8

Finite and discrete probability distributions

To understand the algorithmic aspects of number theory and algebra, and applications such as cryptography, a firm grasp of the basics of probability theory is required. This chapter introduces concepts from probability theory, starting with the basic notions of probability distributions on finite sample spaces, and then continuing with conditional probability and independence, random variables, and expectation. Applications such as "balls and bins," "hash functions," and the "left-over hash lemma" are also discussed. The chapter closes by extending the basic theory to probability distributions on countably infinite sample spaces.

8.1 Basic definitions

Let Ω be a finite, non-empty set. A **probability distribution on** Ω is a function $\mathsf{P} : \Omega \to [0, 1]$ that satisfies the following property:

$$\sum_{\omega \in \Omega} \mathsf{P}(\omega) = 1. \tag{8.1}$$

The set Ω is called the **sample space of** P.

Intuitively, the elements of Ω represent the possible outcomes of a random experiment, where the probability of outcome $\omega \in \Omega$ is $\mathsf{P}(\omega)$. For now, we shall only consider probability distributions on finite sample spaces. Later in this chapter, in §8.10, we generalize this to allow probability distributions on *countably infinite* sample spaces.

Example 8.1. If we think of rolling a fair die, then setting $\Omega := \{1, 2, 3, 4, 5, 6\}$, and $\mathsf{P}(\omega) := 1/6$ for all $\omega \in \Omega$, gives a probability distribution that naturally describes the possible outcomes of the experiment. □

Example 8.2. More generally, if Ω is any non-empty, finite set, and $\mathsf{P}(\omega) := 1/|\Omega|$ for all $\omega \in \Omega$, then P is called the **uniform distribution on** Ω. □

Example 8.3. A coin toss is an example of a **Bernoulli trial**, which in general is an experiment with only two possible outcomes: *success*, which occurs with probability p; and *failure*, which occurs with probability $q := 1 - p$. Of course, *success* and *failure* are arbitrary names, which can be changed as convenient. In the case of a coin, we might associate *success* with the outcome that the coin comes up *heads*. For a fair coin, we have $p = q = 1/2$; for a biased coin, we have $p \neq 1/2$. □

An **event** is a subset $\mathcal{A}$ of Ω, and the **probability of** $\mathcal{A}$ is defined to be

$$\mathsf{P}[\mathcal{A}] := \sum_{\omega \in \mathcal{A}} \mathsf{P}(\omega). \tag{8.2}$$

While an event is simply a subset of the sample space, when discussing the probability of an event (or other properties to be introduced later), the discussion always takes place relative to a particular probability distribution, which may be implicit from context.

For events $\mathcal{A}$ and $\mathcal{B}$, their union $\mathcal{A} \cup \mathcal{B}$ logically represents the event that *either* the event $\mathcal{A}$ *or* the event $\mathcal{B}$ occurs (or both), while their intersection $\mathcal{A} \cap \mathcal{B}$ logically represents the event that *both* $\mathcal{A}$ *and* $\mathcal{B}$ occur. For an event $\mathcal{A}$, we define its complement $\overline{\mathcal{A}} := \Omega \setminus \mathcal{A}$, which logically represents the event that $\mathcal{A}$ does *not* occur.

In working with events, one makes frequent use of the usual rules of Boolean logic. **De Morgan's law** says that for all events $\mathcal{A}$ and $\mathcal{B}$,

$$\overline{\mathcal{A} \cup \mathcal{B}} = \overline{\mathcal{A}} \cap \overline{\mathcal{B}} \text{ and } \overline{\mathcal{A} \cap \mathcal{B}} = \overline{\mathcal{A}} \cup \overline{\mathcal{B}}.$$

We also have the **Boolean distributive law**: for all events $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$,

$$\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C}) \text{ and } \mathcal{A} \cup (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C}).$$

Example 8.4. Continuing with Example 8.1, the event that the die has an odd value is $\mathcal{A} := \{1, 3, 5\}$, and we have $\mathsf{P}[\mathcal{A}] = 1/2$. The event that the die has a value greater than 2 is $\mathcal{B} := \{3, 4, 5, 6\}$, and $\mathsf{P}[\mathcal{B}] = 2/3$. The event that the die has a value that is at most 2 is $\overline{\mathcal{B}} = \{1, 2\}$, and $\mathsf{P}[\overline{\mathcal{B}}] = 1/3$. The event that the value of the die is odd *or* exceeds 2 is $\mathcal{A} \cup \mathcal{B} = \{1, 3, 4, 5, 6\}$, and $\mathsf{P}[\mathcal{A} \cup \mathcal{B}] = 5/6$. The event that the value of the die is odd *and* exceeds 2 is $\mathcal{A} \cap \mathcal{B} = \{3, 5\}$, and $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = 1/3$. □

Example 8.5. If P is the uniform distribution on a set Ω, and $\mathcal{A}$ is a subset of Ω, then $\mathsf{P}[\mathcal{A}] = |\mathcal{A}|/|\Omega|$. □

We next derive some elementary facts about probabilities of certain events, and relations among them. It is clear from the definitions that

$$\mathsf{P}[\emptyset] = 0 \text{ and } \mathsf{P}[\Omega] = 1,$$

and that for every event $\mathcal{A}$, we have

$$\mathsf{P}[\overline{\mathcal{A}}] = 1 - \mathsf{P}[\mathcal{A}].$$

Now consider events $\mathcal{A}$ and $\mathcal{B}$, and their union $\mathcal{A} \cup \mathcal{B}$. We have

$$\mathsf{P}[\mathcal{A} \cup \mathcal{B}] \leq \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}]; \tag{8.3}$$

moreover,

$$\mathsf{P}[\mathcal{A} \cup \mathcal{B}] = \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] \text{ if } \mathcal{A} \text{ and } \mathcal{B} \text{ are disjoint}, \tag{8.4}$$

that is, if $\mathcal{A} \cap \mathcal{B} = \emptyset$. The exact formula for arbitrary events $\mathcal{A}$ and $\mathcal{B}$ is:

$$\mathsf{P}[\mathcal{A} \cup \mathcal{B}] = \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] - \mathsf{P}[\mathcal{A} \cap \mathcal{B}]. \tag{8.5}$$

(8.3), (8.4), and (8.5) all follow from the observation that in the expression

$$\mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] = \sum_{\omega \in \mathcal{A}} \mathsf{P}(\omega) + \sum_{\omega \in \mathcal{B}} \mathsf{P}(\omega),$$

the value $\mathsf{P}(\omega)$ is counted once for each $\omega \in \mathcal{A} \cup \mathcal{B}$, except for those $\omega \in \mathcal{A} \cap \mathcal{B}$, for which $\mathsf{P}(\omega)$ is counted twice.

Example 8.6. Alice rolls two dice, and asks Bob to guess a value that appears on either of the two dice (without looking). Let us model this situation by considering the uniform distribution on $\Omega := \{1, \ldots, 6\} \times \{1, \ldots, 6\}$, where for each pair $(s, t) \in \Omega$, s represents the value of the first die, and t the value of the second.

For $k = 1, \ldots, 6$, let $\mathcal{A}_k$ be the event that the first die is k, and $\mathcal{B}_k$ the event that the second die is k. Let $\mathcal{C}_k = \mathcal{A}_k \cup \mathcal{B}_k$ be the event that k appears on either of the two dice. No matter what value k Bob chooses, the probability that this choice is correct is

$$\begin{aligned} \mathsf{P}[\mathcal{C}_k] &= \mathsf{P}[\mathcal{A}_k \cup \mathcal{B}_k] = \mathsf{P}[\mathcal{A}_k] + \mathsf{P}[\mathcal{B}_k] - \mathsf{P}[\mathcal{A}_k \cap \mathcal{B}_k] \\ &= 1/6 + 1/6 - 1/36 = 11/36, \end{aligned}$$

which is slightly less than the estimate $\mathsf{P}[\mathcal{A}_k] + \mathsf{P}[\mathcal{B}_k]$ obtained from (8.3). □

If $\{\mathcal{A}_i\}_{i \in I}$ is a family of events, indexed by some set I, we can naturally form the union $\bigcup_{i \in I} \mathcal{A}_i$ and intersection $\bigcap_{i \in I} \mathcal{A}_i$. If $I = \emptyset$, then by definition, the union is $\emptyset$, and by special convention, the intersection is the entire sample space Ω. Logically, the union represents the event that *some* $\mathcal{A}_i$ occurs, and the intersection represents the event that *all* the $\mathcal{A}_i$'s occur. De Morgan's law generalizes as follows:

$$\overline{\bigcup_{i \in I} \mathcal{A}_i} = \bigcap_{i \in I} \overline{\mathcal{A}}_i \text{ and } \overline{\bigcap_{i \in I} \mathcal{A}_i} = \bigcup_{i \in I} \overline{\mathcal{A}}_i,$$

and if $\mathcal{B}$ is an event, then the Boolean distributive law generalizes as follows:

$$\mathcal{B} \cap \Big(\bigcup_{i\in I} \mathcal{A}_i\Big) = \bigcup_{i\in I} (\mathcal{B} \cap \mathcal{A}_i) \text{ and } \mathcal{B} \cup \Big(\bigcap_{i\in I} \mathcal{A}_i\Big) = \bigcap_{i\in I} (\mathcal{B} \cup \mathcal{A}_i).$$

We now generalize (8.3), (8.4), and (8.5) from pairs of events to families of events. Let $\{\mathcal{A}_i\}_{i\in I}$ be a finite family of events (i.e., the index set I is finite). Using (8.3), it follows by induction on $|I|$ that

$$\mathsf{P}\Big[\bigcup_{i\in I} \mathcal{A}_i\Big] \leq \sum_{i\in I} \mathsf{P}[\mathcal{A}_i], \tag{8.6}$$

which is known as **Boole's inequality** (and sometimes called the **union bound**). Analogously, using (8.4), it follows by induction on $|I|$ that

$$\mathsf{P}\Big[\bigcup_{i\in I} \mathcal{A}_i\Big] = \sum_{i\in I} \mathsf{P}[\mathcal{A}_i] \text{ if } \{\mathcal{A}_i\}_{i\in I} \text{ is pairwise disjoint}, \tag{8.7}$$

that is, if $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for all $i, j \in I$ with $i \neq j$. We shall refer to (8.7) as **Boole's equality**. Both (8.6) and (8.7) are invaluable tools in calculating or estimating the probability of an event $\mathcal{A}$ by breaking $\mathcal{A}$ up into a family $\{\mathcal{A}_i\}_{i\in I}$ of smaller, and hopefully simpler, events, whose union is $\mathcal{A}$. We shall make frequent use of them.

The generalization of (8.5) is messier. Consider first the case of three events, $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$. We have

$$\begin{aligned}\mathsf{P}[\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}] = \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] + \mathsf{P}[\mathcal{C}] - \mathsf{P}[\mathcal{A} \cap \mathcal{B}] - \mathsf{P}[\mathcal{A} \cap \mathcal{C}] - \mathsf{P}[\mathcal{B} \cap \mathcal{C}] \\ + \mathsf{P}[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}].\end{aligned}$$

Thus, starting with the sum of the probabilities of the individual events, we have to subtract a "correction term" that consists of the sum of probabilities of all intersections of pairs of events; however, this is an "over-correction," and we have to correct the correction by adding back in the probability of the intersection of all three events. The general statement is as follows:

Theorem 8.1 (Inclusion/exclusion principle). *Let $\{\mathcal{A}_i\}_{i\in I}$ be a finite family of events. Then*

$$\mathsf{P}\Big[\bigcup_{i\in I} \mathcal{A}_i\Big] = \sum_{\emptyset\subsetneq J\subseteq I} (-1)^{|J|-1}\, \mathsf{P}\Big[\bigcap_{j\in J} \mathcal{A}_j\Big],$$

the sum being over all non-empty subsets J of I.

Proof. For $\omega \in \Omega$ and $\mathcal{B} \subseteq \Omega$, define $\delta_\omega[\mathcal{B}] := 1$ if $\omega \in \mathcal{B}$, and $\delta_\omega[\mathcal{B}] := 0$ if $\omega \notin \mathcal{B}$. As a function of ω, $\delta_\omega[\mathcal{B}]$ is simply the characteristic function of $\mathcal{B}$. One may easily verify that for all $\omega \in \Omega$, $\mathcal{B} \subseteq \Omega$, and $\mathcal{C} \subseteq \Omega$, we have $\delta_\omega[\overline{\mathcal{B}}] = 1 - \delta_\omega[\mathcal{B}]$ and $\delta_\omega[\mathcal{B} \cap \mathcal{C}] = \delta_\omega[\mathcal{B}]\delta_\omega[\mathcal{C}]$. It is also easily seen that for every $\mathcal{B} \subseteq \Omega$, we have $\sum_{\omega\in\Omega} \mathsf{P}(\omega)\delta_\omega[\mathcal{B}] = \mathsf{P}[\mathcal{B}]$.

Let $\mathcal{A} := \bigcup_{i\in I} \mathcal{A}_i$, and for $J \subseteq I$, let $\mathcal{A}_J := \bigcap_{j\in J} \mathcal{A}_j$. For every $\omega \in \Omega$,

$$1 - \delta_\omega[\mathcal{A}] = \delta_\omega[\overline{\mathcal{A}}] = \delta_\omega\Big[\bigcap_{i\in I} \overline{\mathcal{A}}_i\Big] = \prod_{i\in I} \delta_\omega[\overline{\mathcal{A}}_i] = \prod_{i\in I}(1 - \delta_\omega[\mathcal{A}_i])$$
$$= \sum_{J\subseteq I} (-1)^{|J|} \prod_{j\in J} \delta_\omega[\mathcal{A}_j] = \sum_{J\subseteq I} (-1)^{|J|} \delta_\omega[\mathcal{A}_J],$$

and so

$$\delta_\omega[\mathcal{A}] = \sum_{\emptyset\subsetneq J\subseteq I} (-1)^{|J|-1} \delta_\omega[\mathcal{A}_J]. \tag{8.8}$$

Multiplying (8.8) by $\mathsf{P}(\omega)$, and summing over all $\omega \in \Omega$, we have

$$\mathsf{P}[\mathcal{A}] = \sum_{\omega\in\Omega} \mathsf{P}(\omega)\delta_\omega[\mathcal{A}] = \sum_{\omega\in\Omega} \mathsf{P}(\omega) \sum_{\emptyset\subsetneq J\subseteq I} (-1)^{|J|-1} \delta_\omega[\mathcal{A}_J]$$
$$= \sum_{\emptyset\subsetneq J\subseteq I} (-1)^{|J|-1} \sum_{\omega\in\Omega} \mathsf{P}(\omega)\delta_\omega[\mathcal{A}_J] = \sum_{\emptyset\subsetneq J\subseteq I} (-1)^{|J|-1}\, \mathsf{P}[\mathcal{A}_J]. \quad \square$$

One can also state the inclusion/exclusion principle in a slightly different way, splitting the sum into terms with $|J| = 1$, $|J| = 2$, etc., as follows:

$$\mathsf{P}\Big[\bigcup_{i\in I} \mathcal{A}_i\Big] = \sum_{i\in I} \mathsf{P}[\mathcal{A}_i] \;+\; \sum_{k=2}^{|I|} (-1)^{k-1} \sum_{\substack{J\subseteq I \\ |J|=k}} \mathsf{P}\Big[\bigcap_{j\in J} \mathcal{A}_j\Big],$$

where the last sum in this formula is taken over all subsets J of I of size k.

We next consider a useful way to "glue together" probability distributions. Suppose one conducts two physically separate and unrelated random experiments, with each experiment modeled separately as a probability distribution. What we would like is a way to combine these distributions, obtaining a single probability distribution that models the two experiments as one grand experiment. This can be accomplished in general, as follows.

Let $\mathsf{P}_1 : \Omega_1 \to [0, 1]$ and $\mathsf{P}_2 : \Omega_2 \to [0, 1]$ be probability distributions. Their **product distribution** $\mathsf{P} := \mathsf{P}_1\,\mathsf{P}_2$ is defined as follows:

$$\begin{aligned} \mathsf{P} : \quad & \Omega_1 \times \Omega_2 \to [0, 1] \\ & (\omega_1, \omega_2) \mapsto \mathsf{P}_1(\omega_1)\,\mathsf{P}_2(\omega_2). \end{aligned}$$

It is easily verified that P is a probability distribution on the sample space $\Omega_1 \times \Omega_2$:

$$\sum_{\omega_1,\omega_2} \mathsf{P}(\omega_1, \omega_2) = \sum_{\omega_1,\omega_2} \mathsf{P}_1(\omega_1)\,\mathsf{P}_2(\omega_2) = \Big(\sum_{\omega_1} \mathsf{P}_1(\omega_1)\Big)\Big(\sum_{\omega_2} \mathsf{P}_2(\omega_2)\Big) = 1 \cdot 1 = 1.$$

More generally, if $\mathsf{P}_i : \Omega_i \to [0, 1]$, for $i = 1, \ldots, n$, are probability distributions,

then their product distribution is $\mathsf{P} := \mathsf{P}_1 \cdots \mathsf{P}_n$, where

$$\begin{aligned} \mathsf{P} : \quad & \Omega_1 \times \cdots \times \Omega_n \to [0,1] \\ & (\omega_1, \ldots, \omega_n) \mapsto \mathsf{P}_1(\omega_1) \cdots \mathsf{P}_n(\omega_n). \end{aligned}$$

If $\mathsf{P}_1 = \mathsf{P}_2 = \cdots = \mathsf{P}_n$, then we may write $\mathsf{P} = \mathsf{P}_1^n$. It is clear from the definitions that if each P_i is the uniform distribution on Ω_i, then P is the uniform distribution on $\Omega_1 \times \cdots \times \Omega_n$.

Example 8.7. We can view the probability distribution P in Example 8.6 as P_1^2, where P_1 is the uniform distribution on $\{1, \ldots, 6\}$. □

Example 8.8. Suppose we have a coin that comes up *heads* with some probability p, and *tails* with probability $q := 1 - p$. We toss the coin n times, and record the outcomes. We can model this as the product distribution $\mathsf{P} = \mathsf{P}_1^n$, where P_1 is the distribution of a Bernoulli trial (see Example 8.3) with success probability p, and where we identify *success* with *heads*, and *failure* with *tails*. The sample space Ω of P is the set of all 2^n tuples $\omega = (\omega_1, \ldots, \omega_n)$, where each ω_i is either *heads* or *tails*. If the tuple ω has k *heads* and $n - k$ *tails*, then $\mathsf{P}(\omega) = p^k q^{n-k}$, regardless of the positions of the *heads* and *tails* in the tuple.

For each $k = 0, \ldots, n$, let $\mathcal{A}_k$ be the event that our coin comes up *heads* exactly k times. As a set, $\mathcal{A}_k$ consists of all those tuples in the sample space with exactly k *heads*, and so

$$|\mathcal{A}_k| = \binom{n}{k},$$

from which it follows that

$$\mathsf{P}[\mathcal{A}_k] = \binom{n}{k} p^k q^{n-k}.$$

If our coin is a fair coin, so that $p = q = 1/2$, then P is the uniform distribution on Ω, and for each $k = 0, \ldots, n$, we have

$$\mathsf{P}[\mathcal{A}_k] = \binom{n}{k} 2^{-n}. \quad \square$$

Suppose $\mathsf{P} : \Omega \to [0,1]$ is a probability distribution. The **support** of P is defined to be the set $\{\omega \in \Omega : \mathsf{P}(\omega) \neq 0\}$. Now consider another probability distribution $\mathsf{P}' : \Omega' \to [0,1]$. Of course, these two distributions are equal if and only if $\Omega = \Omega'$ and $\mathsf{P}(\omega) = \mathsf{P}'(\omega)$ for all $\omega \in \Omega$. However, it is natural and convenient to have a more relaxed notion of equality. We shall say that P and P' are **essentially equal** if the restriction of P to its support is equal to the restriction of P' to its support. For example, if P is the probability distribution on $\{1, 2, 3, 4\}$ that assigns probability

1/3 to 1, 2, and 3, and probability 0 to 4, we may say that P is essentially the uniform distribution on $\{1, 2, 3\}$.

EXERCISE 8.1. Show that $\mathsf{P}[\mathcal{A} \cap \mathcal{B}]\,\mathsf{P}[\mathcal{A} \cup \mathcal{B}] \leq \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\mathcal{B}]$ for all events $\mathcal{A}, \mathcal{B}$.

EXERCISE 8.2. Suppose $\mathcal{A}, \mathcal{B}, \mathcal{C}$ are events such that $\mathcal{A} \cap \overline{\mathcal{C}} = \mathcal{B} \cap \overline{\mathcal{C}}$. Show that $|\mathsf{P}[\mathcal{A}] - \mathsf{P}[\mathcal{B}]| \leq \mathsf{P}[\mathcal{C}]$.

EXERCISE 8.3. Let m be a positive integer, and let $\alpha(m)$ be the probability that a number chosen at random from $\{1, \ldots, m\}$ is divisible by either 4, 5, or 6. Write down an exact formula for $\alpha(m)$, and also show that $\alpha(m) = 14/30 + O(1/m)$.

EXERCISE 8.4. This exercise asks you to generalize Boole's inequality (8.6), proving **Bonferroni's inequalities**. Let $\{\mathcal{A}_i\}_{i \in I}$ be a finite family of events, where $n := |I|$. For $m = 0, \ldots, n$, define

$$\alpha_m := \sum_{k=1}^{m} (-1)^{k-1} \sum_{\substack{J \subseteq I \\ |J| = k}} \mathsf{P}\Big[\bigcap_{j \in J} \mathcal{A}_j\Big].$$

Also, define

$$\alpha := \mathsf{P}\Big[\bigcup_{i \in I} \mathcal{A}_i\Big].$$

Show that $\alpha \leq \alpha_m$ if m is odd, and $\alpha \geq \alpha_m$ if m is even. Hint: use induction on n.

8.2 Conditional probability and independence

Let P be a probability distribution on a sample space Ω.

For a given event $\mathcal{B} \subseteq \Omega$ with $\mathsf{P}[\mathcal{B}] \neq 0$, and for $\omega \in \Omega$, let us define

$$\mathsf{P}(\omega \mid \mathcal{B}) := \begin{cases} \mathsf{P}(\omega)/\,\mathsf{P}[\mathcal{B}] & \text{if } \omega \in \mathcal{B}, \\ 0 & \text{otherwise.} \end{cases}$$

Viewing $\mathcal{B}$ as fixed, the function $\mathsf{P}(\cdot \mid \mathcal{B})$ is a new probability distribution on the sample space Ω, called the **conditional distribution (derived from P) given $\mathcal{B}$**.

Intuitively, $\mathsf{P}(\cdot \mid \mathcal{B})$ has the following interpretation. Suppose a random experiment produces an outcome according to the distribution P. Further, suppose we learn that the event $\mathcal{B}$ has occurred, but nothing else about the outcome. Then the distribution $\mathsf{P}(\cdot \mid \mathcal{B})$ assigns new probabilities to all possible outcomes, reflecting the partial knowledge that the event $\mathcal{B}$ has occurred.

For a given event $\mathcal{A} \subseteq \Omega$, its probability with respect to the conditional distribution given $\mathcal{B}$ is

$$\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = \sum_{\omega \in \mathcal{A}} \mathsf{P}(\omega \mid \mathcal{B}) = \frac{\mathsf{P}[\mathcal{A} \cap \mathcal{B}]}{\mathsf{P}[\mathcal{B}]}.$$

The value $\mathsf{P}[\mathcal{A} \mid \mathcal{B}]$ is called the **conditional probability of $\mathcal{A}$ given $\mathcal{B}$**. Again, the intuition is that this is the probability that the event $\mathcal{A}$ occurs, given the partial knowledge that the event $\mathcal{B}$ has occurred.

For events $\mathcal{A}$ and $\mathcal{B}$, if $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\mathcal{B}]$, then $\mathcal{A}$ and $\mathcal{B}$ are called **independent** events. If $\mathsf{P}[\mathcal{B}] \neq 0$, one easily sees that $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = \mathsf{P}[\mathcal{A}]$; intuitively, independence means that the partial knowledge that event $\mathcal{B}$ has occurred does not affect the likelihood that $\mathcal{A}$ occurs.

Example 8.9. Suppose P is the uniform distribution on Ω, and that $\mathcal{B} \subseteq \Omega$ with $\mathsf{P}[\mathcal{B}] \neq 0$. Then the conditional distribution given $\mathcal{B}$ is essentially the uniform distribution on $\mathcal{B}$. □

Example 8.10. Consider again Example 8.4, where $\mathcal{A}$ is the event that the value on the die is odd, and $\mathcal{B}$ is the event that the value of the die exceeds 2. Then as we calculated, $\mathsf{P}[\mathcal{A}] = 1/2$, $\mathsf{P}[\mathcal{B}] = 2/3$, and $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = 1/3$; thus, $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\mathcal{B}]$, and we conclude that $\mathcal{A}$ and $\mathcal{B}$ are independent. Indeed, $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = (1/3)/(2/3) = 1/2 = \mathsf{P}[\mathcal{A}]$; intuitively, given the partial knowledge that the value on the die exceeds 2, we know it is equally likely to be either 3, 4, 5, or 6, and so the conditional probability that it is odd is $1/2$.

