ such that $\mathbb{P}(\{\omega\}) = p$ for all $\omega \in \Omega$. Because $1 = \sum_{\omega\in\Omega} \mathbb{P}(\{\omega\}) = \mathrm{card}(\Omega)p$, it is clear that p must be equal to $\frac{1}{\mathrm{card}(\Omega)}$. When this is the case, $\mathbb{P}$ is said to be the **uniform** probability measure on Ω. Obviously, when $\mathbb{P}$ is the uniform probability measure,

$$\mathbb{P}(A) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)} \quad \text{for all } A \subseteq \Omega. \tag{1.2.6}$$

Thus, the computation of $\mathbb{P}(A)$ comes down to the combinatorial problem of determining how many elements A contains.

Modeling a *fair* coin tossing game of length N (i.e., the coin is unbiased and the game ends after the Nth toss) is a good example of the preceding. In this model, one takes $\Omega = \{0,1\}^N$, the set of all maps $\omega : \{1,\dots,N\} \longrightarrow \{0,1\}$. The map ω records the history of the game: $\omega(n) = 1$ or $\omega(n) = 0$ depending on whether the coin came up heads or tails on the nth toss. If the coin is unbiased, then it is reasonable to expect that any history is just as likely as any other, in which case $\mathbb{P}(\{\omega\}) = 2^{-N}$, since $\mathrm{card}(\Omega) = 2^N$.

It is important to observe that this model of coin tossing has a crucial *homogeneity* property. Namely, given an M-element set $S \subseteq \{1,\dots,N\}$ and $\Gamma \subseteq \{0,1\}^S$,

$$\mathbb{P}\big(\{\omega \in \{0,1\}^N : \omega \restriction S \in \Gamma\}\big) = \mathbb{P}\big(\{\omega \in \{0,1\}^M : \omega \in \Gamma\}\big). \tag{1.2.7}$$

Indeed, there are $2^{N-M}\mathrm{card}(\Gamma)$ elements $\omega \in \{0,1\}^N$ such that $\omega \restriction S \in \Gamma$, and there are $\mathrm{card}(\Gamma)$ such $\omega \in \{0,1\}^M$. Hence, the left-hand side equals $\frac{2^{N-M}\mathrm{card}(\Gamma)}{2^N} = \frac{\mathrm{card}(\Gamma)}{2^M}$, which is equal to the right-hand side. As a consequence, when $\Gamma \subseteq \{0,1\}^S$, the number $\mathbb{P}\big(\{\omega \in \{0,1\}^N : \omega \restriction S \in \Gamma\}\big)$ is the same for all N such that $S \subseteq \{1,\dots,N\}$. In particular, by the preceding considerations, if $A \subseteq \{0,1\}^M$ and $N > M$, then

$$\mathbb{P}(A) = \mathbb{P}(\{\omega \in \{0,1\}^N : \omega \restriction \{1,\dots,M\} \in A\}\big).$$

To develop some feeling for this model, consider the event consisting of those games in which precisely m of the coins come up heads. Equivalently, if

$$S_n(\omega) \equiv \sum_{m=1}^{n} \omega(m), \tag{1.2.8}$$

then we are looking at the event $\{S_N = m\}$. Obviously, $\{S_N = m\} = \emptyset$ unless $0 \le m \le N$. To compute $\operatorname{card}(\{S_N = m\})$ when $0 \le m \le N$, observe that $S_N(\omega) = m$ if and only if there are precisely m tosses on which heads occurred. Thus, the number of such ω's is the same as the number of ways in which one can choose the m tosses on which the heads appeared. Since there is a total of N tosses, this is tantamount to counting the number of ways of choosing m elements from a set of size N. Hence, $\operatorname{card}(A_m)$ is the **binomial coefficient**

$$\binom{N}{m} = \frac{N(N-1)\cdots(N-m+1)}{m!} = \frac{N!}{m!(N-m)!},$$

which is sometimes called N *choose* m, and therefore the probability that, when it is tossed N times, a fair coin will come up heads exactly m times is[4]

$$\mathbb{P}(S_N = m) = \begin{cases} 2^{-N}\binom{N}{m} & \text{for } 0 \le m \le N, \\ 0 & \text{otherwise.} \end{cases} \tag{1.2.9}$$

As a consequence of the homogeneity property discussed above, we know that if we replaced our sample space by $\{0,1\}^{N'}$ for some $N' > N$ and did the same calculation in the sample space $\{0,1\}^{N'}$, (1.2.9) would still hold.

Now suppose that two players toss a fair coin N times and that player 1 wins a dollar from player 2 each time the coin comes up heads and he pays player 2 a dollar each time a tail occurs. Given $k \in \mathbb{Z}$ with $|k| \le N$, consider the event that, at the end of the game, player 1, depending on whether $k \ge 0$ or $k < 0$, gains or loses $|k|$ dollars. Equivalently, since $\sum_{n=1}^N \omega(n)$ is the number of times player 1 wins and $\sum_{n=1}^N (1-\omega(n))$ is the number of times he loses, we are looking at the event $\{W_N = k\}$ where

$$W_N(\omega) \equiv \sum_{n=1}^{N} \omega(n) - \sum_{n=1}^{N} \big(1-\omega(n)\big) = \sum_{n=1}^{N} \big(2\omega(n)-1\big) = 2S_N(\omega) - N. \tag{1.2.10}$$

Since $S_N = \frac{W_N+N}{2}$ and therefore $\{W_N = k\} = \big\{S_N = \frac{N+k}{2}\big\}$, we conclude

[4] Here, and elsewhere, I will use $\mathbb{P}(F \in \Gamma)$ to abbreviate $\mathbb{P}(\{\omega : F(\omega) \in \Gamma\})$.

from (1.2.9) that

$$\text{(1.2.11)}\qquad \mathbb{P}(W_N = k) = \begin{cases} 2^{-N}\binom{N}{\frac{N+k}{2}} & \text{if } |k| \le N \text{ and } \frac{N+k}{2} \in \mathbb{Z}, \\ 0 & \text{otherwise.} \end{cases}$$

For a more challenging example, consider the event $C_{k,\ell}$ that player 1 ends up with a gain of ℓ dollars and that, at some time during the game, he had a gain of k dollars. Before trying to compute $\mathbb{P}(C_{k,\ell})$, it is helpful to introduce a more dynamic formulation. Set $W_0(\omega) = 0$ and, for $1 \le n \le N$, $W_n(\omega) = \sum_{m=1}^{n}(2\omega(m)-1)$. Then $W_n(\omega)$ represents player 1's net gain after n tosses of the coin. Alternatively, one can think of $\{W_n(\omega) : 1 \le n \le N\}$ as an N-step **random walk** that starts at 0 and, at each time $1 \le n \le N$, moves forward or backward 1 step depending on whether the nth toss came up heads or tails. Thus, if

$$\zeta_N^{\{k\}}(\omega) \equiv \inf\{0 \le n \le N : W_n(\omega) = k\},$$

with the understanding that the infimum over the empty set is $+\infty$, then $\zeta_N^{\{k\}}(\omega)$ is the first time that the walk $\{W_n(\omega) : 1 \le n \le N\}$ gets to k. Equivalently, $\zeta_N^{\{k\}}(\omega) = \infty$ if $W_n(\omega) \ne k$ for any $0 \le n \le N$ and $\zeta_N^{\{k\}}(\omega) = n$ for some $0 \le n \le N$ if $W_n(\omega) = k$ and $W_m(\omega) \ne k$ for $0 \le m < n$. In terms of these quantities, $C_{k,\ell} = \{\zeta_N^{\{k\}} \le N \text{ and } W_N = \ell\}$. Because $W_n(\omega) - W_{n-1}(\omega) = \pm 1$ for all $1 \le n \le N$, if $W_N(\omega) = \ell$, then the walk $\{W_n(\omega) : 0 \le n \le N\}$ must pass through all the integers between 0 and ℓ. That is, if $W_N(\omega) = \ell$, then $\zeta_N^{\{k\}}(\omega) \le N$, depending on whether $\ell \ge 0$ or $\ell \le 0$, for all $0 \le k \le \ell$ or all $\ell \le k \le 0$. Hence, if $0 \le k \le \ell$ or $\ell \le k \le 0$, then $C_{k,\ell} = \{W_N = \ell\}$.

In order to handle the cases when either $k \ge 0$ and $\ell < k$ or $k \le 0$ and $\ell > k$, consider the map $R^{(k)} : \{0,1\}^N \longrightarrow \{0,1\}^N$ given by

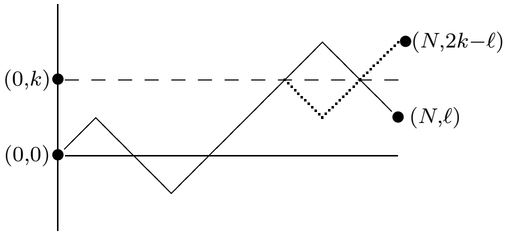
$$R^{(k)}\omega(n) = \begin{cases} \omega(n) & \text{if } n \le N \wedge \zeta_N^{\{k\}}(\omega), \\ 1 - \omega(n) & \text{if } \zeta_N^{\{k\}}(\omega) < n \le N. \end{cases}$$

Then, $W_n(R^{(k)}\omega) = W_{\zeta_N^{\{k\}}(\omega)}(\omega) - \big(W_n(\omega) - W_{\zeta_N^{\{k\}}(\omega)}(\omega)\big)$ when $\zeta_N^{\{k\}}(\omega) < n \le N$, and therefore

$$W_n\big(R^{(k)}\omega\big) = \begin{cases} W_n(\omega) & \text{if } 0 \le n \le N \wedge \zeta_N^{\{k\}}(\omega), \\ 2k - W_n(\omega) & \text{if } \zeta_N^{\{k\}}(\omega) < n \le N. \end{cases}$$

Equivalently, thinking in terms of random walks, $\{W_n(R^{(k)}\omega) : 0 \le n \le N\}$ is the random walk obtained by *reflecting* $\{W_n(\omega) : 0 \le n \le N\}$ at time

$\zeta_N^{\{k\}}(\omega)$. Here is a picture of a path and its reflection at level k:



Original path —, Reflected path $\cdots$

In particular, $\zeta_N^{\{k\}} \circ R^{(k)} = \zeta_N^{\{k\}}$, from which it is clear that $R^{(k)} \circ R^{(k)}(\omega) = \omega$ and therefore that $R^{(k)}$ is both one-to-one and onto. In addition, for each ℓ, $R^{(k)} : C_{k,\ell} \longrightarrow C_{k,2k-\ell}$, and so, for each ℓ, $R^{(k)}$ is a one-to-one map from $C_{k,\ell}$ onto $C_{k,2k-\ell}$.

Now suppose that $k \geq 0$ and that $\ell < k$. Then, by the preceding, $\mathrm{card}(C_{k,\ell}) = \mathrm{card}(C_{k,2k-\ell})$, and, because $2k - \ell \geq k$,

$$C_{k,2k-\ell} = \{W_N = 2k - \ell\}.$$

Hence $\mathrm{card}(C_{k,\ell}) = \mathrm{card}(\{W_N = 2k - \ell\})$ when $k \geq 0$ and $\ell < k$, and the same reasoning shows that this equality holds as well when $k \leq 0$ and $\ell > k$. Summarizing these results, we have now shown that

$$\text{(1.2.12)} \qquad \mathbb{P}(\zeta_N^{\{k\}} \leq N \ \&\ W_N = \ell) = \begin{cases} \mathbb{P}(W_N = \ell) & \text{if } k\ell \geq 0 \ \&\ |k| \leq |\ell|, \\ \mathbb{P}(W_N = 2k-\ell) & \text{otherwise.} \end{cases}$$

Again, it is important to observe that, by the homogeneity property of coin tossing,

$$\begin{aligned} &\mathbb{P}\big(\{\omega \in \{0,1\}^{N'} : \zeta_{N'}^{\{k\}}(\omega) \leq N \ \&\ W_N(\omega) = \ell\}\big) \\ &\quad = \mathbb{P}\big(\{\omega \in \{0,1\}^{N} : \zeta_{N}^{\{k\}}(\omega) \leq N \ \&\ W_N(\omega) = \ell\}\big) \end{aligned}$$

if $N' > N$.

1.2.3. Tournaments[5]**.** Graph theory provides a rich venue in which to think about coin tossing. A **graph** is a pair (V, E) consisting of a set V of points, known as the **vertices** v, and a set E of pairs $\{v, w\}$, called **edges**, of not necessarily distinct vertices v and w. The edge $\{v, w\}$ is thought of as a bond or connection between v and w.

If V has M elements, then there are $\binom{M}{2} = \frac{M(M-1)}{2}$ possible edges connecting distinct vertices and $M + \binom{M}{2} = \frac{M(M+1)}{2}$ possible edges if vertices are allowed to be connected to themselves. A **complete graph** (V, E) is one for which E contains all possible edges between distinct vertices and none connecting a vertex to itself. Thus E has $\binom{M}{2}$ elements if (V, E) is complete.

[5] The material in this subsection is derived from Alon and Spencer's book [**1**].

Given a complete graph (V, E), a **tournament** is an ordering of the vertices in the edges. That is, starting from $E = \big\{\{v, w\},\ v \neq w\big\}$, one constructs a tournament T by replacing each unordered pair $\{v, w\}$ by either the ordered pair (v, w) or by the ordered pair (w, v). The origin of the terminology should be clear: V is thought of as the players in a competition in which every player plays a game against every other player, and a tournament is the record of which of the two players won in each of the games. With this model in mind, we will say that v *dominates* w if $(v, w) \in T$.

Clearly, there are $N \equiv 2^{\binom{M}{2}}$ tournaments. Moreover, there is a natural isometry between $\Omega = \{0, 1\}^N$ and the set of all tournaments. Namely, assign each edge a number, and choose a reference tournament T_0. Then, for a given $\omega \in \Omega$, determine the tournament $T(\omega)$ so that, for each n, its nth edge has the same or opposite ordering as the nth edge in T_0 according to whether $\omega(n) = 0$ or $\omega(n) = 1$. Equivalently, thinking in terms of coin tossing, $T(\omega)$ is constructed from T_0 by keeping or reversing the order of the nth edge depending on whether the nth toss comes up tails or heads. Thus, one can create a *random tournament* by flipping a fair coin N times, thereby putting the uniform probability on the set of tournaments. In other words, the probability of a tournament having a certain property P is the probability that the uniform probability measure $\mathbb{P}$ on $\{0, 1\}^N$ assigns to the set of $\omega \in \{0, 1\}^N$ for which $T(\omega)$ has property P.

A powerful method for proving that there exist graphs possessing a particular property is to consider random graphs and show that a random graph will have that property with positive probability. To see this method in action, say that a tournament T has property P_k if, for every subset $S \subseteq V$ of k vertices, there is a $v \in V \setminus S$ such that $(v, w) \in T$ for all $w \in S$. Phrased more picturesquely, a tournament has property P_k if, for every k-member subset of players, there is a player who beats all of its members. Put that way, a natural question is how many players must there be in order for there to exist a tournament with property P_k. Obviously, M must be larger than k, but it is less clear how much larger it must be. Using random tournaments, P. Erdős showed that *if* $\binom{M}{k}(1 - 2^{-k})^{M-k} < 1$, *then there exists a tournament with property* P_k.

To carry out Erdős's line of reasoning, let S be a k-element subset of vertices. We begin by computing the number of tournaments with the property Q_S that *no* player beats all the players in S, or, equivalently, the number of T's such that, for each $v \notin S$, $(w, v) \in T$ for some $w \in S$. To do this, let N_d be the number of tournaments with property Q_S when $d = M - k$. When $d = 1$, there are $2^{\binom{k}{2}}$ orderings of the edges between elements of S. Moreover, there are 2^k ways to order the edges between the $v \notin S$ to the w's in S, but only 1 of these orderings has the property that v dominates every $w \in S$.

Hence, $N_1 = 2^{\binom{k}{2}}(2^k - 1)$. When $d = 2$, there are two elements v_1 and v_2 which are not in S. To count the number of tournaments with property Q_S, first choose one of the N_1 tournaments with vertices $S \cup \{v_1\}$ having property Q_S. Next, complete the construction of a tournament having property Q_S with vertices $S \cup \{v_1, v_2\}$ by choosing one of the two orderings of the edges between v_1 and v_2 and then choosing one of the $2^k - 1$ orderings of the edges between v_2 and the w's in S so that at least one w dominates v_2. Hence, $N_2 = N_1 2(2^k - 1)$. More generally, if $V \setminus S = \{v_1, \dots, v_{d+1}\}$, there are N_d tournaments having property Q_S with vertices $S \cup \{v_1, \dots, v_d\}$, and for each such tournament there are $2^d(2^k - 1)$ orderings of the edges between v_{d+1} and the vertices in $S \cup \{v_1, \dots, v_d\}$ which result in a tournament having property Q_S with vertices in V. Thus, $N_{d+1} = N_d 2^d(2^k - 1)$.

Starting from the preceding, an easy induction argument shows that, when $\mathrm{card}(V) = M > k$, there are $2^{\binom{k}{2}} 2^{\sum_{j=1}^{M-k-1} j}(2^k - 1)^{M-k}$ tournaments with property Q_S. Hence, the probability of such a tournament is

$$\frac{2^{\binom{k}{2}} 2^{\frac{(M-k)(M-k-1)}{2}}(2^k - 1)^{M-k}}{2^{\binom{M}{2}}}.$$

Noting that $\frac{(M-k)(M-k-1)}{2} = \binom{M-k}{2}$ and that $\binom{M}{2} - \binom{k}{2} - \binom{M-k}{2}$ is $k(M-k)$, we see that this probability simplifies to $(1 - 2^{-k})^{M-k}$.

To complete Erdős's argument, for each k-element $S \subseteq V$, let A_S be the event that a tournament has property Q_S. Then the event that a tournament does not have property P_k is the union over S of the events A_S. Since $\mathbb{P}(A_S) = (1 - 2^{-k})^{M-k}$ for each S and there are $\binom{M}{k}$ S's, it follows from (1.1.7) that the probability of a tournament not having property P_k is less than or equal to $\binom{M}{k}(1 - 2^{-k})^{M-k}$. Hence, if $\binom{M}{k}(1 - 2^{-k})^{M-k} < 1$, then, with positive probability, there is a tournament with property P_k, and therefore there is at least one such tournament.

1.2.4. Symmetric Random Walk. As I said in § 1.2.2, the sequence $\{W_n(\omega) : 0 \le n \le N\}$ can be thought of as a random walk. In fact, because we are dealing with fair coins, the random walk considered in § 1.2.2 is said to be a **symmetric random walk** because, at each step, it is equally likely to move in either direction.

