# 6
# Finite and discrete probability distributions

This chapter introduces concepts from discrete probability theory. We begin with a discussion of finite probability distributions, and then towards the end of the chapter we discuss the more general notion of a discrete probability distribution.

## 6.1 Finite probability distributions: basic definitions

A **finite probability distribution** $\mathbf{D} = (\mathcal{U}, \mathsf{P})$ is a *finite*, non-empty set $\mathcal{U}$, together with a function $\mathsf{P}$ that maps $u \in \mathcal{U}$ to $\mathsf{P}[u] \in [0, 1]$, such that

$$\sum_{u \in \mathcal{U}} \mathsf{P}[u] = 1. \tag{6.1}$$

The set $\mathcal{U}$ is called the **sample space** and the function $\mathsf{P}$ is called the **probability function**.

Intuitively, the elements of $\mathcal{U}$ represent the possible outcomes of a random experiment, where the probability of outcome $u \in \mathcal{U}$ is $\mathsf{P}[u]$.

Up until §6.10, we shall use the phrase "probability distribution" to mean "finite probability distribution."

***Example* 6.1.** If we think of rolling a fair die, then $\mathcal{U} := \{1, 2, 3, 4, 5, 6\}$, and $\mathsf{P}[u] := 1/6$ for all $u \in \mathcal{U}$ gives a probability distribution describing the possible outcomes of the experiment. □

***Example* 6.2.** More generally, if $\mathcal{U}$ is a finite set, and $\mathsf{P}[u] = 1/|\mathcal{U}|$ for all $u \in \mathcal{U}$, then $\mathbf{D}$ is called the **uniform distribution on $\mathcal{U}$**. □

***Example* 6.3.** A coin flip is an example of a **Bernoulli trial**, which is in general an experiment with only two possible outcomes: *success*, which occurs with probability $p$, and *failure*, which occurs with probability $q := 1 - p$. □

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

$$\mathsf{P}[\mathcal{A}] := \sum_{u \in \mathcal{A}} \mathsf{P}[u]. \tag{6.2}$$

Thus, we extend the domain of definition of $\mathsf{P}$ from outcomes $u \in \mathcal{U}$ to events $\mathcal{A} \subseteq \mathcal{U}$.

For an event $\mathcal{A} \subseteq \mathcal{U}$, let $\overline{\mathcal{A}}$ denote the complement of $\mathcal{A}$ in $\mathcal{U}$. We have $\mathsf{P}[\emptyset] = 0$, $\mathsf{P}[\mathcal{U}] = 1$, $\mathsf{P}[\overline{\mathcal{A}}] = 1 - \mathsf{P}[\mathcal{A}]$.

For any events $\mathcal{A}, \mathcal{B} \subseteq \mathcal{U}$, if $\mathcal{A} \subseteq \mathcal{B}$, then $\mathsf{P}[\mathcal{A}] \leq \mathsf{P}[\mathcal{B}]$. Also, for any events $\mathcal{A}, \mathcal{B} \subseteq \mathcal{U}$, we have

$$\mathsf{P}[\mathcal{A} \cup \mathcal{B}] = \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] - \mathsf{P}[\mathcal{A} \cap \mathcal{B}] \leq \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}]; \tag{6.3}$$

in particular, if $\mathcal{A}$ and $\mathcal{B}$ are disjoint, then

$$\mathsf{P}[\mathcal{A} \cup \mathcal{B}] = \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}]. \tag{6.4}$$

More generally, for any events $\mathcal{A}_1, \ldots, \mathcal{A}_n \subseteq \mathcal{U}$ we have

$$\mathsf{P}[\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n] \leq \mathsf{P}[\mathcal{A}_1] + \cdots + \mathsf{P}[\mathcal{A}_n], \tag{6.5}$$

and if the $\mathcal{A}_i$ are pairwise disjoint, then

$$\mathsf{P}[\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n] = \mathsf{P}[\mathcal{A}_1] + \cdots + \mathsf{P}[\mathcal{A}_n]. \tag{6.6}$$

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

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

We also have the distributive law: for events $\mathcal{A}, \mathcal{B}, \mathcal{C}$, we have

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

In some applications and examples, it is more natural to use the logical "or" connective "$\vee$" in place of "$\cup$," and the logical "and" connective "$\wedge$" in place of "$\cap$."

***Example* 6.4.** Continuing with Example 6.1, the probability of an "odd roll" $\mathcal{A} = \{1, 3, 5\}$ is 1/2. □

***Example* 6.5.** More generally, if $\mathbf{D}$ is the uniform distribution on a set $\mathcal{U}$ of cardinality $n$, and $\mathcal{A}$ is a subset of $\mathcal{U}$ of cardinality $k$, then $\mathsf{P}[\mathcal{A}] = k/n$. □

***Example* 6.6.** 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 $\{(x, y) : x, y = 1, \dots, 6\}$, where $x$ represents the value of the first die, and $y$ the value of the second.

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

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

If $\mathbf{D}_1 = (\mathcal{U}_1, \mathsf{P}_1)$ and $\mathbf{D}_2 = (\mathcal{U}_2, \mathsf{P}_2)$ are probability distributions, we can form the **product distribution** $\mathbf{D} = (\mathcal{U}, \mathsf{P})$, where $\mathcal{U} := \mathcal{U}_1 \times \mathcal{U}_2$, and $\mathsf{P}[(u_1, u_2)] := \mathsf{P}_1[u_1]\mathsf{P}_2[u_2]$. It is easy to verify that the product distribution is also a probability distribution. Intuitively, the elements $(u_1, u_2)$ of $\mathcal{U}_1 \times \mathcal{U}_2$ denote the possible outcomes of two separate and independent experiments.

More generally, if $\mathbf{D}_i = (\mathcal{U}_i, \mathsf{P}_i)$ for $i = 1, \dots, n$, we can define the product distribution $\mathbf{D} = (\mathcal{U}, \mathsf{P})$, where $\mathcal{U} := \mathcal{U}_1 \times \cdots \times \mathcal{U}_n$, and $\mathsf{P}[(u_1, \dots, u_n)] := \mathsf{P}[u_1] \dots \mathsf{P}[u_n]$.

***Example* 6.7.** We can view the probability distribution in Example 6.6 as the product of two copies of the uniform distribution on $\{1, \dots, 6\}$. □

***Example* 6.8.** Consider the product distribution of $n$ copies of a Bernoulli trial (see Example 6.3), with associated success probability $p$ and failure probability $q := 1 - p$. An element of the sample space is an $n$-tuple of success/failure values. Any such tuple that contains, say, $k$ successes and $n - k$ failures, occurs with probability $p^k q^{n-k}$, regardless of the particular positions of the successes and failures. □

EXERCISE 6.1. This exercise asks you to recast previously established results in terms of probability theory.