However, consider the event C that the value on the die exceeds 3. We have $\mathsf{P}[C] = 1/2$ and $\mathsf{P}[\mathcal{A} \cap C] = 1/6 \neq 1/4$, from which we conclude that $\mathcal{A}$ and C are *not* independent. Indeed, $\mathsf{P}[\mathcal{A} \mid C] = (1/6)/(1/2) = 1/3 \neq \mathsf{P}[\mathcal{A}]$; intuitively, given the partial knowledge that the value on the die exceeds 3, we know it is equally likely to be either 4, 5, or 6, and so the conditional probability that it is odd is just $1/3$, and not $1/2$. □

Example 8.11. In Example 8.6, suppose that Alice tells Bob the sum of the two dice before Bob makes his guess. The following table is useful for visualizing the situation:

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	2	3	4	5	6

For example, suppose Alice tells Bob the sum is 4. Then what is Bob's best strategy

in this case? Let $\mathcal{D}_\ell$ be the event that the sum is ℓ, for $\ell = 2, \ldots, 12$, and consider the conditional distribution given $\mathcal{D}_4$. This conditional distribution is essentially the uniform distribution on the set $\{(1,3),(2,2),(3,1)\}$. The numbers 1 and 3 both appear in two pairs, while the number 2 appears in just one pair. Therefore,

$$\mathsf{P}[\mathcal{C}_1 \mid \mathcal{D}_4] = \mathsf{P}[\mathcal{C}_3 \mid \mathcal{D}_4] = 2/3,$$

while

$$\mathsf{P}[\mathcal{C}_2 \mid \mathcal{D}_4] = 1/3$$

and

$$\mathsf{P}[\mathcal{C}_4 \mid \mathcal{D}_4] = \mathsf{P}[\mathcal{C}_5 \mid \mathcal{D}_4] = \mathsf{P}[\mathcal{C}_6 \mid \mathcal{D}_4] = 0.$$

Thus, if the sum is 4, Bob's best strategy is to guess either 1 or 3, which will be correct with probability $2/3$.

Similarly, if the sum is 5, then we consider the conditional distribution given $\mathcal{D}_5$, which is essentially the uniform distribution on $\{(1,4),(2,3),(3,2),(4,1)\}$. In this case, Bob should choose one of the numbers $k = 1, \ldots, 4$, each of which will be correct with probability $\mathsf{P}[\mathcal{C}_k \mid \mathcal{D}_5] = 1/2$. □

Suppose $\{\mathcal{B}_i\}_{i \in I}$ is a finite, pairwise disjoint family of events, whose union is Ω. Now consider an arbitrary event $\mathcal{A}$. Since $\{\mathcal{A} \cap \mathcal{B}_i\}_{i \in I}$ is a pairwise disjoint family of events whose union is $\mathcal{A}$, Boole's equality (8.7) implies

$$\mathsf{P}[\mathcal{A}] = \sum_{i \in I} \mathsf{P}[\mathcal{A} \cap \mathcal{B}_i]. \tag{8.9}$$

Furthermore, if each $\mathcal{B}_i$ occurs with non-zero probability (so that, in particular, $\{\mathcal{B}_i\}_{i \in I}$ is a partition of Ω), then we have

$$\mathsf{P}[\mathcal{A}] = \sum_{i \in I} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]\,\mathsf{P}[\mathcal{B}_i]. \tag{8.10}$$

If, in addition, $\mathsf{P}[\mathcal{A}] \neq 0$, then for each $j \in I$, we have

$$\mathsf{P}[\mathcal{B}_j \mid \mathcal{A}] = \frac{\mathsf{P}[\mathcal{A} \cap \mathcal{B}_j]}{\mathsf{P}[\mathcal{A}]} = \frac{\mathsf{P}[\mathcal{A} \mid \mathcal{B}_j]\,\mathsf{P}[\mathcal{B}_j]}{\sum_{i \in I} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]\,\mathsf{P}[\mathcal{B}_i]}. \tag{8.11}$$

Equations (8.9) and (8.10) are sometimes called the **law of total probability**, while equation (8.11) is known as **Bayes' theorem**. Equation (8.10) (resp., (8.11)) is useful for computing or estimating $\mathsf{P}[\mathcal{A}]$ (resp., $\mathsf{P}[\mathcal{B}_j \mid \mathcal{A}]$) by conditioning on the events $\mathcal{B}_i$.

Example 8.12. Let us continue with Example 8.11, and compute Bob's overall probability of winning, assuming he follows an optimal strategy. If the sum is 2 or 12, clearly there is only one sensible choice for Bob to make, and it will certainly

be correct. If the sum is any other number ℓ, and there are N_ℓ pairs in the sample space that sum to that number, then there will always be a value that appears in exactly 2 of these N_ℓ pairs, and Bob should choose such a value (see the diagram in Example 8.11). Indeed, this is achieved by the simple rule of choosing the value 1 if $\ell \leq 7$, and the value 6 if $\ell > 7$. This is an optimal strategy for Bob, and if C is the event that Bob wins following this strategy, then by total probability (8.10), we have

$$\mathsf{P}[C] = \sum_{\ell=2}^{12} \mathsf{P}[C \mid D_\ell]\,\mathsf{P}[D_\ell].$$

Moreover,

$$\mathsf{P}[C \mid D_2]\,\mathsf{P}[D_2] = 1 \cdot \frac{1}{36} = \frac{1}{36}, \quad \mathsf{P}[C \mid D_{12}]\,\mathsf{P}[D_{12}] = 1 \cdot \frac{1}{36} = \frac{1}{36},$$

and for $\ell = 3, \ldots, 11$, we have

$$\mathsf{P}[C \mid D_\ell]\,\mathsf{P}[D_\ell] = \frac{2}{N_\ell} \cdot \frac{N_\ell}{36} = \frac{1}{18}.$$

Therefore,

$$\mathsf{P}[C] = \frac{1}{36} + \frac{1}{36} + \frac{9}{18} = \frac{10}{18}. \quad \square$$

Example 8.13. Suppose that the rate of incidence of disease X in the overall population is 1%. Also suppose that there is a test for disease X; however, the test is not perfect: it has a 5% false positive rate (i.e., 5% of healthy patients test positive for the disease), and a 2% false negative rate (i.e., 2% of sick patients test negative for the disease). A doctor gives the test to a patient and it comes out positive. How should the doctor advise his patient? In particular, what is the probability that the patient actually has disease X, given a positive test result?

Amazingly, many trained doctors will say the probability is 95%, since the test has a false positive rate of 5%. However, this conclusion is completely wrong.

Let A be the event that the test is positive and let B be the event that the patient has disease X. The relevant quantity that we need to estimate is $\mathsf{P}[B \mid A]$; that is, the probability that the patient has disease X, given a positive test result. We use Bayes' theorem to do this:

$$\mathsf{P}[B \mid A] = \frac{\mathsf{P}[A \mid B]\,\mathsf{P}[B]}{\mathsf{P}[A \mid B]\,\mathsf{P}[B] + \mathsf{P}[A \mid \overline{B}]\,\mathsf{P}[\overline{B}]} = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} \approx 0.17.$$

Thus, the chances that the patient has disease X given a positive test result are just 17%. The correct intuition here is that it is much more likely to get a false positive than it is to actually have the disease.

Of course, the real world is a bit more complicated than this example suggests:

the doctor may be giving the patient the test because other risk factors or symptoms may suggest that the patient is more likely to have the disease than a random member of the population, in which case the above analysis does not apply. □

Example 8.14. This example is based on the TV game show "Let's make a deal," which was popular in the 1970's. In this game, a contestant chooses one of three doors. Behind two doors is a "zonk," that is, something amusing but of little or no value, such as a goat, and behind one of the doors is a "grand prize," such as a car or vacation package. We may assume that the door behind which the grand prize is placed is chosen at random from among the three doors, with equal probability. After the contestant chooses a door, the host of the show, Monty Hall, always reveals a zonk behind one of the two doors not chosen by the contestant. The contestant is then given a choice: either stay with his initial choice of door, or switch to the other unopened door. After the contestant finalizes his decision on which door to choose, that door is opened and he wins whatever is behind it. The question is, which strategy is better for the contestant: to stay or to switch?

Let us evaluate the two strategies. If the contestant always stays with his initial selection, then it is clear that his probability of success is exactly $1/3$.

Now consider the strategy of always switching. Let $\mathcal{B}$ be the event that the contestant's initial choice was correct, and let $\mathcal{A}$ be the event that the contestant wins the grand prize. On the one hand, if the contestant's initial choice was correct, then switching will certainly lead to failure (in this case, Monty has two doors to choose from, but his choice does not affect the outcome). Thus, $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = 0$. On the other hand, suppose that the contestant's initial choice was incorrect, so that one of the zonks is behind the initially chosen door. Since Monty reveals the other zonk, switching will lead with certainty to success. Thus, $\mathsf{P}[\mathcal{A} \mid \overline{\mathcal{B}}] = 1$. Furthermore, it is clear that $\mathsf{P}[\mathcal{B}] = 1/3$. So using total probability (8.10), we compute

$$\mathsf{P}[\mathcal{A}] = \mathsf{P}[\mathcal{A} \mid \mathcal{B}]\,\mathsf{P}[\mathcal{B}] + \mathsf{P}[\mathcal{A} \mid \overline{\mathcal{B}}]\,\mathsf{P}[\overline{\mathcal{B}}] = 0 \cdot (1/3) + 1 \cdot (2/3) = 2/3.$$

Thus, the "stay" strategy has a success probability of $1/3$, while the "switch" strategy has a success probability of $2/3$. So it is better to switch than to stay.

Of course, real life is a bit more complicated. Monty did not always reveal a zonk and offer a choice to switch. Indeed, if Monty *only* revealed a zonk when the contestant had chosen the correct door, then switching would certainly be the wrong strategy. However, if Monty's choice itself was a random decision made independently of the contestant's initial choice, then switching is again the preferred strategy. □

We next generalize the notion of independence from pairs of events to families of events. Let $\{\mathcal{A}_i\}_{i \in I}$ be a finite family of events. For a given positive integer k,

we say that the family $\{\mathcal{A}_i\}_{i\in I}$ is **k-wise independent** if the following holds:

$$\mathsf{P}\Big[\bigcap_{j\in J}\mathcal{A}_j\Big] = \prod_{j\in J}\mathsf{P}[\mathcal{A}_j] \text{ for all } J \subseteq I \text{ with } |J| \leq k.$$

The family $\{\mathcal{A}_i\}_{i\in I}$ is called **pairwise independent** if it is 2-wise independent. Equivalently, pairwise independence means that for all $i, j \in I$ with $i \neq j$, we have $\mathsf{P}[\mathcal{A}_i \cap \mathcal{A}_j] = \mathsf{P}[\mathcal{A}_i]\,\mathsf{P}[\mathcal{A}_j]$, or put yet another way, that for all $i, j \in I$ with $i \neq j$, the events $\mathcal{A}_i$ and $\mathcal{A}_j$ are independent.

The family $\{\mathcal{A}_i\}_{i\in I}$ is called **mutually independent** if it is k-wise independent for all positive integers k. Equivalently, mutual independence means that

$$\mathsf{P}\Big[\bigcap_{j\in J}\mathcal{A}_j\Big] = \prod_{j\in J}\mathsf{P}[\mathcal{A}_j] \text{ for all } J \subseteq I.$$

If $n := |I| > 0$, mutual independence is equivalent to n-wise independence; moreover, if $0 < k \leq n$, then $\{\mathcal{A}_i\}_{i\in I}$ is k-wise independent if and only if $\{\mathcal{A}_j\}_{j\in J}$ is mutually independent for every $J \subseteq I$ with $|J| = k$.

In defining independence, the choice of the index set I plays no real role, and we can rename elements of I as convenient.

Example 8.15. Suppose we toss a fair coin three times, which we formally model using the uniform distribution on the set of all 8 possible outcomes of the three coin tosses: (*heads*, *heads*, *heads*), (*heads*, *heads*, *tails*), etc., as in Example 8.8. For $i = 1, 2, 3$, let $\mathcal{A}_i$ be the event that the ith toss comes up *heads*. Then $\{\mathcal{A}_i\}_{i=1}^3$ is a mutually independent family of events, where each individual $\mathcal{A}_i$ occurs with probability $1/2$.

Now let $\mathcal{B}_{12}$ be the event that the first and second tosses agree (i.e., both *heads* or both *tails*), let $\mathcal{B}_{13}$ be the event that the first and third tosses agree, and let $\mathcal{B}_{23}$ be the event that the second and third tosses agree. Then the family of events $\mathcal{B}_{12}, \mathcal{B}_{13}, \mathcal{B}_{23}$ is pairwise independent, but not mutually independent. Indeed, the probability that any given individual event occurs is $1/2$, and the probability that any given pair of events occurs is $1/4$; however, the probability that all three events occur is also $1/4$, since if any two events occur, then so does the third. □

We close this section with some simple facts about independence of events and their complements.

Theorem 8.2. *If $\mathcal{A}$ and $\mathcal{B}$ are independent events, then so are $\mathcal{A}$ and $\overline{\mathcal{B}}$.*

Proof. We have

$$\begin{aligned}\mathsf{P}[\mathcal{A}] &= \mathsf{P}[\mathcal{A} \cap \mathcal{B}] + \mathsf{P}[\mathcal{A} \cap \overline{\mathcal{B}}] \quad \text{(by total probability (8.9))} \\ &= \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\mathcal{B}] + \mathsf{P}[\mathcal{A} \cap \overline{\mathcal{B}}] \quad \text{(since } \mathcal{A} \text{ and } \mathcal{B} \text{ are independent).}\end{aligned}$$

Therefore,

$$\mathsf{P}[\mathcal{A} \cap \overline{\mathcal{B}}] = \mathsf{P}[\mathcal{A}] - \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\mathcal{B}] = \mathsf{P}[\mathcal{A}](1 - \mathsf{P}[\mathcal{B}]) = \mathsf{P}[\mathcal{A}]\,\mathsf{P}[\overline{\mathcal{B}}]. \quad \square$$

This theorem implies that

$$\begin{aligned} \mathcal{A} \text{ and } \mathcal{B} \text{ are independent} &\iff \mathcal{A} \text{ and } \overline{\mathcal{B}} \text{ are independent} \\ &\iff \overline{\mathcal{A}} \text{ and } \mathcal{B} \text{ " " } \\ &\iff \overline{\mathcal{A}} \text{ and } \overline{\mathcal{B}} \text{ " " } . \end{aligned}$$

The following theorem generalizes this result to families of events. It says that if a family of events is k-wise independent, then the family obtained by complementing any number of members of the given family is also k-wise independent.

Theorem 8.3. *Let $\{\mathcal{A}_i\}_{i\in I}$ be a finite, k-wise independent family of events. Let J be a subset of I, and for each $i \in I$, define $\mathcal{A}'_i := \mathcal{A}_i$ if $i \in J$, and $\mathcal{A}'_i := \overline{\mathcal{A}}_i$ if $i \notin J$. Then $\{\mathcal{A}'_i\}_{i\in I}$ is also k-wise independent.*

Proof. It suffices to prove the theorem for the case where $J = I \setminus \{d\}$, for an arbitrary $d \in I$: this allows us to complement any single member of the family that we wish, without affecting independence; by repeating the procedure, we can complement any number of them.

To this end, it will suffice to show the following: if $J \subseteq I$, $|J| < k$, $d \in I \setminus J$, and $\mathcal{A}_J := \bigcap_{j\in J} \mathcal{A}_j$, we have

$$\mathsf{P}[\overline{\mathcal{A}}_d \cap \mathcal{A}_J] = (1 - \mathsf{P}[\mathcal{A}_d]) \prod_{j\in J} \mathsf{P}[\mathcal{A}_j]. \tag{8.12}$$

Using total probability (8.9), along with the independence hypothesis (twice), we have

$$\begin{aligned} \prod_{j\in J} \mathsf{P}[\mathcal{A}_j] = \mathsf{P}[\mathcal{A}_J] &= \mathsf{P}[\mathcal{A}_d \cap \mathcal{A}_J] + \mathsf{P}[\overline{\mathcal{A}}_d \cap \mathcal{A}_J] \\ &= \mathsf{P}[\mathcal{A}_d] \cdot \prod_{j\in J} \mathsf{P}[\mathcal{A}_j] + \mathsf{P}[\overline{\mathcal{A}}_d \cap \mathcal{A}_J], \end{aligned}$$

from which (8.12) follows immediately. $\square$

EXERCISE 8.5. For events $\mathcal{A}_1, \ldots, \mathcal{A}_n$, define $\alpha_1 := \mathsf{P}[\mathcal{A}_1]$, and for $i = 2, \ldots, n$, define $\alpha_i := \mathsf{P}[\mathcal{A}_i \mid \mathcal{A}_1 \cap \cdots \cap \mathcal{A}_{i-1}]$ (assume that $\mathsf{P}[\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_{n-1}] \neq 0$). Show that $\mathsf{P}[\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_n] = \alpha_1 \cdots \alpha_n$.

EXERCISE 8.6. Let $\mathcal{B}$ be an event, and let $\{\mathcal{B}_i\}_{i\in I}$ be a finite, pairwise disjoint family of events whose union is $\mathcal{B}$. Generalizing the law of total probability

(equations (8.9) and (8.10)), show that for every event $\mathcal{A}$, we have $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = \sum_{i \in I} \mathsf{P}[\mathcal{A} \cap \mathcal{B}_i]$, and if $\mathsf{P}[\mathcal{B}] \neq 0$ and $I^* := \{i \in I : \mathsf{P}[\mathcal{B}_i] \neq 0\}$, then

$$\mathsf{P}[\mathcal{A} \mid \mathcal{B}]\,\mathsf{P}[\mathcal{B}] = \sum_{i \in I^*} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]\,\mathsf{P}[\mathcal{B}_i].$$

Also show that if $\mathsf{P}[\mathcal{A} \mid \mathcal{B}_i] \leq \alpha$ for each $i \in I^*$, then $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] \leq \alpha$.

EXERCISE 8.7. Let $\mathcal{B}$ be an event with $\mathsf{P}[\mathcal{B}] \neq 0$, and let $\{\mathcal{C}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union contains $\mathcal{B}$. Again, generalizing the law of total probability, show that for every event $\mathcal{A}$, if $I^* := \{i \in I : \mathsf{P}[\mathcal{B} \cap \mathcal{C}_i] \neq 0\}$, then we have

$$\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = \sum_{i \in I^*} \mathsf{P}[\mathcal{A} \mid \mathcal{B} \cap \mathcal{C}_i]\,\mathsf{P}[\mathcal{C}_i \mid \mathcal{B}].$$

EXERCISE 8.8. Three fair coins are tossed. Let $\mathcal{A}$ be the event that at least two coins are *heads*. Let $\mathcal{B}$ be the event that the number of *heads* is odd. Let $\mathcal{C}$ be the event that the third coin is *heads*. Are $\mathcal{A}$ and $\mathcal{B}$ independent? $\mathcal{A}$ and $\mathcal{C}$? $\mathcal{B}$ and $\mathcal{C}$?

EXERCISE 8.9. Consider again the situation in Example 8.11, but now suppose that Alice only tells Bob the value of the sum of the two dice modulo 6. Describe an optimal strategy for Bob, and calculate his overall probability of winning.

EXERCISE 8.10. Consider again the situation in Example 8.13, but now suppose that the patient is visiting the doctor because he has symptom Y. Furthermore, it is known that everyone who has disease X exhibits symptom Y, while 10% of the population overall exhibits symptom Y. Assuming that the accuracy of the test is not affected by the presence of symptom Y, how should the doctor advise his patient should the test come out positive?

EXERCISE 8.11. This exercise develops an alternative proof, based on probability theory, of Theorem 2.11. Let n be a positive integer and consider an experiment in which a number a is chosen uniformly at random from $\{0, \ldots, n-1\}$. If $n = p_1^{e_1} \cdots p_r^{e_r}$ is the prime factorization of n, let $\mathcal{A}_i$ be the event that a is divisible by p_i, for $i = 1, \ldots, r$.

(a) Show that $\varphi(n)/n = \mathsf{P}[\overline{\mathcal{A}}_1 \cap \cdots \cap \overline{\mathcal{A}}_r]$, where φ is Euler's phi function.

(b) Show that if $J \subseteq \{1, \ldots, r\}$, then

$$\mathsf{P}\Big[\bigcap_{j \in J} \mathcal{A}_j\Big] = 1 \Big/ \prod_{j \in J} p_j.$$

Conclude that $\{\mathcal{A}_i\}_{i=1}^r$ is mutually independent, and that $\mathsf{P}[\mathcal{A}_i] = 1/p_i$ for each $i = 1, \ldots, r$.

(c) Using part (b), deduce that

$$\mathsf{P}[\overline{\mathcal{A}}_1 \cap \cdots \cap \overline{\mathcal{A}}_r] = \prod_{i=1}^{r}(1 - 1/p_i).$$

(d) Combine parts (a) and (c) to derive the result of Theorem 2.11 that

$$\varphi(n) = n \prod_{i=1}^{r}(1 - 1/p_i).$$

8.3 Random variables

It is sometimes convenient to associate a real number, or other mathematical object, with each outcome of a random experiment. The notion of a random variable formalizes this idea.

Let P be a probability distribution on a sample space Ω. A **random variable** X is a function $X : \Omega \to S$, where S is some set, and we say that X **takes values in** S. We do not require that the values taken by X are real numbers, but if this is the case, we say that X is **real valued**. For $s \in S$, "$X = s$" denotes the event $\{\omega \in \Omega : X(\omega) = s\}$. It is immediate from this definition that

$$\mathsf{P}[X = s] = \sum_{\omega \in X^{-1}(\{s\})} \mathsf{P}(\omega).$$

More generally, for any predicate ϕ on S, we may write "$\phi(X)$" as shorthand for the event $\{\omega \in \Omega : \phi(X(\omega))\}$. When we speak of the **image** of X, we simply mean its image in the usual function-theoretic sense, that is, the set $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$. While a random variable is simply a function on the sample space, any discussion of its properties always takes place relative to a particular probability distribution, which may be implicit from context.

One can easily combine random variables to define new random variables. Suppose $X_1, \ldots, X_n$ are random variables, where $X_i : \Omega \to S_i$ for $i = 1, \ldots, n$. Then $(X_1, \ldots, X_n)$ denotes the random variable that maps $\omega \in \Omega$ to $(X_1(\omega), \ldots, X_n(\omega)) \in S_1 \times \cdots \times S_n$. If $f : S_1 \times \cdots \times S_n \to T$ is a function, then $f(X_1, \ldots, X_n)$ denotes the random variable that maps $\omega \in \Omega$ to $f(X_1(\omega), \ldots, X_n(\omega))$. If f is applied using a special notation, the same notation may be applied to denote the resulting random variable; for example, if X and Y are random variables taking values in a set S, and $\star$ is a binary operation on S, then $X \star Y$ denotes the random variable that maps $\omega \in \Omega$ to $X(\omega) \star Y(\omega) \in S$.

Let X be a random variable whose image is S. The variable X determines a probability distribution $\mathsf{P}_X : S \to [0, 1]$ on the set S, where $\mathsf{P}_X(s) := \mathsf{P}[X = s]$ for

each $s \in S$. We call P_X the **distribution of** X. If P_X is the uniform distribution on S, then we say that X is **uniformly distributed over** S.

Suppose X and Y are random variables that take values in a set S. If $\mathsf{P}[X = s] = \mathsf{P}[Y = s]$ for all $s \in S$, then the distributions of X and Y are essentially equal even if their images are not identical.

Example 8.16. Again suppose we roll two dice, and model this experiment as the uniform distribution on $\Omega := \{1, \ldots, 6\} \times \{1, \ldots, 6\}$. We can define the random variable X that takes the value of the first die, and the random variable Y that takes the value of the second; formally, X and Y are functions on Ω, where

$$X(s,t) := s \text{ and } Y(s,t) := t \text{ for } (s,t) \in \Omega.$$

For each value $s \in \{1, \ldots, 6\}$, the event $X = s$ is $\{(s,1), \ldots, (s,6)\}$, and so $\mathsf{P}[X = s] = 6/36 = 1/6$. Thus, X is uniformly distributed over $\{1, \ldots, 6\}$. Likewise, Y is uniformly distributed over $\{1, \ldots, 6\}$, and the random variable (X, Y) is uniformly distributed over Ω. We can also define the random variable $Z := X + Y$, which formally is the function on the sample space defined by

$$Z(s,t) := s + t \text{ for } (s,t) \in \Omega.$$

The image of Z is $\{2, \ldots, 12\}$, and its distribution is given by the following table:

u	2	3	4	5	6	7	8	9	10	11	12
$\mathsf{P}[Z = u]$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

. □

Example 8.17. If $\mathcal{A}$ is an event, we may define a random variable X as follows: $X := 1$ if the event $\mathcal{A}$ occurs, and $X := 0$ otherwise. The variable X is called the **indicator variable for** $\mathcal{A}$. Formally, X is the function that maps $\omega \in \mathcal{A}$ to 1, and $\omega \in \Omega \setminus \mathcal{A}$ to 0; that is, X is simply the characteristic function of $\mathcal{A}$. The distribution of X is that of a Bernoulli trial: $\mathsf{P}[X = 1] = \mathsf{P}[\mathcal{A}]$ and $\mathsf{P}[X = 0] = 1 - \mathsf{P}[\mathcal{A}]$.

It is not hard to see that $1 - X$ is the indicator variable for $\overline{\mathcal{A}}$. Now suppose $\mathcal{B}$ is another event, with indicator variable Y. Then it is also not hard to see that XY is the indicator variable for $\mathcal{A} \cap \mathcal{B}$, and that $X + Y - XY$ is the indicator variable for $\mathcal{A} \cup \mathcal{B}$; in particular, if $\mathcal{A} \cap \mathcal{B} = \emptyset$, then $X + Y$ is the indicator variable for $\mathcal{A} \cup \mathcal{B}$. □

Example 8.18. Consider again Example 8.8, where we have a coin that comes up *heads* with probability p, and *tails* with probability $q := 1 - p$, and we toss it n times. For each $i = 1, \ldots, n$, let $\mathcal{A}_i$ be the event that the ith toss comes up *heads*, and let X_i be the corresponding indicator variable. Let us also define $X := X_1 + \cdots + X_n$, which represents the total number of tosses that come up *heads*. The image of X is $\{0, \ldots, n\}$. By the calculations made in Example 8.8, for each $k = 0, \ldots, n$, we

have

$$\mathsf{P}[X = k] = \binom{n}{k} p^k q^{n-k}.$$

The distribution of the random variable X is called a **binomial distribution**. Such a distribution is parameterized by the success probability p of the underlying Bernoulli trial, and by the number of times n the trial is repeated. □

Uniform distributions are very nice, simple distributions. It is therefore good to have simple criteria that ensure that certain random variables have uniform distributions. The next theorem provides one such criterion. We need a definition: if S and T are finite sets, then we say that a given function $f : S \to T$ is a **regular function** if every element in the image of f has the same number of pre-images under f.