When one thinks in terms of random walks, a host of questions comes to mind, an interesting one of which is with what probability a walk will pass through k by time N. That is, one is asking what $\mathbb{P}(\zeta_N^{\{k\}} \le N)$ is. To find the answer, first suppose that $k \ge 1$. Since $\{\zeta_N^{\{k\}} \le N\}$ is the union over $\ell \in \mathbb{Z}$ of the mutually disjoint events $\{\zeta_N^{\{k\}} \le N \text{ and} W_N = \ell\}$, we know from (1.2.12) that

$$\begin{aligned}\mathbb{P}(\zeta_N^{\{k\}} \le N) &= \sum_\ell \mathbb{P}(\zeta_N^{\{k\}} \le N \text{ and } W_N = \ell)\\ &= \sum_{\ell \ge k} \mathbb{P}(W_N = \ell) + \sum_{\ell < k} \mathbb{P}(W_N = 2k - \ell) = \mathbb{P}(W_N \ge k) + \mathbb{P}(W_N > k).\end{aligned}$$

Since $\mathbb{P}(\zeta_N^{\{-k\}} \le N) = \mathbb{P}(\zeta_N^{\{k\}} \le N)$ and $\mathbb{P}(W_N = -k) = \mathbb{P}(W_N = k)$, we now know that

$$\mathbb{P}(\zeta_N^{\{k\}} \le N) = \mathbb{P}\big(W_N(\omega) \ge |k|\big) + \mathbb{P}\big(W_N > |k|\big). \tag{1.2.13}$$

Because its derivation is an application of the reflection map $R^{(k)}$, equation (1.2.13) is often called the **reflection principle** for symmetric random walks.

Starting from (1.2.13), one sees that, for $k \ge 1$,

$$\mathbb{P}\big(\zeta_N^{\{k\}} > N\big) = 1 - \mathbb{P}(W_N \ge k) - \mathbb{P}(W_N > k) = \mathbb{P}(W_N \le k) - \mathbb{P}(W_N \ge k).$$

Hence, since $\mathbb{P}(\zeta_N^{\{-k\}} > N) = \mathbb{P}(\zeta_N^{\{k\}} > N)$ and $\mathbb{P}(W_N \ge k) = \mathbb{P}(W_N \le -k)$, we have the following corollary of (1.2.13):

$$\mathbb{P}\big(\zeta_N^{\{k\}} > N\big) = \mathbb{P}\big(-|k| < W_N \le |k|\big). \tag{1.2.14}$$

It is interesting to know that a symmetric random walk will eventually visit every integer point. That is,

$$\lim_{N\to\infty} \mathbb{P}\big(\zeta_N^{\{k\}} > N\big) = 0 \quad \text{for all } k \in \mathbb{Z}. \tag{1.2.15}$$

In view of (1.2.14), this comes down to showing that, for every $0 < k \le N$,

$$\mathbb{P}(-k < W_N \le k) = \sum_{-k<\ell\le k} \mathbb{P}(W_N = \ell) = 2^{-N} \sum_{-k<\ell\le k} \binom{N}{\frac{N+\ell}{2}} \longrightarrow 0$$

as $N \to \infty$, and obviously this reduces to showing that $2^{-N}\binom{N}{\ell} \longrightarrow 0$ for every $\ell \in \mathbb{Z}$. To check this, first note that, for any $1 \le m \le n$,

$$\binom{n}{m-1} \le \binom{n}{m} \iff m \le \frac{n+1}{2}. \tag{1.2.16}$$

Hence, it suffices to show that $2^{-2N}\binom{2N}{N}$ and $2^{-2N-1}\binom{2N+1}{N+1}$ tend to 0 as $N \to \infty$. Furthermore, since $2^{-2N-1}\binom{2N+1}{N+1} \le 2^{-2N}\binom{2N}{N}$, we need only worry about $2^{-2N}\binom{2N}{N}$. Finally,

$$2^{-2N}\binom{2N}{N} = \frac{\prod_{m=1}^N (2m-1)}{2^N N!} = \prod_{m=1}^N \big(1 - \tfrac{1}{2m}\big),$$

and so it remains to check that $\lim_{N\to\infty}\prod_{m=1}^N\big(1-\frac{1}{2m}\big)=0$. Equivalently, we need to show that $\log\prod_{m=1}^N\big(1-\frac{1}{2m}\big)=\sum_{m=1}^N\log\big(1-\frac{1}{2m}\big)$ tends to $-\infty$ as $N\to\infty$. However, since $\log(1-x)\le -x$ for $x\in[0,1)$,

$$\sum_{m=1}^N\log\big(1-\tfrac{1}{2m}\big)\le-\frac12\sum_{m=1}^N\frac1m\longrightarrow-\infty$$

as $N\to\infty$, and so we have now proved (1.2.15).

1.2.5. De Moivre's Central Limit Theorem. A second result about the long-time behavior of a symmetric random walk is the one, proved originally in the 17th century by A. De Moivre, which eventually led to what is now called the central limit theorem (cf. Theorem 4.1.3). To understand his result, one should first know that the approximate distance traveled by a symmetric random walk after n steps is on the order of $\sqrt{n}$. It is intuitively obvious that, because of all its dithering back and forth, the walk will have gone a distance significantly less than n, but it is not so obvious how much less. A crude estimate can be obtained from the fact that

$$\tfrac12\mathbb{P}\big(|W_N|\ge N^{\frac12}R\big)=\mathbb{P}\big(\pm W_N\ge N^{\frac12}R\big)\le e^{-\frac{R^2}{2}}\quad\text{for }R>0. \tag{1.2.17}$$

One way to prove this is to note that, for any $\alpha\in\mathbb{R}$,

$$\begin{aligned}\sum_k e^{\alpha k}\mathbb{P}(W_N=k)&=\sum_{k=0}^N e^{\alpha(2k-N)}\mathbb{P}(W_N=2k-N)=e^{-\alpha N}\sum_{k=0}^N e^{2\alpha k}\mathbb{P}(S_N=k)\\&=2^{-N}e^{-\alpha N}\sum_{k=0}^N e^{2\alpha k}\binom{N}{k}=2^{-N}e^{-\alpha N}\big(1+e^{2\alpha}\big)^N=\big(\cosh\alpha\big)^N\le e^{\frac{\alpha^2N}{2}},\end{aligned}$$

since, for $n\ge1$, $(2n)!=2^nn!\prod_{k=1}^n(2k-1)\ge2^nn!$, and therefore

$$\cosh x=\sum_{n=0}^\infty\frac{x^{2n}}{(2n)!}\le\sum_{n=0}^\infty\frac{x^{2n}}{2^nn!}=e^{\frac{x^2}{2}}.$$

Hence, for any $\alpha\in\mathbb{R}$,

$$\mathbb{P}\big(W_N\ge N^{\frac12}R\big)\le e^{-\alpha N^{\frac12}R}\sum_{k\ge N^{\frac12}R}e^{\alpha k}\mathbb{P}(W_N=k)=e^{-\alpha N^{\frac12}R+\frac{\alpha^2N}{2}}.$$

By taking $\alpha=\frac{R}{N^{\frac12}}$, we get $\mathbb{P}(W_N\ge N^{\frac12}R)\le e^{-\frac{R^2}{2}}$, and, after combining this with $\mathbb{P}(|W_N|\ge N^{\frac12}R)=2\mathbb{P}(W_N\ge N^{\frac12}R)$, one arrives at (1.2.17).

Knowing that, with probability close to 1, the size of $|W_N|$ is no larger than a large constant times $N^{\frac12}$, it is reasonable to look more closely and to ask about the probability that $\breve{W}_N\equiv N^{-\frac12}W_N$ lies in an interval. The answer to this question was found by De Moivre, who proved the following theorem.[6]

Theorem 1.2.18 (De Moivre)**.** *Referring to the preceding,*

$$\mathbb{P}\big(a < \breve{W}_N \le b\big) \longrightarrow (2\pi)^{-\frac{1}{2}} \int_a^b e^{-\frac{\xi^2}{2}}\, d\xi, \tag{1.2.19}$$

where the convergence is uniform with respect to a, $b \in [-\infty, \infty]$ with $a \le b$.

The first step in our proof of (1.2.19) will be to show that

$$\lim_{N\to\infty} \mathbb{P}\big(0 < \breve{W}_{2N} \le x\big) \longrightarrow (2\pi)^{-\frac{1}{2}} \int_0^x e^{-\frac{\xi^2}{2}}\, d\xi, \tag{$*$}$$

where the convergence is uniform with respect to x in bounded subsets of $(0, \infty)$. To this end, note that, for $k \ge 1$,

$$\begin{aligned}\frac{\mathbb{P}(W_{2N} = 2k)}{\mathbb{P}(W_{2N} = 0)} &= \frac{(N!)^2}{(N+k)!(N-k)!} = \frac{\prod_{j=1}^k (N - j + 1)}{\prod_{j=1}^k (N + j)}\\ &= \frac{\prod_{j=1}^k \big(1 - \frac{j-1}{N}\big)}{\prod_{j=1}^k \big(1 + \frac{j}{N}\big)}.\end{aligned}$$

Write $\log(1 - x) = -x - E(x)$ for $|x| < 1$, and observe that

$$|E(x)| = \left|\sum_{n=2}^\infty \frac{x^n}{n}\right| \le \frac{|x|^2}{2} \sum_{n=0}^\infty |x|^n = \frac{x^2}{2(1 - |x|)}.$$

In particular, $|E(x)| \le x^2$ for $|x| \le \frac{1}{2}$. Now let $R > 0$ be given. If $N \ge 8R^2$ and $1 \le k \le (2N)^{\frac{1}{2}} R$, then

$$\left|\log \frac{\mathbb{P}(W_{2N} = 2k)}{\mathbb{P}(W_{2N} = 0)} + \frac{k^2}{N}\right| \le \frac{2}{N^2} \sum_{j=1}^k j^2 \le \frac{C_R}{N^{\frac{1}{2}}},$$

where $C_R = 4R^3$. Thus, since

$$\mathbb{P}(0 < \breve{W}_{2N} \le x) = \sum_{0 < 2k \le (2N)^{\frac{1}{2}} x} \mathbb{P}(W_{2N} = 2k),$$

we have that

$$e^{-C_R N^{-\frac{1}{2}}} \sum_{0 < 2k \le (2N)^{\frac{1}{2}} x} e^{-\frac{k^2}{N}} \le \frac{\mathbb{P}\big(0 < \breve{W}_{2N} \le x\big)}{\mathbb{P}(W_{2N} = 0)} \le e^{C_R N^{-\frac{1}{2}}} \sum_{0 < 2k \le (2N)^{\frac{1}{2}} x} e^{-\frac{k^2}{N}}$$

[6] In truth, De Moivre proved somewhat less. Specifically, he did not know the constant $\sqrt{2\pi}$ that appears in the asymptotic formula for $n!$. Nonetheless, aside from that constant, in the course of deriving his result, De Moivre derived what we now call Stirling's formula. After looking at De Moivre's work, Stirling provided the missing constant. Although Stirling fully acknowledged De Moivre as the formula's discoverer, Stirling was the more renowned mathematician, and his name has been attached to it ever since.

for $0 < x \le R$ and $N \ge 8R^2$. At the same time,

$$\int_{\sqrt{\frac{2}{N}}k}^{\sqrt{\frac{2}{N}}(k+1)} e^{-\frac{\xi^2}{2}}\,d\xi \le \sqrt{\frac{2}{N}}e^{-\frac{k^2}{N}} \le \int_{\sqrt{\frac{2}{N}}(k-1)}^{\sqrt{\frac{2}{N}}k} e^{-\frac{\xi^2}{2}}\,d\xi,$$

and so we know that, for $x \in (0,R]$ and $N \ge 8R^2$,

$$e^{-C_R N^{-\frac{1}{2}}}\int_{\sqrt{\frac{2}{N}}}^{x} e^{-\frac{\xi^2}{2}}\,d\xi \le \frac{\mathbb{P}\big(0 < \breve{W}_{2N} \le x\big)}{\sqrt{\frac{N}{2}}\mathbb{P}(W_{2N} = 0)} \le e^{C_R N^{-\frac{1}{2}}}\int_0^x e^{-\frac{\xi^2}{2}}\,d\xi,$$

from which it is evident that

$$\frac{\mathbb{P}\big(0 < \breve{W}_{2N} \le x\big)}{\sqrt{\frac{N}{2}}\mathbb{P}(W_{2N} = 0)} \longrightarrow \int_0^x e^{-\frac{\xi^2}{2}}\,d\xi$$

uniformly for $0 < x \le R$. Finally, by Stirling's formula (2.5.12),[7]

$$\sqrt{\tfrac{N}{2}}\mathbb{P}(W_{2N} = 0) = \sqrt{\tfrac{N}{2}}2^{-2N}\binom{2N}{N} = \frac{1}{\sqrt{2\pi}} + \mathcal{O}\big(\tfrac{1}{N}\big),$$

and therefore we have now proved that $(*)$ holds and that the convergence is uniform in $0 < x \le R$.

The next step is to show that

$$(**) \qquad \mathbb{P}\big(\breve{W}_N \le x\big) \longrightarrow (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\,\xi$$

uniformly with respect to x in bounded subsets of $\mathbb{R}$. Because $\mathbb{P}(\breve{W}_{2N} < 0) = \mathbb{P}(W_{2N} < 0) = \mathbb{P}(W_{2N} > 0)$, $2\mathbb{P}(\breve{W}_{2N} < 0) = \mathbb{P}(W_{2N} \ne 0) = 1 - \mathbb{P}(W_{2N} = 0)$ and therefore

$$\mathbb{P}(\breve{W}_{2N} \le 0\big) = \tfrac{1}{2} + \tfrac{1}{2}\mathbb{P}(W_{2N} = 0) \longrightarrow \frac{1}{2} \text{ as } N \to \infty.$$

Combining this with the preceding, we know that, for $x \ge 0$ (cf. (2.5.4)),

$$\begin{aligned}\mathbb{P}\big(\breve{W}_{2N} \le x\big) &= \mathbb{P}\big(\breve{W}_{2N} \le 0\big) + \mathbb{P}\big(0 < \breve{W}_{2N} \le x\big)\\ &\longrightarrow \tfrac{1}{2} + (2\pi)^{-\frac{1}{2}}\int_0^x e^{-\frac{\xi^2}{2}}\,d\xi = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\,d\xi\end{aligned}$$

and therefore

$$\begin{aligned}\mathbb{P}(\breve{W}_{2N} < -x) &= \mathbb{P}(\breve{W}_{2N} > x) = 1 - \mathbb{P}(\breve{W}_{2N} \le x)\\ &\longrightarrow 1 - (2\pi)^{-\frac{1}{2}}\int_x^{\infty} e^{-\frac{\xi^2}{2}}\,d\xi = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{-x} e^{-\frac{\xi^2}{2}}\,d\xi\end{aligned}$$

uniformly with respect to x in bounded subsets of $[0,\infty)$. Since

$$\mathbb{P}(\breve{W}_{2N} \le x) = \mathbb{P}(\breve{W}_{2N} < x) + \mathbb{P}(\breve{W}_{2N} = -x)$$

[7] Here, and elsewhere, $\mathcal{O}(t)$ is used to denote a function such that $t^{-1}\mathcal{O}(t)$ stays bounded as t tends to a limit. Thus here, $N\mathcal{O}(\frac{1}{N})$ stays bounded as $N \to \infty$.

and

$$\mathbb{P}(W_{2N} = -(2N)^{\frac{1}{2}}x) \le \mathbb{P}(W_{2N} = 0) \longrightarrow 0,$$

we now know that $(**)$ with N replaced by $2N$ holds uniformly for x in bounded subsets of $\mathbb{R}$. To remove the restriction to even N's, note that $\mathbb{P}(\tilde{W}_{2N-1} \le x)$ is bounded above by $\mathbb{P}\big(\breve{W}_{2N} \le x + (2N)^{-\frac{1}{2}}\big)$ and below by $\mathbb{P}\big(\breve{W}_{2N} \le x - (2N)^{-\frac{1}{2}}\big)$. Hence, by the preceding uniform convergence result, it follows that $(**)$ holds and that the convergence is uniform with respect to x in bounded subsets of $\mathbb{R}$.

To complete the proof, let $\epsilon > 0$ be given, and choose $R \ge 1$ satisfying $e^{-\frac{R^2}{2}} < \frac{\epsilon}{2}$. Next, choose $M \in \mathbb{Z}^+$ such that

$$\left|\mathbb{P}\big(\breve{W}_N \le x\big) - (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi\right| < \epsilon$$

for all $|x| \le R$ and $N \ge M$. If $x < -R$, then, by (1.2.17), $\mathbb{P}(\breve{W}_N \le x) \le \mathbb{P}(\breve{W}_N \le -R) < \frac{\epsilon}{2}$ for any N. At the same time (cf. (2.5.6))

$$\int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi = \int_{-x}^{\infty} e^{-\frac{\xi^2}{2}}\, d\xi \le \int_{R}^{\infty} e^{-\frac{\xi^2}{2}}\, d\xi \le e^{-\frac{R^2}{2}} < \frac{\epsilon}{2},$$

and so

$$\left|\mathbb{P}\big(\breve{W}_N \le x\big) - (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi\right| \le \epsilon.$$

If $x > R$, then, because (cf. (2.5.4)) $(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} e^{-\frac{\xi^2}{2}}\, d\xi = 1$, for any N,

$$\left|\mathbb{P}\big(\breve{W}_N \le x\big) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi\right| = \left|\mathbb{P}\big(\breve{W}_n > x\big) - \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{\xi^2}{2}}\, d\xi\right| < \epsilon.$$

Hence, for all $N \ge M$ and $x \in \mathbb{R}$,

$$\left|\mathbb{P}\big(\breve{W}_N \le x\big) - (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{\xi^2}{2}}\, d\xi\right| < \epsilon.$$

Finally, simply note that $\mathbb{P}(a < \breve{W}_N \le b) = \mathbb{P}(\breve{W}_N \le b) - \mathbb{P}(\tilde{W}_N \le a)$ in order to get De Moivre's result.

1.2.6. Independent Events. In that we have been dealing with uniform probability measures, it is inevitable that all our computations have involved combinatorics: the probability of an event is the ratio of its cardinality to the cardinality of the sample space. Nonetheless, as we will see in this subsection, some of our calculations would have been simplified if we had made systematic use of the inherent independence properties possessed by the structures under consideration.