(a) Let $k \geq 2$ be an integer, and suppose an integer $n$ is chosen at random from among all $k$-bit integers. Show that the probability that $n$ is prime is $\Theta(1/k)$.

(b) Let $n$ be a positive integer, and suppose that $a$ and $b$ are chosen at random from the set $\{1, \dots, n\}$. Show that the probability that $\gcd(a, b) = 1$ is at least $1/4$.

(c) Let $n$ be a positive integer, and suppose that $a$ is chosen at random from the set $\{1, \dots, n\}$. Show that the probability that $\gcd(a, n) = 1$ is $\Omega(1/\log\log n)$.

Exercise 6.2. 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 6.3. Generalize equation (6.3) by proving the **inclusion/exclusion principle**: for events $\mathcal{A}_1, \ldots, \mathcal{A}_n$, we have

$$\begin{aligned}\mathsf{P}[\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n] &= \sum_i \mathsf{P}[\mathcal{A}_i] - \sum_{i<j} \mathsf{P}[\mathcal{A}_i \cap \mathcal{A}_j] + \\ &\quad \sum_{i<j<k} \mathsf{P}[\mathcal{A}_i \cap \mathcal{A}_j \cap \mathcal{A}_k] - \cdots + (-1)^{n-1}\mathsf{P}[\mathcal{A}_1 \cap \cdots \cap \mathcal{A}_n] \\ &= \sum_{\ell=1}^{n} (-1)^{\ell-1} \sum_{i_1<\cdots<i_\ell} \mathsf{P}[\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_\ell}].\end{aligned}$$

Exercise 6.4. Show that for events $\mathcal{A}_1, \ldots, \mathcal{A}_n$, we have

$$\mathsf{P}[\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n] \geq \sum_i \mathsf{P}[\mathcal{A}_i] - \sum_{i<j} \mathsf{P}[\mathcal{A}_i \cap \mathcal{A}_j].$$

Exercise 6.5. Generalize inequality (6.5) and the previous exercise by proving **Bonferroni's inequalities**: for events $\mathcal{A}_1, \ldots, \mathcal{A}_n$, and defining

$$e_m := \mathsf{P}[\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n] - \sum_{\ell=1}^{m} (-1)^{\ell-1} \sum_{i_1<\cdots<i_\ell} \mathsf{P}[\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_\ell}]$$

for $m = 1, \ldots, n$, we have $e_m \leq 0$ for odd $m$, and $e_m \geq 0$ for even $m$.

## 6.2 Conditional probability and independence

Let $\mathbf{D} = (\mathcal{U}, \mathsf{P})$ be a probability distribution.

For any event $\mathcal{B} \subseteq \mathcal{U}$ with $\mathsf{P}[\mathcal{B}] \neq 0$ and any $u \in \mathcal{U}$, let us define

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

Viewing $\mathcal{B}$ as fixed, we may view the function $\mathsf{P}[\cdot \mid \mathcal{B}]$ as a new probability function on the sample space $\mathcal{U}$, and this gives rise a new probability distribution $\mathbf{D}_{\mathcal{B}} := (\mathsf{P}[\cdot \mid \mathcal{B}], \mathcal{U})$, called the **conditional distribution given $\mathcal{B}$**.

Intuitively, $\mathbf{D}_{\mathcal{B}}$ has the following interpretation: if a random experiment produces an outcome according to the distribution $\mathbf{D}$, and we learn that the event $\mathcal{B}$ has occurred, then the distribution $\mathbf{D}_{\mathcal{B}}$ assigns new probabilities to all possible outcomes, reflecting the partial knowledge that the event $\mathcal{B}$ has occurred.

As usual, we extend the domain of definition of $\mathsf{P}[\cdot \mid \mathcal{B}]$ from outcomes to events. For any event $\mathcal{A} \subseteq \mathcal{U}$, we have

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

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 $\mathsf{P}[\mathcal{A} \cap \mathcal{B}] = \mathsf{P}[\mathcal{A}] \cdot \mathsf{P}[\mathcal{B}]$, then $\mathcal{A}$ and $\mathcal{B}$ are called **independent** events. If $\mathsf{P}[\mathcal{B}] \neq 0$, a simple calculation shows that $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = \mathsf{P}[\mathcal{A}]$.

A collection $\mathcal{A}_1, \ldots, \mathcal{A}_n$ of events is called **pairwise independent** if $\mathsf{P}[\mathcal{A}_i \cap \mathcal{A}_j] = \mathsf{P}[\mathcal{A}_i]\mathsf{P}[\mathcal{A}_j]$ for all $i \neq j$, and is called **mutually independent** if every subset $\mathcal{A}_{i_1}, \ldots, \mathcal{A}_{i_k}$ of the collection satisfies

$$\mathsf{P}[\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_k}] = \mathsf{P}[\mathcal{A}_{i_1}] \cdots \mathsf{P}[\mathcal{A}_{i_k}].$$

***Example* 6.9.** In Example 6.6, suppose that Alice tells Bob the sum of the two dice before Bob makes his guess. For example, suppose Alice tells Bob the sum is 4. Then what is Bob's best strategy in this case? Let $\mathcal{S}_z$ be the event that the sum is $z$, for $z = 2, \ldots, 12$, and consider the conditional probability distribution given $\mathcal{S}_4$. This is the uniform distribution on the three pairs $(1,3), (2,2), (3,1)$. The numbers 1 and 3 both appear in two pairs, while the number 2 appears in just one pair. Therefore,

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

while

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

and

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

Thus, if the sum is 4, Bob's best strategy is to guess either 1 or 3.

Note that the events $\mathcal{A}_1$ and $\mathcal{B}_2$ are independent, while the events $\mathcal{A}_1$ and $\mathcal{S}_4$ are not. □

***Example* 6.10.** Suppose we toss three fair coins. Let $\mathcal{A}_1$ be the event that the first coin is "heads," let $\mathcal{A}_2$ be the event that the second coin is "heads," and let $\mathcal{A}_3$ be the event that the third coin is "heads." Then the collection of events $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ is mutually independent.