Theorem 8.4. *Suppose $f : S \to T$ is a surjective, regular function, and that X is a random variable that is uniformly distributed over S. Then $f(X)$ is uniformly distributed over T.*

Proof. The assumption that f is surjective and regular implies that for every $t \in T$, the set $S_t := f^{-1}(\{t\})$ has size $|S|/|T|$. So, for each $t \in T$, working directly from the definitions, we have

$$\begin{aligned}\mathsf{P}[f(X) = t] &= \sum_{\omega \in X^{-1}(S_t)} \mathsf{P}(\omega) = \sum_{s \in S_t} \sum_{\omega \in X^{-1}(\{s\})} \mathsf{P}(\omega) = \sum_{s \in S_t} \mathsf{P}[X = s] \\ &= \sum_{s \in S_t} 1/|S| = (|S|/|T|)/|S| = 1/|T|. \quad \square\end{aligned}$$

As a corollary, we have:

Theorem 8.5. *Suppose that $\rho : G \to G'$ is a surjective homomorphism of finite abelian groups G and G', and that X is a random variable that is uniformly distributed over G. Then $\rho(X)$ is uniformly distributed over G'.*

Proof. It suffices to show that ρ is regular. Recall that the kernel K of ρ is a subgroup of G, and that for every $g' \in G'$, the set $\rho^{-1}(\{g'\})$ is a coset of K (see Theorem 6.19); moreover, every coset of K has the same size (see Theorem 6.14). These facts imply that ρ is regular. □

Example 8.19. Let us continue with Example 8.16. Recall that for a given integer a, and positive integer n, $[a]_n \in \mathbb{Z}_n$ denotes the residue class of a modulo n. Let us define $X' := [X]_6$ and $Y' := [Y]_6$. It is not hard to see that both X' and Y' are uniformly distributed over $\mathbb{Z}_6$, while (X', Y') is uniformly distributed over $\mathbb{Z}_6 \times \mathbb{Z}_6$. Let us define $Z' := X' + Y'$ (where addition here is in $\mathbb{Z}_6$). We claim that Z' is

uniformly distributed over $\mathbb{Z}_6$. This follows immediately from the fact that the map that sends $(a, b) \in \mathbb{Z}_6 \times \mathbb{Z}_6$ to $a + b \in \mathbb{Z}_6$ is a surjective group homomorphism (see Example 6.45). Further, we claim that (X', Z') is uniformly distributed over $\mathbb{Z}_6 \times \mathbb{Z}_6$. This follows immediately from the fact that the map that sends $(a, b) \in \mathbb{Z}_6 \times \mathbb{Z}_6$ to $(a, a + b) \in \mathbb{Z}_6 \times \mathbb{Z}_6$ is a surjective group homomorphism (indeed, it is a group isomorphism). □

Let X be a random variable whose image is S. Let $\mathcal{B}$ be an event with $\mathsf{P}[\mathcal{B}] \neq 0$. The **conditional distribution of X given $\mathcal{B}$** is defined to be the distribution of X *relative to the conditional distribution* $\mathsf{P}(\cdot|\mathcal{B})$, that is, the distribution $\mathsf{P}_{X|\mathcal{B}} : S \to [0, 1]$ defined by $\mathsf{P}_{X|\mathcal{B}}(s) := \mathsf{P}[X = s \mid \mathcal{B}]$ for $s \in S$.

Suppose X and Y are random variables, with images S and T, respectively. We say X and Y are **independent** if for all $s \in S$ and all $t \in T$, the events $X = s$ and $Y = t$ are independent, which is to say,

$$\mathsf{P}[(X = s) \cap (Y = t)] = \mathsf{P}[X = s]\,\mathsf{P}[Y = t].$$

Equivalently, X and Y are independent if and only if the distribution of (X, Y) is essentially equal to the product of the distribution of X and the distribution of Y. As a special case, if X is uniformly distributed over S, and Y is uniformly distributed over T, then X and Y are independent if and only if (X, Y) is uniformly distributed over $S \times T$.

Independence can also be characterized in terms of conditional probabilities. From the definitions, it is immediate that X and Y are independent if and only if for all values t taken by Y with non-zero probability, we have

$$\mathsf{P}[X = s \mid Y = t] = \mathsf{P}[X = s]$$

for all $s \in S$; that is, the conditional distribution of X given $Y = t$ is the same as the distribution of X. From this point of view, an intuitive interpretation of independence is that information about the value of one random variable does not reveal any information about the value of the other.

Example 8.20. Let us continue with Examples 8.16 and 8.19. The random variables X and Y are independent: each is uniformly distributed over $\{1, \ldots, 6\}$, and (X, Y) is uniformly distributed over $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$. Let us calculate the conditional distribution of X given $Z = 4$. We have $\mathsf{P}[X = s \mid Z = 4] = 1/3$ for $s = 1, 2, 3$, and $\mathsf{P}[X = s \mid Z = 4] = 0$ for $s = 4, 5, 6$. Thus, the conditional distribution of X given $Z = 4$ is essentially the uniform distribution on $\{1, 2, 3\}$. Let us calculate the conditional distribution of Z given $X = 1$. We have $\mathsf{P}[Z = u \mid X = 1] = 1/6$ for $u = 2, \ldots, 7$, and $\mathsf{P}[Z = u \mid X = 1] = 0$ for $u = 8, \ldots, 12$. Thus, the conditional distribution of Z given $X = 1$ is essentially the uniform distribution on $\{2, \ldots, 7\}$. In particular, it is clear that X and Z are

not independent. The random variables X' and Y' are independent, as are X' and Z': each of X', Y', and Z' is uniformly distributed over $\mathbb{Z}_6$, and each of (X', Y') and (X', Z') is uniformly distributed over $\mathbb{Z}_6 \times \mathbb{Z}_6$. □

We now generalize the notion of independence to families of random variables. Let $\{X_i\}_{i \in I}$ be a finite family of random variables. Let us call a corresponding family of values $\{s_i\}_{i \in I}$ an **assignment** to $\{X_i\}_{i \in I}$ if s_i is in the image of X_i for each $i \in I$. For a given positive integer k, we say that the family $\{X_i\}_{i \in I}$ is **k-wise independent** if for every assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$, the family of events $\{X_i = s_i\}_{i \in I}$ is k-wise independent.

The notions of pairwise and mutual independence for random variables are defined following the same pattern that was used for events. The family $\{X_i\}_{i \in I}$ is called **pairwise independent** if it is 2-wise independent, which means that for all $i, j \in I$ with $i \neq j$, the variables X_i and X_j are independent. The family $\{X_i\}_{i \in I}$ is called **mutually independent** if it is k-wise independent for all positive integers k. Equivalently, and more explicitly, mutual independence means that for every assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$, we have

$$\mathsf{P}\Big[\bigcap_{j \in J} (X_j = s_j)\Big] = \prod_{j \in J} \mathsf{P}[X_j = s_j] \text{ for all } J \subseteq I. \tag{8.13}$$

If $n := |I| > 0$, mutual independence is equivalent to n-wise independence; moreover, if $0 < k \leq n$, then $\{X_i\}_{i \in I}$ is k-wise independent if and only if $\{X_j\}_{j \in J}$ is mutually independent for every $J \subseteq I$ with $|J| = k$.

Example 8.21. Returning again to Examples 8.16, 8.19, and 8.20, we see that the family of random variables X', Y', Z' is pairwise independent, but not mutually independent; for example,

$$\mathsf{P}\big[(X' = [0]_6) \cap (Y' = [0]_6) \cap (Z' = [0]_6)\big] = 1/6^2,$$

but

$$\mathsf{P}\big[X' = [0]_6\big] \cdot \mathsf{P}\big[Y' = [0]_6\big] \cdot \mathsf{P}\big[Z' = [0]_6\big] = 1/6^3. \ \square$$

Example 8.22. Suppose $\{\mathcal{A}_i\}_{i \in I}$ is a finite family of events. Let $\{X_i\}_{i \in I}$ be the corresponding family of indicator variables, so that for each $i \in I$, $X_i = 1$ if $\mathcal{A}_i$ occurs, and $X_i = 0$, otherwise. Theorem 8.3 immediately implies that for every positive integer k, $\{\mathcal{A}_i\}_{i \in I}$ is k-wise independent if and only if $\{X_i\}_{i \in I}$ is k-wise independent. □

Example 8.23. Consider again Example 8.15, where we toss a fair coin 3 times. For $i = 1, 2, 3$, let X_i be the indicator variable for the event $\mathcal{A}_i$ that the ith toss comes up *heads*. Then $\{X_i\}_{i=1}^3$ is a mutually independent family of random variables. Let Y_{12} be the indicator variable for the event $\mathcal{B}_{12}$ that tosses 1 and 2 agree;

similarly, let Y_{13} be the indicator variable for the event $\mathcal{B}_{13}$, and Y_{23} the indicator variable for $\mathcal{B}_{23}$. Then the family of random variables Y_{12}, Y_{13}, Y_{23} is pairwise independent, but not mutually independent. □

We next present a number of useful tools for establishing independence.

Theorem 8.6. *Let X be a random variable with image S, and Y be a random variable with image T. Further, suppose that $f : S \to [0,1]$ and $g : T \to [0,1]$ are functions such that*

$$\sum_{s \in S} f(s) = \sum_{t \in T} g(t) = 1, \tag{8.14}$$

and that for all $s \in S$ and $t \in T$, we have

$$\mathsf{P}[(X = s) \cap (Y = t)] = f(s)g(t). \tag{8.15}$$

Then X and Y are independent, the distribution of X is f, and the distribution of Y is g.

Proof. Since $\{Y = t\}_{t \in T}$ is a partition of the sample space, making use of total probability (8.9), along with (8.15) and (8.14), we see that for all $s \in S$, we have

$$\mathsf{P}[X = s] = \sum_{t \in T} \mathsf{P}[(X = s) \cap (Y = t)] = \sum_{t \in T} f(s)g(t) = f(s) \sum_{t \in T} g(t) = f(s).$$

Thus, the distribution of X is indeed f. Exchanging the roles of X and Y in the above argument, we see that the distribution of Y is g. Combining this with (8.15), we see that X and Y are independent. □

The generalization of Theorem 8.6 to families of random variables is a bit messy, but the basic idea is the same:

Theorem 8.7. *Let $\{X_i\}_{i \in I}$ be a finite family of random variables, where each X_i has image S_i. Also, let $\{f_i\}_{i \in I}$ be a family of functions, where for each $i \in I$, $f_i : S_i \to [0,1]$ and $\sum_{s_i \in S_i} f_i(s_i) = 1$. Further, suppose that*

$$\mathsf{P}\Big[\bigcap_{i \in I} (X_i = s_i)\Big] = \prod_{i \in I} f_i(s_i)$$

for each assignment $\{s_i\}_{i \in I}$ to $\{X_i\}_{i \in I}$. Then the family $\{X_i\}_{i \in I}$ is mutually independent, and for each $i \in I$, the distribution of X_i is f_i.

Proof. To prove the theorem, it suffices to prove the following statement: for every subset J of I, and every assignment $\{s_j\}_{j \in J}$ to $\{X_j\}_{j \in J}$, we have

$$\mathsf{P}\Big[\bigcap_{j \in J} (X_j = s_j)\Big] = \prod_{j \in J} f_j(s_j).$$

Moreover, it suffices to prove this statement for the case where $J = I \setminus \{d\}$, for an arbitrary $d \in I$: this allows us to eliminate any one variable from the family, without affecting the hypotheses, and by repeating this procedure, we can eliminate any number of variables.

Thus, let $d \in I$ be fixed, let $J := I \setminus \{d\}$, and let $\{s_j\}_{j\in J}$ be a fixed assignment to $\{X_j\}_{j\in J}$. Then, since $\{X_d = s_d\}_{s_d\in S_d}$ is a partition of the sample space, we have

$$\mathsf{P}\Big[\bigcap_{j\in J}(X_j = s_j)\Big] = \mathsf{P}\Big[\bigcup_{s_d\in S_d}\Big(\bigcap_{i\in I}(X_i = s_i)\Big)\Big] = \sum_{s_d\in S_d}\mathsf{P}\Big[\bigcap_{i\in I}(X_i = s_i)\Big]$$
$$= \sum_{s_d\in S_d}\prod_{i\in I} f_i(s_i) = \prod_{j\in J} f_j(s_j)\cdot\sum_{s_d\in S_d} f_d(s_d) = \prod_{j\in J} f_j(s_j). \quad \square$$

This theorem has several immediate consequences. First of all, mutual independence may be more simply characterized:

Theorem 8.8. *Let $\{X_i\}_{i\in I}$ be a finite family of random variables. Suppose that for every assignment $\{s_i\}_{i\in I}$ to $\{X_i\}_{i\in I}$, we have*

$$\mathsf{P}\Big[\bigcap_{i\in I}(X_i = s_i)\Big] = \prod_{i\in I}\mathsf{P}[X_i = s_i].$$

Then $\{X_i\}_{i\in I}$ is mutually independent.

Theorem 8.8 says that to check for mutual independence, we only have to consider the index set $J = I$ in (8.13). Put another way, it says that a family of random variables $\{X_i\}_{i=1}^n$ is mutually independent if and only if the distribution of $(X_1, \ldots, X_n)$ is essentially equal to the product of the distributions of the individual X_i's.

Based on the definition of mutual independence, and its characterization in Theorem 8.8, the following is also immediate:

Theorem 8.9. *Suppose $\{X_i\}_{i=1}^n$ is a family of random variables, and that m is an integer with $0 < m < n$. Then the following are equivalent:*

(i) $\{X_i\}_{i=1}^n$ is mutually independent;

(ii) $\{X_i\}_{i=1}^m$ is mutually independent, $\{X_i\}_{i=m+1}^n$ is mutually independent, and the two variables $(X_1, \ldots, X_m)$ and $(X_{m+1}, \ldots, X_n)$ are independent.

The following is also an immediate consequence of Theorem 8.7 (it also follows easily from Theorem 8.4).

Theorem 8.10. *Suppose that $X_1, \ldots, X_n$ are random variables, and that $S_1, \ldots, S_n$ are finite sets. Then the following are equivalent:*

(i) $(X_1, \ldots, X_n)$ is uniformly distributed over $S_1 \times \cdots \times S_n$;

(ii) $\{X_i\}_{i=1}^n$ *is mutually independent, with each* X_i *uniformly distributed over* S_i.

Another immediate consequence of Theorem 8.7 is the following:

Theorem 8.11. *Suppose* P *is the product distribution* $\mathsf{P}_1 \cdots \mathsf{P}_n$, *where each* P_i *is a probability distribution on a sample space* Ω_i, *so that the sample space of* P *is* $\Omega = \Omega_1 \times \cdots \times \Omega_n$. *For each* $i = 1, \ldots, n$, *let* X_i *be the random variable that projects on the* i*th coordinate, so that* $X_i(\omega_1, \ldots, \omega_n) = \omega_i$. *Then* $\{X_i\}_{i=1}^n$ *is mutually independent, and for each* $i = 1, \ldots, n$, *the distribution of* X_i *is* P_i.

Theorem 8.11 is often used to synthesize independent random variables "out of thin air," by taking the product of appropriate probability distributions. Other arguments may then be used to prove the independence of variables derived from these.

Example 8.24. Theorem 8.11 immediately implies that in Example 8.18, the family of indicator variables $\{X_i\}_{i=1}^n$ is mutually independent. □

The following theorem gives us yet another way to establish independence.

Theorem 8.12. *Suppose* $\{X_i\}_{i=1}^n$ *is a mutually independent family of random variables. Further, suppose that for* $i = 1, \ldots, n$, $Y_i := g_i(X_i)$ *for some function* g_i. *Then* $\{Y_i\}_{i=1}^n$ *is mutually independent.*

Proof. It suffices to prove the theorem for $n = 2$. The general case follows easily by induction, using Theorem 8.9. For $i = 1, 2$, let t_i be any value in the image of Y_i, and let $S_i' := g_i^{-1}(\{t_i\})$. We have

$$\begin{aligned}
\mathsf{P}[(Y_1 = t_1) \cap (Y_2 = t_2)] &= \mathsf{P}\Big[\Big(\bigcup_{s_1 \in S_1'} (X_1 = s_1)\Big) \cap \Big(\bigcup_{s_2 \in S_2'} (X_2 = s_2)\Big)\Big] \\
&= \mathsf{P}\Big[\bigcup_{s_1 \in S_1'} \bigcup_{s_2 \in S_2'} \Big((X_1 = s_1) \cap (X_2 = s_2)\Big)\Big] \\
&= \sum_{s_1 \in S_1'} \sum_{s_2 \in S_2'} \mathsf{P}[(X_1 = s_1) \cap (X_2 = s_2)] \\
&= \sum_{s_1 \in S_1'} \sum_{s_2 \in S_2'} \mathsf{P}[X_1 = s_1]\,\mathsf{P}[X_2 = s_2] \\
&= \Big(\sum_{s_1 \in S_1'} \mathsf{P}[X_1 = s_1]\Big)\Big(\sum_{s_2 \in S_2'} \mathsf{P}[X_2 = s_2]\Big) \\
&= \mathsf{P}\Big[\bigcup_{s_1 \in S_1'} (X_1 = s_1)\Big]\,\mathsf{P}\Big[\bigcup_{s_2 \in S_2'} (X_2 = s_2)\Big] = \mathsf{P}[Y_1 = t_1]\,\mathsf{P}[Y_2 = t_2]. \quad \square
\end{aligned}$$

As a special case of the above theorem, if each g_i is the characteristic function for some subset S_i' of the image of X_i, then $X_1 \in S_1', \ldots, X_n \in S_n'$ form a mutually independent family of events.

The next theorem is quite handy in proving the independence of random variables in a variety of algebraic settings.

Theorem 8.13. *Suppose that G is a finite abelian group, and that W is a random variable uniformly distributed over G. Let Z be another random variable, taking values in some finite set U, and suppose that W and Z are independent. Let $\sigma : U \to G$ be some function, and define $Y := W + \sigma(Z)$. Then Y is uniformly distributed over G, and Y and Z are independent.*

Proof. Consider any fixed values $t \in G$ and $u \in U$. Evidently, the events $(Y = t) \cap (Z = u)$ and $(W = t - \sigma(u)) \cap (Z = u)$ are the same, and therefore, because W and Z are independent, we have

$$\mathsf{P}[(Y = t) \cap (Z = u)] = \mathsf{P}[W = t - \sigma(u)]\,\mathsf{P}[Z = u] = \frac{1}{|G|}\,\mathsf{P}[Z = u]. \tag{8.16}$$

Since this holds for every $u \in U$, making use of total probability (8.9), we have

$$\mathsf{P}[Y = t] = \sum_{u \in U} \mathsf{P}[(Y = t) \cap (Z = u)] = \frac{1}{|G|} \sum_{u \in U} \mathsf{P}[Z = u] = \frac{1}{|G|}.$$

Thus, Y is uniformly distributed over G, and by (8.16), Y and Z are independent. (This conclusion could also have been deduced directly from (8.16) using Theorem 8.6—we have repeated the argument here.) □

Note that in the above theorem, we make no assumption about the distribution of Z, or any properties of the function σ.

Example 8.25. Theorem 8.13 may be used to justify the security of the **one-time pad** encryption scheme. Here, the variable W represents a random, secret key—the "pad"—that is shared between Alice and Bob; U represents a space of possible messages; Z represents a "message source," from which Alice draws her message according to some distribution; finally, the function $\sigma : U \to G$ represents some invertible "encoding transformation" that maps messages into group elements.

To encrypt a message drawn from the message source, Alice encodes the message as a group element, and then adds the pad. The variable $Y := W + \sigma(Z)$ represents the resulting ciphertext. Since $Z = \sigma^{-1}(Y - W)$, when Bob receives the ciphertext, he decrypts it by subtracting the pad, and converting the resulting group element back into a message. Because the message source Z and ciphertext Y are independent, an eavesdropping adversary who learns the value of Y does not learn

anything about Alice's message: for any particular ciphertext t, the conditional distribution of Z given $Y = t$ is the same as the distribution of Z.

The term "one time" comes from the fact that a given encryption key should be used only once; otherwise, security may be compromised. Indeed, suppose the key is used a second time, encrypting a message drawn from a second source Z'. The second ciphertext is represented by the random variable $Y' := W + \sigma(Z')$. In general, the random variables (Z, Z') and (Y, Y') will not be independent, since $Y - Y' = \sigma(Z) - \sigma(Z')$. To illustrate this more concretely, suppose Z is uniformly distributed over a set of 1000 messages, Z' is uniformly distributed over a set of two messages, say, $\{u'_1, u'_2\}$, and that Z and Z' are independent. Now, without any further information about Z, an adversary would have at best a 1-in-a-1000 chance of guessing its value. However, if he sees that $Y = t$ and $Y' = t'$, for particular values $t, t' \in G$, then he has a 1-in-2-chance, since the value of Z is equally likely to be one of just two messages, namely, $u_1 := \sigma^{-1}(t - t' + \sigma(u'_1))$ and $u_2 := \sigma^{-1}(t - t' + \sigma(u'_2))$; more formally, the conditional distribution of Z given $(Y = t) \cap (Y' = t')$ is essentially the uniform distribution on $\{u_1, u_2\}$.

In practice, it is convenient to define the group G to be the group of all bit strings of some fixed length, with bit-wise exclusive-or as the group operation. The encoding function σ simply "serializes" a message as a bit string. □

Example 8.26. Theorem 8.13 may also be used to justify a very simple type of **secret sharing**. A colorful, if militaristic, motivating scenario is the following. To launch a nuclear missile, two officers who carry special keys must insert their keys simultaneously into the "authorization device" (at least, that is how it works in Hollywood). In the digital version of this scenario, an authorization device contains a secret, digital "launch code," and each officer holds a digital "share" of this code, so that (i) individually, each share reveals no information about the launch code, but (ii) collectively, the two shares may be combined in a simple way to derive the launch code. Thus, to launch the missile, both officers must input their shares into the authorization device; hardware in the authorization device combines the two shares, and compares the resulting code against the launch code it stores—if they match, the missile flies.

In the language of Theorem 8.13, the launch code is represented by the random variable Z, and the two shares by W and $Y := W + \sigma(Z)$, where (as in the previous example) $\sigma : U \to G$ is some simple, invertible encoding function. Because W and Z are independent, information about the share W leaks no information about the launch code Z; likewise, since Y and Z are independent, information about Y leaks no information about Z. However, by combining both shares, the launch code is easily constructed as $Z = \sigma^{-1}(Y - W)$. □

Example 8.27. Let k be a positive integer. This example shows how we can take a mutually independent family of k random variables, and, from it, construct a much larger, k-wise independent family of random variables.

Let p be a prime, with $p \geq k$. Let $\{H_i\}_{i=0}^{k-1}$ be a mutually independent family of random variables, each of which is uniformly distributed over $\mathbb{Z}_p$. Let us set $H := (H_0, \ldots, H_{k-1})$, which, by assumption, is uniformly distributed over $\mathbb{Z}_p^{\times k}$. For each $s \in \mathbb{Z}_p$, we define the function $\rho_s : \mathbb{Z}_p^{\times k} \to \mathbb{Z}_p$ as follows: for $r = (r_0, \ldots, r_{k-1}) \in \mathbb{Z}_p^{\times k}$, $\rho_s(r) := \sum_{i=0}^{k-1} r_i s^i$; that is, $\rho_s(r)$ is the value obtained by evaluating the polynomial $r_0 + r_1 X + \cdots + r_{k-1} X^{k-1} \in \mathbb{Z}_p[X]$ at the point s.

Each $s \in \mathbb{Z}_p$ defines a random variable $\rho_s(H) = H_0 + H_1 s + \cdots + H_{k-1} s^{k-1}$. We claim that the family of random variables $\{\rho_s(H)\}_{s \in \mathbb{Z}_p}$ is k-wise independent, with each individual $\rho_s(H)$ uniformly distributed over $\mathbb{Z}_p$. By Theorem 8.10, it suffices to show the following: for all distinct points $s_1, \ldots, s_k \in \mathbb{Z}_p$, the random variable $W := (\rho_{s_1}(H), \ldots, \rho_{s_k}(H))$ is uniformly distributed over $\mathbb{Z}_p^{\times k}$. So let $s_1, \ldots, s_k$ be fixed, distinct elements of $\mathbb{Z}_p$, and define the function

$$\begin{aligned} \rho : \quad & \mathbb{Z}_p^{\times k} \to \mathbb{Z}_p^{\times k} \\ & r \mapsto (\rho_{s_1}(r), \ldots, \rho_{s_k}(r)). \end{aligned} \tag{8.17}$$

Thus, $W = \rho(H)$, and by Lagrange interpolation (Theorem 7.15), the function ρ is a bijection; moreover, since H is uniformly distributed over $\mathbb{Z}_p^{\times k}$, so is W.

Of course, the field $\mathbb{Z}_p$ may be replaced by an arbitrary finite field. □

Example 8.28. Consider again the secret sharing scenario of Example 8.26. Suppose at the critical moment, one of the officers is missing in action. The military planners would perhaps like a more flexible secret sharing scheme; for example, perhaps shares of the launch code should be distributed to three officers, in such a way that no single officer can authorize a launch, but any two can. More generally, for positive integers k and ℓ, with $\ell \geq k + 1$, the scheme should distribute shares among ℓ officers, so that no coalition of k (or fewer) officers can authorize a launch, yet any coalition of $k + 1$ officers can. Using the construction of the previous example, this is easily achieved, as follows.

Let us model the secret launch code as a random variable Z, taking values in a finite set U. Assume that p is prime, with $p \geq \ell$, and that $\sigma : U \to \mathbb{Z}_p$ is a simple, invertible encoding function. To construct the shares, we make use of random variables $H_0, \ldots, H_{k-1}$, where each H_i is uniformly distributed over $\mathbb{Z}_p$, and the family of random variables $H_0, \ldots, H_{k-1}, Z$ is mutually independent. For each $s \in \mathbb{Z}_p$, we define the random variable

$$Y_s := H_0 + H_1 s + \cdots + H_{k-1} s^{k-1} + \sigma(Z) s^k.$$

We can pick any subset $S \subseteq \mathbb{Z}_p$ of size ℓ that we wish, so that for each $s \in S$, an officer gets the secret share Y_s (along with the public value s).