A pair of events A and B are said to be independent of one another if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. More generally, the events $A_1, \dots, A_\ell$ are said to be **mutually independent** if

$$\mathbb{P}\left(\bigcap_{i\in F} A_i\right) = \prod_{i\in F} \mathbb{P}(A_i) \quad \text{for all } \emptyset \neq F \subseteq \{1, \dots, \ell\}. \tag{1.2.20}$$

To understand the origin of this terminology, one should think about spaces, like $\{0,1\}^N$, whose elements are built out of components and assume that the choice of one component has no bearing on the choice of the other components. For example, when $\Omega = \{0,1\}^N$ and $S \subseteq \{1, \dots, N\}$, specifying the restriction $\omega \restriction S$ of ω to S does not prejudice the properties of $\omega \restriction S\complement$. Thus, if $M = \operatorname{card}(S)$ and, for some choice of $\Gamma_S \subseteq \{0,1\}^S$ and $\Gamma_{S\complement} \subseteq \{0,1\}^{S\complement}$, $A_S = \{\omega : \omega \restriction S \in \Gamma_S\}$ is an event which depends only on properties of $\omega \restriction S$ and $A_{S\complement} = \{\omega : \omega \restriction S \in \Gamma_{S\complement}\}$ is an event which is entirely determined by the properties of $\omega \restriction S\complement$, then $\operatorname{card}(A_S) = 2^{N-M}\operatorname{card}(\Gamma_S)$, $\operatorname{card}(A_{S\complement}) = 2^M \operatorname{card}(\Gamma_{S\complement})$, and $\operatorname{card}(A_S \cap A_{S\complement}) = \operatorname{card}(\Gamma_S)\operatorname{card}(\Gamma_{S\complement})$. Hence, when $\mathbb{P}$ is the uniform probability measure on this Ω, then

$$\mathbb{P}(A_S \cap A_{S\complement}) = \frac{\operatorname{card}(\Gamma_S)\operatorname{card}(\Gamma_{S\complement})}{2^N} = \frac{\operatorname{card}(\Gamma_S)}{2^M}\frac{\operatorname{card}(\Gamma_{S\complement})}{2^{N-M}} = \mathbb{P}(A_S)\mathbb{P}(A_{S\complement}),$$

and so A_S is independent of $A_{S\complement}$. More generally, the same argument shows that if $\{S_1, \dots, S_\ell\}$ is a partition of $\{1, \dots, N\}$ (i.e., the S_i's are mutually disjoint and their union is $\{1, \dots, N\}$), then, for any choice of $\Gamma_i \subseteq \{0,1\}^{S_i}$, the events $A_{S_i} = \{\omega \in \{0,1\}^N : \omega \restriction S_i \in \Gamma_i\}$ are mutually independent under the uniform probability measure on $\{0,1\}^N$. On the other hand, when A and B are events both of whose descriptions impose restrictions on $\omega(n)$ for some of the same n's, then, even though they may be independent of each other, there is no obvious reason for A to be independent of B under the uniform probability measure. For example, take $N = 2$, $A = \{\omega \in \{0,1\}^2 : \omega(1) = 0\}$, and, for $k \in \{0,1,2\}$, $B_k = \{\omega \in \{0,1\}^2 : \omega(1) + \omega(2) = k\}$. Then, under the uniform probability measure, $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B_k) = \frac{1}{4}$ if $k \in \{0,2\}$, $\mathbb{P}(B_1) = \frac{1}{2}$, $\mathbb{P}(A \cap B_k) = \frac{1}{4}$ if $k \in \{0,1\}$, and $\mathbb{P}(A \cap B_2) = 0$. Hence, A will not be independent of either B_0 or B_2, but, by accident, it will be independent of B_1.

The tournament question in § 1.2.3 provides a typical example of the power of independence considerations to facilitate computations. For each pair of distinct vertices v and w, let $A(v,w)$ be the event that v dominates w. Then the $A(v,w)$'s are mutually independent and $\mathbb{P}\big(A(v,w)\big) = \frac{1}{2}$. Given a subset S of vertices, the event A_S that no vertex dominates all vertices in S is equal to $\bigcap_{v\in V\setminus S} B_v$, where $B_v \equiv \bigcup_{w\in S} A(w,v)$, and so $\mathbb{P}(A_S) =$

$\prod_{v\in V\setminus S}\mathbb{P}(B_v)$. Finally, since $B_v = \big(\bigcap_{w\in S} A(v,w)\big)\complement$,

$$\mathbb{P}(B_v) = 1 - \prod_{w\in S}\mathbb{P}\big(A(v,w)\big) = 1 - 2^{-k},$$

and therefore $\mathbb{P}(A_S) = (1-2^{-k})^{M-k}$, as we saw before.

To describe a second, more interesting, example of the use of independence, return to the setting in § 1.2.4, where we discussed symmetric random walks. Set

$$A_N^{\pm} = \{\omega \in \{0,1\}^N : \pm W_n(\omega) > 0 \text{ for } 1 \le n \le N\}.$$

Obviously, $\mathbb{P}(A_N^-) = \mathbb{P}(A_N^+)$, and $A_N^+ = \{\omega \in \{0,1\}^N : \omega(1) = 1\} \cap B_{N-1}^+$, where

$$B_{N-1}^+ \equiv \left\{\omega \in \{0,1\}^N : \sum_{m=2}^{n}\big(2\omega(n)-1\big) \ge 0 \text{ for } 2 \le n \le N\right\}.$$

Since B_{N-1}^+ depends only on $\omega \restriction \{2,\dots,N\}$, it is independent of the event $\{\omega \in \{0,1\}^N : \omega(1) = 1\}$, and so $\mathbb{P}(A_N^+) = \frac{1}{2}\mathbb{P}(B_{N-1}^+)$. In addition, if $\Gamma_{N-1} = \big\{\omega \in \{0,1\}^{N-1} : \zeta_{N-1}^{\{-1\}}(\omega) > N-1\big\}$, then

$$B_{N-1}^+ = \big\{\omega \in \{0,1\}^N : \omega \restriction \{2,\dots,N\} \in \Gamma_{N-1}\big\},$$

and so, by homogeneity and (1.2.14), $\mathbb{P}(B_{N-1}^+) = \mathbb{P}(0 \le W_{N-1} \le 1)$. Thus, if

$$A_N = A_N^+ \cup A_N^- = \{W_n \ne 0 \text{ for } 1 \le n \le N\},$$

then $\mathbb{P}(A_N) = \mathbb{P}(0 \le W_{N-1} \le 1)$. If N is even, then, because $W_{N+1}(\omega)$ cannot be 0 and $W_N(\omega)$ cannot be 1, $A_N = A_{N+1}$ and so

$$\mathbb{P}(A_N) = \mathbb{P}(0 \le W_N \le 1) = \mathbb{P}(W_N = 0).$$

If N is odd, then $A_N = A_{N-1}$, and so in general we have the remarkable equation[8]

$$\mathbb{P}\big(W_n \ne 0 \text{ for } 1 \le n \le N\big) = \mathbb{P}\big(W_{2\lfloor\frac{N}{2}\rfloor} = 0\big) = 2^{-2\lfloor\frac{N}{2}\rfloor}\binom{2\lfloor\frac{N}{2}\rfloor}{\lfloor\frac{N}{2}\rfloor}, \tag{1.2.21}$$

where $\lfloor t\rfloor = \max\{n \in \mathbb{Z} : n \le t\}$ is the **integer part** of $t \in \mathbb{R}$.

Another way to interpret (1.2.21) is in terms of the **time of first return** $\rho_N^{(1)}(\omega)$, given by (remember the infimum over the empty set is $+\infty$)

$$\rho_N^{(1)}(\omega) \equiv \inf\{1 \le n \le N : W_n(\omega) = 0\}, \tag{1.2.22}$$

[8] We take $\binom{0}{0} = 1$.

of the walk to 0. Obviously, if $\rho_N^{(1)}(\omega) < \infty$, then $\rho_N^{(1)}(\omega)$ is an even number. In addition, if $1 \le r \le \frac{N}{2}$, then $\{\rho_N^{(1)} > 2r\} = \{W_n \neq 0 \text{ for } 1 \le n \le 2r\}$, and therefore, by (1.2.21),

$$\mathbb{P}\big(\rho_N^{(1)} > 2r\big) = \mathbb{P}\big(W_{2r} = 0\big) = 2^{-2r}\binom{2r}{r}. \tag{1.2.23}$$

In particular, since, as we have already shown, $\mathbb{P}(W_{2r} = 0) \longrightarrow 0$ as $r \to \infty$,

$$\lim_{N\to\infty} \mathbb{P}\big(\rho_N^{(1)} \le N\big) = 1, \tag{1.2.24}$$

which, because it says that the walk will eventually return to the place where it starts, is called the **recurrence** property of the symmetric random walk on $\mathbb{Z}$. In addition, since

$$\mathbb{P}(\rho_N^{(1)} = 2r) = \mathbb{P}\big(\rho_N^{(1)} > 2(r-1)\big) - \mathbb{P}\big(\rho_N^{(1)} > 2r\big),$$

$$\mathbb{P}\big(\rho_N^{(1)} = 2r\big) = \frac{2^{-2r}}{2r-1}\binom{2r}{r} = \frac{\mathbb{P}(W_{2r} = 0)}{2r-1}. \tag{1.2.25}$$

Related to the preceding is another important relationship, known as a **renewal equation**, between the time of first return and the probability that the walk is at 0. Namely, suppose that $W_{2M}(\omega) = 0$. Then $\rho_N^{(1)}(\omega) \le 2M$ and, if $\rho_N^{(1)}(\omega) = 2r$, then $W_{2M}(2M) - W_{2r}(\omega) = \sum_{2r<m\le 2M}(2\omega(m)-1) = 0$. Hence,

$$\{W_{2M} = 0\} = \bigcup_{r=1}^{M}\{\rho_N^{(1)} = 2r\} \cap \left\{\omega : \sum_{2r<m\le 2M}\big(2\omega(m)-1\big) = 0\right\}.$$

Since $\{\omega : \rho_N^{(1)}(\omega) = 2r\}$ depends only on $\omega \restriction \{1, \dots, 2r\}$ whereas

$$\left\{\omega : \sum_{2r<m\le 2M}\big(2\omega(m)-1\big) = 0\right\}$$

depends only on $\omega \restriction \{2r+1, \dots, 2M\}$, these events are independent. Furthermore, by the same homogeneity argument as we used above,

$$\mathbb{P}\left(\left\{\omega : \sum_{2r<m\le 2M}\big(2\omega(m)-1\big) = 0\right\}\right) = \mathbb{P}\big(W_{2(M-r)} = 0\big).$$

Hence

$$\mathbb{P}\big(W_{2M} = 0\big) = \sum_{r=1}^{M}\mathbb{P}\big(\rho_N^{(1)} = 2r\big)\mathbb{P}\big(W_{2(M-r)} = 0\big). \tag{1.2.26}$$

The reason why (1.2.26) is called a *renewal equation* is that it reflects the fact that, upon returning to its starting point, the walk begins again afresh.

Of course, one can argue that, in view of (1.2.25), (1.2.26) is simply the purely combinatorial identity

$$\binom{2M}{M} = \sum_{r=1}^{M} \frac{1}{2r-1}\binom{2r}{r}\binom{2(M-r)}{M-r}.$$

On the other hand, attempting to verify this identity by purely combinatorial means may increase one's appreciation of probabilistic reasoning.

1.2.7. The Arc Sine Law. Thinking of $\{W_0(\omega), \dots, W_{2M}(\omega)\}$ as the discrete time path of a random walk, construct the continuous time path $\{W_t(\omega) : t \in [0, 2M]\}$ by piecewise linear interpolation. That is,

$W_t(\omega) = (n-t)W_{n-1}(\omega) + (t-n+1)W_n(\omega)$ for $n-1 \le t \le n$ and $1 \le n \le 2M$.

Next, set

$$\mathcal{T}_{[0,2M]}(\omega) \equiv \int_0^{2M} \mathbf{1}_{[0,\infty)}\big(W_t(\omega)\big)\,dt,$$

where, for any set S, I use $\mathbf{1}_S$ to denote the **indicator function** of S. That is, $\mathbf{1}_S(x) = 1$ if $x \in S$ and $\mathbf{1}_S(x) = 0$ if $x \notin S$. Since, $t \rightsquigarrow \mathbf{1}_{[0,\infty)}(W_t(\omega))$ has at most a finite number of discontinuities in the interval $[0, 2M]$, the preceding integral is well-defined as a Riemann integral. In fact, for each $1 \le n \le M$, $W_t(\omega) > 0$ for all $t \in (2n-2, 2n)$ if $W_{2n-1}(\omega) > 0$ and $W_t(\omega) < 0$ for all $t \in (2n-2, 2n)$ if $W_{2n-1}(\omega) \le 0$, and so

$$\mathcal{T}_{[0,2M]}(\omega) = 2\sum_{n=1}^{M} \mathbf{1}_{(0,\infty)}\big(W_{2n-1}(\omega)\big),$$

twice the number of $1 \le n \le M$ for which $W_{2n-1}(\omega) > 0$.

To understand why $\mathcal{T}_{[0,2M]}(\omega)$ is a quantity of interest, think again about a gambling game with two contestants who toss a fair coin $2M$ times, waiting one second between tosses. Further, suppose that one of them wins a dollar each time a head turns up and the other wins a dollar each time a tail turns up. Then $W_n(\omega)$ represents the net gain or loss of the player who wins on heads, and so $\mathcal{T}_{[0,2M]}(\omega)$ represents the amount of time that that player is not behind. Those who have not done a lot of gambling often believe that *the law of averages* predicts that each player should be ahead about half the time. In other words, their naïve prediction is that, with high probability, the value of $\mathcal{T}_{[0,2M]}(\omega)$ will be in a neighborhood of M. On the other hand, habitual gamblers know that this prediction is false and that acting on it can lead to tragic consequences.

To show that our model of coin tossing reflects the experience of habitual gamblers, set $B_m^{2M} = \{\omega \in \{0,1\}^{2M} : \mathcal{T}_{[0,2M]}(\omega) = m\}$ for $0 \le m \le 2M$.

Because $\mathcal{T}_{[0,2M]}$ is even, $B_m^{2M} = \emptyset$ when m is odd. Also, since

$$B_{2M}^{2M} = \{\zeta_{2M}^{\{-1\}} > 2M\} \quad \text{and} \quad B_0^{2M} = \{\zeta_{2M}^{(1)} > 2M\},$$

(1.2.14) implies that

$$(*) \qquad \mathbb{P}(B_0^{2M}) = \mathbb{P}(B_{2M}^{2M}) = \mathbb{P}(W_{2M} = 0).$$

Now assume that $1 \le m < M$. Then $B_{2m}^{2M} \subseteq \{\rho_{2M}^{(1)} < 2M\}$, and so

$$\mathbb{P}(B_{2m}^{2M}) = \sum_{1\le r<M} \mathbb{P}(R_{2r} \cap B_{2m}^{2M}) = \sum_{1\le r<M} \mathbb{P}(R_{2r}^+ \cap B_{2m}^{2M}) + \sum_{1\le r<M} \mathbb{P}(R_{2r}^- \cap B_{2m}^{2M}),$$

where $R_{2r} = \{\rho_{2M}^{(1)} = 2r\}$ and $R_{2k}^\pm = R_{2r} \cap \{W(1) = \pm 1\}$.

Observe that $\omega \in R_{2r}^+ \implies W_{2r}(\omega) = 0$ and $\mathcal{T}_{[0,2r]}(\omega) = 2r$, and therefore $\omega \in R_{2r}^+ \cap B_{2m}^{2M}$ if and only if $r \le m$ and $\omega \in R_{2r}^+ \cap C_{2(m-r)}^{2M}(2r)$, where $C_{2n}^{2M}(2r)$ is the set of ω's for which $\omega \restriction \{2r+1, \dots, 2M\} \in B_{2(m-r)}^{2(M-r)}$. Further, $C_{2(m-r)}^{2M}(2r)$ is independent of R_{2r}^+ and, by homogeneity, $\mathbb{P}\big(C_{2(m-r)}^{2M}(r)\big) = \mathbb{P}(B_{2(m-r)}^{2(M-r)})$. Hence

$$\sum_{1\le r<M} \mathbb{P}(R_{2r}^+ \cap B_{2m}^{2M}) = \sum_{r=1}^{m} \mathbb{P}(R_{2r}^+)\mathbb{P}(B_{2(m-r)}^{2(M-r)}).$$

Next observe that $\omega \in R_{2r}^- \implies W_{2r}(\omega) = 0$ and $\mathcal{T}_{[0,2r]}(\omega) = 0$. Hence, $R_{2r}^- \cap B_{2m}^{2M} \ne \emptyset$ if and only if $r \le M - m$ and $R_{2r}^- \cap B_{2m}^{2M} = R_{2r}^- \cap C_{2m}^{2M}(2r)$. Starting from here and reasoning as in the preceding paragraph, we find that

$$\sum_{1\le r<M} \mathbb{P}(R_{2r}^- \cap B_{2m}^{2M}) = \sum_{r=1}^{M-m} \mathbb{P}(R_{2r}^-)\mathbb{P}(B_{2m}^{2(M-r)}).$$

Finally, since

$$R_{2r}^\pm = \{\pm W_n > 0 \text{ for } 1 \le n < 2r \text{ and } W_{2r} = 0\},$$

$\mathbb{P}(R_{2r}^+) = \mathbb{P}(R_{2r}^-)$, and therefore $2\mathbb{P}(R_{2r}^\pm) = \mathbb{P}(R_{2r}^+) + \mathbb{P}(R_{2r}^-) = \mathbb{P}(R_{2r})$. Putting this together with the results in the preceding paragraphs, we have shown that, for $1 \le m < M$,

$$(**) \qquad \mathbb{P}(B_{2m}^{2M}) = \frac{1}{2}\sum_{r=1}^{m} \mathbb{P}(R_{2r})\mathbb{P}(B_{2(m-r)}^{2(M-r)}) + \frac{1}{2}\sum_{r=1}^{M-m} \mathbb{P}(R_{2r})\mathbb{P}(B_{2m}^{2(M-r)}).$$

We can now show that

$$\begin{aligned} \mathbb{P}\big(\mathcal{T}_{[0,2M]} = 2m\big) &= \mathbb{P}\big(W_{2m} = 0\big)\mathbb{P}\big(W_{2(M-m)} = 0\big) \\ &= 4^{-M}\binom{2m}{m}\binom{2(M-m)}{M-m} \end{aligned} \tag{1.2.27}$$

for all $M \geq 1$ and $0 \leq m \leq M$. Indeed, it is clear that $(*)$ together with $(**)$ completely determines the numbers $\mathbb{P}(B_{2m}^{2M})$ for all $M \geq 1$ and $0 \leq m \leq M$. Thus, all that we have to do is verify that the numbers u_m^M on the right-hand side of (1.2.27) satisfy $(*)$ and $(**)$. There is nothing to do when $m \in \{0, 2M\}$. To prove it when $1 \leq m < M$, observe that

$$\{W_{2m} = 0 \;\&\; W_{2M} = 0\} = \{W_{2m} = 0\} \cap \{W_{2(M-m)} - W_{2m} = 0\}$$

and therefore that $u_m^M = \mathbb{P}(W_{2m} = 0 \,\&\, W_{2M} = 0)$. Hence, by the same reasoning as we used to derive the renewal equation (1.2.26),

$$u_m^M = \sum_{r=1}^{m} \mathbb{P}(R_{2r})\mathbb{P}\big(W_{2(m-r)} = 0 \;\&\; W_{2(M-r)} = 0\big) = \sum_{r=1}^{m} \mathbb{P}(R_{2r}) u_{m-r}^{M-r}.$$

Similarly, because

$$\{W_{2(M-m)} = 0 \;\&\; W_{2M} = 0\} = \{W_{2(M-m)} = 0\} \cap \{W_{2M} - W_{2(M-m)} = 0\},$$

$u_m^M = \sum_{r=1}^{M-m} \mathbb{P}(R_{2r}) u_m^{M-r}$, and so $\{u_m^M : 1 \leq m < M\}$ satisfies $(**)$.