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

Suppose we have a collection $\mathcal{B}_1, \ldots, \mathcal{B}_n$ of events that partitions $\mathcal{U}$, such that each event $\mathcal{B}_i$ occurs with non-zero probability. Then it is easy to see that for any event $\mathcal{A}$,

$$\mathsf{P}[\mathcal{A}] = \sum_{i=1}^{n} \mathsf{P}[\mathcal{A} \cap \mathcal{B}_i] = \sum_{i=1}^{n} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i] \cdot \mathsf{P}[\mathcal{B}_i]. \tag{6.7}$$

Furthermore, if $\mathsf{P}[\mathcal{A}] \neq 0$, then for any $j = 1, \ldots, n$, we have

$$\mathsf{P}[\mathcal{B}_j \mid \mathcal{A}] = \frac{\mathsf{P}[\mathcal{A} \cap \mathcal{B}_j]}{\mathsf{P}[\mathcal{A}]} = \frac{\mathsf{P}[\mathcal{A} \mid \mathcal{B}_j]\mathsf{P}[\mathcal{B}_j]}{\sum_{i=1}^{n} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]\mathsf{P}[\mathcal{B}_i]}. \tag{6.8}$$

This equality, known as **Bayes' theorem**, lets us compute the conditional probability $\mathsf{P}[\mathcal{B}_j \mid \mathcal{A}]$ in terms of the conditional probabilities $\mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]$.

The equation (6.7) is useful for computing or estimating probabilities by conditioning on specific events $\mathcal{B}_i$ (i.e., by considering the conditional probability distribution given $\mathcal{B}_i$) in such a way that the conditional probabilities $\mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]$ are easy to compute or estimate. Also, if we want to compute a conditional probability $\mathsf{P}[\mathcal{A} \mid \mathcal{C}]$, we can do so by partitioning $\mathcal{C}$ into events $\mathcal{B}_1, \ldots, \mathcal{B}_n$, where each $\mathcal{B}_i$ occurs with non-zero probability, and use the following simple fact:

$$\mathsf{P}[\mathcal{A} \mid \mathcal{C}] = \sum_{i=1}^{n} \mathsf{P}[\mathcal{A} \mid \mathcal{B}_i]\mathsf{P}[\mathcal{B}_i]/\mathsf{P}[\mathcal{C}]. \tag{6.9}$$

***Example* 6.11.** 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 the chosen door. 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. That is, $\mathsf{P}[\mathcal{A} \mid \mathcal{B}] = 0$. 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. That is, $\mathsf{P}[\mathcal{A} \mid \overline{\mathcal{B}}] = 1$. Furthermore, it is clear that $\mathsf{P}[\mathcal{B}] = 1/3$. So 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 independent of the contestant's initial choice, then switching is again the preferred strategy. □

***Example* 6.12.** 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 $\mathcal{A}$ be the event that the test is positive and let $\mathcal{B}$ be the event that the patient has disease $X$. The relevant quantity that we need to estimate is $\mathsf{P}[\mathcal{B} \mid \mathcal{A}]$; that is, the probability that the patient has disease $X$, given a positive test result. We use Bayes' theorem to do this:

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

Thus, the chances that the patient has disease $X$ given a positive test result is 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. □

Exercise 6.6. Consider again the situation in Example 6.12, 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 6.7. Suppose we roll two dice, and let $(x, y)$ denote the outcome (as in Example 6.6). For each of the following pairs of events $\mathcal{A}$ and $\mathcal{B}$, determine if they are independent or not:

(a) $\mathcal{A}$: $x = y$; $\mathcal{B}$: $y = 1$.

(b) $\mathcal{A}$: $x \geq y$; $\mathcal{B}$: $y = 1$.

(c) $\mathcal{A}$: $x \geq y$; $\mathcal{B}$: $y^2 = 7y - 6$.

(d) $\mathcal{A}$: $xy = 6$; $\mathcal{B}$: $y = 3$.

Exercise 6.8. Let $\mathcal{C}$ be an event that occurs with non-zero probability, and let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be a partition of $\mathcal{C}$, such that each event $\mathcal{B}_i$ occurs with non-zero probability. Let $\mathcal{A}$ be an event and let $p$ be a real number with $0 \leq p \leq 1$. Suppose that for each $i = 1, \ldots, n$, the conditional probability of $\mathcal{A}$ given $\mathcal{B}_i$ is $\leq p$ (resp., $<, =, >, \geq p$). Show that the conditional probability of $\mathcal{A}$ given $\mathcal{C}$ is also $\leq p$ (resp., $<, =, >, \geq p$).

Exercise 6.9. Show that if two events $\mathcal{A}$ and $\mathcal{B}$ are independent, then so are $\mathcal{A}$ and $\overline{\mathcal{B}}$. More generally, show that if $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are mutually independent, then so are $\mathcal{A}'_1, \ldots, \mathcal{A}'_n$, where each $\mathcal{A}'_i$ denotes either $\mathcal{A}_i$ or $\overline{\mathcal{A}}_i$.

Exercise 6.10. This exercise develops an alternative proof, based on probability theory, of Theorem 2.14. Let $n > 1$ be an integer and consider an experiment in which a number $a$ is chosen 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

$$\phi(n)/n = \mathsf{P}[\overline{\mathcal{A}}_1 \cap \cdots \cap \overline{\mathcal{A}}_r],$$

where $\phi$ is Euler's phi function.

(b) Show that if $i_1, \ldots, i_\ell$ are distinct indices between 1 and $r$, then

$$\mathsf{P}[\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_\ell}] = \frac{1}{p_{i_1} \cdots p_{i_\ell}}.$$

Conclude that the events $\mathcal{A}_i$ are mutually independent, and $\mathsf{P}[\mathcal{A}_i] = 1/p_i$.

(c) Using part (b) and the result of the previous exercise, show 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 2.14 that

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

## 6.3 Random variables

Let $\mathbf{D} = (\mathcal{U}, \mathsf{P})$ be a probability distribution.

It is sometimes convenient to associate a real number, or other mathematical object, with each outcome $u \in \mathcal{U}$. Such an association is called a **random variable**; more formally, a random variable $X$ is a function from $\mathcal{U}$ into a set $\mathcal{X}$. If $\mathcal{X}$ is a subset of the real numbers, then $X$ is called a **real random variable**. When we speak of the **image** of $X$, we simply mean its image in the usual function-theoretic sense, that is, the set $X(\mathcal{U}) = \{X(u) : u \in \mathcal{U}\}$.

One may define any number of random variables on a given probability distribution. If $X : \mathcal{U} \to \mathcal{X}$ is a random variable, and $f : \mathcal{X} \to \mathcal{Y}$ is a function, then $f(X) := f \circ X$ is also a random variable.