First, we show how any coalition of $k+1$ officers can reconstruct the launch code from their collection of shares, say, $Y_{s_1}, \ldots, Y_{s_{k+1}}$. This is easily done by means of the Lagrange interpolation formula (again, Theorem 7.15). Indeed, we only need to recover the high-order coefficient, $\sigma(Z)$, which we can obtain via the formula

$$\sigma(Z) = \sum_{i=1}^{k+1} \frac{Y_{s_i}}{\prod_{j \neq i}(s_i - s_j)}.$$

Second, we show that no coalition of k officers learn anything about the launch code, even if they pool their shares. Formally, this means that if $s_1, \ldots, s_k$ are fixed, distinct points, then $Y_{s_1}, \ldots, Y_{s_k}, Z$ form a mutually independent family of random variables. This is easily seen, as follows. Define $H := (H_0, \ldots, H_{k-1})$, and $W := \rho(H)$, where $\rho : \mathbb{Z}_p^{\times k} \to \mathbb{Z}_p^{\times k}$ is as defined in (8.17), and set $Y := (Y_{s_1}, \ldots, Y_{s_k})$. Now, by hypothesis, H and Z are independent, and H is uniformly distributed over $\mathbb{Z}_p^{\times k}$. As we noted in Example 8.27, ρ is a bijection, and hence, W is uniformly distributed over $\mathbb{Z}_p^{\times k}$; moreover (by Theorem 8.12), W and Z are independent. Observe that $Y = W + \sigma'(Z)$, where σ' maps $u \in U$ to $(\sigma(u)s_1^k, \ldots, \sigma(u)s_k^k) \in \mathbb{Z}_p^{\times k}$, and so applying Theorem 8.13 (with the group $\mathbb{Z}_p^{\times k}$, the random variables W and Z, and the function σ'), we see that Y and Z are independent, where Y is uniformly distributed over $\mathbb{Z}_p^{\times k}$. From this, it follows (using Theorems 8.9 and 8.10) that the family of random variables $Y_{s_1}, \ldots, Y_{s_k}, Z$ is mutually independent, with each Y_{s_i} uniformly distributed over $\mathbb{Z}_p$.

Finally, we note that when $k = 1$, $\ell = 2$, and $S = \{0, 1\}$, this construction degenerates to the construction in Example 8.26 (with the additive group $\mathbb{Z}_p$). □

EXERCISE 8.12. Suppose X and X' are random variables that take values in a set S and that have *essentially* the same distribution. Show that if $f : S \to T$ is a function, then $f(X)$ and $f(X')$ have essentially the same distribution.

EXERCISE 8.13. Let $\{X_i\}_{i=1}^n$ be a family of random variables, and let S_i be the image of X_i for $i = 1, \ldots, n$. Show that $\{X_i\}_{i=1}^n$ is mutually independent if and only if for each $i = 2, \ldots, n$, and for all $s_1 \in S_1, \ldots, s_i \in S_i$, we have

$$\mathsf{P}[X_i = s_i \mid (X_1 = s_1) \cap \cdots \cap (X_{i-1} = s_{i-1})] = \mathsf{P}[X_i = s_i].$$

EXERCISE 8.14. Suppose that $\rho : G \to G'$ is a surjective group homomorphism, where G and G' are finite abelian groups. Show that if $g', h' \in G'$, and X and Y are independent random variables, where X is uniformly distributed over $\rho^{-1}(\{g'\})$, and Y takes values in $\rho^{-1}(\{h'\})$, then $X+Y$ is uniformly distributed over $\rho^{-1}(\{g' + h'\})$.

EXERCISE 8.15. Suppose X and Y are random variables, where X takes values in S, and Y takes values in T. Further suppose that Y' is uniformly distributed over T, and that (X, Y) and Y' are independent. Let ϕ be a predicate on $S \times T$. Show that $\mathsf{P}[\phi(X, Y) \cap (Y = Y')] = \mathsf{P}[\phi(X, Y)]/|T|$.

EXERCISE 8.16. Let X and Y be independent random variables, where X is uniformly distributed over a set S, and Y is uniformly distributed over a set $T \subseteq S$. Define a third random variable Z as follows: if $X \in T$, then $Z := X$; otherwise, $Z := Y$. Show that Z is uniformly distributed over T.

EXERCISE 8.17. Let n be a positive integer, and let X be a random variable, uniformly distributed over $\{0, \ldots, n-1\}$. For each positive divisor d of n, let us define the random variable $X_d := X \bmod d$. Show that:

(a) if d is a divisor of n, then the variable X_d is uniformly distributed over $\{0, \ldots, d-1\}$;

(b) if $d_1, \ldots, d_k$ are divisors of n, then $\{X_{d_i}\}_{i=1}^k$ is mutually independent if and only if $\{d_i\}_{i=1}^k$ is pairwise relatively prime.

EXERCISE 8.18. Suppose X and Y are random variables, each uniformly distributed over $\mathbb{Z}_2$, but not necessarily independent. Show that the distribution of (X, Y) is the same as the distribution of $(X + 1, Y + 1)$.

EXERCISE 8.19. Let $I := \{1, \ldots, n\}$, where $n \geq 2$, let $B := \{0, 1\}$, and let G be a finite abelian group, with $|G| > 1$. Suppose that $\{X_{ib}\}_{(i,b) \in I \times B}$ is a mutually independent family of random variables, each uniformly distributed over G. For each $\beta = (b_1, \ldots, b_n) \in B^{\times n}$, let us define the random variable $Y_\beta := X_{1b_1} + \cdots + X_{nb_n}$. Show that each Y_β is uniformly distributed over G, and that $\{Y_\beta\}_{\beta \in B^{\times n}}$ is 3-wise independent, but not 4-wise independent.

8.4 Expectation and variance

Let P be a probability distribution on a sample space Ω. If X is a real-valued random variable, then its **expected value**, or **expectation**, is

$$\mathsf{E}[X] := \sum_{\omega \in \Omega} X(\omega)\, \mathsf{P}(\omega). \tag{8.18}$$

If S is the image of X, and if for each $s \in S$ we group together the terms in (8.18) with $X(\omega) = s$, then we see that

$$\mathsf{E}[X] = \sum_{s \in S} s\, \mathsf{P}[X = s]. \tag{8.19}$$

From (8.19), it is clear that $\mathsf{E}[X]$ depends only on the distribution of X: if X' is another random variable with the same (or essentially the same) distribution as X, then $\mathsf{E}[X] = \mathsf{E}[X']$.

More generally, suppose X is an arbitrary random variable (not necessarily real valued) whose image is S, and f is a real-valued function on S. Then again, if for each $s \in S$ we group together the terms in (8.18) with $X(\omega) = s$, we see that

$$\mathsf{E}[f(X)] = \sum_{s \in S} f(s)\,\mathsf{P}[X = s]. \tag{8.20}$$

We make a few trivial observations about expectation, which the reader may easily verify. First, if X is equal to a constant c (i.e., $X(\omega) = c$ for every $\omega \in \Omega$), then $\mathsf{E}[X] = \mathsf{E}[c] = c$. Second, if X and Y are random variables such that $X \geq Y$ (i.e., $X(\omega) \geq Y(\omega)$ for every $\omega \in \Omega$), then $\mathsf{E}[X] \geq \mathsf{E}[Y]$. Similarly, if $X > Y$, then $\mathsf{E}[X] > \mathsf{E}[Y]$.

In calculating expectations, one rarely makes direct use of (8.18), (8.19), or (8.20), except in rather trivial situations. The next two theorems develop tools that are often quite effective in calculating expectations.

Theorem 8.14 (Linearity of expectation). *If X and Y are real-valued random variables, and a is a real number, then*

$$\mathsf{E}[X + Y] = \mathsf{E}[X] + \mathsf{E}[Y] \quad \text{and} \quad \mathsf{E}[aX] = a\,\mathsf{E}[X].$$

Proof. It is easiest to prove this using the defining equation (8.18) for expectation. For $\omega \in \Omega$, the value of the random variable $X+Y$ at ω is by definition $X(\omega)+Y(\omega)$, and so we have

$$\begin{aligned} \mathsf{E}[X + Y] &= \sum_{\omega} (X(\omega) + Y(\omega))\,\mathsf{P}(\omega) \\ &= \sum_{\omega} X(\omega)\,\mathsf{P}(\omega) + \sum_{\omega} Y(\omega)\,\mathsf{P}(\omega) \\ &= \mathsf{E}[X] + \mathsf{E}[Y]. \end{aligned}$$

For the second part of the theorem, by a similar calculation, we have

$$\mathsf{E}[aX] = \sum_{\omega} (aX(\omega))\,\mathsf{P}(\omega) = a \sum_{\omega} X(\omega)\,\mathsf{P}(\omega) = a\,\mathsf{E}[X]. \quad \square$$

More generally, the above theorem implies (using a simple induction argument) that if $\{X_i\}_{i \in I}$ is a finite family of real-valued random variables, then we have

$$\mathsf{E}\Big[\sum_{i \in I} X_i\Big] = \sum_{i \in I} \mathsf{E}[X_i]. \tag{8.21}$$

So we see that expectation is linear; however, expectation is not in general multiplicative, except in the case of independent random variables:

Theorem 8.15. *If X and Y are independent, real-valued random variables, then* $\mathsf{E}[XY] = \mathsf{E}[X]\,\mathsf{E}[Y]$.

Proof. It is easiest to prove this using (8.20), with the function $f(s,t) := st$ applied to the random variable (X, Y). We have

$$\begin{aligned}\mathsf{E}[XY] &= \sum_{s,t} st\,\mathsf{P}[(X = s) \cap (Y = t)] \\ &= \sum_{s,t} st\,\mathsf{P}[X = s]\,\mathsf{P}[Y = t] \\ &= \Big(\sum_{s} s\,\mathsf{P}[X = s]\Big)\Big(\sum_{t} t\,\mathsf{P}[Y = t]\Big) \\ &= \mathsf{E}[X]\,\mathsf{E}[Y]. \quad \square\end{aligned}$$

More generally, the above theorem implies (using a simple induction argument) that if $\{X_i\}_{i \in I}$ is a finite, mutually independent family of real-valued random variables, then

$$\mathsf{E}\Big[\prod_{i \in I} X_i\Big] = \prod_{i \in I} \mathsf{E}[X_i]. \tag{8.22}$$

The following simple facts are also sometimes quite useful in calculating expectations:

Theorem 8.16. *Let X be a $0/1$-valued random variable. Then* $\mathsf{E}[X] = \mathsf{P}[X = 1]$.

Proof. $\mathsf{E}[X] = 0 \cdot \mathsf{P}[X = 0] + 1 \cdot \mathsf{P}[X = 1] = \mathsf{P}[X = 1]$. $\square$

Theorem 8.17. *If X is a random variable that takes only non-negative integer values, then*

$$\mathsf{E}[X] = \sum_{i \geq 1} \mathsf{P}[X \geq i].$$

Note that since X has a finite image, the sum appearing above is finite.

Proof. Suppose that the image of X is contained in $\{0, \ldots, n\}$, and for $i = 1, \ldots, n$, let X_i be the indicator variable for the event $X \geq i$. Then $X = X_1 + \cdots + X_n$, and by linearity of expectation and Theorem 8.16, we have

$$\mathsf{E}[X] = \sum_{i=1}^{n} \mathsf{E}[X_i] = \sum_{i=1}^{n} \mathsf{P}[X \geq i]. \quad \square$$

Let X be a real-valued random variable with $\mu := \mathsf{E}[X]$. The **variance** of X is $\mathsf{Var}[X] := \mathsf{E}[(X - \mu)^2]$. The variance provides a measure of the spread or dispersion of the distribution of X around its expected value. Note that since $(X - \mu)^2$ takes only non-negative values, variance is always non-negative.

Theorem 8.18. *Let X be a real-valued random variable, with $\mu := \mathsf{E}[X]$, and let a and b be real numbers. Then we have*

(i) $\mathsf{Var}[X] = \mathsf{E}[X^2] - \mu^2$,

(ii) $\mathsf{Var}[aX] = a^2 \mathsf{Var}[X]$, *and*

(iii) $\mathsf{Var}[X + b] = \mathsf{Var}[X]$.

Proof. For part (i), observe that

$$\begin{aligned}\mathsf{Var}[X] &= \mathsf{E}[(X - \mu)^2] = \mathsf{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathsf{E}[X^2] - 2\mu\,\mathsf{E}[X] + \mathsf{E}[\mu^2] = \mathsf{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathsf{E}[X^2] - \mu^2,\end{aligned}$$

where in the third equality, we used the fact that expectation is linear, and in the fourth equality, we used the fact that $\mathsf{E}[c] = c$ for constant c (in this case, $c = \mu^2$).

For part (ii), observe that

$$\begin{aligned}\mathsf{Var}[aX] &= \mathsf{E}[a^2X^2] - \mathsf{E}[aX]^2 = a^2\,\mathsf{E}[X^2] - (a\mu)^2 \\ &= a^2(\mathsf{E}[X^2] - \mu^2) = a^2\,\mathsf{Var}[X],\end{aligned}$$

where we used part (i) in the first and fourth equality, and the linearity of expectation in the second.

Part (iii) follows by a similar calculation:

$$\begin{aligned}\mathsf{Var}[X + b] &= \mathsf{E}[(X + b)^2] - (\mu + b)^2 \\ &= (\mathsf{E}[X^2] + 2b\mu + b^2) - (\mu^2 + 2b\mu + b^2) \\ &= \mathsf{E}[X^2] - \mu^2 = \mathsf{Var}[X]. \quad \square\end{aligned}$$

The following is an immediate consequence of part (i) of Theorem 8.18, and the fact that variance is always non-negative:

Theorem 8.19. *If X is a real-valued random variable, then $\mathsf{E}[X^2] \geq \mathsf{E}[X]^2$.*

Unlike expectation, the variance of a sum of random variables is not equal to the sum of the variances, unless the variables are *pairwise independent*:

Theorem 8.20. *If $\{X_i\}_{i \in I}$ is a finite, pairwise independent family of real-valued random variables, then*

$$\mathsf{Var}\Big[\sum_{i \in I} X_i\Big] = \sum_{i \in I} \mathsf{Var}[X_i].$$

Proof. We have

$$\begin{aligned}
\mathsf{Var}\Big[\sum_{i\in I} X_i\Big] &= \mathsf{E}\Big[\Big(\sum_{i\in I} X_i\Big)^2\Big] - \Big(\mathsf{E}\Big[\sum_{i\in I} X_i\Big]\Big)^2 \\
&= \sum_{i\in I} \mathsf{E}[X_i^2] + \sum_{\substack{i,j\in I \\ i\neq j}} \big(\mathsf{E}[X_i X_j] - \mathsf{E}[X_i]\,\mathsf{E}[X_j]\big) - \sum_{i\in I} \mathsf{E}[X_i]^2 \\
&\qquad \text{(by linearity of expectation and rearranging terms)} \\
&= \sum_{i\in I} \mathsf{E}[X_i^2] - \sum_{i\in I} \mathsf{E}[X_i]^2 \\
&\qquad \text{(by pairwise independence and Theorem 8.15)} \\
&= \sum_{i\in I} \mathsf{Var}[X_i]. \quad \square
\end{aligned}$$

Corresponding to Theorem 8.16, we have:

Theorem 8.21. *Let X be a $0/1$-valued random variable, with $p := \mathsf{P}[X = 1]$ and $q := \mathsf{P}[X = 0] = 1 - p$. Then $\mathsf{Var}[X] = pq$.*

Proof. We have $\mathsf{E}[X] = p$ and $\mathsf{E}[X^2] = \mathsf{P}[X^2 = 1] = \mathsf{P}[X = 1] = p$. Therefore,

$$\mathsf{Var}[X] = \mathsf{E}[X^2] - \mathsf{E}[X]^2 = p - p^2 = p(1-p) = pq. \quad \square$$

Let $\mathcal{B}$ be an event with $\mathsf{P}[\mathcal{B}] \neq 0$, and let X be a real-valued random variable. We define the **conditional expectation of X given $\mathcal{B}$**, denoted $\mathsf{E}[X \mid \mathcal{B}]$, to be the expected value of the X *relative to the conditional distribution* $\mathsf{P}(\cdot \mid \mathcal{B})$, so that

$$\mathsf{E}[X \mid \mathcal{B}] = \sum_{\omega\in\Omega} X(\omega)\,\mathsf{P}(\omega \mid \mathcal{B}) = \mathsf{P}[\mathcal{B}]^{-1} \sum_{\omega\in\mathcal{B}} X(\omega)\,\mathsf{P}(\omega).$$

Analogous to (8.19), if S is the image of X, we have

$$\mathsf{E}[X \mid \mathcal{B}] = \sum_{s\in S} s\,\mathsf{P}[X = s \mid \mathcal{B}]. \tag{8.23}$$

Furthermore, suppose I is a finite index set, and $\{\mathcal{B}_i\}_{i\in I}$ is a partition of the sample space, where each $\mathcal{B}_i$ occurs with non-zero probability. If for each $i \in I$ we group together the terms in (8.18) with $\omega \in \mathcal{B}_i$, we obtain the **law of total expectation**:

$$\mathsf{E}[X] = \sum_{i\in I} \mathsf{E}[X \mid \mathcal{B}_i]\,\mathsf{P}[\mathcal{B}_i]. \tag{8.24}$$

Example 8.29. Let X be uniformly distributed over $\{1, \ldots, m\}$. Let us compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$. We have

$$\mathsf{E}[X] = \sum_{s=1}^{m} s \cdot \frac{1}{m} = \frac{m(m+1)}{2} \cdot \frac{1}{m} = \frac{m+1}{2}.$$

We also have

$$\mathsf{E}[X^2] = \sum_{s=1}^{m} s^2 \cdot \frac{1}{m} = \frac{m(m+1)(2m+1)}{6} \cdot \frac{1}{m} = \frac{(m+1)(2m+1)}{6}.$$

Therefore,

$$\mathsf{Var}[X] = \mathsf{E}[X^2] - \mathsf{E}[X]^2 = \frac{m^2-1}{12}. \quad \square$$

Example 8.30. Let X denote the value of a roll of a die. Let $\mathcal{A}$ be the event that X is even. Then the conditional distribution of X given $\mathcal{A}$ is essentially the uniform distribution on $\{2, 4, 6\}$, and hence

$$\mathsf{E}[X \mid \mathcal{A}] = \frac{2+4+6}{3} = 4.$$

Similarly, the conditional distribution of X given $\overline{\mathcal{A}}$ is essentially the uniform distribution on $\{1, 3, 5\}$, and so

$$\mathsf{E}[X \mid \overline{\mathcal{A}}] = \frac{1+3+5}{3} = 3.$$

Using the law of total expectation, we can compute the expected value of X as follows:

$$\mathsf{E}[X] = \mathsf{E}[X \mid \mathcal{A}]\,\mathsf{P}[\mathcal{A}] + \mathsf{E}[X \mid \overline{\mathcal{A}}]\,\mathsf{P}[\overline{\mathcal{A}}] = 4 \cdot \frac{1}{2} + 3 \cdot \frac{1}{2} = \frac{7}{2},$$

which agrees with the calculation in the previous example. $\square$

Example 8.31. Let X be a random variable with a binomial distribution, as in Example 8.18, that counts the number of successes among n Bernoulli trials, each of which succeeds with probability p. Let us compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$. We can write X as the sum of indicator variables, $X = \sum_{i=1}^{n} X_i$, where X_i is the indicator variable for the event that the ith trial succeeds; each X_i takes the value 1 with probability p and 0 with probability $q := 1 - p$, and the family of random variables $\{X_i\}_{i=1}^{n}$ is mutually independent (see Example 8.24). By Theorems 8.16 and 8.21, we have $\mathsf{E}[X_i] = p$ and $\mathsf{Var}[X_i] = pq$ for $i = 1, \ldots, n$. By linearity of expectation, we have

$$\mathsf{E}[X] = \sum_{i=1}^{n} \mathsf{E}[X_i] = np.$$

By Theorem 8.20, and the fact that $\{X_i\}_{i=1}^{n}$ is mutually independent (and hence pairwise independent), we have

$$\mathsf{Var}[X] = \sum_{i=1}^{n} \mathsf{Var}[X_i] = npq. \quad \square$$

Example 8.32. Our proof of Theorem 8.1 could be elegantly recast in terms of indicator variables. For $\mathcal{B} \subseteq \Omega$, let $X_{\mathcal{B}}$ be the indicator variable for $\mathcal{B}$, so that $X_{\mathcal{B}}(\omega) = \delta_\omega[\mathcal{B}]$ for each $\omega \in \Omega$. Equation (8.8) then becomes

$$X_{\mathcal{A}} = \sum_{\emptyset \subsetneq J \subseteq I} (-1)^{|J|-1} X_{\mathcal{A}_J},$$

and by Theorem 8.16 and linearity of expectation, we have

$$\mathsf{P}[\mathcal{A}] = \mathsf{E}[X_{\mathcal{A}}] = \sum_{\emptyset \subsetneq J \subseteq I} (-1)^{|J|-1} \mathsf{E}[X_{\mathcal{A}_J}] = \sum_{\emptyset \subsetneq J \subseteq I} (-1)^{|J|-1} \mathsf{P}[X_{\mathcal{A}_J}]. \quad \square$$

EXERCISE 8.20. Suppose X is a real-valued random variable. Show that $|\mathsf{E}[X]| \leq \mathsf{E}[|X|] \leq \mathsf{E}[X^2]^{1/2}$.

EXERCISE 8.21. Suppose X and Y take non-negative real values, and that $Y \leq c$ for some constant c. Show that $\mathsf{E}[XY] \leq c\,\mathsf{E}[X]$

EXERCISE 8.22. Let X be a 0/1-valued random variable. Show that $\mathsf{Var}[X] \leq 1/4$.

EXERCISE 8.23. Let $\mathcal{B}$ be an event with $\mathsf{P}[\mathcal{B}] \neq 0$, and let $\{\mathcal{B}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union is $\mathcal{B}$. Generalizing the law of total expectation (8.24), show that for every real-valued random variable X, if $I^* := \{i \in I : \mathsf{P}[\mathcal{B}_i] \neq 0\}$, then we have

$$\mathsf{E}[X \mid \mathcal{B}]\,\mathsf{P}[\mathcal{B}] = \sum_{i \in I^*} \mathsf{E}[X \mid \mathcal{B}_i]\,\mathsf{P}[\mathcal{B}_i].$$

Also show that if $\mathsf{E}[X \mid \mathcal{B}_i] \leq \alpha$ for each $i \in I^*$, then $\mathsf{E}[X \mid \mathcal{B}] \leq \alpha$.

EXERCISE 8.24. Let $\mathcal{B}$ be an event with $\mathsf{P}[\mathcal{B}] \neq 0$, and let $\{C_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union contains $\mathcal{B}$. Again, generalizing the law of total expectation, show that for every real-valued random variable X, if $I^* := \{i \in I : \mathsf{P}[\mathcal{B} \cap C_i] \neq 0\}$, then we have

$$\mathsf{E}[X \mid \mathcal{B}] = \sum_{i \in I^*} \mathsf{E}[X \mid \mathcal{B} \cap C_i]\,\mathsf{P}[C_i \mid \mathcal{B}].$$

EXERCISE 8.25. This exercise makes use of the notion of *convexity* (see §A8).

(a) Prove **Jensen's inequality**: if f is convex on an interval, and X is a random variable taking values in that interval, then $\mathsf{E}[f(X)] \geq f(\mathsf{E}[X])$. Hint: use induction on the size of the image of X. (Note that Theorem 8.19 is a special case of this, with $f(s) := s^2$.)

(b) Using part (a), show that if X takes non-negative real values, and α is a positive number, then $\mathsf{E}[X^\alpha] \geq \mathsf{E}[X]^\alpha$ if $\alpha \geq 1$, and $\mathsf{E}[X^\alpha] \leq \mathsf{E}[X]^\alpha$ if $\alpha \leq 1$.

(c) Using part (a), show that if X takes positive real values, then $\mathsf{E}[X] \geq e^{\mathsf{E}[\log X]}$.

(d) Using part (c), derive the **arithmetic/geometric mean inequality**: for all positive numbers $x_1, \ldots, x_n$, we have

$$(x_1 + \cdots + x_n)/n \geq (x_1 \cdots x_n)^{1/n}.$$

EXERCISE 8.26. For real-valued random variables X and Y, their **covariance** is defined as $\mathsf{Cov}[X, Y] := \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]$. Show that:

(a) if X, Y, and Z are real-valued random variables, and a is a real number, then $\mathsf{Cov}[X + Y, Z] = \mathsf{Cov}[X, Z] + \mathsf{Cov}[Y, Z]$ and $\mathsf{Cov}[aX, Z] = a\,\mathsf{Cov}[X, Z]$;

(b) if $\{X_i\}_{i \in I}$ is a finite family of real-valued random variables, then

$$\mathsf{Var}\Big[\sum_{i \in I} X_i\Big] = \sum_{i \in I} \mathsf{Var}[X_i] + \sum_{\substack{i,j \in I \\ i \neq j}} \mathsf{Cov}[X_i, X_j].$$

EXERCISE 8.27. Let $f : [0, 1] \to \mathbb{R}$ be a function that is "nice" in the following sense: for some constant c, we have $|f(s) - f(t)| \leq c|s - t|$ for all $s, t \in [0, 1]$. This condition is implied, for example, by the assumption that f has a derivative that is bounded in absolute value by c on the interval $[0, 1]$. For each positive integer n, define the polynomial $B_{n,f} := \sum_{k=0}^{n} \binom{n}{k} f(k/n) T^k (1 - T)^{n-k} \in \mathbb{R}[T]$. Show that $|B_{n,f}(p) - f(p)| \leq c/2\sqrt{n}$ for all positive integers n and all $p \in [0, 1]$. Hint: let X be a random variable with a binomial distribution that counts the number of successes among n Bernoulli trials, each of which succeeds with probability p, and begin by observing that $B_{n,f}(p) = \mathsf{E}[f(X/n)]$. The polynomial $B_{n,f}$ is called the nth **Bernstein approximation** to f, and this result proves a classical result that any "nice" function can approximated to arbitrary precision by a polynomial of sufficiently high degree.