Using the second equation in (1.2.27), one can easily check that

$$\mathbb{P}\big(\mathcal{T}_{[0,2M]} = 2(m-1)\big) \geq \mathbb{P}\big(\mathcal{T}_{[0,2M]} = 2m\big) \iff m \leq \frac{M+1}{2}.$$

Hence, the closer m is to M, the *less* likely it is that the amount of time the player who wins on heads will be ahead for $2m$ seconds. Equivalently, if $\overline{\mathcal{T}}_{[0,2M]}(\omega) = \frac{1}{2M}\mathcal{T}_{[0,2M]}(\omega)$ is the average time that he is ahead, then the least likely values for $\overline{\mathcal{T}}_{[0,2M]}(\omega)$ are those near $\frac{1}{2}$.

As another application of (1.2.27) and Stirling's formula, one can show that

$$\lim_{M\to\infty} \mathbb{P}\big(a < \overline{\mathcal{T}}_{[0,2M]} \leq b\big) = \tfrac{2}{\pi}\big(\arcsin\sqrt{b} - \arcsin\sqrt{a}\big) \tag{1.2.28}$$

for all $0 \leq a < b \leq 1$, a famous result known as the **arc sine law**. Indeed, by Stirling's formula (2.5.12),

$$2^{-n}\binom{2n}{n} = \frac{1}{\sqrt{\pi n}}\Big(1 + \mathcal{O}\big(\tfrac{1}{n}\big)\Big).$$

Hence, if $0 < a < b < 1$, then, because $B_m^{2M} = \emptyset$ when m is odd, for sufficiently large M

$$\begin{aligned}\mathbb{P}\big(a < \overline{\mathcal{T}}_{[0,2M]} \leq b\big) &= \sum_{aM<m\leq bM} \mathbb{P}(B_{2m}^{2M}) \\ &= \big(1 + \mathcal{O}\left(\tfrac{1}{M}\right)\big)\,\frac{1}{\pi} \sum_{aM<m\leq bM} \frac{1}{\sqrt{m(M-m)}}.\end{aligned}$$

Next write

$$\sum_{aM<m\le bM}\frac{1}{\sqrt{m(M-m)}}=M\sum_{aM<m\le bM}\left(\tfrac{m}{M}(1-\tfrac{m}{M})\right)^{-\frac{1}{2}},$$

think of the right-hand side as a Riemann sum, and conclude that

$$\begin{aligned}\lim_{M\to\infty}\sum_{aM<m\le bM}\frac{1}{\sqrt{m(M-m)}}&=\int_a^b t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}\,dt\\&=2\big(\arcsin\sqrt{b}-\arcsin\sqrt{a}\big).\end{aligned}$$

Thus (1.2.28) is proved for $0<a<b<1$. To prove it when $0=a<b<1$, note that, for any $0<\epsilon<b$, by the preceding,

$$\varlimsup_{M\to\infty}\left|\mathbb{P}\big(0\le\overline{\mathcal{T}}_{[0,2M]}\le b\big)-\tfrac{2}{\pi}\arcsin\sqrt{b}\right|\le\varlimsup_{M\to\infty}\mathbb{P}\big(\overline{\mathcal{T}}_{[0,2M]}<\epsilon\big)+\tfrac{2}{\pi}\arcsin\sqrt{\epsilon}.$$

Hence, it suffices to show that

$$\lim_{\epsilon\searrow 0}\varlimsup_{M\to\infty}\mathbb{P}\big(\overline{\mathcal{T}}_{[0,2M]}<\epsilon\big)=0.$$

But

$$\begin{aligned}&\mathbb{P}\big(\overline{\mathcal{T}}_{[0,2M]}<\epsilon\big)\le\mathbb{P}\big(\overline{\mathcal{T}}_{[0,2M]}<\epsilon\text{ or }\overline{\mathcal{T}}_{[0,2M]}>1-\epsilon\big)\\&\quad=1-\mathbb{P}\big(\epsilon\le\overline{\mathcal{T}}_{[0,2M]}\le 1-\epsilon\big)\longrightarrow 1-\tfrac{2}{\pi}\arcsin\sqrt{1-\epsilon}+\tfrac{2}{\pi}\arcsin\sqrt{\epsilon},\end{aligned}$$

which tends to 0 as $\epsilon\searrow 0$. The argument when $b=1$ is essentially the same.

1.2.8. Conditional Probability. A key concept in probability theory is that of *conditioning*. Namely, knowing that an event A occurs, one wants to compute the probability that an event B occurs. The mathematical interpretation of such a computation is that the **conditional probability** $\mathbb{P}(B|A)$ of B given A is the ratio $\frac{\mathbb{P}(B\cap A)}{\mathbb{P}(A)}$.[9] To understand the origin of this interpretation, suppose that $\mathbb{P}$ is the uniform probability measure on some finite space Ω. Then $\mathbb{P}(B|A)$ is the proportion of the set A which lies in B. Equivalently, $\mathbb{P}(B|A)$ is the probability that the uniform probability measure on A assigns to $B\cap A$, the portion of B inside A.

In order to be useful, the computation of a conditional probability must be done without having to compute the ratio. That is, in most applications, one wants to use conditioning to compute the probability of the intersection of events A and B as the product $\mathbb{P}(B|A)\mathbb{P}(A)$ of the conditional probability of B given A and the probability of A. For example, in the derivation of (1.2.26), we could have argued that the conditional probability of the event

[9] Until further notice, we will always condition with respect to events of positive probability.

$\{W_{2M} = 0\}$ given the event $\{\rho_{2M}^{(1)} = 2r\}$ must be $\mathbb{P}(W_{2(M-r)} = 0)$ since, knowing that $\rho_{2M}^{(1)}(\omega) = 2r$, $W_{2M}(\omega) = 0$ is equivalent to saying that half the tosses $2r+1$ through $2M$ are heads, and that probability is the same as the probability that half the tosses 1 through $2(M-r)$ are heads. Hence, we have shown that $\mathbb{P}(W_{2M} = 0 \,|\, \rho_{2M}^{(1)} = 2r) = \mathbb{P}(W_{2(M-r)} = 0)$, and so

$$\mathbb{P}(\{\rho_{2M}^{(1)} = 2r\} \cap \{W_{2M} = 0\}) = \mathbb{P}(\rho_{2M}^{(1)} = 2r)\mathbb{P}(W_{2(M-r)} = 0).$$

Equation (1.2.26) now can be viewed as an example of **Bayes's formula**, which is the simple observation that if $\{A_1, \dots, A_L\}$ is a partition of the sample space into events of positive probability, then for any event B

$$P(B) = \sum_{\ell=1}^{L} \mathbb{P}(B|A_\ell)\mathbb{P}(A_\ell). \tag{1.2.29}$$

Indeed, (1.2.29) is nothing but $\mathbb{P}(B) = \sum_{\ell=1}^{L} \mathbb{P}(B \cap A_\ell)$ written in terms of conditional probabilities.

The art of successful conditioning requires that one make a judicious choice of the event on which one is conditioning. In the preceding, $\{\rho_{2M}^{(1)} = 2r\}$ was a good choice because we can write

$$\{\rho_{2M}^{(1)} = 2r\} \cap \{W_{2M} = 0\}$$

as the intersection of $\{\rho_{2M}^{(1)} = 2r\}$ with a set, namely $\{W_{2M} - W_{2r} = 0\}$, of which it is independent. More generally, if $A \cap B = A \cap C$, where C is independent of A, then $\mathbb{P}(B|A) = \mathbb{P}(C)$. To give another example of the same reasoning, recall the argument which eventually led to (1.2.27). The key step can be thought of as the computation of $\mathbb{P}(B_{2m}^{2M}|R_{2r})$. However, $R_{2r} \cap B_{2m}^{2M}$ cannot be written as the intersection of R_{2r} with an event of which it is independent. For this reason, we chose to condition with respect to $R_{2r}^{\pm}$ instead. What we found is that $R_{2r}^{+} \cap B_{2m}^{2M} = R_{2r}^{+} \cap C_{2(m-r)}^{2(M-r)}(2r)$ and $R_{2r}^{-} = R_{2r}^{-} \cap C_{2m}^{2(M-r)}(2r)$, and since $C_{2m}^{2(M-r)}(2r)$ is independent of R_{2r}^{+} and $C_{2m}^{2(M-r)}(2r)$ is independent of R_{2r}^{-}, this means that

$$\mathbb{P}(B_{2m}^{2M}|R_{2r}^{+}) = \mathbb{P}\big(C_{2(m-r)}^{2(M-r)}(2r)\big) \quad \text{and} \quad \mathbb{P}(B_{2m}^{2M}|R_{2r}^{-}) = \mathbb{P}\big(C_{2m}^{2M}(2r)\big).$$

Finally, because $\mathbb{P}(R_{2r}^{\pm}) = \frac{1}{2}\mathbb{P}(R_{2r})$ and $\mathbb{P}(C_{2n}^{2M}(2r)) = \mathbb{P}(B_{2n}^{2(M-r)})$, this leads to $\mathbb{P}(R_{2r}^{\pm}|R_{2r}) = \frac{1}{2}$ and

$$\mathbb{P}(B_{2m}^{2M}|R_{2r}) = \tfrac{1}{2}\mathbb{P}(B_{2(m-r)}^{2(M-r)}) + \tfrac{1}{2}\mathbb{P}(B_{2m}^{2(M-r)})$$

after one applies the easily verified equality

$$\begin{gathered}\mathbb{P}(B|A) = \mathbb{P}(B|A_1)\mathbb{P}(A_1|A) + \mathbb{P}(B|A_2)\mathbb{P}(A_2|A) \\ \text{if } A_1 \cap A_2 = \emptyset \text{ and } A = A_1 \cup A_2.\end{gathered} \tag{1.2.30}$$

Exercises for § 1.2[10]

Exercise 1.2.31. Although I have seldom had an occasion to use it in one of my own classes, in probability classes that attract large audiences a standard opening gambit is to find out if there are any students who share a birth date. If the class has at least twenty-five students and they were all born in the same year, then there is a surprisingly good chance that two or more of them were born on the same date, and if there are fifty or more students, then it is nearly certain that this will be the case. To understand why, consider the space $\Omega = \{1, \dots, N\}^M$ with the uniform probability measure. Here N is thought of as the number of days (ignore leap years) in the year and M as the number of students in the class. Then the quantity of interest is the probability of the event $B = \{\omega : \omega(m) = \omega(n) \text{ for some } m \neq n\}$. By a trivial pigeonhole argument, $\mathbb{P}(B) = 1$ if $M \geq N$. If $M < N$, show that

$$\mathbb{P}(B) = 1 - \prod_{m=1}^{M-1} \Big(1 - \frac{m}{N}\Big) \geq 1 - e^{-\frac{M(M-1)}{2N}}.$$

Exercise 1.2.32. Suppose that an illiterate secretary is assigned the task of putting N invitations into N envelopes. Each invitation is to a different individual, and on each envelope is the address of the individual to whom the corresponding invitation is to be sent. What is the probability that no individual will receive the correct invitation? To model this problem mathematically, let Π_N denote the group of permutations of $\{1, \dots, N\}$, and take $\mathbb{P}$ to be the uniform probability measure on Π_N. Then the quantity of interest is $\mathbb{P}\big(\{\pi : \pi(n) \neq n \text{ for any } 1 \leq n \leq N\}\big)$. To compute this, for $\emptyset \neq F \subseteq \{1, \dots, N\}$, set $A_F = \{\pi : \pi(n) = n \text{ for } n \in F\}$, and show that $\mathbb{P}(A_F) = \frac{(N-r)!}{N!}$ if $\operatorname{card}(F) = r$. Use this together with Exercise 1.1.9 to show that

$$\mathbb{P}\big(\{\pi : \pi(n) = n \text{ for some } 1 \leq n \leq N\}\big) = -\sum_{r=1}^{N} \frac{(-1)^r}{r!}$$

and therefore that

$$\mathbb{P}\big(\{\pi : \pi(n) \neq n \text{ for any } 1 \leq n \leq N\}\big) = \sum_{r=0}^{N} \frac{(-1)^r}{r!}.$$

Note that there are two rather surprising aspects of this result. Namely, one might expect that this probability should decrease monotonically to 0 as $N \to \infty$. However, the result shows that the probability is larger for $2N$ than it is for $2N - 1$ and that, as $N \to \infty$, it tends to $\frac{1}{e}$, not 0.

[10] For a much more comprehensive and challenging list of exercises of this sort, see either [**3**] or [**8**]. The second of these contains a helpful selection of worked problems.

Exercise 1.2.33. Let $\mathbb{P}$ be the uniform probability measure on $\{0,1\}^N$.

(**i**) If (cf. (1.2.8))

$$\sigma_k(\omega) = \inf\{n:\ 1 \le n \le N \text{ and } S_n(\omega) = k\} \quad \text{for } 1 \le k \le N,$$

show that $\mathbb{P}(\sigma_k = n) = 2^{-n}\binom{n-1}{k-1}$ for $1 \le k \le n \le N$.

(**ii**) Show that, for $1 \le k < \ell \le N$ and $k \le m < n \le N$,

$$\mathbb{P}(\sigma_\ell = n | \sigma_k = m) = \mathbb{P}(\sigma_{\ell-k} = n - m),$$

and use Bayes's formula together with (**i**) to conclude that

$$\binom{n-1}{\ell-1} = \sum_{m=k}^{n-\ell+k} \binom{n-m-1}{\ell-k-1}\binom{m-1}{k-1}.$$

Exercise 1.2.34. Refer to § 1.2.3. Given a complete graph (V, E) and two colors (e.g., blue and yellow), a two-coloring of its edges is any coloring of its edges with the two colors. That is, each edge is assigned one of the two colors. Say that a coloring is monotone if all the edges are assigned the same color. Assuming that all colorings are equally likely and that $1 \le k \le N \equiv \text{card}(V)$, show that with probability no more than

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

will the coloring of some k-vertex subgraph be monotone. Use this to conclude that if $\binom{N}{k} < 2^{\frac{k(k-1)}{2}-1}$, then there will exist a coloring of (V, E) for which the coloring of no k-vertex subgraph will be monotone.

Exercise 1.2.35. Again take $\mathbb{P}$ to be the uniform probability measure on $\{0,1\}^N$. Assume that $N = 2M$, and define

$$L_{2M}(\omega) = \max\{n:\ 0 \le n \le 2M \text{ and } W_{2n}(\omega) = 0\}$$

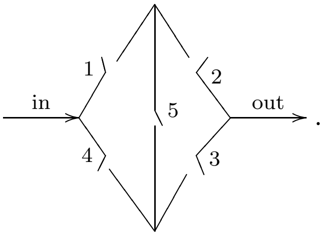
to be the last time the random walk visits 0 before or at time $2M$. Show that

$$\begin{aligned}\mathbb{P}(L_{2M} = 2m) &= \mathbb{P}(W_{2m} = 0)\mathbb{P}\big(\rho^{(1)}_{2(M-m)} > 2(M-m)\big)\\ &= \mathbb{P}(W_{2m} = 0)\mathbb{P}(W_{2(M-m)} = 0).\end{aligned}$$

Conclude that $\mathbb{P}(L_{2M} = 2m) = \mathbb{P}(\mathcal{T}_{[0,2M]} = 2m)$ and therefore that, as $M \to \infty$, $\mathbb{P}(L_{2M} \le 2Mx) \longrightarrow \frac{2}{\pi}\arcsin\sqrt{x}$ for $x \in [0,1]$.

Exercise 1.2.36. Consider a coin tossing game in which two coins, C_1 and C_2, are used. C_1 comes up heads with probability p_1 and C_2 with probability $p_2 \ne p_1$. On the first toss, C_1 is used. Thereafter, if a head comes up on toss n, then C_1 is used on toss $n+1$, and if tails comes up on toss n, then C_2 is used on toss $n+1$. Let P_n be the probability of a head coming up on toss n. Show that $P_{n+1} = p_1 P_n + p_2(1 - P_n)$, and use this to see that $P_n = \frac{p_2}{1+\Delta} + \left(p_1 - \frac{p_2}{1+\Delta}\right)(-\Delta)^{n-1}$, where $\Delta = p_2 - p_1$.

Exercise 1.2.37. Consider the network



.

Assume that each switch closes with probability $p \in (0,1)$ and remains open with probability $(1-p)$ and that the switches are independent of one another. What is the probability that a signal will pass through the network? Also, what is the conditional probability that switch 5 is open given that a signal passes through?

Exercise 1.2.38. Let $\mathbb{P}$ be a probability on Ω, and suppose that $A_1, \dots, A_N$ are events.

(**i**) If the A_n's are mutually independent, show that $\mathbb{P}(B_1 \cap \cdots \cap B_N) = \prod_{n=1}^N \mathbb{P}(B_n)$ if $B_n \in \{A_n, A_n\complement, \Omega\}$ for each $1 \le n \le N$.

(**ii**) If the A_n's form a partition of Ω and each one has positive probability, show that for any B of positive probability,

$$\mathbb{P}(A_m|B) = \frac{\mathbb{P}(B|A_m)\mathbb{P}(A_m)}{\sum_{n=1}^N \mathbb{P}(B|A_n)\mathbb{P}(A_n)} \quad \text{for } 1 \le m \le N. \tag{1.2.39}$$

Like the one in (1.2.29), the expression in (1.2.39) is called **Bayes's formula**.