***Example* 6.13.** Suppose we flip $n$ fair coins. Then we may define a random variable $X$ that maps each outcome to a bit string of length $n$, where a "head" is encoded as a 1-bit, and a "tail" is encoded as a 0-bit. We may define another random variable $Y$ that is the number of "heads." The variable $Y$ is a real random variable. □

***Example* 6.14.** 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 $\mathcal{A}$**. Conversely, if $Y$ is any 0/1-valued random variable, we can define the event $\mathcal{B}$ to be the subset of all possible outcomes that lead to $Y = 1$, and $Y$ is the indicator variable for the event $\mathcal{B}$. Thus, we can work with either events or indicator variables, whichever is more natural and convenient. $\Box$

Let $X : \mathcal{U} \to \mathcal{X}$ be a random variable. For $x \in \mathcal{X}$, we write "$X = x$" as shorthand for the event $\{u \in \mathcal{U} : X(u) = x\}$. More generally, for any predicate $\phi$, we may write "$\phi(X)$" as shorthand for the event $\{u \in \mathcal{U} : \phi(X(u))\}$.

A random variable $X$ defines a probability distribution on its image $\mathcal{X}$, where the probability associated with $x \in \mathcal{X}$ is $\mathsf{P}[X = x]$. We call this the **distribution of $X$**. For two random variables $X, Y$ defined on a probability distribution, $Z := (X, Y)$ is also a random variable whose distribution is called the **joint distribution of $X$ and $Y$**.

If $X$ is a random variable, and $\mathcal{A}$ is an event with non-zero probability, then the **conditional distribution of $X$ given $\mathcal{A}$** is a probability distribution on the image $\mathcal{X}$ of $X$, where the probability associated with $x \in \mathcal{X}$ is $\mathsf{P}[X = x \mid \mathcal{A}]$.

We say two random variables $X, Y$ are **independent** if for all $x$ in the image of $X$ and all $y$ in the image of $Y$, the events $X = x$ and $Y = y$ are independent, which is to say,

$$\mathsf{P}[X = x \wedge Y = y] = \mathsf{P}[X = x]\mathsf{P}[Y = y].$$

Equivalently, $X$ and $Y$ are independent if and only if their joint distribution is equal to the product of their individual distributions. Alternatively, $X$ and $Y$ are independent if and only if for all values $x$ taken by $X$ with non-zero probability, the conditional distribution of $Y$ given the event $X = x$ is the same as the distribution of $Y$.

Let $X_1, \ldots, X_n$ be a collection of random variables, and let $\mathcal{X}_i$ be the image of $X_i$ for $i = 1, \ldots, n$. We say $X_1, \ldots, X_n$ are **pairwise independent** if for all $i, j = 1, \ldots, n$ with $i \neq j$, the variables $X_i$ and $X_j$ are independent. We say that $X_1, \ldots, X_n$ are **mutually independent** if for all $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$, we have

$$\mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_n = x_n] = \mathsf{P}[X_1 = x_1] \cdots \mathsf{P}[X_n = x_n].$$

More generally, for $k = 2, \ldots, n$, we say that $X_1, \ldots, X_n$ are **$k$-wise independent** if any $k$ of them are mutually independent.

***Example* 6.15.** We toss three coins, and set $X_i := 0$ if the $i$th coin is "tails," and $X_i := 1$ otherwise. The variables $X_1, X_2, X_3$ are mutually independent. Let us set $Y_{12} := X_1 \oplus X_2$, $Y_{13} := X_1 \oplus X_3$, and $Y_{23} := X_2 \oplus X_3$, where "$\oplus$" denotes "exclusive or," that is, addition modulo 2. Then the variables $Y_{12}, Y_{13}, Y_{23}$ are pairwise independent, but not mutually independent—observe that $Y_{12} \oplus Y_{13} = Y_{23}$. □

The following is a simple but useful fact:

**Theorem 6.1.** *Let $X_i : \mathcal{U} \to \mathcal{X}_i$ be random variables, for $i = 1, \ldots, n$, and suppose that there exist functions $f_i : \mathcal{X}_i \to [0, 1]$, for $i = 1, \ldots, n$, such that*

$$\sum_{x_i \in \mathcal{X}_i} f_i(x_i) = 1 \quad (i = 1 \ldots n),$$

*and*

$$\mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_n = x_n] = f_1(x_1) \cdots f_n(x_n)$$

*for all $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$. Then for any subset of distinct indices $i_1, \ldots, i_\ell \in \{1, \ldots, n\}$, we have*

$$\mathsf{P}[X_{i_1} = x_{i_1} \wedge \cdots \wedge X_{i_\ell} = x_{i_\ell}] = f_{i_1}(x_{i_1}) \cdots f_{i_\ell}(x_{i_\ell})$$

*for all $x_{i_1} \in \mathcal{X}_{i_1}, \ldots, x_{i_\ell} \in \mathcal{X}_{i_\ell}$.*

*Proof.* To prove the theorem, it will suffice to show that we can "eliminate" a single variable, say $X_n$, meaning that for all $x_1, \ldots, x_{n-1}$, we have

$$\mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_{n-1} = x_{n-1}] = f_1(x_1) \cdots f_{n-1}(x_{n-1}).$$

Having established this, we may then proceed to eliminate any number of variables (the ordering of the variables is clearly irrelevant).

We have

$$\begin{aligned}
&\mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_{n-1} = x_{n-1}] \\
&\qquad = \sum_{x_n \in \mathcal{X}_n} \mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_{n-1} = x_{n-1} \wedge X_n = x_n] \\
&\qquad = \sum_{x_n \in \mathcal{X}_n} f_1(x_1) \cdots f_{n-1}(x_{n-1}) f_n(x_n) \\
&\qquad = f_1(x_2) \cdots f_{n-1}(x_{n-1}) \sum_{x_n \in \mathcal{X}_n} f_n(x_n) \\
&\qquad = f_1(x_1) \cdots f_{n-1}(x_{n-1}). \quad \square
\end{aligned}$$

The following three theorems are immediate consequences of the above theorem:

**Theorem 6.2.** *Let $X_i : \mathcal{U} \to \mathcal{X}_i$ be random variables, for $i = 1, \ldots, n$, such that*

$$\mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_n = x_n] = \frac{1}{|\mathcal{X}_1|} \cdots \frac{1}{|\mathcal{X}_n|} \quad (\text{for all } x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n).$$

*Then the variables $X_i$ are mutually independent with each $X_i$ uniformly distributed over $\mathcal{X}_i$.*

**Theorem 6.3.** *If $X_1, \ldots, X_n$ are mutually independent random variables, then they are $k$-wise independent for all $k = 2, \ldots, n$.*

**Theorem 6.4.** *If $\mathbf{D}_i = (\mathcal{U}_i, \mathsf{P}_i)$ are probability distributions for $i = 1, \ldots, n$, then the projection functions $\pi_i : \mathcal{U}_1 \times \cdots \times \mathcal{U}_n \to \mathcal{U}_i$, where $\pi_i(u_1, \ldots, u_n) = u_i$, are mutually independent random variables on the product distribution $\mathbf{D}_1 \times \cdots \times \mathbf{D}_n$.*

We also have:

**Theorem 6.5.** *If $X_1, \ldots, X_n$ are mutually independent random variables, and $g_1, \ldots, g_n$ are functions, then $g_1(X_1), \ldots, g_n(X_n)$ are also mutually independent random variables.*

*Proof.* The proof is a straightforward, if somewhat tedious, calculation. For $i = 1, \ldots, n$, let $y_i$ be some value in the image of $g_i(X_i)$, and let $\mathcal{X}_i := g_i^{-1}(\{y_i\})$. We have

$$\begin{aligned}
&\mathsf{P}[g_1(X_1) = y_1 \wedge \cdots \wedge g_n(X_n) = y_n] \\
&\quad = \mathsf{P}\Big[\big(\bigvee_{x_1 \in \mathcal{X}_1} X_1 = x_1\big) \wedge \cdots \wedge \big(\bigvee_{x_n \in \mathcal{X}_n} X_n = x_n\big)\Big] \\
&\quad = \mathsf{P}\Big[\bigvee_{x_1 \in \mathcal{X}_1} \cdots \bigvee_{x_n \in \mathcal{X}_n} (X_1 = x_1 \wedge \cdots \wedge X_n = x_n)\Big] \\
&\quad = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \mathsf{P}[X_1 = x_1 \wedge \cdots \wedge X_n = x_n] \\
&\quad = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \mathsf{P}[X_1 = x_1] \cdots \mathsf{P}[X_n = x_n] \\
&\quad = \Big(\sum_{x_1 \in \mathcal{X}_1} \mathsf{P}[X_1 = x_1]\Big) \cdots \Big(\sum_{x_n \in \mathcal{X}_n} \mathsf{P}[X_n = x_n]\Big)
\end{aligned}$$

$$= \mathsf{P}\Big[\bigvee_{x_1 \in \mathcal{X}_1} X_1 = x_1\Big] \cdots \mathsf{P}\Big[\bigvee_{x_n \in \mathcal{X}_n} X_n = x_n\Big]$$
$$= \mathsf{P}[g_1(X_1) = y_1] \cdots \mathsf{P}[g_n(X_n) = y_n]. \quad \square$$

***Example* 6.16.** If we toss $n$ dice, and let $X_i$ denote the value of the $i$th die for $i = 1, \ldots, n$, then the $X_i$ are mutually independent random variables. If we set $Y_i := X_i^2$ for $i = 1, \ldots, n$, then the $Y_i$ are also mutually independent random variables. □

***Example* 6.17.** This example again illustrates the notion of pairwise independence. Let $X$ and $Y$ be independent and uniformly distributed over $\mathbb{Z}_p$, where $p$ is a prime. For $a \in \mathbb{Z}_p$, let $Z_a := aX + Y$. Then we claim that each $Z_a$ is uniformly distributed over $\mathbb{Z}_p$, and that the collection of random variables $\{Z_a : a \in \mathbb{Z}_p\}$ is pairwise independent.

To prove this claim, let $a, b \in \mathbb{Z}_p$ with $a \neq b$, and consider the map $f_{a,b} : \mathbb{Z}_p \times \mathbb{Z}_p \to \mathbb{Z}_p \times \mathbb{Z}_p$ that sends $(x, y)$ to $(ax+y, bx+y)$. It is easy to see that $f_{a,b}$ is injective; indeed, if $ax + y = ax' + y'$ and $bx + y = bx' + y'$, then subtracting these two equations, we obtain $(a - b)x = (a - b)x'$, and since $a - b \neq [0]_p$, it follows that $x = x'$, which also implies $y = y'$. Since $f_{a,b}$ is injective, it must be a bijection from $\mathbb{Z}_p \times \mathbb{Z}_p$ onto itself. Thus, since $(X, Y)$ is uniformly distributed over $\mathbb{Z}_p \times \mathbb{Z}_p$, so is $(Z_a, Z_b) = (aX + Y, bX + Y)$. So for all $z, z' \in \mathbb{Z}_p$, we have

$$\mathsf{P}[Z_a = z \wedge Z_b = z'] = 1/p^2,$$

and so the claim follows from Theorem 6.2.

Note that the variables $Z_a$ are not 3-wise independent, since the value of any two determines the value of all the rest (verify). □

***Example* 6.18.** We can generalize the previous example as follows. Let $X_1, \ldots, X_t, Y$ be mutually independent and uniformly distributed over $\mathbb{Z}_p$, where $p$ is prime, and for $a_1, \ldots, a_t \in \mathbb{Z}_p$, let $Z_{a_1,\ldots,a_t} := a_1X_1 + \cdots + a_tX_t + Y$. We claim that each $Z_{a_1,\ldots,a_t}$ is uniformly distributed over $\mathbb{Z}_p$, and that the collection of all such $Z_{a_1,\ldots,a_t}$ is pairwise independent.

To prove this claim, it will suffice (by Theorem 6.2) to prove that for all

$$a_1, \ldots, a_t,\ b_1, \ldots, b_t,\ z, z' \in \mathbb{Z}_p,$$

subject to $(a_1, \ldots, a_t) \neq (b_1, \ldots, b_t)$, we have

$$\mathsf{P}[Z_{a_1,\ldots,a_t} = z \wedge Z_{b_1,\ldots,b_t} = z'] = 1/p^2. \tag{6.10}$$

Since $(a_1, \ldots, a_t) \neq (b_1, \ldots, b_t)$, we know that $a_j \neq b_j$ for some $j = 1, \ldots, t$. Let us assume that $a_1 \neq b_1$ (the argument for $j > 1$ is analogous).

We first show that for all $x_2, \ldots, x_t \in \mathbb{Z}_p$, we have

$$\mathsf{P}[Z_{a_1,\ldots,a_t} = z \wedge Z_{b_1,\ldots,b_t} = z' \mid X_2 = x_2 \wedge \cdots \wedge X_t = x_t] = 1/p^2. \qquad (6.11)$$

To prove (6.11), consider the conditional probability distribution given $X_2 = x_2 \wedge \cdots \wedge X_t = x_t$. In this conditional distribution, we have

$$Z_{a_1,\ldots,a_t} = a_1X_1 + Y + c \text{ and } Z_{b_1,\ldots,b_t} = b_1X_1 + Y + d,$$

where

$$c := a_2x_2 + \cdots + a_tx_t \text{ and } d := b_2x_2 + \cdots + b_tx_t,$$

and $X_1$ and $Y$ are independent and uniformly distributed over $\mathbb{Z}_p$ (this follows from the mutual independence of $X_1, \ldots, X_t, Y$ before conditioning). By the result of the previous example, $(a_1X_1 + Y, b_1X_1 + Y)$ is uniformly distributed over $\mathbb{Z}_p \times \mathbb{Z}_p$, and since the function sending $(x, y) \in \mathbb{Z}_p \times \mathbb{Z}_p$ to $(x+c, y+d) \in \mathbb{Z}_p \times \mathbb{Z}_p$ is a bijection, it follows that $(a_1X_1+Y+c, b_1X_1+Y+d)$ is uniformly distributed over $\mathbb{Z}_p \times \mathbb{Z}_p$. That proves (6.11).