EXERCISE 8.28. Consider again the game played between Alice and Bob in Example 8.11. Suppose that to play the game, Bob must place a one dollar bet. However, after Alice reveals the sum of the two dice, Bob may elect to double his bet. If Bob's guess is correct, Alice pays him his bet, and otherwise Bob pays Alice his bet. Describe an optimal playing strategy for Bob, and calculate his expected winnings.

EXERCISE 8.29. A die is rolled repeatedly until it comes up "1," or until it is rolled n times (whichever comes first). What is the expected number of rolls of the die?

8.5 Some useful bounds

In this section, we present several theorems that can be used to bound the probability that a random variable deviates from its expected value by some specified amount.

Theorem 8.22 (Markov's inequality). *Let X be a random variable that takes only non-negative real values. Then for every $\alpha > 0$, we have*

$$\mathsf{P}[X \geq \alpha] \leq \mathsf{E}[X]/\alpha.$$

Proof. We have

$$\mathsf{E}[X] = \sum_s s\,\mathsf{P}[X = s] \;=\; \sum_{s<\alpha} s\,\mathsf{P}[X = s] + \sum_{s\geq\alpha} s\,\mathsf{P}[X = s],$$

where the summations are over elements s in the image of X. Since X takes only non-negative values, all of the terms are non-negative. Therefore,

$$\mathsf{E}[X] \geq \sum_{s\geq\alpha} s\,\mathsf{P}[X = s] \;\geq\; \sum_{s\geq\alpha} \alpha\,\mathsf{P}[X = s] \;=\; \alpha\,\mathsf{P}[X \geq \alpha]. \;\square$$

Markov's inequality may be the only game in town when nothing more about the distribution of X is known besides its expected value. However, if the variance of X is also known, then one can get a better bound.

Theorem 8.23 (Chebyshev's inequality). *Let X be a real-valued random variable, with $\mu := \mathsf{E}[X]$ and $\nu := \mathsf{Var}[X]$. Then for every $\alpha > 0$, we have*

$$\mathsf{P}[|X - \mu| \geq \alpha] \leq \nu/\alpha^2.$$

Proof. Let $Y := (X - \mu)^2$. Then Y is always non-negative, and $\mathsf{E}[Y] = \nu$. Applying Markov's inequality to Y, we have

$$\mathsf{P}[|X - \mu| \geq \alpha] = \mathsf{P}[Y \geq \alpha^2] \leq \nu/\alpha^2. \;\square$$

An important special case of Chebyshev's inequality is the following. Suppose that $\{X_i\}_{i\in I}$ is a finite, non-empty, pairwise independent family of real-valued random variables, each with the same distribution. Let μ be the common value of $\mathsf{E}[X_i]$, ν be the common value of $\mathsf{Var}[X_i]$, and $n := |I|$. Set

$$\overline{X} := \frac{1}{n}\sum_{i\in I} X_i.$$

The variable $\overline{X}$ is called the **sample mean** of $\{X_i\}_{i\in I}$. By the linearity of expectation, we have $\mathsf{E}[\overline{X}] = \mu$, and since $\{X_i\}_{i\in I}$ is pairwise independent, it follows from

Theorem 8.20 (along with part (ii) of Theorem 8.18) that $\mathsf{Var}[\overline{X}] = \nu/n$. Applying Chebyshev's inequality, for every $\varepsilon > 0$, we have

$$\mathsf{P}[|\overline{X} - \mu| \geq \varepsilon] \leq \frac{\nu}{n\varepsilon^2}. \tag{8.25}$$

The inequality (8.25) says that for all $\varepsilon > 0$, and for all $\delta > 0$, there exists n_0 (depending on ε and δ, as well as the variance ν) such that $n \geq n_0$ implies

$$\mathsf{P}[|\overline{X} - \mu| \geq \varepsilon] \leq \delta. \tag{8.26}$$

In words:

> *As n gets large, the sample mean closely approximates the expected value μ with high probability.*

This fact, known as the **law of large numbers**, justifies the usual intuitive interpretation given to expectation.

Let us now examine an even more specialized case of the above situation, where each X_i is a $0/1$-valued random variable, taking the value 1 with probability p, and 0 with probability $q := 1 - p$. By Theorems 8.16 and 8.21, the X_i's have a common expected value p and variance pq. Therefore, by (8.25), for every $\varepsilon > 0$, we have

$$\mathsf{P}[|\overline{X} - p| \geq \varepsilon] \leq \frac{pq}{n\varepsilon^2}. \tag{8.27}$$

The bound on the right-hand side of (8.27) decreases linearly in n. If one makes the stronger assumption that the family $\{X_i\}_{i \in I}$ is *mutually independent* (so that $X := \sum_i X_i$ has a binomial distribution), one can obtain a much better bound that decreases *exponentially* in n:

Theorem 8.24 (Chernoff bound). *Let $\{X_i\}_{i \in I}$ be a finite, non-empty, and mutually independent family of random variables, such that each X_i is 1 with probability p and 0 with probability $q := 1 - p$. Assume that $0 < p < 1$. Also, let $n := |I|$ and $\overline{X}$ be the sample mean of $\{X_i\}_{i \in I}$. Then for every $\varepsilon > 0$, we have:*

(i) $\mathsf{P}[\overline{X} - p \geq \varepsilon] \leq e^{-n\varepsilon^2/2q}$;

(ii) $\mathsf{P}[\overline{X} - p \leq -\varepsilon] \leq e^{-n\varepsilon^2/2p}$;

(iii) $\mathsf{P}[|\overline{X} - p| \geq \varepsilon] \leq 2e^{-n\varepsilon^2/2}$.

Proof. First, we observe that (ii) follows directly from (i) by replacing X_i by $1 - X_i$ and exchanging the roles of p and q. Second, we observe that (iii) follows directly from (i) and (ii). Thus, it suffices to prove (i).

Let $\alpha > 0$ be a parameter, whose value will be determined later. Define the random variable $Z := e^{\alpha n(\overline{X} - p)}$. Since the function $x \mapsto e^{\alpha n x}$ is strictly increasing, we have $\overline{X} - p \geq \varepsilon$ if and only if $Z \geq e^{\alpha n \varepsilon}$. By Markov's inequality, it follows that

$$\mathsf{P}[\overline{X} - p \geq \varepsilon] = \mathsf{P}[Z \geq e^{\alpha n \varepsilon}] \leq \mathsf{E}[Z] e^{-\alpha n \varepsilon}. \tag{8.28}$$

So our goal is to bound $\mathsf{E}[Z]$ from above.

For each $i \in I$, define the random variable $Z_i := e^{\alpha(X_i - p)}$. Observe that $Z = \prod_{i \in I} Z_i$, that $\{Z_i\}_{i \in I}$ is a mutually independent family of random variables (see Theorem 8.12), and that for each $i \in I$, we have

$$\mathsf{E}[Z_i] = e^{\alpha(1-p)}p + e^{\alpha(0-p)}q = pe^{\alpha q} + qe^{-\alpha p}.$$

It follows that

$$\mathsf{E}[Z] = \mathsf{E}\Big[\prod_{i \in I} Z_i\Big] = \prod_{i \in I} \mathsf{E}[Z_i] = (pe^{\alpha q} + qe^{-\alpha p})^n.$$

We will prove below that

$$pe^{\alpha q} + qe^{-\alpha p} \leq e^{\alpha^2 q/2}. \tag{8.29}$$

From this, it follows that

$$\mathsf{E}[Z] \leq e^{\alpha^2 qn/2}. \tag{8.30}$$

Combining (8.30) with (8.28), we obtain

$$\mathsf{P}[\overline{X} - p \geq \varepsilon] \leq e^{\alpha^2 qn/2 - \alpha n\varepsilon}. \tag{8.31}$$

Now we choose the parameter α so as to minimize the quantity $\alpha^2 qn/2 - \alpha n\varepsilon$. The optimal value of α is easily seen to be $\alpha = \varepsilon/q$, and substituting this value of α into (8.31) yields (i).

To finish the proof of the theorem, it remains to prove the inequality (8.29). Let

$$\beta := pe^{\alpha q} + qe^{-\alpha p}.$$

We want to show that $\beta \leq e^{\alpha^2 q/2}$, or equivalently, that $\log \beta \leq \alpha^2 q/2$. We have

$$\beta = e^{\alpha q}(p + qe^{-\alpha}) = e^{\alpha q}(1 - q(1 - e^{-\alpha})),$$

and taking logarithms and applying parts (i) and (ii) of §A1, we obtain

$$\log \beta = \alpha q + \log(1 - q(1 - e^{-\alpha})) \leq \alpha q - q(1 - e^{-\alpha}) = q(e^{-\alpha} + \alpha - 1) \leq q\alpha^2/2.$$

This establishes (8.29) and completes the proof of the theorem. □

Thus, the Chernoff bound is a quantitatively superior version of the law of large numbers, although its range of application is clearly more limited.

Example 8.33. Suppose we toss a fair coin 10,000 times. The expected number of *heads* is 5,000. What is an upper bound on the probability α that we get 6,000 or more *heads*? Using Markov's inequality, we get $\alpha \leq 5/6$. Using Chebyshev's inequality, and in particular, the inequality (8.27), we get

$$\alpha \leq \frac{1/4}{10^4 10^{-2}} = \frac{1}{400}.$$

Finally, using the Chernoff bound, we obtain

$$\alpha \leq e^{-10^4 10^{-2}/2(0.5)} = e^{-100} \approx 10^{-43.4}. \quad \square$$

EXERCISE 8.30. With notation and assumptions as in Theorem 8.24, and with $p := q := 1/2$, show that there exist constants c_1 and c_2 such that

$$\mathsf{P}[|\overline{X} - 1/2| \geq c_1/\sqrt{n}] \leq 1/2 \quad \text{and} \quad \mathsf{P}[|\overline{X} - 1/2| \geq c_2/\sqrt{n}] \geq 1/2.$$

Hint: for the second inequality, use Exercise 5.16.

EXERCISE 8.31. In each step of a **random walk**, we toss a coin, and move either one unit to the right, or one unit to the left, depending on the outcome of the coin toss. The question is, after n steps, what is our expected distance from the starting point? Let us model this using a mutually independent family of random variables $\{Y_i\}_{i=1}^n$, with each Y_i uniformly distributed over $\{-1, 1\}$, and define $Y := Y_1 + \cdots + Y_n$. Show that the $c_1\sqrt{n} \leq \mathsf{E}[|Y|] \leq c_2\sqrt{n}$, for some constants c_1 and c_2.

EXERCISE 8.32. The goal of this exercise is to prove that with probability very close to 1, a random number between 1 and m has very close to $\log\log m$ prime factors. To prove this result, you will need to use appropriate theorems from Chapter 5. Suppose N is a random variable that is uniformly distributed over $\{1, \ldots, m\}$, where $m \geq 3$. For $i = 1, \ldots, m$, let D_i be the indicator variable for the event that i divides N. Also, define $X := \sum_{p \leq m} D_p$, where the sum is over all primes $p \leq m$, so that X counts the number of distinct primes dividing N. Show that:

(a) $1/i - 1/m < \mathsf{E}[D_i] \leq 1/i$, for each $i = 1, \ldots, m$;

(b) $|\mathsf{E}[X] - \log\log m| \leq c_1$ for some constant c_1;

(c) for all primes p, q, where $p \leq m$, $q \leq m$, and $p \neq q$, we have

$$\mathsf{Cov}[D_p, D_q] \leq \frac{1}{m}\Big(\frac{1}{p} + \frac{1}{q}\Big),$$

where Cov is the covariance, as defined in Exercise 8.26;

(d) $\mathsf{Var}[X] \leq \log\log m + c_2$ for some constant c_2;

(e) for some constant c_3, and for every $\alpha \geq 1$, we have

$$\mathsf{P}\Big[|X - \log\log m| \geq \alpha(\log\log m)^{1/2}\Big] \leq \alpha^{-2}\Big(1 + c_3(\log\log m)^{-1/2}\Big).$$

EXERCISE 8.33. For each positive integer n, let $\tau(n)$ denote the number of positive divisors of n. Suppose that N is uniformly distributed over $\{1, \ldots, m\}$. Show that $\mathsf{E}[\tau(N)] = \log m + O(1)$.

EXERCISE 8.34. You are given three biased coins, where for $i = 1, 2, 3$, coin i comes up *heads* with probability p_i. The coins look identical, and all you know is the following: (1) $|p_1 - p_2| > 0.01$ and (2) either $p_3 = p_1$ or $p_3 = p_2$. Your goal is to determine whether p_3 is equal to p_1, or to p_2. Design a random experiment to determine this. The experiment may produce an incorrect result, but this should happen with probability at most 10^{-12}. Try to use a reasonable number of coin tosses.

EXERCISE 8.35. Consider the following game, parameterized by a positive integer n. One rolls a pair of dice, and records the value of their sum. This is repeated until some value ℓ is recorded n times, and this value ℓ is declared the "winner." It is intuitively clear that 7 is the most likely winner. Let α_n be the probability that 7 does not win. Give a careful argument that $\alpha_n \to 0$ as $n \to \infty$. Assume that the rolls of the dice are mutually independent.

8.6 Balls and bins

This section and the next discuss applications of the theory developed so far.

Our first application is a brief study of "balls and bins." Suppose you throw n balls into m bins. A number of questions naturally arise, such as:

- What is the probability that a *collision* occurs, that is, two balls land in the same bin?
- What is the expected value of the maximum number of balls that land in any one bin?

To formalize these questions, we introduce some notation that will be used throughout this section. Let I be a finite set of size $n > 0$, and S a finite set of size $m > 0$. Let $\{X_i\}_{i \in I}$ be a family of random variables, where each X_i is uniformly distributed over the set S. The idea is that I represents a set of labels for our n balls, S represents the set of m bins, and X_i represents the bin into which ball i lands.

We define C to be the event that a collision occurs; formally, this is the event that $X_i = X_j$ for some $i, j \in I$ with $i \neq j$. We also define M to be the random variable that measures that maximum number of balls in any one bin; formally,

$$M := \max\{N_s : s \in S\},$$

where for each $s \in S$, N_s is the number of balls that land in bin s; that is,

$$N_s := |\{i \in I : X_i = s\}|.$$

The questions posed above can now be stated as the problems of estimating $\mathsf{P}[C]$

and $\mathsf{E}[M]$. However, to estimate these quantities, we have to make some assumptions about the independence of the X_i's. While it is natural to assume that the family of random variables $\{X_i\}_{i\in I}$ is mutually independent, it is also interesting and useful to estimate these quantities under weaker independence assumptions. We shall therefore begin with an analysis under the weaker assumption that $\{X_i\}_{i\in I}$ is *pairwise* independent. We start with a simple observation:

Theorem 8.25. *Suppose $\{X_i\}_{i\in I}$ is pairwise independent. Then for all $i, j \in I$ with $i \neq j$, we have $\mathsf{P}[X_i = X_j] = 1/m$.*

Proof. The event $X_i = X_j$ occurs if and only if $X_i = s$ and $X_j = s$ for some $s \in S$. Therefore,

$$\begin{aligned}\mathsf{P}[X_i = X_j] &= \sum_{s\in S} \mathsf{P}[(X_i = s) \cap (X_j = s)] \ \text{(by Boole's equality (8.7))} \\ &= \sum_{s\in S} 1/m^2 \ \text{(by pairwise independence)} \\ &= 1/m. \ \square\end{aligned}$$

Theorem 8.26. *Suppose $\{X_i\}_{i\in I}$ is pairwise independent. Then*

$$\mathsf{P}[C] \leq \frac{n(n-1)}{2m}.$$

Proof. Let $I^{(2)} := \{J \subseteq I : |J| = 2\}$. Then using Boole's inequality (8.6) and Theorem 8.25, we have

$$\mathsf{P}[C] \leq \sum_{\{i,j\}\in I^{(2)}} \mathsf{P}[X_i = X_j] = \sum_{\{i,j\}\in I^{(2)}} \frac{1}{m} = \frac{|I^{(2)}|}{m} = \frac{n(n-1)}{2m}. \ \square$$

Theorem 8.27. *Suppose $\{X_i\}_{i\in I}$ is pairwise independent. Then*

$$\mathsf{E}[M] \leq \sqrt{n^2/m + n}.$$

Proof. To prove this, we use the fact that $\mathsf{E}[M]^2 \leq \mathsf{E}[M^2]$ (see Theorem 8.19), and that $M^2 \leq Z := \sum_{s\in S} N_s^2$. It will therefore suffice to show that

$$\mathsf{E}[Z] \leq n^2/m + n. \tag{8.32}$$

To this end, for $i \in I$ and $s \in S$, let L_{is} be the indicator variable for the event that ball i lands in bin s (i.e., $X_i = s$), and for $i, j \in I$, let C_{ij} be the indicator variable for the event that balls i and j land in the same bin (i.e., $X_i = X_j$). Observing that

$C_{ij} = \sum_{s \in S} L_{is} L_{js}$, we have

$$Z = \sum_{s \in S} N_s^2 = \sum_{s \in S} \Big(\sum_{i \in I} L_{is} \Big)^2 = \sum_{s \in S} \Big(\sum_{i \in I} L_{is} \Big) \Big(\sum_{j \in I} L_{js} \Big) = \sum_{i,j \in I} \sum_{s \in S} L_{is} L_{js}$$
$$= \sum_{i,j \in I} C_{ij}.$$

For $i, j \in I$, we have $\mathsf{E}[C_{ij}] = \mathsf{P}[X_i = X_j]$ (see Theorem 8.16), and so by Theorem 8.25, we have $\mathsf{E}[C_{ij}] = 1/m$ if $i \neq j$, and clearly, $\mathsf{E}[C_{ij}] = 1$ if $i = j$. By linearity of expectation, we have

$$\mathsf{E}[Z] = \sum_{i,j \in I} \mathsf{E}[C_{ij}] = \sum_{\substack{i,j \in I \\ i \neq j}} \mathsf{E}[C_{ij}] + \sum_{i \in I} \mathsf{E}[C_{ii}] = \frac{n(n-1)}{m} + n \leq n^2/m + n,$$

which proves (8.32). □

We next consider the situation where $\{X_i\}_{i \in I}$ is mutually independent. Of course, Theorem 8.26 is still valid in this case, but with our stronger assumption, we can derive a *lower* bound on $\mathsf{P}[C]$.

Theorem 8.28. *Suppose $\{X_i\}_{i \in I}$ is mutually independent. Then*

$$\mathsf{P}[C] \geq 1 - e^{-n(n-1)/2m}.$$

Proof. Let $\alpha := \mathsf{P}[\overline{C}]$. We want to show $\alpha \leq e^{-n(n-1)/2m}$. We may assume that $I = \{1, \ldots, n\}$ (the labels make no difference) and that $n \leq m$ (otherwise, $\alpha = 0$). Under the hypothesis of the theorem, the random variable $(X_1, \ldots, X_n)$ is uniformly distributed over $S^{\times n}$. Among all m^n sequences $(s_1, \ldots, s_n) \in S^{\times n}$, there are a total of $m(m-1) \cdots (m-n+1)$ that contain no repetitions: there are m choices for s_1, and for any fixed value of s_1, there are $m-1$ choices for s_2, and so on. Therefore

$$\alpha = m(m-1) \cdots (m-n+1)/m^n = \Big(1 - \frac{1}{m}\Big)\Big(1 - \frac{2}{m}\Big) \cdots \Big(1 - \frac{n-1}{m}\Big).$$

Using part (i) of §A1, we obtain

$$\alpha \leq e^{-\sum_{i=1}^{n-1} i/m} = e^{-n(n-1)/2m}. \quad \square$$

Theorem 8.26 implies that if $n(n-1) \leq m$, then the probability of a collision is *at most* $1/2$; moreover, Theorem 8.28 implies that if $n(n-1) \geq (2 \log 2)m$, then the probability of a collision is *at least* $1/2$. Thus, for n near $\sqrt{m}$, the probability of a collision is roughly $1/2$. A colorful illustration of this is the following fact: in a room with 23 or more people, the odds are better than even that two people in the room have birthdays on the same day of the year. This follows by setting $n = 23$ and $m = 365$ in Theorem 8.28. Here, we are ignoring leap years, and the fact that

birthdays are not uniformly distributed over the calendar year (however, any skew in the birthday distribution only increases the odds that two people share the same birthday—see Exercise 8.40 below). Because of this fact, Theorem 8.28 is often called the **birthday paradox** (the "paradox" being the perhaps surprisingly small number of people in the room).

The hypothesis that $\{X_i\}_{i \in I}$ is mutually independent is crucial in Theorem 8.28. Indeed, assuming just pairwise independence, we may have $\mathsf{P}[C] = 1/m$, even when $n = m$ (see Exercise 8.42 below). However, useful, non-trivial lower bounds on $\mathsf{P}[C]$ can still be obtained under assumptions weaker than mutual independence (see Exercise 8.43 below).

Assuming $\{X_i\}_{i \in I}$ is mutually independent, we can get a much sharper upper bound on $\mathsf{E}[M]$ than that provided by Theorem 8.27. For simplicity, we only consider the case where $m = n$; in this case, Theorem 8.27 gives us the bound $\mathsf{E}[M] \leq \sqrt{2n}$ (which cannot be substantially improved assuming only pairwise independence—see Exercise 8.44 below).

Theorem 8.29. *Suppose $\{X_i\}_{i \in I}$ is mutually independent and that $m = n$. Then*

$$\mathsf{E}[M] \leq (1 + o(1)) \frac{\log n}{\log \log n}.$$

Proof. We use Theorem 8.17, which says that $\mathsf{E}[M] = \sum_{k \geq 1} \mathsf{P}[M \geq k]$.

Claim 1. For $k \geq 1$, we have $\mathsf{P}[M \geq k] \leq n/k!$.

To prove Claim 1, we may assume that $k \leq n$ (as otherwise, $\mathsf{P}[M \geq k] = 0$). Let $I^{(k)} := \{J \subseteq I : |J| = k\}$. Now, $M \geq k$ if and only if there is an $s \in S$ and a subset $J \in I^{(k)}$, such that $X_j = s$ for all $j \in J$. Therefore,

$$\begin{aligned}
\mathsf{P}[M \geq k] &\leq \sum_{s \in S} \sum_{J \in I^{(k)}} \mathsf{P}\Big[\bigcap_{j \in J} (X_j = s)\Big] \quad \text{(by Boole's inequality (8.6))} \\
&= \sum_{s \in S} \sum_{J \in I^{(k)}} \prod_{j \in J} \mathsf{P}[X_j = s] \quad \text{(by mutual independence)} \\
&= n \binom{n}{k} n^{-k} \leq n/k!.
\end{aligned}$$

That proves Claim 1.

Of course, Claim 1 is only interesting when $n/k! \leq 1$, since $\mathsf{P}[M \geq k]$ is always at most 1. Define $F(n)$ to be the smallest positive integer k such that $k! \geq n$.

Claim 2. $F(n) \sim \log n / \log \log n$.

To prove this, let us set $k := F(n)$. It is clear that $n \leq k! \leq nk$, and taking

logarithms, $\log n \leq \log k! \leq \log n + \log k$. Moreover, we have

$$\log k! = \sum_{\ell=1}^{k} \log \ell = \int_1^k \log x \, dx + O(\log k) = k \log k - k + O(\log k) \sim k \log k,$$

where we have estimated the sum by an integral (see §A5). Thus,

$$\log n = \log k! + O(\log k) \sim k \log k.$$

Taking logarithms again, we see that

$$\log \log n = \log k + \log \log k + o(1) \sim \log k,$$

and so $\log n \sim k \log k \sim k \log \log n$, from which Claim 2 follows.

Finally, observe that each term in the sequence $\{n/k!\}_{k=1}^{\infty}$ is at most half the previous term. Combining this observation with Claims 1 and 2, and the fact that $\mathsf{P}[M \geq k]$ is always at most 1, we have

$$\begin{aligned} \mathsf{E}[M] &= \sum_{k \geq 1} \mathsf{P}[M \geq k] = \sum_{k \leq F(n)} \mathsf{P}[M \geq k] + \sum_{k > F(n)} \mathsf{P}[M \geq k] \\ &\leq F(n) + \sum_{\ell \geq 1} 2^{-\ell} = F(n) + 1 \sim \log n / \log \log n. \quad \square \end{aligned}$$

EXERCISE 8.36. Let $\alpha_1, \ldots, \alpha_m$ be real numbers that sum to 1. Show that $0 \leq \sum_{s=1}^{m} (\alpha_s - 1/m)^2 = \sum_{s=1}^{m} \alpha_s^2 - 1/m$, and in particular, $\sum_{s=1}^{m} \alpha_s^2 \geq 1/m$.

EXERCISE 8.37. Let X and X' be independent random variables, both having the same distribution on a set S of size m. Show that $\mathsf{P}[X = X'] = \sum_{s \in S} \mathsf{P}[X = s]^2 \geq 1/m$.

EXERCISE 8.38. Suppose that the family of random variables X, Y, Y' is mutually independent, where X has image S, and where Y and Y' have the same distribution on a set T. Let ϕ be a predicate on $S \times T$, and let $\alpha := \mathsf{P}[\phi(X, Y)]$. Show that $\mathsf{P}[\phi(X, Y) \cap \phi(X, Y')] \geq \alpha^2$. In addition, show that if Y and Y' are both uniformly distributed over T, then $\mathsf{P}[\phi(X, Y) \cap \phi(X, Y') \cap (Y \neq Y')] \geq \alpha^2 - \alpha/|T|$.