(**iii**) Assuming that $\mathbb{P}$ is the uniform probability measure on $\{0,1\}^{2M}$, show that, for $0 \le m \le M$,

$$\mathbb{P}(W_{2M} = 0|W_{2m} = 0) = \mathbb{P}(W_{2(M-m)} = 0) = 4^{-M+m}\binom{2(M-m)}{M-m}$$
$$\text{and} \quad \mathbb{P}(W_{2m} = 0|W_{2M} = 0) = \frac{\binom{2m}{m}\binom{2(M-m)}{M-m}}{\binom{2M}{M}}.$$

Exercise 1.2.40. If $k \ge 1$ and $N \ge k$ with $\frac{N+k}{2} \in \mathbb{Z}$, show that

$$\begin{aligned} 2\mathbb{P}(\zeta_N^{\{k\}} = N) &= \mathbb{P}(\zeta_N^{\{k\}} > N - 1 \;\&\; W_{N-1} = k - 1) \\ &= \mathbb{P}(W_{N-1} = k-1) - \mathbb{P}(\zeta_N^{\{k\}} \le N - 1 \;\&\; W_{N-1} = k - 1) \\ &= \mathbb{P}(W_{N-1} = k - 1) - \mathbb{P}(W_{N-1} = k + 1), \end{aligned}$$

and use this to show that $\mathbb{P}(\zeta_N^{\{k\}} = N) = \frac{|k|}{N}\mathbb{P}(W_N = k)$ for all $1 \le |k| \le N$.

Exercise 1.2.41. Say that a sequence of events $\{A_n : n \geq 1\}$ are mutually independent if $\{A_1, \dots, A_N\}$ are mutually independent for all $N \in \mathbb{Z}^+$. Assuming that the events $\{A_n : n \geq 1\}$ are mutually independent events, show that, for each $1 \leq m \leq N$,

$$\mathbb{P}\left(\bigcup_{n=m}^{N} A_n\right) = 1 - \prod_{n=1}^{N}\bigl(1 - \mathbb{P}(A_n)\bigr) \geq 1 - e^{-\sum_{n=m}^{N}\mathbb{P}(A_n)},$$

and conclude that (cf. (1.1.11)) $\mathbb{P}\left(\overline{\lim}_{n\to\infty} A_n\right) = 1$ if $\sum_{n=1}^{\infty}\mathbb{P}(A_n) = \infty$. In conjunction with (1.1.13), this proves that if $\{A_n : n \geq 1\}$ is a sequence of mutually independent events, then

$$\mathbb{P}\left(\overline{\lim_{n\to\infty}} A_n\right) = \begin{cases} 0 & \text{if } \sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty, \\ 1 & \text{if } \sum_{n=1}^{\infty}\mathbb{P}(A_n) = \infty. \end{cases}$$

That is, $\mathbb{P}\left(\overline{\lim}_{n\to\infty} A_n\right)$ is either 0 or 1 according to whether $\sum_{n=1}^{\infty}\mathbb{P}(A_n)$ is finite or infinite. This dichotomy is known as the **Borel–Cantelli lemma**.

1.3. Some Non-Uniform Probability Measures

As I said in the introduction to § 1.2.1, if Ω is finite or countable, then any probability function $p : \Omega \longrightarrow [0,1]$ determines a probability measure $\mathbb{P}$ on Ω by the prescription $\mathbb{P}(A) = \sum_{\omega\in A} p(\omega)$. Thus, even when Ω is finite, there is no need to take $\mathbb{P}(\{\omega\})$ to be the same for all ω. On the other hand, if one is going to have a chance of computing probabilities of non-trivial events, then the assignment of $\mathbb{P}(\{\omega\})$ had better possess some structure, and, as we will see, independence is the sort of structure for which one should be looking.

1.3.1. Random Variables and Their Distributions. As we are about to see, we already have a ready source of non-uniform probability measures. Namely, suppose that Ω is a finite or countable sample space. Then *any* function X on Ω is called a **random variable** no matter where it takes its values. Because Ω is at most countable, so is the image $\text{Image}(X) \equiv \{X(\omega) : \omega \in \Omega\}$ of X. Moreover, if $\mathbb{P}$ is a probability measure on Ω and $p_X(x) = \mathbb{P}(X = x)$ for $x \in \text{Image}(X)$, then p_X is a probability function on $\text{Image}(X)$ and, as such, determines a probability measure μ_X, known as the **distribution** of X under $\mathbb{P}$, on $\text{Image}(X)$. Clearly,

$$\mu_X(\Gamma) = \mathbb{P}(X \in \Gamma) \quad \text{for } \Gamma \subseteq \text{Image}(X). \tag{1.3.1}$$

The considerations in §§ 1.2–1.3 give us several examples of random variables whose distributions arise again and again.

Bernoulli Distribution: Recall from § 1.2.2 the random variable S_n in (1.2.8), which is the number of heads when a fair coin is tossed n times. The distribution of S_n under the uniform probability measure is the probability measure on $\{0, \dots, n\}$ that assigns probability $2^{-n}\binom{n}{m}$ to each $0 \le m \le n$. For historical reasons, this probability measure is called a **Bernoulli measure**.

Discrete Arcsine Distribution: In § 1.2.7 we discussed the random variable $\mathcal{T}_{[0,2M]}$, which was the amount of time that a symmetric random walk spends above 0 during the time interval $[0, 2M]$. What we showed there is that the $\mathcal{T}_{[0,2M]}$ takes its values in $\{2m : 0 \le m \le M\}$ and, in (1.2.27), that

$$\mathbb{P}(\mathcal{T}_{[0,2M]} = 2m) = \mathbb{P}(W_{2m} = 0)\mathbb{P}(W_{2(M-m)} = 0).$$

Hence, there is a probability measure on $\{2m : 0 \le m \le M\}$ which assigns probability $4^{-M}\binom{2m}{m}\binom{2(M-m)}{M-m}$ to $2m$. For reasons that are made clear by (1.2.28), this probability measure is sometimes called the **discrete arcsine measure**. In Exercise 1.2.35, it was shown that L_{2M}, the last time the walk visits 0 at or before time $2M$, also has the discrete arcsine measure as its distribution.

Hitting Distributions: Again refer to § 1.2.4, and, for $N \ge k \ge 1$, consider the random variable $\zeta_N^{\{k\}}$, which is the first time that a random walk of duration N hits k. By the homogeneity property of coin tossing, we know that, for given $k \in \mathbb{Z}$ and $n \ge k$, $\mathbb{P}(\zeta_N^{\{k\}} = n)$ is the same for all $N \ge n$. Thus, by Exercise 1.2.40, $\mathbb{P}(\zeta_N^{\{k\}} = n) = \frac{k}{n}\mathbb{P}(W_n = k)$. Furthermore,

$$\sum_{n=1}^{N} \mathbb{P}(\zeta_N^{\{k\}} = n) = 1 - \mathbb{P}(\zeta_N^{\{k\}} > N) \quad \text{for all } N \ge k,$$

and, by (1.2.15), $\mathbb{P}(\zeta_N^{\{k\}} > N) \longrightarrow 0$ as $N \to \infty$. Hence, for each $k \in \mathbb{Z}^+$, there is a probability measure on $\{k + 2n : n \in \mathbb{N}\}$ that assigns probability $\frac{2^{-k-2n}k}{k+2n}\binom{k+2n}{n+k}$ to $k + 2n$.

First Return Time: A similar example is provided by the time $\rho_N^{(1)}$ that a random walk of duration N first returns to 0. For a symmetric random walk (cf. (1.2.23) and (1.2.25)), we know that $\mathbb{P}(\rho_N^{(1)} = 2r) = \frac{4^{-r}}{2r-1}\binom{2r}{r}$ and that $\mathbb{P}(\rho_N^{(1)} > 2r) = 4^{-r}\binom{2r}{r}$ for $N \ge 2r$. Since, as we saw in the derivation of (1.2.24), $4^{-r}\binom{2r}{r} \longrightarrow 0$ as $r \to \infty$, the same reasoning as we used in the preceding shows that there is a probability measure on $\{2r : r \in \mathbb{Z}^+\}$ that assigns probability $\frac{4^{-r}}{2r-1}\binom{2r}{r}$ to $2r$.

1.3.2. Biased Coins. A quite different source of non-uniform probability comes from replacing the fair coin in § 1.1.1 with a biased one.

Again take the sample space to be $\{0,1\}^N$. In § 1.1.1, we modeled N tosses of a fair coin by putting the uniform probability measure on $\{0,1\}^N$. However, we could have chosen a different description of that model. Namely, we could have said that the tosses are mutually independent and that, on each toss, the chance of a head is the same as that of a tail. More precisely, the uniform probability measure on $\{0,1\}^N$ is the probability measure on $\{0,1\}^N$ with the properties that, for any choice of $\{\eta_n : 1 \le n \le N\} \subseteq \{0,1\}$, the events $\{\omega : \omega(1) = \eta_1\}, \dots, \{\omega : \omega(N) = \eta_N\}$ are mutually independent and that each of these events has probability $\frac{1}{2}$. The advantage of this description is that it lends itself to natural generalizations. To wit, continue assuming that the tosses are mutually independent, but suppose that, instead of a fair one, the coin being tossed may be biased so that on each toss it comes up heads with probability $p \in (0,1)$ and tails with probability $q = 1 - p$. Then, unless $p = \frac{1}{2}$, the corresponding probability measure $\mathbb{P}_p$ on $\{0,1\}^N$ is no longer uniform. Indeed (cf. (1.2.8)),

$$\mathbb{P}_p(\{\omega\}) = p^{S_N(\omega)} q^{N - S_N(\omega)}. \tag{1.3.2}$$

Of course, when $p = \frac{1}{2}$, $\mathbb{P}_{\frac{1}{2}}$ is the uniform probability measure.

Even though $\mathbb{P}_p$ is not uniform, it is locally uniform on sets of the form $\{S_n = m\}$. Indeed, if $A \subseteq \{0,1\}^N$ depends only on $\omega \restriction \{1, \dots, n\}$ for some $1 \le n \le N$ and if $S_n(\omega) = m$ for all $\omega \in A$, then $\mathbb{P}(A) = \text{card}(A) p^m q^{n-m}$. In particular,

$$\mathbb{P}(S_n = m) = \binom{n}{m} p^m q^{n-m} \quad \text{for } n \ge 1 \text{ and } 0 \le m \le n. \tag{1.3.3}$$

This distribution is called the **binomial distribution with parameters (n,p)**, and obviously the Bernoulli measures are binomial distributions corresponding to $p = \frac{1}{2}$. A second way of expressing this local uniformity of $\mathbb{P}_p$ is in terms of conditional probabilities. Namely, if $1 \le n \le N$ and if A and B are events that depend only on $\omega \restriction \{1, \dots, n\}$, then

$$\mathbb{P}_p(B|A) = \frac{\text{card}(B \cap A)}{\text{card}(A)} = \mathbb{P}_{\frac{1}{2}}(B|A) \quad \text{if } S_n \text{ is constant on } A. \tag{1.3.4}$$

Another important fact about $\mathbb{P}_p$ is that it too has the same homogeneity property that $\mathbb{P}_{\frac{1}{2}}$ does. That is, (1.2.7) holds with $\mathbb{P}_p$ replacing $\mathbb{P}$, and so for any M-element subset $S \subseteq \{1, \dots, N\}$ and $\Gamma \subseteq \{0,1\}^M$, $\mathbb{P}_p$, thought of as a probability measure on $\{0,1\}^N$, assigns $\big\{\omega : \omega \restriction S \in \Gamma\big\}$ the same probability that $\mathbb{P}_p$, thought of as a probability measure on $\{0,1\}^M$, assigns to Γ. Finally, it is clear that, for any $A \subseteq \{0,1\}^N$,

$$\begin{gathered} \mathbb{P}_q(A) = \mathbb{P}_q\big(\{\omega \in \{0,1\}^N : \breve{\omega} \in A\}\big) \\ \text{where } \breve{\omega} \in \{0,1\}^N \text{ is given by } \breve{\omega}(n) = 1 - \omega(n) \text{ for } 1 \le n \le N. \end{gathered} \tag{1.3.5}$$

Because of (1.3.4), many of the calculations that we made in § 1.2 for symmetric random walks transfer to **biased random walks**, a random walk corresponding to a biased coin. To see how this is done, note that, since $W_n = k \iff S_n = \frac{n+k}{2}$, (1.3.4) implies that, for any B that depends only on $\omega \restriction \{1, \dots, n\}$,

$$\begin{aligned}\mathbb{P}_p\big(B \cap \{W_n = k\}\big) &= \mathbb{P}_{\frac{1}{2}}\big(B \cap \{W_n = k\}\big)\frac{\mathbb{P}_p(W_n = k)}{\mathbb{P}_{\frac{1}{2}}(W_n = k)} \\ &= \mathbb{P}_{\frac{1}{2}}\big(B \cap \{W_n = k\}\big)2^n p^{\frac{n+k}{2}} q^{\frac{n-k}{2}}.\end{aligned} \tag{1.3.6}$$

As an immediate applications of (1.3.6) and homogeneity, we have

$$\begin{aligned}&\mathbb{P}_p\big(\zeta_N^{\{k\}} \le n \;\&\; W_n = \ell\big) \\ &= \begin{cases} \mathbb{P}_p(W_n = \ell) & \text{if } k\ell \ge 0 \;\&\; |k| \le |\ell|, \\ \left(\frac{p}{q}\right)^{\ell - k} \mathbb{P}_p(W_n = 2k - \ell) & \text{otherwise} \end{cases}\end{aligned} \tag{1.3.7}$$

from (1.2.12),

$$\mathbb{P}_p\big(\rho_N^{(1)} = 2r\big) = \frac{\mathbb{P}_p(W_{2r} = 0)}{2r - 1} = \frac{(pq)^r}{2r - 1}\binom{2r}{r} \quad \text{for } N \ge 2r \tag{1.3.8}$$

from (1.2.25),

$$\mathbb{P}_p(W_{2M} = 0) = \sum_{r=1}^{M} \mathbb{P}_p\big(\rho_N^{(1)} = 2r\big)\mathbb{P}_p\big(W_{2(M-r)} = 0\big) \quad \text{for } N \ge 2M \tag{1.3.9}$$

from (1.2.26), and from Exercise 1.2.38

$$\mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big) = \frac{|k|}{n}\mathbb{P}_p(W_n = k) \quad \text{for } N \ge n \ge |k|. \tag{1.3.10}$$

It should be recognized that, because they involve events to which (1.3.4) does not apply, results like (1.2.14) and (1.2.21) do not admit such simple generalizations to random walks that are not symmetric.

Before closing this discussion, it seems appropriate to mention a famous approximation procedure discovered by Poisson. Namely, given a number $\alpha > 0$ and an $n > \alpha$,

$$\begin{aligned}\mathbb{P}_{\frac{\alpha}{n}}\big(S_n = m\big) &= \frac{n!\alpha^m}{m!(n-m)!n^m}\alpha^m\big(1 - \tfrac{\alpha}{n}\big)^{n-m} \\ &= \frac{\alpha^m}{m!}\left(\prod_{\ell=0}^{m-1}\big(1 - \tfrac{\ell}{n}\big)\right)\big(1 - \tfrac{\alpha}{n}\big)^{n-m}.\end{aligned}$$

Hence

$$\lim_{n\to\infty} \mathbb{P}_{\frac{\alpha}{n}}\big(S_n = m\big) = \frac{\alpha^m e^{-\alpha}}{m!} \quad \text{for each } m \in \mathbb{N}. \tag{1.3.11}$$

This result is known as the **Poisson approximation**, and the limit measure on $\mathbb{N}$ which assigns probability $\frac{\alpha^m e^{-\lambda}}{m!}$ to m is called the **Poisson measure with rate** α. Although a full explanation of the term "rate" will not come until the end of § 6.2.1, it should already be apparent what sort of phenomena are modeled by Poisson measures. Namely, they arise when one has a large number of rare, independent events and one is counting the number of events that occur. For example, if one has a chunk of radioactive material and one counts the number of particles that it emits over a long period of time, then the distribution of that number will be well approximated by a Poisson measure.

1.3.3. Recurrence and Transience of Random Walks. In § 1.2.4 we showed that the symmetric random walk will eventually visit every point in the sense that $\lim_{N\to\infty}\mathbb{P}_{\frac{1}{2}}(\zeta_N^{\{k\}} \le N) = 1$ for every $k \in \mathbb{Z}\setminus\{0\}$, and then, in (1.2.24), we showed that it is recurrent in the sense that $\lim_{N\to\infty}\mathbb{P}_{\frac{1}{2}}(\rho_N^{(1)} \le N) = 1$. We will now investigate the corresponding properties of the random walk under $\mathbb{P}_p$ when $p \ne \frac{1}{2}$. One suspects that, when $p > \frac{1}{2}$, it should be harder for a biased random walk to visit points to the left of 0 and easier for it to visit points to the right. In fact, the same intuition ought to make one think that such a random walk will, with positive probability, drift off to the right and never return to 0. In this subsection, we will verify both of these guesses.