(6.10) now follows easily from (6.11), as follows:

$$\begin{aligned}
&\mathsf{P}[Z_{a_1,\ldots,a_t} = z \wedge Z_{b_1,\ldots,b_t} = z'] \\
&\quad = \sum_{x_2,\ldots,x_t} \mathsf{P}[Z_{a_1,\ldots,a_t} = z \wedge Z_{b_1,\ldots,b_t} = z' \mid X_2 = x_2 \wedge \cdots \wedge X_t = x_t] \cdot \\
&\qquad\qquad \mathsf{P}[X_2 = x_2 \wedge \cdots \wedge X_t = x_t] \\
&\quad = \sum_{x_2,\ldots,x_t} \frac{1}{p^2} \cdot \mathsf{P}[X_2 = x_2 \wedge \cdots \wedge X_t = x_t] \\
&\quad = \frac{1}{p^2} \cdot \sum_{x_2,\ldots,x_t} \mathsf{P}[X_2 = x_2 \wedge \cdots \wedge X_t = x_t] \\
&\quad = \frac{1}{p^2} \cdot 1. \quad \square
\end{aligned}$$

Using other algebraic techniques, there are many ways to construct pairwise and $k$-wise independent families of random variables. Such families play an important role in many areas of computer science.

***Example* 6.19.** Suppose we perform an experiment by executing $n$ Bernoulli trials (see Example 6.3), where each trial succeeds with the same probability $p$, and fails with probability $q := 1 - p$, independently of the outcomes of all the other trials. Let $X$ denote the total number of successes. For $k = 0, \ldots, n$, let us calculate the probability that $X = k$.

To do this, let us introduce indicator variables $X_1, \ldots, X_n$, where for

$i = 1, \ldots, n$, we have $X_i = 1$ if the $i$th trial succeeds, and $X_i = 0$, otherwise. By assumption, the $X_i$ are mutually independent. Then we see that $X = X_1 + \cdots + X_n$. Now, consider a fixed value $k = 0, \ldots, n$. Let $\mathcal{C}_k$ denote the collection of all subsets of $\{1, \ldots, n\}$ of size $k$. For $I \in \mathcal{C}_k$, let $\mathcal{A}_I$ be the event that $X_i = 1$ for all $i \in I$ and $X_i = 0$ for all $i \notin I$. Since the $X_i$ are mutually independent, we see that $\mathsf{P}[\mathcal{A}_I] = p^k q^{n-k}$ (as in Example 6.8). Evidently, the collection of events $\{\mathcal{A}_I\}_{I \in \mathcal{C}_k}$ is a partition of the event that $X = k$. Therefore,

$$\mathsf{P}[X = k] = \sum_{I \in \mathcal{C}_k} \mathsf{P}[\mathcal{A}_I] = \sum_{I \in \mathcal{C}_k} p^k q^{n-k} = |\mathcal{C}_k| p^k q^{n-k}.$$

Finally, since

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

we conclude that

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

The distribution of the random variable $X$ is called a **binomial distribution**. □

Exercise 6.11. Let $X_1, \ldots, X_n$ be random variables, and let $\mathcal{X}_i$ be the image of $X_i$ for $i = 1, \ldots, n$. Show that $X_1, \ldots, X_n$ are mutually independent if and only if for all $i = 2, \ldots, n$, and all $x_1 \in \mathcal{X}_1, \ldots, x_i \in \mathcal{X}_i$, we have

$$\mathsf{P}[X_i = x_i \mid X_{i-1} = x_{i-1} \wedge \cdots \wedge X_1 = x_1] = \mathsf{P}[X_i = x_i].$$

Exercise 6.12. Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be events with corresponding indicator variables $X_1, \ldots, X_n$. Show that the events $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are mutually independent if and only if the random variables $X_1, \ldots, X_n$ are mutually independent. Note: there is actually something non-trivial to prove here, since our definitions for independent events and independent random variables superficially look quite different.

Exercise 6.13. Let $\mathcal{C}$ be an event that occurs with non-zero probability, and let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be a partition of $\mathcal{C}$, such that each event $\mathcal{B}_i$ occurs with non-zero probability. Let $X$ be a random variable whose image is $\mathcal{X}$, and let $\mathbf{D}'$ be a probability distribution on $\mathcal{X}$. Suppose that for each $i = 1, \ldots, n$, the conditional distribution of $X$ given $\mathcal{B}_i$ is equal to $\mathbf{D}'$. Show that the conditional distribution of $X$ given $\mathcal{C}$ is also equal to $\mathbf{D}'$.

Exercise 6.14. Let $n$ be a positive integer, and let $X$ be a random variable whose distribution is uniform over $\{0, \ldots, n-1\}$. For each positive divisor $d$ of $n$, let use define the random variable $X_d := X \bmod d$. Show that for any positive divisors $d_1, \ldots, d_k$ of $n$, the random variables $X_{d_1}, \ldots, X_{d_k}$ are mutually independent if and only if $d_1, \ldots, d_k$ are pairwise relatively prime.

Exercise 6.15. With notation as in the previous exercise, let $n := 30$, and describe the conditional distribution of $X_{15}$ given that $X_6 = 1$.

Exercise 6.16. Let $W, X, Y$ be mutually independent and uniformly distributed over $\mathbb{Z}_p$, where $p$ is prime. For any $a \in \mathbb{Z}_p$, let $Z_a := a^2W + aX + Y$. Show that each $Z_a$ is uniformly distributed over $\mathbb{Z}_p$, and that the collection of all $Z_a$ is 3-wise independent.

Exercise 6.17. Let $X_{ib}$, for $i = 1, \ldots, k$ and $b \in \{0, 1\}$, be mutually independent random variables, each with a uniform distribution on $\{0, 1\}$. For $b_1, \ldots, b_k \in \{0, 1\}$, let us define the random variable

$$Y_{b_1 \cdots b_k} := X_{1b_1} \oplus \cdots \oplus X_{kb_k},$$

where "$\oplus$" denotes "exclusive or." Show that the $2^k$ variables $Y_{b_1 \cdots b_k}$ are pairwise independent, each with a uniform distribution over $\{0, 1\}$.