EXERCISE 8.39. Let $\alpha_1, \ldots, \alpha_m$ be non-negative real numbers that sum to 1. Let $S := \{1, \ldots, m\}$, and for $n = 1, \ldots, m$, let $S^{(n)} := \{\mathcal{T} \subseteq S : |\mathcal{T}| = n\}$, and define

$$P_n(\alpha_1, \ldots, \alpha_m) := \sum_{T \in S^{(n)}} \prod_{t \in T} \alpha_t.$$

Show that $P_n(\alpha_1, \ldots, \alpha_m)$ is maximized when $\alpha_1 = \cdots = \alpha_m = 1/m$. Hint: first argue that if $\alpha_s < \alpha_t$, then for every $\varepsilon \in [0, \alpha_t - \alpha_s]$, replacing the pair (α_s, α_t) by $(\alpha_s + \varepsilon, \alpha_t - \varepsilon)$ does not decrease the value of $P_n(\alpha_1, \ldots, \alpha_m)$.

EXERCISE 8.40. Suppose that $\{X_i\}_{i \in I}$ is a finite, non-empty, mutually independent family of random variables, where each X_i is uniformly distributed over a finite set S. Suppose that $\{Y_i\}_{i \in I}$ is another finite, non-empty, mutually independent family of random variables, where each Y_i has the same distribution and takes values in the set S. Let α be the probability that the X_i's are distinct, and β be the probability that the Y_i's are distinct. Using the previous exercise, show that $\beta \leq \alpha$.

EXERCISE 8.41. Suppose n balls are thrown into m bins. Let $\mathcal{A}$ be the event that there is some bin that is empty. Assuming that the throws are mutually independent, and that $n \geq m(\log m + t)$ for some $t \geq 0$, show that $\mathsf{P}[\mathcal{A}] \leq e^{-t}$.

EXERCISE 8.42. Show that for every prime p, there exists a pairwise independent family of random variables $\{X_i\}_{i \in \mathbb{Z}_p}$, where each X_i is uniformly distributed over $\mathbb{Z}_p$, and yet the probability that all the X_i's are distinct is $1 - 1/p$.

EXERCISE 8.43. Let $\{X_i\}_{i=1}^n$ be a finite, non-empty, *4-wise independent* family of random variables, each uniformly distributed over a set S. Let α be the probability that the X_i's are distinct. For $i, j = 1, \ldots, n$, let C_{ij} be the indicator variable for the event that $X_i = X_j$, and define $K := \{(i, j) : 1 \leq i \leq n - 1,\ i + 1 \leq j \leq n\}$ and $Z := \sum_{(i,j) \in K} C_{ij}$. Show that:

(a) $\{C_{ij}\}_{(i,j) \in K}$ is pairwise independent;

(b) $\mathsf{E}[Z] = n(n-1)/2m$ and $\mathsf{Var}[Z] = (1 - 1/m)\,\mathsf{E}[Z]$;

(c) $\alpha \leq 1/\mathsf{E}[Z]$;

(d) $\alpha \leq 1/2$, provided $n(n-1) \geq 2m$ (hint: Exercise 8.4).

EXERCISE 8.44. Let k be a positive integer, let $n := k^2 - k + 1$, let I and S be sets of size n, and let s_0 be a fixed element of S. Also, let $I^{(k)} := \{J \subseteq I : |J| = k\}$, and let Π be the set of all permutations on S. For each $J \in I^{(k)}$, let f_J be some function that maps J to s_0, and maps $I \setminus J$ injectively into $S \setminus \{s_0\}$. For $\pi \in \Pi$, $J \in I^{(k)}$, and $i \in I$, define $\rho_i(\pi, J) := \pi(f_J(i))$. Finally, let Y be uniformly distributed over $\Pi \times I^{(k)}$, and for $i \in I$, define $X_i := \rho_i(Y)$. Show that $\{X_i\}_{i \in I}$ is pairwise independent, with each X_i uniformly distributed over S, and yet the number of X_i's with the same value is always at least $\sqrt{n}$.

EXERCISE 8.45. Let S be a set of size $m \geq 1$, and let s_0 be an arbitrary, fixed element of S. Let F be a random variable that is uniformly distributed over the set of all m^m functions from S into S. Let us define random variables X_i, for $i = 0, 1, 2, \ldots$, as follows:

$$X_0 := s_0, \quad X_{i+1} := F(X_i) \ \ (i = 0, 1, 2, \ldots).$$

Thus, the value of X_i is obtained by applying the function F a total of i times to the

starting value s_0. Since S has size m, the sequence $\{X_i\}_{i=0}^{\infty}$ must repeat at some point; that is, there exists a positive integer n (with $n \leq m$) such that $X_n = X_i$ for some $i = 0, \ldots, n-1$. Define the random variable Y to be the smallest such value n.

(a) Show that for every $i \geq 0$ and for all $s_1, \ldots, s_i \in S$ such that $s_0, s_1, \ldots, s_i$ are distinct, the conditional distribution of X_{i+1} given the event $(X_1 = s_1) \cap \cdots \cap (X_i = s_i)$ is the uniform distribution on S.

(b) Show that for every integer $n \geq 1$, we have $Y \geq n$ if and only if the random variables $X_0, X_1, \ldots, X_{n-1}$ take on distinct values.

(c) From parts (a) and (b), show that for each $n = 1, \ldots, m$, we have

$$\mathsf{P}[Y \geq n \mid Y \geq n-1] = 1 - (n-1)/m,$$

and conclude that

$$\mathsf{P}[Y \geq n] = \prod_{i=1}^{n-1}(1 - i/m) \leq e^{-n(n-1)/2m}.$$

(d) Using part (c), show that

$$\mathsf{E}[Y] = \sum_{n \geq 1} \mathsf{P}[Y \geq n] \leq \sum_{n \geq 1} e^{-n(n-1)/2m} = O(m^{1/2}).$$

(e) Modify the above argument to show that $\mathsf{E}[Y] = \Omega(m^{1/2})$.

EXERCISE 8.46. The setup for this exercise is identical to that of the previous exercise, except that now, F is uniformly distributed over the set of all $m!$ *permutations* of S.

(a) Show that if $Y = n$, then $X_n = X_0$.

(b) Show that for every $i \geq 0$ and all $s_1, \ldots, s_i \in S$ such that $s_0, s_1, \ldots, s_i$ are distinct, the conditional distribution of X_{i+1} given $(X_1 = s_1) \cap \cdots \cap (X_i = s_i)$ is essentially the uniform distribution on $S \setminus \{s_1, \ldots, s_i\}$.

(c) Show that for each $n = 2, \ldots, m$, we have

$$\mathsf{P}[Y \geq n \mid Y \geq n-1] = 1 - \frac{1}{m-n+2},$$

and conclude that for all $n = 1, \ldots, m$, we have

$$\mathsf{P}[Y \geq n] = \prod_{i=0}^{n-2}\left(1 - \frac{1}{m-i}\right) = 1 - \frac{n-1}{m}.$$

(d) From part (c), show that Y is uniformly distributed over $\{1, \ldots, m\}$, and in particular, $\mathsf{E}[Y] = (m+1)/2$.

8.7 Hash functions

In this section, we apply the tools we have developed thus far to a particularly important area of computer science: the theory and practice of hashing.

Let R, S, and T be finite, non-empty sets. Suppose that for each $r \in R$, we have a function $\Phi_r : S \to T$. We call Φ_r a **hash function (from S to T)**. Elements of R are called **keys**, and if $\Phi_r(s) = t$, we say that s **hashes to** t **under** r.

In applications of hash functions, we are typically interested in what happens when various inputs are hashed under a randomly chosen key. To model such situations, let H be a random variable that is uniformly distributed over R, and for each $s \in S$, let us define the random variable $\Phi_H(s)$, which takes the value $\Phi_r(s)$ when $H = r$.

- We say that the family of hash functions $\{\Phi_r\}_{r \in R}$ is **pairwise independent** if the family of random variables $\{\Phi_H(s)\}_{s \in S}$ is pairwise independent, with each $\Phi_H(s)$ uniformly distributed over T.
- We say that $\{\Phi_r\}_{r \in R}$ is **universal** if

$$\mathsf{P}[\Phi_H(s) = \Phi_H(s')] \leq 1/|T|$$

for all $s, s' \in S$ with $s \neq s'$.

We make a couple of simple observations. First, by Theorem 8.25, if the family of hash functions $\{\Phi_r\}_{r \in R}$ is pairwise independent, then it is universal. Second, by Theorem 8.10, if $|S| > 1$, then $\{\Phi_r\}_{r \in R}$ is pairwise independent if and only if the following condition holds:

> the random variable $(\Phi_H(s), \Phi_H(s'))$ is uniformly distributed over $T \times T$, for all $s, s' \in S$ with $s \neq s'$;

or equivalently,

> $\mathsf{P}[\Phi_H(s) = t \cap \Phi_H(s') = t'] = 1/|T|^2$ for all $s, s' \in S$ with $s \neq s'$, and for all $t, t' \in T$.

Before looking at constructions of pairwise independent and universal families of hash functions, we briefly discuss two important applications.

Example 8.34. Suppose $\{\Phi_r\}_{r \in R}$ is a *universal* family of hash functions from S to T. One can implement a "dictionary" using a so-called **hash table**, which is basically an array A indexed by T, where each entry in A is a list. Entries in the dictionary are drawn from the set S. To insert a word $s \in S$ into the dictionary, s is first hashed to an index t, and then s is appended to the list $A[t]$; likewise, to see if an arbitrary word $s \in S$ is in the dictionary, s is first hashed to an index t, and then the list $A[t]$ is searched for s.

Usually, the set of entries in the dictionary is much smaller than the set S. For

example, S may consist of all bit strings of length up to, say 2048, but the dictionary may contain just a few thousand, or a few million, entries. Also, to be practical, the set T should not be too large.

Of course, all entries in the dictionary could end up hashing to the same index, in which case, looking up a word in the dictionary degenerates into linear search. However, we hope that this does not happen, and that entries hash to indices that are nicely spread out over T. As we will now see, in order to ensure reasonable performance (in an expected sense), T needs to be of size roughly equal to the number of entries in the dictionary,

Suppose we create a dictionary containing n entries. Let $m := |T|$, and let $I \subseteq S$ be the set of entries (so $n = |I|$). These n entries are inserted into the hash table using a randomly chosen hash key, which we model as a random variable H that is uniformly distributed over R. For each $s \in S$, we define the random variable L_s to be the number of entries in I that hash to the same index as s under the key H; that is, $L_s := |\{i \in I : \Phi_H(s) = \Phi_H(i)\}|$. Intuitively, L_s measures the cost of looking up the particular word s in the dictionary. We want to bound $\mathsf{E}[L_s]$. To this end, we write L_s as a sum of indicator variables: $L_s = \sum_{i \in I} C_{si}$, where C_{si} is the indicator variable for the event that $\Phi_H(s) = \Phi_H(i)$. By Theorem 8.16, we have $\mathsf{E}[C_{si}] = \mathsf{P}[\Phi_H(s) = \Phi_H(i)]$; moreover, by the universal property, $\mathsf{E}[C_{si}] \leq 1/m$ if $s \neq i$, and clearly, $\mathsf{E}[C_{si}] = 1$ if $s = i$. By linearity of expectation, we have

$$\mathsf{E}[L_s] = \sum_{i \in I} \mathsf{E}[C_{si}].$$

If $s \notin I$, then each term in the sum is $\leq 1/m$, and so $\mathsf{E}[L_s] \leq n/m$. If $s \in I$, then one term in the sum is 1, and the other $n - 1$ terms are $\leq 1/m$, and so $\mathsf{E}[L_s] \leq 1 + (n-1)/m$. In any case, we have

$$\mathsf{E}[L_s] \leq 1 + n/m.$$

In particular, this means that if $m \geq n$, then the expected cost of looking up any particular word in the dictionary is bounded by a constant. □

Example 8.35. Suppose Alice wants to send a message to Bob in such a way that Bob can be reasonably sure that the message he receives really came from Alice, and was not modified in transit by some malicious adversary. We present a solution to this problem here that works assuming that Alice and Bob share a randomly generated secret key, and that this key is used to authenticate just a single message (multiple messages can be authenticated using multiple keys).

Suppose that $\{\Phi_r\}_{r \in R}$ is a *pairwise independent* family of hash functions from S to T. We model the shared random key as a random variable H, uniformly distributed over R. We also model Alice's message as a random variable X, taking values in the set S. We make no assumption about the distribution of X, but we do

assume that X and H are independent. When Alice sends the message X to Bob, she also sends the "authentication tag" $Y := \Phi_H(X)$. Now, when Bob receives a message X' and tag Y', he checks that $\Phi_H(X') = Y'$; if this holds, he accepts the message X' as authentic; otherwise, he rejects it. Here, X' and Y' are also random variables; however, they may have been created by a malicious adversary who may have even created them after seeing X and Y. We can model such an adversary as a pair of functions $f : S \times T \to S$ and $g : S \times T \to T$, so that $X' := f(X, Y)$ and $Y' := g(X, Y)$. The idea is that after seeing X and Y, the adversary computes X' and Y' and sends X' and Y' to Bob instead of X and Y. Let us say that the adversary *fools* Bob if $\Phi_H(X') = Y'$ and $X' \neq X$. We will show that $\mathsf{P}[\mathcal{F}] \leq 1/m$, where $\mathcal{F}$ is the event that the adversary fools Bob, and $m := |T|$. Intuitively, this bound holds because the pairwise independence property guarantees that after seeing the value of Φ_H at one input, the value of Φ_H at any other input is completely unpredictable, and cannot be guessed with probability any better than $1/m$. If m is chosen to be suitably large, the probability that Bob gets fooled can be made acceptably small. For example, S may consist of all bit strings of length up to, say, 2048, while the set T may be encoded using much shorter bit strings, of length, say, 64. This is nice, as it means that the authentication tags consume very little additional bandwidth.

A straightforward calculation justifies the claim that $\mathsf{P}[\mathcal{F}] \leq 1/m$:

$$\begin{aligned}
\mathsf{P}[\mathcal{F}] &= \sum_{s \in S} \sum_{t \in T} \mathsf{P}\Big[(X = s) \cap (Y = t) \cap \mathcal{F}\Big] \quad \text{(law of total probability (8.9))} \\
&= \sum_{s \in S} \sum_{t \in T} \mathsf{P}\Big[(X = s) \cap (\Phi_H(s) = t) \cap (\Phi_H(f(s,t)) = g(s,t)) \cap \\
&\qquad\qquad (f(s,t) \neq s)\Big] \\
&= \sum_{s \in S} \sum_{t \in T} \mathsf{P}[X = s]\, \mathsf{P}\Big[(\Phi_H(s) = t) \cap (\Phi_H(f(s,t)) = g(s,t)) \cap \\
&\qquad\qquad (f(s,t) \neq s)\Big] \quad \text{(since } X \text{ and } H \text{ are independent)} \\
&\leq \sum_{s \in S} \sum_{t \in T} \mathsf{P}[X = s] \cdot (1/m^2) \quad \text{(since } \{\Phi_r\}_{r \in R} \text{ is pairwise independent)} \\
&= (1/m) \sum_{s \in S} \mathsf{P}[X = s] = 1/m. \quad \square
\end{aligned}$$

We now present several constructions of pairwise independent and universal families of hash functions.

Example 8.36. By setting $k := 2$ in Example 8.27, for each prime p, we immediately get a pairwise independent family of hash functions $\{\Phi_r\}_{r \in R}$ from $\mathbb{Z}_p$ to $\mathbb{Z}_p$,

where $R = \mathbb{Z}_p \times \mathbb{Z}_p$, and for $r = (r_0, r_1) \in R$, the hash function Φ_r is given by

$$\begin{aligned} \Phi_r : \quad & \mathbb{Z}_p \to \mathbb{Z}_p \\ & s \mapsto r_0 + r_1 s. \end{aligned}$$ □

While very simple and elegant, the family of hash functions in Example 8.36 is not very useful in practice. As we saw in Examples 8.34 and 8.35, what we would really like are families of hash functions that hash *long* inputs to *short* outputs. The next example provides us with a pairwise independent family of hash functions that satisfies this requirement.

Example 8.37. Let p be a prime, and let ℓ be a positive integer. Let $S := \mathbb{Z}_p^{\times \ell}$ and $R := \mathbb{Z}_p^{\times(\ell+1)}$. For each $r = (r_0, r_1, \ldots, r_\ell) \in R$, we define the hash function

$$\begin{aligned} \Phi_r : \quad & S \to \mathbb{Z}_p \\ & (s_1, \ldots, s_\ell) \mapsto r_0 + r_1 s_1 + \cdots + r_\ell s_\ell. \end{aligned}$$

We will show that $\{\Phi_r\}_{r \in R}$ is a pairwise independent family of hash functions from S to $\mathbb{Z}_p$. To this end, let H be a random variable uniformly distributed over R. We want to show that for each $s, s' \in S$ with $s \neq s'$, the random variable $(\Phi_H(s), \Phi_H(s'))$ is uniformly distributed over $\mathbb{Z}_p \times \mathbb{Z}_p$. So let $s \neq s'$ be fixed, and define the function

$$\begin{aligned} \rho : \quad & R \to \mathbb{Z}_p \times \mathbb{Z}_p \\ & r \mapsto (\Phi_r(s), \Phi_r(s')). \end{aligned}$$

Because ρ is a group homomorphism, it will suffice to show that ρ is surjective (see Theorem 8.5). Suppose $s = (s_1, \ldots, s_\ell)$ and $s' = (s'_1, \ldots, s'_\ell)$. Since $s \neq s'$, we must have $s_j \neq s'_j$ for some $j = 1, \ldots, \ell$. For this j, consider the function

$$\begin{aligned} \rho' : \quad & R \to \mathbb{Z}_p \times \mathbb{Z}_p \\ & (r_0, r_1, \ldots, r_\ell) \mapsto (r_0 + r_j s_j, r_0 + r_j s'_j). \end{aligned}$$

Evidently, the image of ρ includes the image of ρ', and by Example 8.36, the function ρ' is surjective. □

To use the construction in Example 8.37 in applications where the set of inputs consists of bit strings of a given length, one can naturally split such a bit string up into short bit strings which, when viewed as integers, lie in the set $\{0, \ldots, p-1\}$, and which can in turn be viewed as elements of $\mathbb{Z}_p$. This gives us a natural, injective map from bit strings to elements of $\mathbb{Z}_p^{\times \ell}$. The appropriate choice of the prime p depends on the application. Of course, the requirement that p is prime limits our choice in the size of the output set; however, this is usually not a severe restriction, as Bertrand's postulate (Theorem 5.8) tells us that we can always choose p

to within a factor of 2 of any desired value of the output set size. Nevertheless, the construction in the following example gives us a *universal* (but not pairwise independent) family of hash functions with an output set of any size we wish.

Example 8.38. Let p be a prime, and let m be an arbitrary positive integer. Let us introduce some convenient notation: for $\alpha \in \mathbb{Z}_p$, let $[\![\alpha]\!]_m := [\mathrm{rep}(\alpha)]_m \in \mathbb{Z}_m$ (recall that $\mathrm{rep}(\alpha)$ denotes the unique integer $a \in \{0, \ldots, p-1\}$ such that $\alpha = [a]_p$). Let $R := \mathbb{Z}_p \times \mathbb{Z}_p^*$, and for each $r = (r_0, r_1) \in R$, define the hash function

$$\begin{aligned} \Phi_r : \quad & \mathbb{Z}_p \to \mathbb{Z}_m \\ & s \mapsto [\![r_0 + r_1 s]\!]_m. \end{aligned}$$

Our goal is to show that $\{\Phi_r\}_{r \in R}$ is a universal family of hash functions from $\mathbb{Z}_p$ to $\mathbb{Z}_m$. So let $s, s' \in \mathbb{Z}_p$ with $s \neq s'$, let H_0 and H_1 be independent random variables, with H_0 uniformly distributed over $\mathbb{Z}_p$ and H_1 uniformly distributed over $\mathbb{Z}_p^*$, and let $H := (H_0, H_1)$. Also, let C be the event that $\Phi_H(s) = \Phi_H(s')$. We want to show that $\mathsf{P}[C] \leq 1/m$. Let us define random variables $Y := H_0 + H_1 s$ and $Y' := H_0 + H_1 s'$. Also, let $\hat{s} := s' - s \neq 0$. Then we have

$$\begin{aligned} \mathsf{P}[C] &= \mathsf{P}\Big[[\![Y]\!]_m = [\![Y']\!]_m\Big] \\ &= \mathsf{P}\Big[[\![Y]\!]_m = [\![Y + H_1\hat{s}]\!]_m\Big] \quad (\text{since } Y' = Y + H_1\hat{s}) \\ &= \sum_{\alpha \in \mathbb{Z}_p} \mathsf{P}\Big[\big([\![Y]\!]_m = [\![Y + H_1\hat{s}]\!]_m\big) \cap (Y = \alpha)\Big] \quad (\text{law of total probability (8.9)}) \\ &= \sum_{\alpha \in \mathbb{Z}_p} \mathsf{P}\Big[\big([\![\alpha]\!]_m = [\![\alpha + H_1\hat{s}]\!]_m\big) \cap (Y = \alpha)\Big] \\ &= \sum_{\alpha \in \mathbb{Z}_p} \mathsf{P}\Big[[\![\alpha]\!]_m = [\![\alpha + H_1\hat{s}]\!]_m\Big]\, \mathsf{P}[Y = \alpha] \\ &\qquad (\text{by Theorem 8.13, } Y \text{ and } H_1 \text{ are independent}). \end{aligned}$$

It will suffice to show that

$$\mathsf{P}\Big[[\![\alpha]\!]_m = [\![\alpha + H_1\hat{s}]\!]_m\Big] \leq 1/m \tag{8.33}$$

for each $\alpha \in \mathbb{Z}_p$, since then

$$\mathsf{P}[C] \leq \sum_{\alpha \in \mathbb{Z}_p} (1/m)\, \mathsf{P}[Y = \alpha] = (1/m) \sum_{\alpha \in \mathbb{Z}_p} \mathsf{P}[Y = \alpha] = 1/m.$$

So consider a fixed $\alpha \in \mathbb{Z}_p$. As $\hat{s} \neq 0$ and H_1 is uniformly distributed over $\mathbb{Z}_p^*$, it follows that $H_1\hat{s}$ is uniformly distributed over $\mathbb{Z}_p^*$, and hence $\alpha + H_1\hat{s}$ is uniformly distributed over the set $\mathbb{Z}_p \setminus \{\alpha\}$. Let $M_\alpha := \{\beta \in \mathbb{Z}_p : [\![\alpha]\!]_m = [\![\beta]\!]_m\}$. To prove

(8.33), we need to show that $|M_\alpha \setminus \{\alpha\}| \leq (p-1)/m$. But it is easy to see that $|M_\alpha| \leq \lceil p/m \rceil$, and since M_α certainly contains α, we have

$$|M_\alpha \setminus \{\alpha\}| \leq \left\lceil \frac{p}{m} \right\rceil - 1 \leq \frac{p}{m} + \frac{m-1}{m} - 1 = \frac{p-1}{m}. \quad \square$$

One drawback of the family of hash functions in the previous example is that the prime p may need to be quite large (at least as large as the size of the set of inputs) and so to evaluate a hash function, we have to perform modular multiplication of large integers. In contrast, in Example 8.37, the prime p can be much smaller (only as large as the size of the set of outputs), and so these hash functions can be evaluated much more quickly.

Another consideration in designing families of hash functions is the size of key set. The following example gives a variant of the family in Example 8.37 that uses somewhat a smaller key set (relative to the size of the input), but is only a universal family, and not a pairwise independent family.

Example 8.39. Let p be a prime, and let ℓ be a positive integer. Let $S := \mathbb{Z}_p^{\times(\ell+1)}$ and $R := \mathbb{Z}_p^{\times \ell}$. For each $r = (r_1, \ldots, r_\ell) \in R$, we define the hash function

$$\begin{aligned} \Phi_r : \quad & S \to \mathbb{Z}_p \\ & (s_0, s_1, \ldots, s_\ell) \mapsto s_0 + r_1 s_1 + \cdots + r_\ell s_\ell. \end{aligned}$$

Our goal is to show that $\{\Phi_r\}_{r \in R}$ is a universal family of hash functions from S to $\mathbb{Z}_p$. So let $s, s' \in S$ with $s \neq s'$, and let H be a random variable that is uniformly distributed over R. We want to show that $\mathsf{P}[\Phi_H(s) = \Phi_H(s')] \leq 1/p$. Let $s = (s_0, s_1, \ldots, s_\ell)$ and $s' = (s'_0, s'_1, \ldots, s'_\ell)$, and set $\hat{s}_i := s'_i - s_i$ for $i = 0, 1, \ldots, \ell$. Let us define the function

$$\begin{aligned} \rho : \quad & R \to \mathbb{Z}_p \\ & (r_1, \ldots, r_\ell) \mapsto r_1 \hat{s}_1 + \cdots + r_\ell \hat{s}_\ell. \end{aligned}$$

Clearly, $\Phi_H(s) = \Phi_H(s')$ if and only if $\rho(H) = -\hat{s}_0$. Moreover, ρ is a group homomorphism. There are two cases to consider. In the first case, $\hat{s}_i = 0$ for all $i = 1, \ldots, \ell$; in this case, the image of ρ is $\{0\}$, but $\hat{s}_0 \neq 0$ (since $s \neq s'$), and so $\mathsf{P}[\rho(H) = -\hat{s}_0] = 0$. In the second case, $\hat{s}_i \neq 0$ for some $i = 1, \ldots, \ell$; in this case, the image of ρ is $\mathbb{Z}_p$, and so $\rho(H)$ is uniformly distributed over $\mathbb{Z}_p$ (see Theorem 8.5); thus, $\mathsf{P}[\rho(H) = -\hat{s}_0] = 1/p$. $\square$

One can get significantly smaller key sets, if one is willing to relax the definitions of universal and pairwise independence. Let $\{\Phi_r\}_{r \in R}$ be a family of hash functions from S to T, where $m := |T|$. Let H be a random variable that is uniformly distributed over R. We say that $\{\Phi_r\}_{r \in R}$ is **ε-almost universal** if for all $s, s' \in S$ with $s \neq s'$, we have $\mathsf{P}[\Phi_H(s) = \Phi_H(s')] \leq \varepsilon$. Thus, $\{\Phi_r\}_{r \in R}$ is

universal if and only if it is $1/m$-almost universal. We say that $\{\Phi_r\}_{r\in R}$ is **ε-almost strongly universal** if $\Phi_H(s)$ is uniformly distributed over T for each $s \in S$, and $\mathsf{P}[(\Phi_H(s) = t) \cap (\Phi_H(s') = t')] \leq \varepsilon/m$ for all $s, s' \in S$ with $s \neq s'$ and all $t, t' \in T$. Constructions, properties, and applications of these types of hash functions are developed in some of the exercises below.