By (1.3.10), with $k = 1$, and (1.3.3)

$$\lim_{N\to\infty}\mathbb{P}_p\big(\zeta_N^{\{1\}} \le N\big) = \lim_{N\to\infty}\sum_{n=0}^{N}\frac{\mathbb{P}_p(W_{2n+1} = 1)}{2n+1} = pu(pq),$$

where

$$u(x) = \sum_{n=0}^{\infty}\frac{1}{2n+1}\binom{2n+1}{n+1}x^n \quad \text{for } x \in \big(0, \tfrac{1}{4}\big]. \tag{1.3.12}$$

Because

$$1 = \lim_{N\to\infty}\mathbb{P}_{\frac{1}{2}}\big(\zeta_N^{(1)} \le N\big) = \frac{u\big(\frac{1}{4}\big)}{2},$$

we know that $u(x) \le 2$. In order to get a closed form expression for $u(x)$, observe that

$$\begin{aligned}\frac{1}{2n+1}\binom{2n+1}{n+1} &= \frac{(2n)!}{n!(n+1)!} = 2^n\frac{\prod_{m=1}^n(2m-1)}{(n+1)!} = (-4)^n\frac{\prod_{m=1}^n\big(\frac{1}{2}-m\big)}{(n+1)!} \\ &= 2(-4)^n\frac{\prod_{m=0}^n\big(\frac{1}{2}-m\big)}{(n+1)!} = -\frac{(-4)^{n+1}}{2}\binom{\frac{1}{2}}{n+1},\end{aligned}$$

where, for $r \in \mathbb{R}$ and $m \geq 1$,

$$(1.3.13) \qquad \binom{r}{m} \equiv \frac{\prod_{\ell=0}^{m-1}(r-\ell)}{m!}$$

is the *generalized binomial coefficient*: the coefficient of ξ^m in the Taylor expansion of $\xi \rightsquigarrow (1+\xi)^r$ around 0. Hence,

$$u(x) = -\frac{1}{2x}\sum_{n=0}^{\infty}\binom{\frac{1}{2}}{n+1}(-4x)^{n+1} = -\frac{1}{2x}\sum_{n=1}^{\infty}\binom{\frac{1}{2}}{n}(-4x)^n,$$

and so

$$(1.3.14) \qquad u(x) = \frac{1-\sqrt{1-4x}}{2x} \quad \text{for } x \in \left(0, \tfrac{1}{4}\right].$$

Noting that $pq \leq \frac{1}{4}$ and that $1-4pq = (p+q)^2 - 4pq = (p-q)^2$, we see that $u(pq) = \frac{p\wedge q}{pq} = \frac{1}{p\vee q}$ and therefore that

$$\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{1\}} \leq N\big) = \frac{p}{p\vee q} = \begin{cases} 1 & \text{if } p \geq \frac{1}{2}, \\ \frac{p}{q} & \text{if } p < \frac{1}{2}. \end{cases}$$

Further, since, by (1.3.5), $\mathbb{P}_p(\zeta_N^{\{-k\}} \leq N) = \mathbb{P}_q(\zeta_N^{\{k\}} \leq N)$, we also have that

$$\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{-1\}} \leq N\big) = \frac{q}{p\vee q} = \begin{cases} 1 & \text{if } p \leq \frac{1}{2}, \\ \frac{q}{p} & \text{if } p > \frac{1}{2}. \end{cases}$$

Hence, as predicted, when $p > \frac{1}{2}$, the biased walk will, with probability 1, eventually visit 1, but, with positive probability, it will never visit -1.

Starting from the preceding, we can now show that (cf. (1.2.22))

$$(1.3.15) \qquad \lim_{N\to\infty} \mathbb{P}_p\big(\rho_N^{(1)} \leq N\big) = 2(p\wedge q)$$

and therefore that $\lim_{N\to\infty}\mathbb{P}_p(\rho_N^{(1)} \leq N) < 1$ if $p \neq \frac{1}{2}$. Indeed, if $N > 2$, then

$$\mathbb{P}_p\big(\rho_N^{(1)} \leq N \,\big|\, W_1 = \pm 1\big) = \mathbb{P}_p\big(\zeta_N^{\{\mp 1\}} \leq N-1\big),$$

and so

$$\mathbb{P}_p\big(\rho_N^{(1)} \leq N\big) = p\mathbb{P}_p\big(\zeta_{N-1}^{\{-1\}} \leq N-1\big) + q\mathbb{P}_p\big(\zeta_{N-1}^{\{1\}} \leq N-1\big) \longrightarrow \frac{2pq}{p\vee q} = 2(p\wedge q).$$

In that this says that a biased (i.e., $p \neq \frac{1}{2}$) random walk will, with positive probability, never return to the place where it starts, one says that biased random walks are **transient**.

A similar argument allows us to compute $\lim_{N\to\infty}\mathbb{P}_p(\zeta_N^{\{k\}} \leq N)$ for all $k \neq 0$. Namely, if $k \geq 1$, then

$$\mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N \,\big|\, \zeta_N^{\{k\}} = n\big) = \mathbb{P}_p\big(\zeta_{N-n}^{\{1\}} \leq N-n\big)$$

for all $N \geq k+1$ and $k \leq n \leq N-1$. Hence, by Bayes's formula,

$$\mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) = \sum_{n=k}^{N-1} \mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big)\mathbb{P}_p\big(\zeta_{N-n}^{\{1\}} \leq N-n\big).$$

Therefore, since $\mathbb{P}_p\big(\zeta_k^{(1)} \leq N-n\big) \leq \frac{p}{p\vee q}$ for $1 \leq n \leq N$,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) &\leq \frac{p}{p\vee q} \lim_{N\to\infty} \sum_{n=k}^{N} \mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big)\\ &= \frac{p}{p\vee q} \lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k\}} \leq N\big),\end{aligned}$$

and so

$$\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) \leq \frac{p}{p\vee q} \lim_{N\to\infty} \mathbb{P}_p\big(\zeta^{\{k\}} \leq N\big).$$

At the same time, for each $k < M \leq N$

$$\mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) \geq \sum_{n=k}^{M-1} \mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big)\mathbb{P}_p\big(\zeta_{N-n}^{\{1\}} \leq N-n\big),$$

and therefore

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) &\geq \frac{p}{p\vee q} \sum_{n=k}^{M-1} \mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big)\\ &= \frac{p}{p\vee q} \mathbb{P}_p\big(\zeta_{M-1}^{\{k\}} \leq M-1\big)\end{aligned}$$

for all $k < M \leq N$. Combining this with the preceding, we conclude that

$$\lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k+1\}} \leq N\big) = \frac{p}{p\vee q} \lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k\}} \leq N\big)$$

and therefore, by induction and (1.3.5), that

$$\text{(1.3.16)} \qquad \lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k\}} \leq N\big) = \begin{cases} \left(\frac{p}{p\vee q}\right)^k & \text{for } k \geq 1,\\ \left(\frac{q}{p\vee q}\right)^{-k} & \text{for } k \leq -1.\end{cases}$$

Related to the preceding is the following strengthening of (1.2.24). Given N, use induction to define the mth return time $\rho_N^{(m)}$ on $\{0,1\}^N$ for $N \geq m \geq 2$ by

$$\rho_N^{(m)}(\omega) = \begin{cases} \inf\{n :\, \rho^{(m-1)}(\omega) < n \leq N \;\&\; W_n(\omega) = 0\} & \text{if } \rho_N^{(m-1)}(\omega) < N,\\ \infty & \text{otherwise.}\end{cases}$$

Then, $\mathbb{P}_p(\rho_N^{(m)} \leq N \,|\, \rho^{(m-1)} = n) = \mathbb{P}_p(\rho_{N-n}^{(1)} \leq N-n)$, and so

$$\mathbb{P}_p\big(\rho_N^{(m)} \leq N\big) = \sum_{n=0}^{N-1} \mathbb{P}_p\big(\rho_N^{(m-1)} = n\big)\mathbb{P}_p\big(\rho_{N-n}^{(1)} \leq N-n\big),$$

from which, reasoning in the same way as above, we find that

$$\lim_{N\to\infty} \mathbb{P}_p\big(\rho_N^{(m)} \le N\big) = \big(2(p\wedge q)\big)^m. \tag{1.3.17}$$

In particular, with probability 1, *the symmetric random walk will eventually return to* 0 *arbitrarily often.*

Exercises for § 1.3

Exercise 1.3.18. Assuming that $k\wedge\ell \ge 1$ and $N \ge k+\ell$, show that

$$\mathbb{P}_p\big(\zeta_N^{\{k+\ell\}} = n\big) = \sum_{m=0}^{n} \mathbb{P}_p\big(\zeta_N^{\{k\}} = m\big)\mathbb{P}_p\big(\zeta_{N-m}^{\{\ell\}} = n-m\big) \quad \text{for } N \ge n.$$

Similarly, show that

$$\mathbb{P}_p\big(\rho_N^{(\ell+m)} = r\big) = \sum_{n=0}^{r} \mathbb{P}_p\big(\rho_N^{(\ell)} = n\big)\mathbb{P}_p\big(\rho_{N-n}^{(m)} = r-n\big) \quad \text{for } N \ge r.$$

Exercise 1.3.19. If X and Y are independent random variables, show that

$$\begin{aligned}\mathbb{P}(X+Y=z) &= \sum_{x\in\text{Image}(X)} \mathbb{P}(Y=z-x)\mathbb{P}(X=x)\\ &= \sum_{y\in\text{Image}(Y)} \mathbb{P}(X=z-y)\mathbb{P}(Y=y).\end{aligned}$$

Next, suppose that X and Y are independent, **Poisson, random variables** with rates α and β. That is, X and Y are $\mathbb{N}$-valued and

$$\mathbb{P}(X=m\;\&\,Y=n) = e^{-(\alpha+\beta)}\frac{\alpha^m\beta^n}{m!n!}.$$

Show that $X+Y$ is a Poisson random variable with rate $(\alpha+\beta)$.

Exercise 1.3.20. Show that

$$\sum_{r=1}^{\infty} \frac{1}{2r-1}\binom{2r}{r}x^r = 1-\sqrt{1-4x} \quad \text{for } x\in\big[0,\tfrac14\big].$$

Next show that, for $x\in[0,1]$,

$$\sum_{r=1}^{\infty} \lim_{N\to\infty}\mathbb{P}_p\big(\rho_N^{(1)} = 2r\big)x^r = 1-\sqrt{1-4pqx},$$

and use the second part of Exercise 1.3.18 and induction to conclude that

$$\sum_{r=0}^{\infty} \lim_{N\to\infty}\mathbb{P}_p\big(\rho_N^{(m)} = 2r\big)x^r = \big(1-\sqrt{1-4pqx}\big)^m \quad \text{for } x\in[0,1].$$

Finally, use this to give another derivation of (1.3.17).

Exercise 1.3.21. Show that

$$\sum_{n=0}^{\infty} \lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{1\}} = n\big)x^n = \frac{1-\sqrt{1-4pqx}}{2q} \quad \text{for } x \in [0,1].$$

Using this expression together with the first part of Exercise 1.3.18 and induction on $k \geq 1$, show that

$$\sum_{n=0}^{\infty} \lim_{N\to\infty} \mathbb{P}_p\big(\zeta_N^{\{k\}} = n\big)x^n = \left(\frac{1-\sqrt{1-4pqx}}{2q}\right)^k \quad \text{for } x \in [0,1].$$

Finally, use these considerations to give another derivation of (1.3.16).

1.4. Expectation Values

In § 1.3.1 we used random variables as a source of non-uniform probability measures. In this section we will take their expectation values, and again we will restrict our attention to sample spaces which are finite or countable. To carry out this program, it will be useful to have the following results about series.

Lemma 1.4.1. *Let $\mathcal{I}$ be a finite or countable index set and let*

$$\{a_i : i \in \mathcal{I}\} \cup \{b_i : i \in \mathcal{I}\} \subseteq (-\infty, \infty],$$

and assume that $\sum_{i\in\mathcal{I}}(a_i^- + b_i^-) < \infty$. If either $\alpha, \beta \in [0,\infty)$ or $\alpha, \beta \in \mathbb{R}$ and $\sum_{i\in\mathcal{I}}(|a_i| + |b_i|) < \infty$, then

$$\sum_{i\in\mathcal{I}}(\alpha a_i + \beta b_i) = \alpha \sum_{i\in\mathcal{I}} a_i + \beta \sum_{i\in\mathcal{I}} b_i.$$

Proof. When $\mathcal{I}$ is finite, there is nothing to do. Thus, assume that $\mathcal{I}$ is infinite. In addition, in either case, $\sum_{i\in\mathcal{I}} \alpha a_i = \alpha \sum_{i\in\mathcal{I}} a_i$ and $\sum_{i\in\mathcal{I}} \beta b_i = \beta \sum_{i\in\mathcal{I}} b_i$, and so it is suffices to handle $\alpha = 1 = \beta$.

First suppose that $a_i \wedge b_i \geq 0$ for all $i \in \mathcal{I}$. Then, for any $F \subset\subset \mathcal{I}$,

$$\sum_{i\in F}(a_i + b_i) = \sum_{i\in F} a_i + \sum_{i\in F} b_i \leq \sum_{i\in\mathcal{I}} a_i + \sum_{i\in\mathcal{I}} b_i,$$

and so $\sum_{i\in\mathcal{I}}(a_i + b_i) \leq \sum_{i\in\mathcal{I}} a_i + \sum_{i\in\mathcal{I}} b_i$. At the same time, if $F \subset\subset \mathcal{I}$ and $G \subset\subset \mathcal{I}$, then

$$\sum_{i\in F} a_i + \sum_{i\in G} b_i \leq \sum_{i\in F\cup G}(a_i + b_i) \leq \sum_{i\in\mathcal{I}}(a_i + b_i),$$

and therefore $\sum_{i\in\mathcal{I}} a_i + \sum_{i\in\mathcal{I}} b_i \leq \sum_{i\in\mathcal{I}}(a_i + b_i)$.

Having handled the case when the a_i's and b_i's are non-negative, the other cases can be handled as follows. Set $\mathcal{I}^+ = \{i \in \mathcal{I} : a_i + b_i \geq 0\}$ and

$\mathcal{I}^- = \mathcal{I} \setminus \mathcal{I}^+$. Since $a_i^+ + b_i^+ = (a_i + b_i) + (a_i^- + b_i^-)$, the preceding implies that

$$\sum_{i\in\mathcal{I}^+} (a_i^+ + b_i^+) = \sum_{i\in\mathcal{I}^+} (a_i + b_i) + \sum_{i\in\mathcal{I}^+} (a_i^- + b_i^-)$$

and therefore that

$$\sum_{i\in\mathcal{I}^+} (a_i + b_i) = \sum_{i\in\mathcal{I}^+} a_i^+ + \sum_{i\in\mathcal{I}^+} b_i^+ - \sum_{i\in\mathcal{I}^+} a_i^- - \sum_{i\in\mathcal{I}^+} b_i^-.$$

The same line of reasoning shows that

$$\sum_{i\in\mathcal{I}^-} (a_i + b_i) = \sum_{i\in\mathcal{I}^-} a_i^+ + \sum_{i\in\mathcal{I}^-} b_i^+ - \sum_{i\in\mathcal{I}^-} a_i^- - \sum_{i\in\mathcal{I}^-} b_i^-.$$

Hence, by Lemma 1.2.1, when we add these two and apply (1.2.4), we get the desired result. □

Lemma 1.4.2. *Suppose that $\mathcal{I}$ and $\mathcal{J}$ are finite or countable index sets, and let $\{a_{i,j} : (i,j) \in \mathcal{I}\times\mathcal{J}\} \subseteq (-\infty,\infty]$. If either $a_{i,j} \ge 0$ for all $(i,j)\in\mathcal{I}\times\mathcal{J}$ or $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} |a_{i,j}| < \infty$, then both*

$$\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right) \quad \textit{and} \quad \sum_{j\in\mathcal{J}}\left(\sum_{i\in\mathcal{I}} a_{i,j}\right)$$

converge and are equal to $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} a_{i,j}$.

Proof. There is nothing to do when $\mathcal{I}$ and $\mathcal{J}$ are both finite. In addition, by reversing the roles of i and j, one can reduce the problem to showing that $\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right)$ has the asserted properties.

Now assume that the $a_{i,j}$'s are non-negative. Then, for each $F \subset\subset \mathcal{I}$ and $G \subset\subset \mathcal{J}$,

$$\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} a_{(i,j)} \ge \sum_{(i,j)\in F\times G} a_{(i,j)} = \sum_{i\in F}\left(\sum_{j\in G} a_{i,j}\right).$$

Thus, $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} a_{(i,j)} \ge \sum_{i\in F}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right)$ for all $F \subset\subset \mathcal{I}$, and therefore $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} a_{(i,j)} \ge \sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right)$. At the same time, if $H \subset\subset \mathcal{I}\times\mathcal{J}$ and $F \subset\subset \mathcal{I}$ and $G \subset\subset \mathcal{J}$ are chosen so that $H \subseteq F\times G$, then

$$\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right) \ge \sum_{i\in F}\left(\sum_{j\in G} a_{i,j}\right) = \sum_{(i,j)\in F\times G} a_{i,j} \ge \sum_{(i,j)\in H} a_{i,j},$$

and therefore $\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}} a_{i,j}\right) \ge \sum_{(i,j)\in\mathcal{I}\times\mathcal{J}} a_{i,j}$.

Finally, assume that $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}|a_{i,j}|<\infty$. Then, by the preceding,

$$\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}}|a_{i,j}|\right)<\infty,$$

and so $\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}a_{i,j}$ converges to some $s\in\mathbb{R}$, $\sum_{j\in\mathcal{J}}a_{i,j}$ converges to some $u_i\in\mathbb{R}$ for each i, and $\sum_{i\in\mathcal{I}}|u_i|<\infty$. Hence, by the preceding and Lemmas 1.2.1 and 1.4.1, the following equalities are justified:

$$\begin{aligned}\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}a_{i,j}&=\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}a_{i,j}^{+}-\sum_{(i,j)\in\mathcal{I}\times\mathcal{J}}a_{i,j}^{-}\\&=\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}}a_{i,j}^{+}\right)-\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}}a_{i,j}^{-}\right)\\&=\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}}a_{i,j}^{+}-\sum_{j\in\mathcal{J}}a_{i,j}^{-}\right)=\sum_{i\in\mathcal{I}}\left(\sum_{j\in\mathcal{J}}a_{i,j}\right).\end{aligned}$$ □

Now let $\mathbb{P}$ be a probability measure on a finite or countable sample space Ω. Given a random variable $X:\Omega\longrightarrow[0,\infty]$, the **expected value** $\mathbb{E}^{\mathbb{P}}[X]$ of X with respect to $\mathbb{P}$ is $\sum_{\omega\in\Omega}X(\omega)\mathbb{P}(\{\omega\})$. Given a random variable $X:\Omega\longrightarrow(-\infty,\infty]$, define the non-negative random variables $X^{\pm}$ by $X^{\pm}(\omega)=X(\omega)^{\pm}$ and $|X|=X^{+}+X^{-}$. Then we say that the **expected value of** X **exists** if $\mathbb{E}^{\mathbb{P}}[X^-]<\infty$, in which case we define its expected value $\mathbb{E}^{\mathbb{P}}[X]\equiv\mathbb{E}^{\mathbb{P}}[X^+]-\mathbb{E}^{\mathbb{P}}[X^-]$. Notice that, by Lemma 1.2.1,

$$\mathbb{E}^{\mathbb{P}}[X]=\sum_{\omega\in\Omega}X(\omega)\mathbb{P}(\omega)\quad\text{when }\mathbb{E}^{\mathbb{P}}[X^-]<\infty.\tag{1.4.3}$$

Thus, when Ω is finite and $\mathbb{P}$ is uniform, $\mathbb{E}^{\mathbb{P}}[X]$ is precisely the ordinary average value of X on Ω. When $\mathbb{E}^{\mathbb{P}}[|X|]<\infty$, we say that X is **integrable**.