## 6.4 Expectation and variance

Let $\mathbf{D} = (\mathcal{U}, \mathsf{P})$ be a probability distribution. If $X$ is a real random variable, then its **expected value** is

$$\mathsf{E}[X] := \sum_{u \in \mathcal{U}} X(u) \cdot \mathsf{P}[u]. \tag{6.12}$$

If $\mathcal{X}$ is the image of $X$, we have

$$\mathsf{E}[X] = \sum_{x \in \mathcal{X}} \sum_{u \in X^{-1}(\{x\})} x\mathsf{P}[u] = \sum_{x \in \mathcal{X}} x \cdot \mathsf{P}[X = x]. \tag{6.13}$$

From (6.13), it is clear that $\mathsf{E}[X]$ depends only on the distribution of $X$ (and not on any other properties of the underlying distribution $\mathbf{D}$). More generally, by a similar calculation, one sees that if $X$ is any random variable with image $\mathcal{X}$, and $f$ is a real-valued function on $\mathcal{X}$, then

$$\mathsf{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)\mathsf{P}[X = x]. \tag{6.14}$$

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., $X(u) = c$ for all

$u \in \mathcal{U}$), then $\mathsf{E}[X] = \mathsf{E}[c] = c$. Second, if $X$ takes only non-negative values (i.e., $X(u) \geq 0$ all $u \in \mathcal{U}$), then $\mathsf{E}[X] \geq 0$. Similarly, if $X$ takes only positive values, then $\mathsf{E}[X] > 0$.

A crucial property about expectation is the following:

**Theorem 6.6 (Linearity of expectation).** *For real random variables $X$ and $Y$, and real number $a$, we have*

$$\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y]$$

*and*

$$\mathsf{E}[aX] = a\mathsf{E}[X].$$

*Proof.* It is easiest to prove this using the defining equation (6.12) for expectation. For $u \in \mathcal{U}$, the value of the random variable $X+Y$ at $u$ is by definition $X(u) + Y(u)$, and so we have

$$\begin{aligned} \mathsf{E}[X+Y] &= \sum_{u \in \mathcal{U}} (X(u) + Y(u))\mathsf{P}[u] \\ &= \sum_{u \in \mathcal{U}} X(u)\mathsf{P}[u] + \sum_{u \in \mathcal{U}} Y(u)\mathsf{P}[u] \\ &= \mathsf{E}[X] + \mathsf{E}[Y]. \end{aligned}$$

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

$$\mathsf{E}[aX] = \sum_u (aX(u))\mathsf{P}[u] = a\sum_u X(u)\mathsf{P}[u] = a\mathsf{E}[X]. \quad \square$$

More generally, the above theorem implies (using a simple induction argument) that for any real random variables $X_1, \ldots, X_n$, we have

$$\mathsf{E}[X_1 + \cdots + X_n] = \mathsf{E}[X_1] + \cdots + \mathsf{E}[X_n].$$

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

**Theorem 6.7.** *If $X$ and $Y$ are independent real random variables, then $\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y]$.*

*Proof.* It is easiest to prove this using (6.14). We have

$$\begin{aligned} \mathsf{E}[XY] &= \sum_{x,y} xy\mathsf{P}[X = x \wedge Y = y] \\ &= \sum_{x,y} xy\mathsf{P}[X = x]\mathsf{P}[Y = y] \end{aligned}$$

$$
\begin{aligned}
&= \Big(\sum_x x\mathsf{P}[X = x]\Big)\Big(\sum_y y\mathsf{P}[Y = y]\Big)\\
&= \mathsf{E}[X] \cdot \mathsf{E}[Y]. \quad \Box
\end{aligned}
$$

More generally, the above theorem implies (using a simple induction argument) that if $X_1, \ldots, X_n$ are mutually independent real random variables, then

$$\mathsf{E}[X_1 \cdots X_n] = \mathsf{E}[X_1] \cdots \mathsf{E}[X_n].$$

The following fact is sometimes quite useful:

**Theorem 6.8.** *If $X$ is a random variable that takes values in the set $\{0, 1, \ldots, n\}$, then*

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

*Proof.* For $i = 1, \ldots, n$, define the random variable $X_i$ so that $X_i = 1$ if $X \geq i$ and $X_i = 0$ if $X < i$. Note that $\mathsf{E}[X_i] = 1 \cdot \mathsf{P}[X \geq i] + 0 \cdot \mathsf{P}[X < i] = \mathsf{P}[X \geq i]$. Moreover, $X = X_1 + \cdots + X_n$, and hence

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

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

**Theorem 6.9.** *Let $X$ be a real random variable, and let $a$ and $b$ be real numbers. Then we have*

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

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

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

*Proof.* Let $\mu := \mathsf{E}[X]$. 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 (verify):

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

A simple consequence of part (i) of Theorem 6.9 is that $\mathsf{E}[X^2] \geq (\mathsf{E}[X])^2$.

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 6.10.** *If $X_1, \ldots, X_n$ is a collection of pairwise independent real random variables, then*

$$\mathsf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathsf{Var}[X_i].$$

*Proof.* We have

$$\begin{aligned}\mathsf{Var}\left[\sum_i X_i\right] &= E\left[(\sum_i X_i)^2\right] - \left(\mathsf{E}[\sum_i X_i]\right)^2 \\ &= \sum_i \mathsf{E}[X_i^2] + 2\sum_{\substack{i,j \\ j<i}}(\mathsf{E}[X_iX_j] - \mathsf{E}[X_i]\mathsf{E}[X_j]) - \sum_i \mathsf{E}[X_i]^2 \\ &\quad \text{(by Theorem 6.6 and rearranging terms)} \\ &= \sum_i \mathsf{E}[X_i^2] - \sum_i \mathsf{E}[X_i]^2 \\ &\quad \text{(by pairwise independence and Theorem 6.7)} \\ &= \sum_i \mathsf{Var}[X_i]. \quad \square\end{aligned}$$

For any random variable $X$ and event $\mathcal{B}$, with $\mathsf{P}[\mathcal{B}] \neq 0$, we can define the **conditional expectation of $X$ given $\mathcal{B}$**, denoted $\mathsf{E}[X \mid \mathcal{B}]$, to be the expected value of $X$ in the conditional probability distribution given $\mathcal{B}$. We have

$$\mathsf{E}[X \mid \mathcal{B}] = \sum_{u \in \mathcal{U}} X(u) \cdot \mathsf{P}[u \mid \mathcal{B}] = \sum_{x \in \mathcal{X}} x \mathsf{P}[X = x \mid \mathcal{B}], \tag{6.15}$$

where $\mathcal{X}$ is the image of $X$.