EXERCISE 8.47. For each positive integer n, let I_n denote $\{0, \ldots, n-1\}$. Let m be a power of a prime, ℓ be a positive integer, $S := I_m^{\times \ell}$, and $R := I_{m^2}^{\times(\ell+1)}$. For each $r = (r_0, r_1, \ldots, r_\ell) \in R$, define the hash function

$$\begin{aligned} \Phi_r : \quad & S \to I_m \\ & (s_1, \ldots, s_\ell) \mapsto \Big\lfloor \Big((r_0 + r_1 s_1 + \cdots + r_\ell s_\ell) \bmod m^2 \Big) \Big/ m \Big\rfloor. \end{aligned}$$

Using the result from Exercise 2.13, show that $\{\Phi_r\}_{r\in R}$ is a pairwise independent family of hash functions from S to I_m. Note that on a typical computer, if m is a suitable power of 2, then it is very easy to evaluate these hash functions, using just multiplications, additions, shifts, and masks (no divisions).

EXERCISE 8.48. Let $\{\Phi_r\}_{r\in R}$ be an ε-almost universal family of hash functions from S to T. Also, let H, X, X' be random variables, where H is uniformly distributed over R, and both X and X' take values in S. Moreover, assume H and (X, X') are independent. Show that $\mathsf{P}[\Phi_H(X) = \Phi_H(X')] \leq \mathsf{P}[X = X'] + \varepsilon$.

EXERCISE 8.49. Let $\{\Phi_r\}_{r\in R}$ be an ε-almost universal a family of hash functions from S to T, and let H be a random variable that is uniformly distributed over R. Let I be a subset of S of size $n > 0$. Let C be the event that $\Phi_H(i) = \Phi_H(j)$ for some $i, j \in I$ with $i \neq j$. We define several random variables: for each $t \in T$, $N_t := |\{i \in I : \Phi_H(i) = t\}|$; $M := \max\{N_t : t \in T\}$; for each $s \in S$, $L_s := |\{i \in I : \Phi_H(s) = \Phi_H(i)\}|$. Show that:

(a) $\mathsf{P}[C] \leq \varepsilon n(n-1)/2$;

(b) $\mathsf{E}[M] \leq \sqrt{\varepsilon n^2 + n}$;

(c) for each $s \in S$, $\mathsf{E}[L_s] \leq 1 + \varepsilon n$.

The results of the previous exercise show that for many applications, the ε-almost universal property is good enough, provided ε is suitably small. The next three exercises develop ε-almost universal families of hash functions with very small sets of keys, even when ε is quite small.

EXERCISE 8.50. Let p be a prime, and let ℓ be a positive integer. Let $S := \mathbb{Z}_p^{\times(\ell+1)}$.

For each $r \in \mathbb{Z}_p$, define the hash function

$$\begin{aligned} \Phi_r : \quad & S \to \mathbb{Z}_p \\ & (s_0, s_1, \ldots, s_\ell) \mapsto s_0 + s_1 r + \cdots + s_\ell r^\ell. \end{aligned}$$

Show that $\{\Phi_r\}_{r \in \mathbb{Z}_p}$ is an ℓ/p-almost universal family of hash functions from S to $\mathbb{Z}_p$.

EXERCISE 8.51. Let $\{\Phi_r\}_{r \in R}$ be an ε-almost universal family of hash functions from S to T. Let $\{\Phi'_{r'}\}_{r' \in R'}$ be an ε'-almost universal family of hash functions from S' to T', where $T \subseteq S'$. Show that

$$\{\Phi'_{r'} \circ \Phi_r\}_{(r,r') \in R \times R'}$$

is an $(\varepsilon + \varepsilon')$-almost universal family of hash functions from S to T' (here, "$\circ$" denotes function composition).

EXERCISE 8.52. Let m and ℓ be positive integers, and let $0 < \alpha < 1$. Given these parameters, show how to construct an ε-almost universal family of hash functions $\{\Phi_r\}_{r \in R}$ from $\mathbb{Z}_m^{\times \ell}$ to $\mathbb{Z}_m$, such that

$$\varepsilon \leq (1 + \alpha)/m \quad \text{and} \quad \log|R| = O(\log m + \log \ell + \log(1/\alpha)).$$

Hint: use the previous two exercises, and Example 8.38.

EXERCISE 8.53. Let $\{\Phi_r\}_{r \in R}$ be an ε-almost universal family of hash functions from S to T. Show that $\varepsilon \geq 1/|T| - 1/|S|$.

EXERCISE 8.54. Let $\{\Phi_r\}_{r \in R}$ be a family of hash functions from S to T, with $m := |T|$. Show that:

(a) if $\{\Phi_r\}_{r \in R}$ is ε-almost strongly universal, then it is ε-almost universal;

(b) if $\{\Phi_r\}_{r \in R}$ is pairwise independent, then it is $1/m$-almost strongly universal;

(c) if $\{\Phi_r\}_{r \in R}$ is ε-almost universal, and $\{\Phi'_{r'}\}_{r' \in R'}$ is an ε'-almost strongly universal family of hash functions from S' to T', where $T \subseteq S'$, then $\{\Phi'_{r'} \circ \Phi_r\}_{(r,r') \in R \times R'}$ is an $(\varepsilon + \varepsilon')$-almost strongly universal family of hash functions from S to T'.

EXERCISE 8.55. Show that if an ε-almost strongly universal family of hash functions is used in Example 8.35, then Bob gets fooled with probability at most ε.

EXERCISE 8.56. Show how to construct an ε-almost strongly universal family of hash functions satisfying the same bounds as in Exercise 8.52, under the restriction that m is a prime power.

EXERCISE 8.57. Let p be a prime, and let ℓ be a positive integer. Let $S := \mathbb{Z}_p^{\times \ell}$ and $R := \mathbb{Z}_p \times \mathbb{Z}_p$. For each $(r_0, r_1) \in R$, define the hash function

$$\begin{aligned} \Phi_r : \quad & S \to \mathbb{Z}_p \\ & (s_1, \ldots, s_\ell) \mapsto r_0 + s_1 r_1 + \cdots + s_\ell r_1^\ell. \end{aligned}$$

Show that $\{\Phi_r\}_{r \in R}$ is an ℓ/p-almost strongly universal family of hash functions from S to $\mathbb{Z}_p$.

8.8 Statistical distance

This section discusses a useful measure of "distance" between two random variables. Although important in many applications, the results of this section (and the next) will play only a very minor role in the remainder of the text.

Let X and Y be random variables which both take values in a finite set S. We define the **statistical distance between X and Y** as

$$\Delta[X; Y] := \frac{1}{2} \sum_{s \in S} \big| \mathsf{P}[X = s] - \mathsf{P}[Y = s] \big|.$$

Theorem 8.30. *For random variables X, Y, Z, we have*

(i) $0 \leq \Delta[X; Y] \leq 1$,

(ii) $\Delta[X; X] = 0$,

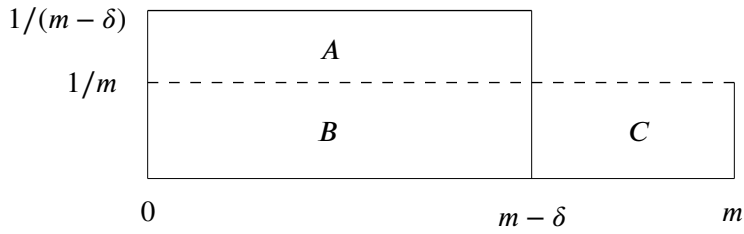
(iii) $\Delta[X; Y] = \Delta[Y; X]$, *and*

(iv) $\Delta[X; Z] \leq \Delta[X; Y] + \Delta[Y; Z]$.

Proof. Exercise. □

It is also clear from the definition that $\Delta[X; Y]$ depends only on the distributions of X and Y, and not on any other properties. As such, we may sometimes speak of the statistical distance between two distributions, rather than between two random variables.

Example 8.40. Suppose X has the uniform distribution on $\{1, \ldots, m\}$, and Y has the uniform distribution on $\{1, \ldots, m - \delta\}$, where $\delta \in \{0, \ldots, m - 1\}$. Let us compute $\Delta[X; Y]$. We could apply the definition directly; however, consider the following graph of the distributions of X and Y:



The statistical distance between X and Y is just $1/2$ times the area of regions A and C in the diagram. Moreover, because probability distributions sum to 1, we must have

$$\text{area of } B + \text{area of } A = 1 = \text{area of } B + \text{area of } C,$$

and hence, the areas of region A and region C are the same. Therefore,

$$\Delta[X; Y] = \text{area of } A = \text{area of } C = \delta/m. \quad \square$$

The following characterization of statistical distance is quite useful:

Theorem 8.31. *Let X and Y be random variables taking values in a set S. For every $S' \subseteq S$, we have*

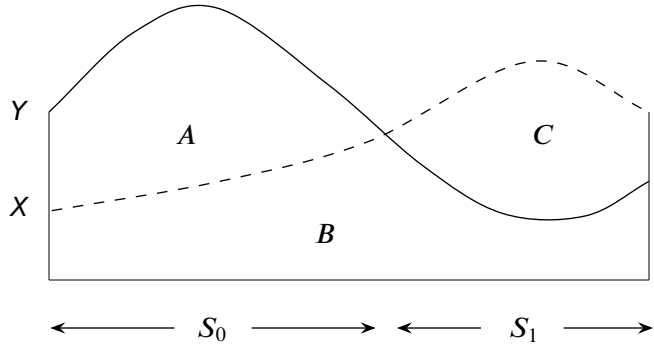
$$\Delta[X; Y] \geq |\mathsf{P}[X \in S'] - \mathsf{P}[Y \in S']|,$$

and equality holds for some $S' \subseteq S$, and in particular, for the set

$$S' := \{s \in S : \mathsf{P}[X = s] < \mathsf{P}[Y = s]\},$$

as well as its complement.

Proof. Suppose we split the set S into two disjoint subsets: the set S_0 consisting of those $s \in S$ such that $\mathsf{P}[X = s] < \mathsf{P}[Y = s]$, and the set S_1 consisting of those $s \in S$ such that $\mathsf{P}[X = s] \geq \mathsf{P}[Y = s]$. Consider the following rough graph of the distributions of X and Y, where the elements of S_0 are placed to the left of the elements of S_1:



Now, as in Example 8.40,

$$\Delta[X;Y] = \text{area of } A = \text{area of } C.$$

Now consider any subset S' of S, and observe that

$$\mathsf{P}[X \in S'] - \mathsf{P}[Y \in S'] = \text{area of } C' - \text{area of } A',$$

where C' is the subregion of C that lies above S', and A' is the subregion of A that lies above S'. It follows that $|\mathsf{P}[X \in S'] - \mathsf{P}[Y \in S']|$ is maximized when $S' = S_0$ or $S' = S_1$, in which case it is equal to $\Delta[X;Y]$. □

We can restate Theorem 8.31 as follows:

$$\Delta[X;Y] = \max\{|\mathsf{P}[\phi(X)] - \mathsf{P}[\phi(Y)]| : \phi \text{ is a predicate on } S\}.$$

This implies that when $\Delta[X;Y]$ is very small, then for *every* predicate ϕ, the events $\phi(X)$ and $\phi(Y)$ occur with almost the same probability. Put another way, there is no "statistical test" that can effectively distinguish between the distributions of X and Y. For many applications, this means that the distribution of X is "for all practical purposes" equivalent to that of Y, and hence in analyzing the behavior of X, we can instead analyze the behavior of Y, if that is more convenient.

Theorem 8.32. *If S and T are finite sets, X and Y are random variables taking values in S, and $f : S \to T$ is a function, then $\Delta[f(X); f(Y)] \leq \Delta[X;Y]$.*

Proof. We have

$$\begin{aligned}\Delta[f(X); f(Y)] &= |\mathsf{P}[f(X) \in T'] - \mathsf{P}[f(Y) \in T']| \text{ for some } T' \subseteq T \\ &\qquad \text{(by Theorem 8.31)} \\ &= |\mathsf{P}[X \in f^{-1}(T')] - \mathsf{P}[Y \in f^{-1}(T')]| \\ &\leq \Delta[X;Y] \text{ (again by Theorem 8.31).}\ \square\end{aligned}$$

Example 8.41. Let X be uniformly distributed over the set $\{0,\ldots,m-1\}$, and let Y be uniformly distributed over the set $\{0,\ldots,n-1\}$, for $n \geq m$. Let $f(t) := t \bmod m$. We want to compute an upper bound on the statistical distance between X and $f(Y)$. We can do this as follows. Let $n = qm - r$, where $0 \leq r < m$, so that $q = \lceil n/m \rceil$. Also, let Z be uniformly distributed over $\{0,\ldots,qm-1\}$. Then $f(Z)$ is uniformly distributed over $\{0,\ldots,m-1\}$, since every element of $\{0,\ldots,m-1\}$ has the same number (namely, q) of pre-images under f which lie in the set $\{0,\ldots,qm-1\}$. Since statistical distance depends only on the distributions of the random variables, by the previous theorem, we have

$$\Delta[X; f(Y)] = \Delta[f(Z); f(Y)] \leq \Delta[Z;Y],$$

and as we saw in Example 8.40,

$$\Delta[Z; Y] = r/qm < 1/q \leq m/n.$$

Therefore,

$$\Delta[X; f(Y)] < m/n. \quad \square$$

We close this section with two useful theorems.

Theorem 8.33. *Suppose X, Y, and Z are random variables, where X and Z are independent, and Y and Z are independent. Then $\Delta[X, Z; Y, Z] = \Delta[X, Y]$.*

Note that $\Delta[X, Z; Y, Z]$ is shorthand for $\Delta[(X, Z); (Y, Z)]$.

Proof. Suppose X and Y take values in a finite set S, and Z takes values in a finite set T. From the definition of statistical distance,

$$\begin{aligned}
2\Delta[X, Z; Y, Z] &= \sum_{s,t} \Big| \mathsf{P}[(X = s) \cap (Z = t)] - \mathsf{P}[(Y = s) \cap (Z = t)] \Big| \\
&= \sum_{s,t} \Big| \mathsf{P}[X = s]\,\mathsf{P}[Z = t] - \mathsf{P}[Y = s]\,\mathsf{P}[Z = t] \Big| \\
&\quad \text{(by independence)} \\
&= \sum_{s,t} \mathsf{P}[Z = t] \Big| \mathsf{P}[X = s] - \mathsf{P}[Y = s] \Big| \\
&= \Big(\sum_{t} \mathsf{P}[Z = t] \Big) \Big(\sum_{s} \Big| \mathsf{P}[X = s] - \mathsf{P}[Y = s] \Big| \Big) \\
&= 1 \cdot 2\Delta[X; Y]. \quad \square
\end{aligned}$$

Theorem 8.34. *Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be random variables, where $\{X_i\}_{i=1}^n$ is mutually independent, and $\{Y_i\}_{i=1}^n$ is mutually independent. Then we have*

$$\Delta[X_1, \ldots, X_n; Y_1, \ldots, Y_n] \leq \sum_{i=1}^{n} \Delta[X_i; Y_i].$$

Proof. Since $\Delta[X_1, \ldots, X_n; Y_1, \ldots, Y_n]$ depends only on the individual distributions of the random variables $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, without loss of generality, we may assume that $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are independent, so that $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ form a mutually independent family of random variables. We introduce random variables $Z_0, \ldots, Z_n$, defined as follows:

$$\begin{aligned}
Z_0 &:= (X_1, \ldots, X_n), \\
Z_i &:= (Y_1, \ldots, Y_i, X_{i+1}, \ldots, X_n) \quad \text{for } i = 1, \ldots, n-1, \text{ and} \\
Z_n &:= (Y_1, \ldots, Y_n).
\end{aligned}$$

By definition, $\Delta[X_1, \ldots, X_n; Y_1, \ldots, Y_n] = \Delta[Z_0; Z_n]$. Moreover, by part (iv) of Theorem 8.30, we have $\Delta[Z_0; Z_n] \leq \sum_{i=1}^{n} \Delta[Z_{i-1}; Z_i]$. Now consider any fixed index $i = 1, \ldots, n$. By Theorem 8.33, we have

$$\begin{aligned}\Delta[Z_{i-1}; Z_i] &= \Delta[\, X_i,\ (Y_1, \ldots, Y_{i-1}, X_{i+1}, \ldots, X_n);\\ &\qquad\qquad Y_i,\ (Y_1, \ldots, Y_{i-1}, X_{i+1}, \ldots, X_n)]\\ &= \Delta[X_i; Y_i].\end{aligned}$$

The theorem now follows immediately. □

The technique used in the proof of the previous theorem is sometimes called a **hybrid argument**, as one considers the sequence of "hybrid" random variables $Z_0, Z_1, \ldots, Z_n$, and shows that the distance between each consecutive pair of variables is small.

EXERCISE 8.58. Let X and Y be independent random variables, each uniformly distributed over $\mathbb{Z}_p$, where p is prime. Calculate $\Delta[X, Y; X, XY]$.

EXERCISE 8.59. Let n be an integer that is the product of two distinct primes of the same bit length. Let X be uniformly distributed over $\mathbb{Z}_n$, and let Y be uniformly distributed over $\mathbb{Z}_n^*$. Show that $\Delta[X; Y] \leq 3n^{-1/2}$.

EXERCISE 8.60. Let X and Y be 0/1-valued random variables. Show that

$$\Delta[X; Y] = |\mathsf{P}[X = 1] - \mathsf{P}[Y = 1]|.$$

EXERCISE 8.61. Let S be a finite set, and consider any function $\phi : S \to \{0, 1\}$. Let B be a random variable uniformly distributed over $\{0, 1\}$, and for $b = 0, 1$, let X_b be a random variable taking values in S, and assume that X_b and B are independent. Show that

$$|\mathsf{P}[\phi(X_B) = B] - \tfrac{1}{2}| = \tfrac{1}{2}|\mathsf{P}[\phi(X_0) = 1] - \mathsf{P}[\phi(X_1) = 1]| \leq \tfrac{1}{2}\Delta[X_0; X_1].$$

EXERCISE 8.62. Let X, Y be random variables taking values in a finite set S. For an event $\mathcal{B}$ that occurs with non-zero probability, define the **conditional statistical distance**

$$\Delta[X; Y \mid \mathcal{B}] := \frac{1}{2} \sum_{s \in S} \Big| \mathsf{P}[X = s \mid \mathcal{B}] - \mathsf{P}[Y = s \mid \mathcal{B}] \Big|.$$

Let $\{\mathcal{B}_i\}_{i \in I}$ be a finite, pairwise disjoint family of events whose union is $\mathcal{B}$. Show that

$$\Delta[X; Y \mid \mathcal{B}]\, \mathsf{P}[\mathcal{B}] \leq \sum_{\mathsf{P}[\mathcal{B}_i] \neq 0} \Delta[X; Y \mid \mathcal{B}_i]\, \mathsf{P}[\mathcal{B}_i].$$

EXERCISE 8.63. Let $\{\Phi_r\}_{r\in R}$ be a family of hash functions from S to T, with $m := |T|$. We say $\{\Phi_r\}_{r\in R}$ is **ε-variationally universal** if $\Phi_H(s)$ is uniformly distributed over T for each $s \in S$, and $\Delta[\Phi_H(s'); Y \mid \Phi_H(s) = t] \leq \varepsilon$ for each $s, s' \in S$ with $s \neq s'$ and each $t \in T$; here, H and Y are independent random variables, with H uniformly distributed over R, and Y uniformly distributed over T. Show that:

(a) if $\{\Phi_r\}_{r\in R}$ is pairwise independent, then it is 0-variationally universal;

(b) if $\{\Phi_r\}_{r\in R}$ is ε-variationally universal, then it is $(1/m + \varepsilon)$-almost strongly universal;

(c) if $\{\Phi_r\}_{r\in R}$ is ε-almost universal, and $\{\Phi'_{r'}\}_{r'\in R'}$ is an ε'-variationally universal family of hash functions from S' to T', where $T \subseteq S'$, then $\{\Phi'_{r'} \circ \Phi_r\}_{(r,r')\in R\times R'}$ is an $(\varepsilon + \varepsilon')$-variationally universal family of hash functions from S to T'.

EXERCISE 8.64. Let $\{\Phi_r\}_{r\in R}$ be a family hash functions from S to T such that (i) each Φ_r maps S injectively into T, and (ii) there exists $\varepsilon \in [0, 1]$ such that $\Delta[\Phi_H(s); \Phi_H(s')] \leq \varepsilon$ for all $s, s' \in S$, where H is uniformly distributed over R. Show that $|R| \geq (1 - \varepsilon)|S|$.

EXERCISE 8.65. Let X and Y be random variables that take the same value unless a certain event $\mathcal{F}$ occurs (i.e., $X(\omega) = Y(\omega)$ for all $\omega \in \overline{\mathcal{F}}$). Show that $\Delta[X; Y] \leq \mathsf{P}[\mathcal{F}]$.

EXERCISE 8.66. Let X and Y be random variables taking values in the interval $[0, t]$. Show that $|\mathsf{E}[X] - \mathsf{E}[Y]| \leq t \cdot \Delta[X; Y]$.

EXERCISE 8.67. Show that Theorem 8.33 is not true if we drop the independence assumptions.

EXERCISE 8.68. Let S be a set of size $m \geq 1$. Let F be a random variable that is uniformly distributed over the set of all functions from S into S. Let G be a random variable that is uniformly distributed over the set of all permutations of S. Let $s_1, \ldots, s_n$ be distinct, fixed elements of S. Show that

$$\Delta[F(s_1), \ldots, F(s_n); G(s_1), \ldots, G(s_n)] \leq \frac{n(n-1)}{2m}.$$

EXERCISE 8.69. Let m be a large integer. Consider three random experiments. In the first, we generate a random integer X_1 between 1 and m, and then a random integer Y_1 between 1 and X_1. In the second, we generate a random integer X_2 between 2 and m, and then generate a random integer Y_2 between 1 and X_2. In the third, we generate a random integer X_3 between 2 and m, and then a random integer Y_3

between 2 and X_3. Show that $\Delta[X_1, Y_1; X_2, Y_2] = O(1/m)$ and $\Delta[X_2, Y_2; X_3, Y_3] = O(\log m/m)$, and conclude that $\Delta[X_1, Y_1; X_3, Y_3] = O(\log m/m)$.

8.9 Measures of randomness and the leftover hash lemma (∗)

In this section, we discuss different ways to measure "how random" the distribution of a random variable is, and relations among them.

Let X be a random variable taking values in a finite set S of size m. We define three measures of randomness:

1. the **collision probability of** X is $\sum_{s \in S} \mathsf{P}[X = s]^2$;
2. the **guessing probability of** X is $\max\{\mathsf{P}[X = s] : s \in S\}$;
3. the **distance of** X **from uniform on** S is $\frac{1}{2} \sum_{s \in S} |\mathsf{P}[X = s] - 1/m|$.

Suppose X has collision probability β, guessing probability γ, and distance δ from uniform on S. If X' is another random variable with the same distribution as X, where X and X' independent, then $\beta = \mathsf{P}[X = X']$ (see Exercise 8.37). If Y is a random variable that is uniformly distributed over S, then $\delta = \Delta[X; Y]$. If X itself is uniformly distributed over S, then $\beta = \gamma = 1/m$, and $\delta = 0$. The quantity $\log_2(1/\gamma)$ is sometimes called the **min entropy** of X, and the quantity $\log_2(1/\beta)$ is sometimes called the **Renyi entropy** of X.

We first state some easy inequalities:

Theorem 8.35. *Suppose X is a random variable that takes values in a finite set S of size m. If X has collision probability β, guessing probability γ, and distance δ from uniform on S, then:*

(i) $\beta \geq 1/m$;

(ii) $\gamma^2 \leq \beta \leq \gamma \leq 1/m + \delta$.

Proof. Part (i) is immediate from Exercise 8.37. The other inequalities are left as easy exercises. □

This theorem implies that the collision and guessing probabilities are minimal for the uniform distribution, which perhaps agrees with one's intuition.

While the above theorem implies that β and γ are close to $1/m$ when δ is small, the following theorem provides a converse:

Theorem 8.36. *Suppose X is a random variable that takes values in a finite set S of size m. If X has collision probability β, and distance δ from uniform on S, then* $\delta \leq \frac{1}{2}\sqrt{m\beta - 1}$.

Proof. We may assume that $\delta > 0$, since otherwise the theorem is already true, simply from the fact that $\beta \geq 1/m$.

For $s \in S$, let $p_s := \mathsf{P}[X = s]$. We have $\delta = \frac{1}{2}\sum_s |p_s - 1/m|$, and hence $1 = \sum_s q_s$, where $q_s := |p_s - 1/m|/2\delta$. So we have

$$\begin{aligned}
\frac{1}{m} &\leq \sum_s q_s^2 \quad \text{(by Exercise 8.36)} \\
&= \frac{1}{4\delta^2} \sum_s (p_s - 1/m)^2 \\
&= \frac{1}{4\delta^2} \Big(\sum_s p_s^2 - 1/m \Big) \quad \text{(again by Exercise 8.36)} \\
&= \frac{1}{4\delta^2} (\beta - 1/m),
\end{aligned}$$

from which the theorem follows immediately. □

We are now in a position to state and prove a very useful result which, intuitively, allows us to convert a "low quality" source of randomness into a "high quality" source of randomness, making use of an almost universal family of hash functions (see end of §8.7).