Starting from (1.4.3) and applying the facts that we already have about series, it is clear that

$$\mathbb{E}^{\mathbb{P}}[X]\le\mathbb{E}^{\mathbb{P}}[Y]\quad\text{if }\mathbb{E}^{\mathbb{P}}[X^-]<\infty\text{ and }X\le Y,\tag{1.4.4}$$

and, for $\alpha,\beta\in\mathbb{R}$,

$$\mathbb{E}^{\mathbb{P}}[\alpha X+\beta Y]=\alpha\mathbb{E}^{\mathbb{P}}[X]+\beta\mathbb{E}^{\mathbb{P}}[Y]\tag{1.4.5}$$

if either $\alpha\wedge\beta\ge0$ and $\mathbb{E}^{\mathbb{P}}[X^-]\vee\mathbb{E}^{\mathbb{P}}[Y^-]<\infty$ or $\mathbb{E}^{\mathbb{P}}[|X|]\vee\mathbb{E}^{\mathbb{P}}[|Y|]<\infty$. Note that if $A\subseteq\Omega$ is an event, then $\mathbf{1}_AX^{\pm}\le X^{\pm}$ and so, by (1.4.4), $\mathbb{E}^{\mathbb{P}}[\mathbf{1}_AX]$ exists if $\mathbb{E}^{\mathbb{P}}[X]$ does, and $\mathbf{1}_AX$ is integrable if X is. In the future, when $\mathbb{E}^{\mathbb{P}}[X]$ exists, I will use the notation $\mathbb{E}^{\mathbb{P}}[X,\,A]$ to denote $\mathbb{E}^{\mathbb{P}}[\mathbf{1}_AX]$. Obviously,

$$\mathbb{E}^{\mathbb{P}}[X,\,A]=\sum_{\omega\in A}X(\omega)\mathbb{P}(\{\omega\})\quad\text{when }\mathbb{E}^{\mathbb{P}}[X^-]<\infty.$$

By using Lemma 1.4.2, we see that

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[X] &= \sum_{\omega\in\Omega}\left(\sum_{x\in\mathrm{Image}(X)} X(\omega)\mathbf{1}_{\{x\}}\big(X(\omega)\big)\mathbb{P}(\{\omega\})\right)\\ &= \sum_{x\in\mathrm{Image}(X)} x\left(\sum_{\{\omega:\, X(\omega)=x\}}\mathbb{P}(\{\omega\})\right),\end{aligned}$$

and so

$$\mathbb{E}^{\mathbb{P}}[X] = \sum_{x\in\mathrm{Image}(X)} x\mathbb{P}(X=x) \quad \text{when } \mathbb{E}^{\mathbb{P}}[X^-]<\infty. \tag{1.4.6}$$

In other words, when it exists, $\mathbb{E}^{\mathbb{P}}[X]$ is the weighted average of the values of X, the weight assigned to each value being the probability that X takes that value; and, for this reason, $\mathbb{E}^{\mathbb{P}}[X]$ is often called the **mean value** of X.

Equation (1.4.6) generalizes in the following way. Suppose that X is a random variable with values in some space E and that $f : E \longrightarrow [0,\infty)$, and set $Y = f\circ X$, the composition of f with X. Then, by (1.4.6), $\mathbb{E}^{\mathbb{P}}[Y]$ equals

$$\begin{aligned}\sum_{y\in\mathrm{Image}(Y)} y\mathbb{P}(Y=y) &= \sum_{y\in\mathrm{Image}(Y)}\left(\sum_{x\in\mathrm{Image}(X)} y\mathbf{1}_{\{y\}}\big(f(x)\big)\mathbb{P}(X=x)\right)\\ &= \sum_{x\in\mathrm{Image}(X)} f(x)\left(\sum_{y\in\mathrm{Image}(Y)}\mathbf{1}_{\{y\}}\big(f(x)\big)\right)\mathbb{P}(X=x)\\ &= \sum_{x\in\mathrm{Image}(X)} f(x)\mathbb{P}(X=x).\end{aligned}$$

More generally, if $f : E \longrightarrow (-\infty,\infty]$, then, by applying the preceding to $(f\circ X)^+$ and $(f\circ X)^-$, we have that

$$\mathbb{E}^{\mathbb{P}}[f\circ X] = \sum_{x\in\mathrm{Image}(X)} f(x)\mathbb{P}(X=x) \quad \text{if } \mathbb{E}^{\mathbb{P}}\big[(f\circ X)^-\big]<\infty. \tag{1.4.7}$$

As a weighted average of the size of X, one should expect that $\mathbb{E}^{\mathbb{P}}[|X|]$ can be used to estimate the probability that $|X|$ is large. To see that this is the case, note that, for any $R>0$, $R\mathbf{1}_{\{X\ge R\}} \le \mathbf{1}_{\{X\ge R\}}X \le |X|$ and therefore, by (1.4.4), that

$$\mathbb{P}(X\ge R) \le \frac{1}{R}\mathbb{E}^{\mathbb{P}}[X,\, X\ge R] \le \frac{1}{R}\mathbb{E}^{\mathbb{P}}\big[|X|\big] \quad \text{for } R>0, \tag{1.4.8}$$

an inequality which is known as **Markov's inequality**. Simple as it is, Markov's inequality is the origin of a great many estimates in probability theory. In fact, although we did not say so at the time, we used Markov's inequality in our derivation of (1.2.17). Namely, in the language of expectation values, if $\Omega = \{0,1\}^N$ and $\mathbb{P} = \mathbb{P}_{\frac{1}{2}}$ is the uniform probability measure on $\{0,1\}^N$, then, for $\alpha \in \mathbb{R}$,

$$\mathbb{E}^{\mathbb{P}}\big[e^{\alpha W_N}\big] = e^{-\alpha N}\mathbb{E}^{\mathbb{P}}\big[e^{2\alpha S_N}\big] = 2^{-N}e^{-\alpha N}\sum_{n=1}^{N}\binom{N}{n}e^{2\alpha n}$$
$$= 2^{-N}e^{-\alpha N}\big(e^{2\alpha}+1\big)^N = \big(\cosh\alpha\big)^N.$$

Hence, if $\alpha \geq 0$ and $R > 0$, then, by Markov's inequality,

$$\mathbb{P}\big(W_N \geq R\big) = \mathbb{P}\big(e^{\alpha W_N} \geq e^{\alpha R}\big) \leq e^{-\alpha R}\mathbb{E}^{\mathbb{P}}\big[e^{\alpha W_N}\big] \leq e^{-\alpha R}\big(\cosh\alpha\big)^N,$$

and it was from this inequality that (1.2.17) was an easy consequence.

Closely related to the preceding are the following considerations. Suppose that X is an $\mathbb{R}$-valued random variable for which X^2 is integrable. Since $|X| \leq \frac{1+X^2}{2}$, X is also integrable. For any $\alpha \in \mathbb{R}$,[11]

$$0 \leq \mathbb{E}^{\mathbb{P}}\big[(X-\alpha)^2\big] = \mathbb{E}^{\mathbb{P}}[X^2] - 2\alpha\mathbb{E}^{\mathbb{P}}[X] + \alpha^2.$$

In particular, because the right-hand side achieves its unique minimum at $\alpha = \mathbb{E}^{\mathbb{P}}[X]$, we see that

$$\text{(1.4.9)} \qquad \operatorname{Var}(X) \equiv \mathbb{E}^{\mathbb{P}}\Big[\big(X - \mathbb{E}^{\mathbb{P}}[X]\big)^2\Big] = \mathbb{E}^{\mathbb{P}}[X^2] - \mathbb{E}^{\mathbb{P}}[X]^2$$

is the minimum value of $\alpha \rightsquigarrow \mathbb{E}^{\mathbb{P}}\big[(X-\alpha)^2\big]$. In other words, when one uses $\mathbb{E}^{\mathbb{P}}\big[(X-\alpha)^2\big]$ to measure the difference between the random variable X and the constant α, $\mathbb{E}^{\mathbb{P}}[X]$ is the one and only choice of α which is closest to X. Thus, in this sense, $\operatorname{Var}(X)$ measures how much X differs from a constant, and for this reason it is called the **variance** of X. Notice that, by Markov's inequality, for any $R > 0$,

$$\mathbb{P}\big(\big|X - \mathbb{E}^{\mathbb{P}}[X]\big| \geq R\big) = \mathbb{P}\big(\big|X - \mathbb{E}^{\mathbb{P}}[X]\big|^2 \geq R^2\big) \leq R^{-2}\mathbb{E}^{\mathbb{P}}\Big[\big(X - \mathbb{E}^{\mathbb{P}}[X]\big)^2\Big],$$

and therefore

$$\text{(1.4.10)} \qquad \mathbb{P}\big(\big|X - \mathbb{E}^{\mathbb{P}}[X]\big| \geq R\big) \leq \frac{\operatorname{Var}(X)}{R^2} \quad \text{for } R > 0.$$

Inequality (1.4.10) is called **Chebychev's inequality**.

Before completing this introduction to expectation values, it should be pointed out that there is an alternative way to think about the "expected value" of an $\mathbb{R}$-valued random variable X. Namely, if one interprets "expected" as being synonymous with "typical," then one might say that an equally good candidate would be a number $\gamma \in \mathbb{R}$ such that X is equally likely to be larger or smaller than γ. That is, $\mathbb{P}(X \geq \gamma) = \frac{1}{2} = \mathbb{P}(X \leq \gamma)$.

[11] In the following and elsewhere, if $\alpha \in \mathbb{R}$, then I will use α to denote the random variable $\alpha\mathbf{1}_{\Omega}$.

However, after a moment's thought, one realizes that there either may be no such γ or that there may be more than one such γ. For example,

$$\mathbb{P}_{\frac{1}{2}}(S_2 \geq \gamma) \geq \tfrac{1}{2} \implies \gamma \leq 1 \implies \mathbb{P}_{\frac{1}{2}}(S_2 \geq \gamma) \geq \tfrac{3}{4},$$

whereas $\mathbb{P}_{\frac{1}{2}}(S_1 \geq \gamma) = \frac{1}{2} = \mathbb{P}_{\frac{1}{2}}(S_1 \leq \gamma)$ for all $\gamma \in (0,1)$. To eliminate the existence problem, one looks for a $\gamma \in \mathbb{R}$ such that

$$\mathbb{P}(X \geq \gamma) \wedge \mathbb{P}(X \leq \gamma) \geq \frac{1}{2} \tag{1.4.11}$$

and calls such a γ a **median** of X.

That every $\mathbb{R}$-valued random variable has a median is easy to check. Indeed, define

$$\alpha = \inf\{x : \mathbb{P}(X \leq x) \geq \tfrac{1}{2}\} \quad \text{and} \quad \beta = \sup\{x : \mathbb{P}(X \geq x) \geq \tfrac{1}{2}\}.$$

Because $\mathbb{P}(X \in \mathbb{R}) = 1$, (1.1.6) implies there exists an $R > 0$ such that $\mathbb{P}(X \geq -R) > \frac{1}{2}$ and $\mathbb{P}(X \geq R) < \frac{1}{2}$. Hence $\alpha, \beta \in [-R, R]$. Furthermore, since $\{X \leq \alpha + \frac{1}{n}\} \searrow \{X \leq \alpha\}$, (1.1.5) implies that

$$\mathbb{P}(X \leq \alpha) = \lim_{n\to\infty} \mathbb{P}\big(X \leq \alpha + \tfrac{1}{n}\big) \geq \tfrac{1}{2}.$$

Similarly, since $\{X \geq \alpha - \frac{1}{n}\} \nearrow \{X < \alpha\}$,

$$\mathbb{P}(X < \alpha) = \lim_{n\to\infty} \mathbb{P}\big(X \geq \alpha - \tfrac{1}{n}\big) \leq \tfrac{1}{2},$$

and therefore $\mathbb{P}(X \geq \alpha) = 1 - \mathbb{P}(X < \alpha) \geq \frac{1}{2}$. Hence, α is a median of X. The same sort of reasoning shows that β is also a median of X. In addition, since $\mathbb{P}(X \geq \alpha) \geq \frac{1}{2}$, we also know that $\alpha \leq \beta$. Finally, it is obvious that any median must lie between α and β, and knowing that α and β are both medians, it is clear that every $\gamma \in [\alpha, \beta]$ is a median. In other words, X always admits a median, but, in general, it will admit an entire, non-trivial interval of them. See Exercise 1.4.22 for a variational characterization of medians.

1.4.1. Some Elementary Examples.

There is no universally applicable procedure for computing expectation values any more than there is one for computing probabilities. Indeed, the probability of an event A is the expectation value of its indicator function $\mathbf{1}_A$, and so any universal technique for computing expectation values would be a technique for computing probabilities. Nonetheless, as is often the case, there are advantages to thinking in terms of the more general problem, the one of computing expectation values, even if our primary goal is to compute probabilities.

We will devote this subsection to a few elementary examples, and we begin by taking a hands-on approach. Our first example is the computation of (cf. (1.2.8)) $\mathbb{E}^{\mathbb{P}_p}[S_n]$ and $\mathrm{Var}_p(S_n)$, the variance of S_n under $\mathbb{P}_p$. From (1.3.3) and (1.4.6),

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_p}\big[S_n\big] &= \sum_{m=0}^{n} m\binom{n}{m}p^m q^{n-m} = np\sum_{m=1}^{n}\binom{n-1}{m-1}p^{m-1}q^{n-m}\\ &= np\sum_{m=0}^{n-1}\binom{n-1}{m}p^m q^{n-1-m} = np.\end{aligned}$$

Similarly, from (1.4.3), if $n \geq 1$,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_p}\big[S_n^2\big] &= \sum_{m=0}^{n} m^2\binom{n}{m}p^m q^{n-m}\\ &= \sum_{m=0}^{n} m(m-1)\binom{n}{m}p^m q^{n-m} + \sum_{m=0}^{n} m\binom{n}{m}p^m q^{n-m}.\end{aligned}$$

We already know that the last of these sums equals np. If $n = 1$, the second to last sum is 0, and if $n \geq 2$, it equals

$$\begin{aligned}n(n-1)p^2\sum_{m=2}^{n}\binom{n-2}{m-2}p^{m-2}q^{n-m} &= n(n-1)p^2\sum_{m=0}^{n-2}\binom{n-2}{m}p^m q^{n-2-m}\\ &= n(n-1)p^2.\end{aligned}$$

Hence

$$\mathbb{E}^{\mathbb{P}_p}\big[S_n^2\big] = n(n-1)p^2 + np = np(np - p + 1) = npq + (np)^2,$$

and so $\mathrm{Var}_p(S_n) = npq$.

In the preceding computation, we took minimal advantage of structure. In particular, we did not take any advantage of the fact that S_n is a sum and apply (1.4.5). If we had, we would have immediately seen that

$$\mathbb{E}^{\mathbb{P}_p}[S_n] = \sum_{m=1}^{n}\mathbb{E}^{\mathbb{P}_p}\big[\omega(m)\big] = \sum_{m=1}^{N}\mathbb{P}_p\big(\omega(m) = 1\big) = np$$

and that, because $\omega(m_1)$ is independent of $\omega(m_2)$ when $m_1 \neq m_2$,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\big[S_n^2\big] &= \sum_{m=1}^{n}\mathbb{E}^{\mathbb{P}_p}\big[\omega(m)^2\big] + \sum_{1\leq m_1\neq m_2\leq n}\mathbb{E}^{\mathbb{P}_p}\big[\omega(m_1)\omega(m_2)\big]\\ &= \sum_{m=1}^{n}\mathbb{P}_p\big(\omega(m) = 1\big) + \sum_{1\leq m_1\neq m_2\leq n}\mathbb{P}_p\big(\omega(m_1) = 1\big)\mathbb{P}_p\big(\omega(m_2) = 1\big)\\ &= np + n(n-1)p^2 = npq + (np)^2.\end{aligned}$$

On the basis of these computations, we can easily compute the expected value and variance of W_n under $\mathbb{P}_p$. Indeed, $W_n = 2S_n - n$, and so, by (1.4.5), $\mathbb{E}^{\mathbb{P}_p}[W_n] = 2np - n = np - nq = n(p-q)$ and

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_p}[W_n^2] = 4\mathbb{E}^{\mathbb{P}_p}\big[S_n^2\big] - 4n\mathbb{E}^{\mathbb{P}_p}[S_n] + n^2 &= 4npq + 4(np)^2 - 4n^2p + n^2 \\ &= 4npq + n^2(1-4pq) = 4npq + \big(n(p-q)\big)^2.\end{aligned}$$

Hence, $\mathrm{Var}_p(W_n) = 4npq$.

1.4.2. Independence and Moment Generating Functions. Although hands-on procedures work, one suspects that there must be more clever approaches. One such approach is to begin by computing the **moment generating function**

$$g_X(\lambda) \equiv \mathbb{E}^{\mathbb{P}}\big[e^{\lambda X}\big] \quad \text{for } \lambda \in \mathbb{R}. \tag{1.4.12}$$

Of course, in general, $g_X(\lambda)$ will be infinite for all $\lambda \neq 0$. On the other hand, if X is non-negative, it will be finite for $\lambda \in (-\infty, 0]$, and it will be finite for all $\lambda \in \mathbb{R}$ if X is bounded (i.e., $\mathbb{P}(|X| \le R) = 1$ for some $R \in (0,\infty)$).

To understand the virtue, as well as the origin, of the name of moment generating functions, again consider S_n under $\mathbb{P}_p$. Obviously

$$\mathbb{E}^{\mathbb{P}_p}\big[e^{\lambda S_n}\big] = \sum_{m=0}^{n} e^{\lambda m}\binom{n}{m}p^m q^{n-m} = \sum_{m=0}^{n}\binom{n}{m}(pe^{\lambda})^m q^{n-m} = \big(pe^{\lambda}+q\big)^n.$$

Hence, $\mathbb{E}^{\mathbb{P}_p}[e^{\lambda S_n}]$ is a smooth function of λ and[12]

$$\partial_\lambda^k \mathbb{E}^{\mathbb{P}_p}\big[e^{\lambda S_n}\big] = \sum_{m=0}^{n} m^k e^{\lambda m}\binom{n}{m}p^m q^{n-m} = \mathbb{E}^{\mathbb{P}_p}\big[S_n^k e^{\lambda S_n}\big].$$

In particular,

$$\mathbb{E}^{\mathbb{P}_p}\big[S_n^k\big] = \partial_\lambda^k\big(pe^{\lambda}+q\big)^n\big|_{\lambda=0},$$

from which it is an easy exercise to recover the results proved in § 1.4.1. More generally, because $\mathbb{E}^{\mathbb{P}}[X^k]$ is called the kth **moment** of X and since, under appropriate conditions, these moments can be computed by differentiating g_X at 0, it is reasonable to think of g_X as *generating* them.