Theorem 8.37 (Leftover hash lemma). *Let $\{\Phi_r\}_{r \in R}$ be a $(1+\alpha)/m$-almost universal family of hash functions from S to T, where $m := |T|$. Let H and X be independent random variables, where H is uniformly distributed over R, and X takes values in S. If β is the collision probability of X, and δ' is the distance of $(H, \Phi_H(X))$ from uniform on $R \times T$, then $\delta' \leq \frac{1}{2}\sqrt{m\beta + \alpha}$.*

Proof. Let β' be the collision probability of $(H, \Phi_H(X))$. Our goal is to bound β' from above, and then apply Theorem 8.36 to the random variable $(H, \Phi_H(X))$. To this end, let $\ell := |R|$, and suppose H' and X' are random variables, where H' has the same distribution as H, X' has the same distribution as X, and H, H', X, X' form a mutually independent family of random variables. Then we have

$$\begin{aligned}
\beta' &= \mathsf{P}[(H = H') \cap (\Phi_H(X) = \Phi_{H'}(X'))] \\
&= \mathsf{P}[(H = H') \cap (\Phi_H(X) = \Phi_H(X'))] \\
&= \frac{1}{\ell} \mathsf{P}[\Phi_H(X) = \Phi_H(X')] \quad \text{(a special case of Exercise 8.15)} \\
&\leq \frac{1}{\ell} (\mathsf{P}[X = X'] + (1+\alpha)/m) \quad \text{(by Exercise 8.48)} \\
&= \frac{1}{\ell m} (m\beta + 1 + \alpha).
\end{aligned}$$

The theorem now follows immediately from Theorem 8.36. □

In the previous theorem, if $\{\Phi_r\}_{r \in R}$ is a universal family of hash functions, then

we can take $\alpha = 0$. However, it is convenient to allow $\alpha > 0$, as this allows for the use of families with a smaller key set (see Exercise 8.52).

Example 8.42. Suppose $S := \{0,1\}^{\times 1000}$, $T := \{0,1\}^{\times 64}$, and that $\{\Phi_r\}_{r \in R}$ is a universal family of hash functions from S to T. Suppose X and H are independent random variables, where X is uniformly distributed over some subset S' of S of size $\geq 2^{160}$, and H is uniformly distributed over R. Then the collision and guessing probabilities of X are at most 2^{-160}, and so the leftover hash lemma (with $\alpha = 0$) says that the distance of $(H, \Phi_H(X))$ from uniform on $R \times T$ is δ', where $\delta' \leq \frac{1}{2}\sqrt{2^{64} 2^{-160}} = 2^{-49}$. By Theorem 8.32, it follows that the distance of $\Phi_H(X)$ from uniform on T is at most $\delta' \leq 2^{-49}$. □

The leftover hash lemma allows one to convert "low quality" sources of randomness into "high quality" sources of randomness. Suppose that to conduct an experiment, we need to sample a random variable Y whose distribution is uniform on a set T of size m, or at least, its distance from uniform on T is sufficiently small. However, we may not have direct access to a source of "real" randomness whose distribution looks anything like that of the desired uniform distribution, but rather, only to a "low quality" source of randomness. For example, one could model various characteristics of a person's typing at the keyboard, or perhaps various characteristics of the internal state of a computer (both its software and hardware) as a random process. We cannot say very much about the probability distributions associated with such processes, but perhaps we can conservatively estimate the collision or guessing probabilities associated with these distributions. Using the leftover hash lemma, we can hash the output of this random process, using a suitably generated random hash function. The hash function acts like a "magnifying glass": it "focuses" the randomness inherent in the "low quality" source distribution onto the set T, obtaining a "high quality," nearly uniform, distribution on T.

Of course, this approach requires a random hash function, which may be just as difficult to generate as a random element of T. The following theorem shows, however, that we can at least use the same "magnifying glass" many times over, with the statistical distance from uniform of the output distribution increasing linearly in the number of applications of the hash function.

Theorem 8.38. *Let $\{\Phi_r\}_{r \in R}$ be a $(1 + \alpha)/m$-almost universal family of hash functions from S to T, where $m := |T|$. Let $H, X_1, \ldots, X_n$ be random variables, where H is uniformly distributed over R, each X_i takes values in S, and $H, X_1, \ldots, X_n$ form a mutually independent family of random variables. If β is an upper bound on the collision probability of each X_i, and δ' is the distance of $(H, \Phi_H(X_1), \ldots, \Phi_H(X_n))$ from uniform on $R \times T^{\times n}$, then $\delta' \leq \frac{1}{2} n \sqrt{m\beta + \alpha}$.*

Proof. Let $Y_1, \ldots, Y_n$ be random variables, each uniformly distributed over T, and assume that $H, X_1, \ldots, X_n, Y_1, \ldots, Y_n$ form a mutually independent family of random variables. We shall make a hybrid argument (as in the proof of Theorem 8.34). Define random variables $Z_0, Z_1, \ldots, Z_n$ as follows:

$$\begin{aligned} Z_0 &:= (H, \Phi_H(X_1), \ldots, \Phi_H(X_n)), \\ Z_i &:= (H, Y_1, \ldots, Y_i, \Phi_H(X_{i+1}), \ldots, \Phi_H(X_n)) \quad \text{for } i = 1, \ldots, n-1, \text{ and} \\ Z_n &:= (H, Y_1, \ldots, Y_n). \end{aligned}$$

We have

$$\begin{aligned} \delta' &= \Delta[Z_0; Z_n] \\ &\leq \sum_{i=1}^{n} \Delta[Z_{i-1}; Z_i] \quad \text{(by part (iv) of Theorem 8.30)} \\ &\leq \sum_{i=1}^{n} \Delta[\, H, Y_1, \ldots, Y_{i-1}, \Phi_H(X_i), X_{i+1}, \ldots, X_n; \\ &\qquad\qquad H, Y_1, \ldots, Y_{i-1}, \quad Y_i, \quad X_{i+1}, \ldots, X_n \,] \\ &\quad \text{(by Theorem 8.32)} \\ &= \sum_{i=1}^{n} \Delta[H, \Phi_H(X_i); H, Y_i] \quad \text{(by Theorem 8.33)} \\ &\leq \tfrac{1}{2} n \sqrt{m\beta + \alpha} \quad \text{(by Theorem 8.37).} \quad \square \end{aligned}$$

Another source of "low quality" randomness arises in certain cryptographic applications, where we have a "secret value" X, which is a random variable that takes values in a set S, and which has small collision or guessing probability. We want to derive from X a "secret key" whose distance from uniform on some specified "key space" T is small. Typically, T is the set of all bit strings of some given length, as in Example 8.25. Theorem 8.38 allows us to do this using a "public" hash function—generated at random once and for all, published for all to see, and used over and over to derive secret keys as needed. However, to apply this theorem, it is crucial that the secret values (and the hash key) are mutually independent.

EXERCISE 8.70. Consider again the situation in Theorem 8.37. Suppose that $T = \{0, \ldots, m-1\}$, but that we would rather have a nearly uniform distribution on $T' = \{0, \ldots, m'-1\}$, for some $m' < m$. While it may be possible to work with a different family of hash functions, we do not have to if m is large enough with respect to m', in which case we can just use the value $Y' := \Phi_H(X) \bmod m'$. Show that the distance of (H, Y') from uniform on $R \times T'$ is at most $\frac{1}{2}\sqrt{m\beta + \alpha} + m'/m$.

EXERCISE 8.71. Let $\{\Phi_r\}_{r \in R}$ be a $(1 + \alpha)/m$-almost universal family of hash functions from S to T, where $m := |T|$. Suppose H, X, Y, Z are random variables, where H is uniformly distributed over R, X takes values in S, Y is uniformly distributed over T, and U is the set of values taken by Z with non-zero probability. Assume that the family of random variables H, Y, (X, Z) is mutually independent.

(a) For $u \in U$, define $\beta(u) := \sum_{s \in S} \mathsf{P}[X = s \mid Z = u]^2$. Also, let $\beta' := \sum_{u \in U} \beta(u) \mathsf{P}[Z = u]$. Show that $\Delta[H, \Phi_H(X), Z; H, Y, Z] \leq \frac{1}{2}\sqrt{m\beta' + \alpha}$.

(b) Suppose that X is uniformly distributed over a subset S' of S, and that $Z = f(X)$ for some function $f : S \to U$. Show that $\Delta[H, \Phi_H(X), Z; H, Y, Z] \leq \frac{1}{2}\sqrt{m|U|/|S'| + \alpha}$.

8.10 Discrete probability distributions

In addition to working with probability distributions over finite sample spaces, one can also work with distributions over infinite sample spaces. If the sample space is *countable*, that is, either finite or *countably infinite* (see §A3), then the distribution is called a **discrete probability distribution**. We shall not consider any other types of probability distributions in this text. The theory developed in §§8.1–8.5 extends fairly easily to the countably infinite setting, and in this section, we discuss how this is done.

8.10.1 Basic definitions

To say that the sample space Ω is countably infinite simply means that there is a bijection f from the set of positive integers onto Ω; thus, we can enumerate the elements of Ω as $\omega_1, \omega_2, \omega_3, \ldots$, where $\omega_i := f(i)$.

As in the finite case, a **probability distribution on** Ω is a function $\mathsf{P} : \Omega \to [0, 1]$, where all the probabilities sum to 1, which means that the infinite series $\sum_{i=1}^{\infty} \mathsf{P}(\omega_i)$ converges to one. Luckily, the convergence properties of an infinite series whose terms are all non-negative is invariant under a reordering of terms (see §A6), so it does not matter how we enumerate the elements of Ω.

Example 8.43. Suppose we toss a fair coin repeatedly until it comes up *heads*, and let k be the total number of tosses. We can model this experiment as a discrete probability distribution P, where the sample space consists of the set of all positive integers: for each positive integer k, $\mathsf{P}(k) := 2^{-k}$. We can check that indeed $\sum_{k=1}^{\infty} 2^{-k} = 1$, as required.

One may be tempted to model this experiment by setting up a probability distribution on the sample space of all infinite sequences of coin tosses; however, this sample space is not countably infinite, and so we cannot construct a discrete

probability distribution on this space. While it is possible to extend the notion of a probability distribution to such spaces, this would take us too far afield. □

Example 8.44. More generally, suppose we repeatedly execute a Bernoulli trial until it succeeds, where each execution succeeds with probability $p > 0$ independently of the previous trials, and let k be the total number of trials executed. Then we associate the probability $\mathsf{P}(k) := q^{k-1}p$ with each positive integer k, where $q := 1 - p$, since we have $k - 1$ failures before the one success. One can easily check that these probabilities sum to 1. Such a distribution is called a **geometric distribution**. □

Example 8.45. The series $\sum_{k=1}^{\infty} 1/k^3$ converges to some positive number c. Therefore, we can define a probability distribution on the set of positive integers, where we associate with each $k \geq 1$ the probability $1/ck^3$. □

As in the finite case, an event is an arbitrary subset $\mathcal{A}$ of Ω. The probability $\mathsf{P}[\mathcal{A}]$ of $\mathcal{A}$ is defined as the sum of the probabilities associated with the elements of $\mathcal{A}$. This sum is treated as an infinite series when $\mathcal{A}$ is infinite. This series is guaranteed to converge, and its value does not depend on the particular enumeration of the elements of $\mathcal{A}$.

Example 8.46. Consider the geometric distribution discussed in Example 8.44, where p is the success probability of each Bernoulli trial, and $q := 1 - p$. For a given integer $i \geq 1$, consider the event $\mathcal{A}$ that the number of trials executed is at least i. Formally, $\mathcal{A}$ is the set of all integers greater than or equal to i. Intuitively, $\mathsf{P}[\mathcal{A}]$ should be q^{i-1}, since we perform at least i trials if and only if the first $i - 1$ trials fail. Just to be sure, we can compute

$$\mathsf{P}[\mathcal{A}] = \sum_{k \geq i} \mathsf{P}(k) = \sum_{k \geq i} q^{k-1}p = q^{i-1}p \sum_{k \geq 0} q^k = q^{i-1}p \cdot \frac{1}{1-q} = q^{i-1}. \quad \square$$

It is an easy matter to check that all the statements and theorems in §8.1 carry over *verbatim* to the case of countably infinite sample spaces. Moreover, Boole's inequality (8.6) and equality (8.7) are also valid for countably infinite families of events:

Theorem 8.39. *Suppose $\mathcal{A} := \bigcup_{i=1}^{\infty} \mathcal{A}_i$, where $\{\mathcal{A}_i\}_{i=1}^{\infty}$ is an infinite sequence of events. Then*

(i) $\mathsf{P}[\mathcal{A}] \leq \sum_{i=1}^{\infty} \mathsf{P}[\mathcal{A}_i]$, *and*

(ii) $\mathsf{P}[\mathcal{A}] = \sum_{i=1}^{\infty} \mathsf{P}[\mathcal{A}_i]$ *if* $\{\mathcal{A}_i\}_{i=1}^{\infty}$ *is pairwise disjoint.*

Proof. As in the proof of Theorem 8.1, for $\omega \in \Omega$ and $\mathcal{B} \subseteq \Omega$, define $\delta_\omega[\mathcal{B}] := 1$ if $\omega \in \mathcal{B}$, and $\delta_\omega[\mathcal{B}] := 0$ if $\omega \notin \mathcal{B}$. First, suppose that $\{\mathcal{A}_i\}_{i=1}^{\infty}$ is pairwise disjoint.

Evidently, $\delta_\omega[\mathcal{A}] = \sum_{i=1}^{\infty} \delta_\omega[\mathcal{A}_i]$ for each $\omega \in \Omega$, and so

$$\begin{aligned}\mathsf{P}[\mathcal{A}] &= \sum_{\omega\in\Omega} \mathsf{P}(\omega)\delta_\omega[\mathcal{A}] = \sum_{\omega\in\Omega} \mathsf{P}(\omega) \sum_{i=1}^{\infty} \delta_\omega[\mathcal{A}_i] \\ &= \sum_{i=1}^{\infty} \sum_{\omega\in\Omega} \mathsf{P}(\omega)\delta_\omega[\mathcal{A}_i] = \sum_{i=1}^{\infty} \mathsf{P}[\mathcal{A}_i],\end{aligned}$$

where we use the fact that we may reverse the order of summation in an infinite double summation of non-negative terms (see §A7). That proves (ii), and (i) follows from (ii), applied to the sequence $\{\mathcal{A}'_i\}_{i=1}^{\infty}$, where $\mathcal{A}'_i := \mathcal{A}_i \setminus \bigcup_{j=1}^{i-1} \mathcal{A}_i$, as $\mathsf{P}[\mathcal{A}] = \sum_{i=1}^{\infty} \mathsf{P}[\mathcal{A}'_i] \leq \sum_{i=1}^{\infty} \mathsf{P}[\mathcal{A}_i]$. □

8.10.2 Conditional probability and independence

All of the definitions and results in §8.2 carry over *verbatim* to the countably infinite case. The law of total probability (equations (8.9) and (8.10)), as well as Bayes' theorem (8.11), extend to families of events $\{\mathcal{B}_i\}_{i\in I}$ indexed by any countably infinite set I. The definitions of independent families of events (k-wise and mutually) extend *verbatim* to infinite families.

8.10.3 Random variables

All of the definitions and results in §8.3 carry over *verbatim* to the countably infinite case. Note that the image of a random variable may be either finite or countably infinite. The definitions of independent families of random variables (k-wise and mutually) extend *verbatim* to infinite families.

8.10.4 Expectation and variance

We define the expected value of a real-valued random variable X exactly as in (8.18); that is, $\mathsf{E}[X] := \sum_\omega X(\omega)\,\mathsf{P}(\omega)$, but where this sum is now an infinite series. If this series converges absolutely (see §A6), then we say that X has **finite expectation**, or that $\mathsf{E}[X]$ is **finite**. In this case, the series defining $\mathsf{E}[X]$ converges to the same finite limit, regardless of the ordering of the terms.

If $\mathsf{E}[X]$ is not finite, then under the right conditions, $\mathsf{E}[X]$ may still exist, although its value will be $\pm\infty$. Consider first the case where X takes only non-negative values. In this case, if $\mathsf{E}[X]$ is not finite, then we naturally define $\mathsf{E}[X] := \infty$, as the series defining $\mathsf{E}[X]$ diverges to ∞, regardless of the ordering of the terms. In the general case, we may define random variables X^+ and X^-, where

$$X^+(\omega) := \max\{0, X(\omega)\} \quad \text{and} \quad X^-(\omega) := \max\{0, -X(\omega)\},$$

so that $X = X^+ - X^-$, and both X^+ and X^- take only non-negative values. Clearly, X has finite expectation if and only if both X^+ and X^- have finite expectation. Now suppose that $\mathsf{E}[X]$ is not finite, so that one of $\mathsf{E}[X^+]$ or $\mathsf{E}[X^-]$ is infinite. If $\mathsf{E}[X^+] = \mathsf{E}[X^-] = \infty$, then we say that $\mathsf{E}[X]$ **does not exist**; otherwise, we define $\mathsf{E}[X] := \mathsf{E}[X^+] - \mathsf{E}[X^-]$, which is $\pm\infty$; in this case, the series defining $\mathsf{E}[X]$ diverges to $\pm\infty$, regardless of the ordering of the terms.

Example 8.47. Let X be a random variable whose distribution is as in Example 8.45. Since the series $\sum_{k=1}^{\infty} 1/k^2$ converges and the series $\sum_{k=1}^{\infty} 1/k$ diverges, the expectation $\mathsf{E}[X]$ is finite, while $\mathsf{E}[X^2] = \infty$. One may also verify that the random variable $(-1)^X X^2$ has no expectation. □

All of the results in §8.4 carry over essentially unchanged, although one must pay some attention to "convergence issues."

If $\mathsf{E}[X]$ exists, then we can regroup the terms in the series $\sum_{\omega} X(\omega)\,\mathsf{P}(\omega)$, without affecting its value. In particular, equation (8.19) holds provided $\mathsf{E}[X]$ exists, and equation (8.20) holds provided $\mathsf{E}[f(X)]$ exists.

Theorem 8.14 still holds, under the additional hypothesis that $\mathsf{E}[X]$ and $\mathsf{E}[Y]$ are finite. Equation (8.21) also holds, provided the individual expectations $\mathsf{E}[X_i]$ are finite. More generally, if $\mathsf{E}[X]$ and $\mathsf{E}[Y]$ exist, then $\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y]$, unless $\mathsf{E}[X] = \infty$ and $\mathsf{E}[Y] = -\infty$, or $\mathsf{E}[X] = -\infty$ and $\mathsf{E}[Y] = \infty$. Also, if $\mathsf{E}[X]$ exists, then $\mathsf{E}[aX] = a\,\mathsf{E}[X]$, unless $a = 0$ and $\mathsf{E}[X] = \pm\infty$.

One might consider generalizing (8.21) to countably infinite families of random variables. To this end, suppose $\{X_i\}_{i=1}^{\infty}$ is an infinite sequence of real-valued random variables. The random variable $X := \sum_{i=1}^{\infty} X_i$ is well defined, provided the series $\sum_{i=1}^{\infty} X_i(\omega)$ converges for each $\omega \in \Omega$. One might hope that $\mathsf{E}[X] = \sum_{i=1}^{\infty} \mathsf{E}[X_i]$; however, this is not in general true, even if the individual expectations, $\mathsf{E}[X_i]$, are non-negative, and even if the series defining X converges absolutely for each ω; nevertheless, it is true when the X_i's are non-negative:

Theorem 8.40. *Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of random variables. Suppose that for each $i \geq 1$, X_i takes non-negative values only, and has finite expectation. Also suppose that $\sum_{i=1}^{\infty} X_i(\omega)$ converges for each $\omega \in \Omega$, and define $X := \sum_{i=1}^{\infty} X_i$. Then we have*

$$\mathsf{E}[X] = \sum_{i=1}^{\infty} \mathsf{E}[X_i].$$

Proof. This is a calculation just like the one made in the proof of Theorem 8.39, where, again, we use the fact that we may reverse the order of summation in an

infinite double summation of non-negative terms:

$$\mathsf{E}[X] = \sum_{\omega\in\Omega} \mathsf{P}(\omega)X(\omega) = \sum_{\omega\in\Omega} \mathsf{P}(\omega) \sum_{i=1}^{\infty} X_i(\omega)$$
$$= \sum_{i=1}^{\infty} \sum_{\omega\in\Omega} \mathsf{P}(\omega)X_i(\omega) = \sum_{i=1}^{\infty} \mathsf{E}[X_i]. \quad \square$$

Theorem 8.15 holds under the additional hypothesis that $\mathsf{E}[X]$ and $\mathsf{E}[Y]$ are finite. Equation (8.22) also holds, provided the individual expectations $\mathsf{E}[X_i]$ are finite. Theorem 8.16 still holds, of course. Theorem 8.17 also holds, but where now the sum may be infinite; it can be proved using essentially the same argument as in the finite case, combined with Theorem 8.40.

Example 8.48. Suppose X is a random variable with a geometric distribution, as in Example 8.44, with an associated success probability p and failure probability $q := 1 - p$. As we saw in Example 8.46, for every integer $i \geq 1$, we have $\mathsf{P}[X \geq i] = q^{i-1}$. We may therefore apply the infinite version of Theorem 8.17 to easily compute the expected value of X:

$$\mathsf{E}[X] = \sum_{i=1}^{\infty} \mathsf{P}[X \geq i] = \sum_{i=1}^{\infty} q^{i-1} = \frac{1}{1-q} = \frac{1}{p}. \quad \square$$

Example 8.49. To illustrate that Theorem 8.40 does not hold in general, consider the geometric distribution on the positive integers, where $\mathsf{P}(j) = 2^{-j}$ for $j \geq 1$. For $i \geq 1$, define the random variable X_i so that $X_i(i) = 2^i$, $X_i(i+1) = -2^{i+1}$, and $X_i(j) = 0$ for all $j \notin \{i, i+1\}$. Then $\mathsf{E}[X_i] = 0$ for all $i \geq 1$, and so $\sum_{i\geq 1} \mathsf{E}[X_i] = 0$. Now define $X := \sum_{i\geq 1} X_i$. This is well defined, and in fact $X(1) = 2$, while $X(j) = 0$ for all $j > 1$. Hence $\mathsf{E}[X] = 1$. $\square$

The variance $\mathsf{Var}[X]$ of X exists only when $\mu := \mathsf{E}[X]$ is finite, in which case it is defined as usual as $\mathsf{E}[(X-\mu)^2]$, which may be either finite or infinite. Theorems 8.18, 8.19, and 8.20 hold provided all the relevant expectations and variances are finite.

The definition of conditional expectation carries over verbatim. Equation (8.23) holds, provided $\mathsf{E}[X \mid \mathcal{B}]$ exists, and the law of total expectation (8.24) holds, provided $\mathsf{E}[X]$ exists. The law of total expectation also holds for a countably infinite partition $\{\mathcal{B}_i\}_{i\in I}$, provided $\mathsf{E}[X]$ exists, and each of the conditional expectations $\mathsf{E}[X \mid \mathcal{B}_i]$ is finite.

8.10.5 Some useful bounds

All of the results in this section hold, provided the relevant expectations and variances are finite.

EXERCISE 8.72. Let $\{\mathcal{A}_i\}_{i=1}^{\infty}$ be a family of events, such that $\mathcal{A}_i \subseteq \mathcal{A}_{i+1}$ for each $i \geq 1$, and let $\mathcal{A} := \bigcup_{i=1}^{\infty} \mathcal{A}_i$. Show that $\mathsf{P}[\mathcal{A}] = \lim_{i\to\infty} \mathsf{P}[\mathcal{A}_i]$.

EXERCISE 8.73. Generalize Exercises 8.6, 8.7, 8.23, and 8.24 to the discrete setting, allowing a countably infinite index set I.

EXERCISE 8.74. Suppose X is a random variable taking positive integer values, and that for some real number q, with $0 \leq q \leq 1$, and for all integers $i \geq 1$, we have $\mathsf{P}[X \geq i] = q^{i-1}$. Show that X has a geometric distribution with associated success probability $p := 1 - q$.

EXERCISE 8.75. This exercise extends Jensen's inequality (see Exercise 8.25) to the discrete setting. Suppose that f is a convex function on an interval I. Let X be a random variable whose image is a countably infinite subset of I, and assume that both $\mathsf{E}[X]$ and $\mathsf{E}[f(X)]$ are finite. Show that $\mathsf{E}[f(X)] \geq f(\mathsf{E}[X])$. Hint: use continuity.

EXERCISE 8.76. A gambler plays a simple game in a casino: with each play of the game, the gambler may bet any number m of dollars; a fair coin is tossed, and if it comes up *heads*, the casino pays m dollars to the gambler, and otherwise, the gambler pays m dollars to the casino. The gambler plays the game repeatedly, using the following strategy: he initially bets a dollar, and with each subsequent play, he doubles his bet; if he ever wins, he quits and goes home; if he runs out of money, he also goes home; otherwise, he plays again. Show that if the gambler has an infinite amount of money, then his expected winnings are one dollar, and if he has a finite amount of money, his expected winnings are zero.

8.11 Notes

The idea of sharing a secret via polynomial evaluation and interpolation (see Example 8.28) is due to Shamir [90].

Our Chernoff bound (Theorem 8.24) is one of a number of different types of bounds that appear in the literature under the rubric of "Chernoff bound."

Universal and pairwise independent hash functions, with applications to hash tables and message authentication codes, were introduced by Carter and Wegman [25, 105]. The notions of ε-almost universal and ε-almost strongly universal

hashing were developed by Stinson [101]. The notion of ε-variationally universal hashing (see Exercise 8.63) is from Krovetz and Rogaway [57].

The leftover hash lemma (Theorem 8.37) was originally stated and proved by Impagliazzo, Levin, and Luby [48], who use it to obtain an important result in the theory of cryptography. Our proof of the leftover hash lemma is loosely based on one by Impagliazzo and Zuckermann [49], who also present further applications.