Of course, moment generating functions are no *deus ex machina*. Aside from technical questions (addressed in Theorem 1.4.16) about justifying the differentiation of them to compute moments, there is a more basic problem: that of computing them at all. To understand why moment generating functions are often easier to compute than the moments themselves, we will repeat the preceding calculation, this time taking advantage of independence.

[12] I use ∂_x to denote differentiation with respect to x.

Namely,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_p}\big[e^{\lambda S_n}\big] &= \sum_{\omega\in\{0,1\}^n} e^{\lambda\sum_{m=1}^n \omega(m)}\mathbb{P}_p(\{\omega\}) = \sum_{\omega\in\{0,1\}^n}\prod_{m=1}^n e^{\lambda\omega(m)}p^{\omega(m)}q^{1-\omega(m)}\\ &= \prod_{m=1}^n \mathbb{E}^{\mathbb{P}_p}\big[e^{\lambda\omega(m)}\big] = \big(pe^{\lambda}+q\big)^n,\end{aligned}$$

the point being that, because the events $\{\omega(1)=\pm1\},\dots,\{\omega(n)=\pm1\}$ are mutually independent,

$$\mathbb{P}_p(\{\omega\}) = \prod_{m=1}^n \mathbb{P}_p\big(\{\omega' : \omega'(m)=\omega(m)\}\big) = \prod_{m=1}^n p^{\omega(m)}q^{1-\omega(m)}$$

and therefore

$$\mathbb{E}^{\mathbb{P}_p}\left[\prod_{m=1}^n e^{\lambda\omega(m)}\right] = \prod_{m=1}^n \mathbb{E}^{\mathbb{P}_p}\big[e^{\lambda\omega(m)}\big].$$

To generalize the above line of reasoning, let $\mathbb{P}$ be a probability measure on a finite or countable sample space Ω. We will say that the random variables $X_1,\dots,X_n$ on Ω are mutually **independent** under $\mathbb{P}$ if, for all $x_1,\dots,x_n\in\mathbb{R}$, the events $\{X_1=x_1\},\dots,\{X_n=x_n\}$ are mutually independent under $\mathbb{P}$. Hence, if $X_1,\dots,X_n$ are mutually independent random variables with values in some space E and if $f_1,\dots,f_n$ are non-negative functions on E, then, by (1.4.7),

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}\left[\prod_{m=1}^n f_m\circ X_m\right] &= \sum_{(x_1,\dots,x_n)}\prod_{m=1}^n f_m(x_m)\mathbb{P}(X_m=x_m)\\ &= \prod_{m=1}^n\sum_{x_m} f_m(x_m)\mathbb{P}(X_m=x_m),\end{aligned}$$

and so

$$\mathbb{E}^{\mathbb{P}}\left[\prod_{m=1}^n f_m\circ X_m\right] = \prod_{m=1}^n \mathbb{E}^{\mathbb{P}}\big[f_m(X_m)\big]. \tag{1.4.13}$$

Starting from (1.4.13), it is clear that for any $f_1,\dots,f_n$ on E into $\mathbb{R}$ such that $f_1\circ X_1,\dots,f_n\circ X_n$ are integrable, $\prod_{m=1}^n f_m\circ X_m$ is again integrable and (1.4.13) continues to hold.

As an immediate consequence of (1.4.13), we see that if $X_1, \dots, X_n$ are mutually independent, $(-\infty, \infty]$-valued random variables, then

$$g_S(\lambda) = \prod_{m=1}^{n} g_{X_m}(\lambda) \quad \text{where } S = \sum_{m=1}^{n} X_m, \tag{1.4.14}$$

and this important fact is one of the major reasons why moment generating functions are often easier to compute than one might expect.

1.4.3. Basic Convergence Results. For many applications, it is important to know under what conditions one can say that $\mathbb{E}^{\mathbb{P}}[X_n] \longrightarrow \mathbb{E}^{\mathbb{P}}[X]$ when one knows that $X_n(\omega) \longrightarrow X(\omega)$ for each $\omega \in \Omega$. For instance, one needs such results when one attempts to justify the interchange, used in the computation of moments from the moment generating function, of differentiation with the taking of expectation values.

Theorem 1.4.15. *Let $\mathbb{P}$ be a probability measure on a finite or countable sample space Ω, and let $\{X_n : n \geq 1\}$ be a sequence of $(-\infty, \infty]$-valued random variables on Ω with $\mathbb{E}^{\mathbb{P}}[X_n^-] < \infty$ for all $n \geq 1$.*

(**i**) *If $0 \leq X_n(\omega) \nearrow X(\omega)$ for each $\omega \in \Omega$, then $\mathbb{E}^{\mathbb{P}}[X_n] \nearrow \mathbb{E}^{\mathbb{P}}[X]$.*

(**ii**) *If $X_n \geq 0$ for all $n \geq 1$, then*

$$\mathbb{E}^{\mathbb{P}}\left[\varliminf_{n\to\infty} X_n\right] \leq \varliminf_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n].$$

(**iii**) *If there exists an integrable random variable Y such that $X_n \leq Y$ for all $n \geq 1$ and $\varlimsup_{n\to\infty} X_n$ is integrable, then*

$$\varlimsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n] \leq \mathbb{E}^{\mathbb{P}}\left[\varlimsup_{n\to\infty} X_n\right].$$

(**iv**) *If there exists an integrable random variable Y such that $|X_n| \leq Y$ for all $n \geq 1$ and $X_n(\omega) \longrightarrow X(\omega)$ for each $\omega \in \Omega$, then X is integrable and*

$$\left|\mathbb{E}^{\mathbb{P}}[X_n] - \mathbb{E}^{\mathbb{P}}[X]\right| \leq \mathbb{E}^{\mathbb{P}}\big[|X_n - X|\big] \longrightarrow 0.$$

Proof. (**i**) First note that, by (1.4.4), $0 \leq \mathbb{E}^{\mathbb{P}}[X_n] \leq \mathbb{E}^{\mathbb{P}}[X_{n+1}] \leq \mathbb{E}^{\mathbb{P}}[X]$ for all $n \geq 1$. Thus $L \equiv \lim_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n]$ exists in $[0, \infty]$ and is dominated by $\mathbb{E}^{\mathbb{P}}[X]$. To complete the proof, note that, for any $A \subset\subset \Omega$, $\mathbb{E}^{\mathbb{P}}[X,\, A] = \lim_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n,\, A] \leq L$, and therefore (cf. Lemma 1.2.1)

$$\mathbb{E}^{\mathbb{P}}[X] = \sup_{A \subset\subset \Omega} \mathbb{E}^{\mathbb{P}}[X,\, A] \leq L.$$

(**ii**) Set $Y_n(\omega) = \inf\{X_m(\omega) : m \geq n\}$ for $n \geq 1$. Then, $X_n \geq Y_n \nearrow \varliminf_{n\to\infty} X_n$, and therefore, by (**i**) and (1.4.4),

$$\varliminf_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n] \geq \lim_{n\to\infty} \mathbb{E}^{\mathbb{P}}[Y_n] = \mathbb{E}^{\mathbb{P}}\left[\varliminf_{n\to\infty} X_n\right].$$

(**iii**) Set $Z_n = Y - X_n$. Then, by (**ii**) applied to $\{Z_n : n \geq 1\}$,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}}[Y] + \mathbb{E}^{\mathbb{P}}\left[-\varlimsup_{n\to\infty} X_n\right] &= \mathbb{E}^{\mathbb{P}}\left[Y - \varlimsup_{n\to\infty} X_n\right] = \mathbb{E}^{\mathbb{P}}\left[\varliminf_{n\to\infty} Z_n\right] \\ &\leq \varliminf_{n\to\infty} \mathbb{E}^{\mathbb{P}}[Y - X_n] = \mathbb{E}^{\mathbb{P}}[Y] - \varlimsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n].\end{aligned}$$

Hence, $\varlimsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n] \leq -\mathbb{E}^{\mathbb{P}}\left[-\varlimsup_{n\to\infty} X_n\right]$, and so $\varlimsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}[X_n] \leq \mathbb{E}^{\mathbb{P}}\left[\varlimsup_{n\to\infty} X_n\right]$.

(**iv**) Since $|X| \leq |Y|$, it is clear that X is integrable. In addition, since $\pm(X_n - X) \leq |X_n - X|$ and therefore $\pm\mathbb{E}^{\mathbb{P}}[X_n - X] \leq \mathbb{E}^{\mathbb{P}}\big[|X_n - X|\big]$, we know that $\big|\mathbb{E}^{\mathbb{P}}[X_n] - \mathbb{E}^{\mathbb{P}}[X]\big| \leq \mathbb{E}^{\mathbb{P}}\big[|X_n - X|\big]$. Now set $Z_n = |X_n - X|$. Then $0 \leq Z_n \leq 2Y$ and $Z_n(\omega) \longrightarrow 0$ for all $\omega \in \Omega$. Hence, by (**iii**), we know that $\varlimsup_{n\to\infty} \mathbb{E}^{\mathbb{P}}\big[|X_n - X|\big] = 0$. □

The results in Theorem 1.4.15 are the countable sample space version of three famous convergence results (cf. Theorem 2.4.12) in Lebesgue's theory of integration. The result in (**i**) is known as the **monotone convergence theorem**, those in (**ii**) and (**iii**) are called **Fatou's lemma**, and the one in (**iv**) is **Lebesgue's dominated convergence theorem**.

As an application of these, we give an important result about computing moments from moment generating functions. Here and elsewhere, $e^{-\infty}$ is taken to be 0.

Theorem 1.4.16. *If $-\infty \leq a < b \leq \infty$ and $g_X(\lambda) < \infty$ for $\lambda \in (a,b)$, then (cf.* (1.4.12)*) g_X is smooth (i.e., infinitely differentiable) on (a,b) and*

$$\mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X}\big] = \partial_\lambda^k g_X(\lambda) \quad \textit{for } k \in \mathbb{N} \textit{ and } \lambda \in (a,b).$$

In addition, if X is bounded below (i.e., $\mathbb{P}(X \geq -R) = 1$ for some $R \in (0,\infty)$), then

$$\mathbb{E}^{\mathbb{P}}\big[X^k,\, X < \infty\big] = \lim_{\lambda\nearrow 0} \partial_\lambda^k g_X(\lambda) \quad \textit{for each } k \in \mathbb{N}.$$

Proof. To prove the first assertion, we will work by induction on $k \in \mathbb{N}$, and when $k = 0$ there is nothing to do. Now let $k \geq 1$, and assume the result for $k - 1$. Given $\lambda \in (a,b)$, choose $\delta > 0$ satisfying $[\lambda - 2\delta, \lambda + 2\delta] \subseteq (a,b)$. Then for any $h \in [-2\delta, 2\delta]$,

$$\frac{\partial_\lambda^{k-1} g_X(\lambda + h) - \partial_\lambda^{k-1} g_X(\lambda)}{h} = \mathbb{E}^{\mathbb{P}}\left[X^k \int_0^1 e^{(\lambda + th)X}\, dt\right].$$

Since, for any $h \in \mathbb{R}$ with $|h| < \delta$,

$$|x^k|e^{(\lambda+h)x} \le |x^k|e^{-\delta|x|}e^{\lambda x+2\delta|x|} \le \left(\frac{k}{e\delta}\right)^k \left(e^{(\lambda+2\delta)x} + e^{(\lambda-2\delta)x}\right),$$

$$\left|X^k \int_0^1 e^{(\lambda+th)X}\,dt\right| \le \left(\frac{k}{e\delta}\right)^k \left(e^{(\lambda+2\delta)X} + e^{(\lambda-2\delta)X}\right)$$

for such h. Hence, by Lebesgue's dominated convergence theorem,

$$\lim_{h\to 0} \frac{\partial_\lambda^{k-1} g_X(\lambda+h) - \partial_\lambda^{k-1} g_X(\lambda)}{h} = \mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X}\big].$$

Given the first part, the second part comes down to showing that $\mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X}\big] \longrightarrow \mathbb{E}^{\mathbb{P}}\big[X^k,\, X < \infty\big]$ as $\lambda \nearrow 0$. To this end, write

$$\mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X}\big] = \mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X},\, X \in [-R, 0)\big] + \mathbb{E}^{\mathbb{P}}\big[X^k e^{\lambda X},\, X \in [0,\infty)\big].$$

By Lebesgue's dominated convergence theorem, the first term on the right tends to $\mathbb{E}^{\mathbb{P}}\big[X^k,\, X < 0\big]$ and, by the monotone convergence theorem, the second term tends to $\mathbb{E}^{\mathbb{P}}\big[X^k,\, X \in [0,\infty)\big]$ as $\lambda \nearrow 0$. Thus, the sum tends to $\mathbb{E}^{\mathbb{P}}\big[X^k,\, X < \infty\big]$. □

Exercises for § 1.4

Exercise 1.4.17. If X takes its values in $\mathbb{N}$, show that

$$\mathbb{E}^{\mathbb{P}}[X] = \sum_{n=0}^{\infty} \mathbb{P}(X > n) \quad \text{and} \quad \mathbb{E}^{\mathbb{P}}[X^2] = \sum_{n=0}^{\infty} (2n+1)\mathbb{P}(X > n).$$

Exercise 1.4.18. If X is a **Poisson random variable** with rate α (i.e., X is $\mathbb{N}$-valued and $\mathbb{P}(X = n) = \frac{\alpha^n e^{-\alpha}}{n!}$ for $n \in \mathbb{N}$), show that $g_X(\lambda) = e^{\alpha(e^\lambda - 1)}$ and conclude that $\mathbb{E}^{\mathbb{P}}[X] = \alpha = \mathrm{Var}(X)$.

Exercise 1.4.19. Let $p \in (0,1)$ and $q = 1 - p$, suppose that X is a $\mathbb{Z}^+$-valued random variable with $\mathbb{P}(X = n) = qp^{n-1}$ for $n \in \mathbb{Z}^+$, and show that $g_X(\lambda) = \frac{qe^\lambda}{1-pe^\lambda}$ for $\lambda < -\log p$, $\mathbb{E}^{\mathbb{P}}[X] = \frac{1}{q}$, and $\mathrm{Var}(X) = \frac{p}{q^2}$. Also, show that such a random variable arises when one asks what the probability is that a tail occurs for the first time on the nth toss of a coin.

Exercise 1.4.20. Assume that X is an $\mathbb{R}$-valued random variable for which there exists a $\delta > 0$ such that $g_X(\lambda) < \infty$ whenever $\lambda \in (-\delta, \delta)$. If $\Lambda_X(\lambda) = \log g_X(\lambda)$, show that

$$\mathbb{E}^{\mathbb{P}}[X] = \frac{d\Lambda_X}{d\lambda}(0) \quad \text{and} \quad \mathrm{Var}(X) = \frac{d^2\Lambda_X}{d\lambda^2}(0).$$

More generally, the derivatives of Λ_X are called the **cumulants** of X.

Exercise 1.4.21. It should be clear that a major difference between a median of a random variable and its expected value is that a median ignores "outliers" whereas the expected value can be influenced by them. Thus, one should expect that the difference between a median and the expected value can be estimated in terms of the variance. To verify this, let X be an $\mathbb{R}$-valued random variable for which X^2 is integrable, and suppose that γ is a median of X. Show that, for any $x \in \mathbb{R}$, $(\gamma - x)^2 \le 2\mathbb{E}^{\mathbb{P}}\big[(X-x)^2\big]$, and conclude that $|\gamma - \mathbb{E}^{\mathbb{P}}[X]| \le \sqrt{2\mathrm{Var}(X)}$.

Exercise 1.4.22. Suppose that X is a $\mathbb{Z}$-valued random variable.

(**i**) Show that if $\alpha \in \mathbb{R} \setminus \mathbb{Z}$ is a median of X and if $m \in \mathbb{Z}$ is determined by $m < \alpha < m+1$, then $\mathbb{P}(X \le m) = \frac{1}{2} = \mathbb{P}(X \ge m+1)$ and therefore, for every $\beta \in [m, m+1]$, $\mathbb{P}(X \le \beta) = \frac{1}{2} = \mathbb{P}(X \ge \beta)$. In particular, conclude that there exist integers $m_1 \le m_2$ such that, for any $\alpha \in \mathbb{R}$, α is a median of X if and only if $m_1 \le \alpha \le m_2$.

(**ii**) Assume that X is integrable and let $m \in \mathbb{Z}$. If $m \le \alpha \le m+1$, show that

$$\mathbb{E}^{\mathbb{P}}\big[|X-\alpha|\big] - \mathbb{E}^{\mathbb{P}}\big[|X-m|\big] = (\alpha - m)\big(1 - 2\mathbb{P}(X \ge m+1)\big).$$

If $m-1 \le \alpha \le m$, show that

$$\mathbb{E}^{\mathbb{P}}\big[|X-\alpha|\big] - \mathbb{E}^{\mathbb{P}}\big[|X-m|\big] = (\alpha - m)\big(1 - 2\mathbb{P}(X \ge m)\big).$$

(**iii**) Again, assume that X is integrable, and let m_1 and m_2 be as in (**i**). Using (**ii**), show that

$$\mathbb{E}^{\mathbb{P}}\big[|X-\alpha|\big] \ge \mathbb{E}^{\mathbb{P}}\big[|X-m_1|\big] \quad \text{for all } \alpha \in \mathbb{R}$$

and that equality holds if and only if $\alpha \in [m_1, m_2]$. Conclude from this that $\alpha \in \mathbb{R}$ is a median of X if and only if

$$\mathbb{E}^{\mathbb{P}}\big|[X-\alpha|\big] = \min\big\{\mathbb{E}^{\mathbb{P}}\big[|X-\beta|\big] : \beta \in \mathbb{R}\big\}.$$

It can be shown (cf. Exercise 1.4.3 in my book [**9**]) that this variational characterization works for general $\mathbb{R}$-valued random variables, not just for integer-valued random variables.

Comments on Chapter 1

As I have said, when the sample space is finite, the engine behind probabilistic computations is combinatorics. Nonetheless, when these computations are phrased in probabilitic terms, many techniques for solving them become more transparent and better motivated. In addition, because it does not require the sometimes cumbersome machinery required to do probability theory in uncountable sample spaces, finite and countable state spaces provide a testing ground for results that one would like to prove. To wit, people working in statistical mechanics or quantum field theory make systematic use of finite analogs of the physical situations in which they are interested. In fact, they sometimes are able to prove the results they want by making finite approximations and then passing to a limit. Three of the most intriguing examples of this procedure are percolation theory, the Ising model, and Euclidean quantum field theory. Also, results are often first discovered in a finite setting and only later are shown to be more general. Both De Moivre's theorem and the arc sine law are examples.