If $\mathcal{B}_1, \ldots, \mathcal{B}_n$ is a collection of events that partitions $\mathcal{U}$, where each $\mathcal{B}_i$ occurs with non-zero probability, then it follows from the definitions that

$$\mathsf{E}[X] = \sum_{i=1}^{n} \mathsf{E}[X \mid \mathcal{B}_i] \mathsf{P}[\mathcal{B}_i]. \tag{6.16}$$

***Example* 6.20.** Let $X$ be uniformly distributed over $\{1, \ldots, n\}$. Let us compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$. We have

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

We also have

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

Therefore,

$$\mathsf{Var}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2 = \frac{n^2-1}{12}. \quad \Box$$

***Example* 6.21.** Let $X$ denote the value of a die toss. Let $\mathcal{A}$ be the event that $X$ is even. Then in the conditional probability space given $\mathcal{A}$, we see that $X$ is uniformly distributed over $\{2, 4, 6\}$, and hence

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

Similarly, in the conditional probability space given $\overline{\mathcal{A}}$, we see that $X$ is uniformly distributed over $\{1, 3, 5\}$, and hence

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

We can compute the expected value of $X$ using these conditional expectations; indeed, we have

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

which agrees with the calculation in previous example. □

***Example* 6.22.** Suppose that a random variable $X$ takes the value 1 with probability $p$, and 0 with probability $q := 1 - p$. The distribution of $X$ is that of a Bernoulli trial, as discussed in Example 6.3. Let us compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$. We have

$$\mathsf{E}[X] = 1 \cdot p + 0 \cdot q = p.$$

We also have

$$\mathsf{E}[X^2] = 1^2 \cdot p + 0^2 \cdot q = p.$$

Therefore,

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

***Example* 6.23.** Suppose that $X_1, \ldots, X_n$ are mutually independent random variables such that each $X_i$ takes the value 1 with probability $p$ and 0 with probability $q := 1 - p$. Let us set $X := X_1 + \cdots + X_n$. Note that the distribution of each $X_i$ is that of a Bernoulli trial, as in Example 6.3, and the distribution of $X$ is a binomial distribution, as in Example 6.19. By the previous example, we have $\mathsf{E}[X_i] = p$ and $\mathsf{Var}[X_i] = pq$ for $i = 1, \ldots, n$. Let us compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$. By Theorem 6.6, we have

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

and by Theorem 6.10, and the fact that the $X_i$ are mutually independent, we have

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

Exercise 6.18. A casino offers you the following four dice games. In each game, you pay 15 dollars to play, and two dice are rolled. In the first game, the house pays out four times the value of the first die (in dollars). In the second, the house pays out twice the sum of the two dice. In the third, the house pays the square of the first. In the fourth, the house pays the product of the two dice. Which game should you play? That is, which game maximizes your expected winnings?

Exercise 6.19. Suppose $X$ and $Y$ are independent real random variables such that $\mathsf{E}[X] = \mathsf{E}[Y]$. Show that

$$\mathsf{E}[(X - Y)^2] = \mathsf{Var}[X] + \mathsf{Var}[Y].$$

Exercise 6.20. Show that the variance of any 0/1-valued random variable is at most 1/4.

Exercise 6.21. A die is tossed repeatedly until it comes up "1," or until it is tossed $n$ times (whichever comes first). What is the expected number of tosses of the die?

Exercise 6.22. Suppose that 20 percent of the students who took a certain test were from school $A$ and the average of their scores on the test was 65. Also, suppose that 30 percent of the students were from school $B$ and the average of their scores was 85. Finally, suppose that the remaining 50 percent of the students were from school $C$ and the average of their scores was 72. If a student is selected at random from the entire group that took the test, what is the expected value of his score?

Exercise 6.23. An urn contains $r \geq 0$ red balls and $b \geq 1$ black balls. Consider the following experiment. At each step in the experiment, a single ball is removed from the urn, randomly chosen from among all balls that remain in the urn: if a black ball is removed, the experiment halts, and if a red ball is removed, the experiment continues (without returning the red ball to the urn). Show that the expected number of steps performed is $(r+b+1)/(b+1)$.

## 6.5 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 6.11 (Markov's inequality).** *Let $X$ be a random variable that takes only non-negative real values. Then for any $t > 0$, we have*

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

*Proof.* We have

$$\mathsf{E}[X] = \sum_x x\mathsf{P}[X = x] \;=\; \sum_{x<t} x\mathsf{P}[X = x] + \sum_{x \geq t} x\mathsf{P}[X = x].$$

Since $X$ takes only non-negative values, all of the terms in the summation are non-negative. Therefore,

$$\mathsf{E}[X] \geq \sum_{x \geq t} x\mathsf{P}[X = x] \;\geq\; \sum_{x \geq t} t\mathsf{P}[X = x] \;=\; t\mathsf{P}[X \geq t]. \quad \square$$

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 6.12 (Chebyshev's inequality).** *Let $X$ be a real random variable. Then for any $t > 0$, we have*

$$\mathsf{P}[|X - \mathsf{E}[X]| \geq t] \leq \mathsf{Var}[X]/t^2.$$

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

$$\mathsf{P}[|X - \mathsf{E}[X]| \geq t] = \mathsf{P}[Y \geq t^2] \leq \mathsf{Var}[X]/t^2. \quad \square$$

An important special case of Chebyshev's inequality is the following. Suppose that $X_1, \ldots, X_n$ are *pairwise independent* real random variables, each with the same distribution. Let $\mu$ be the common value of $\mathsf{E}[X_i]$ and $\nu$ the common value of $\mathsf{Var}[X_i]$. Set

$$\overline{X} := \frac{1}{n}(X_1 + \cdots + X_n).$$

The variable $\overline{X}$ is called the **sample mean** of $X_1, \ldots, X_n$. By the linearity of expectation, we have $\mathsf{E}[\overline{X}] = \mu$, and since the $X_i$ are pairwise independent, it follows from Theorem 6.10 (along with part (ii) of Theorem 6.9) that $\mathsf{Var}[\overline{X}] = \nu/n$. Applying Chebyshev's inequality, for any $\epsilon > 0$, we have

$$\mathsf{P}[|\overline{X} - \mu| \geq \epsilon] \leq \frac{\nu}{n\epsilon^2}. \tag{6.17}$$

The inequality (6.17) says that for all $\epsilon > 0$, and for all $\delta > 0$, there exists $n_0$ (depending on $\epsilon$ and $\delta$, as well as the variance $\nu$) such that $n \geq n_0$ implies

$$\mathsf{P}[|\overline{X} - \mu| \geq \epsilon] \leq \delta. \tag{6.18}$$

In words:

> *As $n$ gets large, the sample mean closely approximates the expected value $\mu$ 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. Suppose that $X_1, \ldots, X_n$ are pairwise independent random variables, each of which takes the value 1 with probability $p$, and 0 with probability $q := 1-p$. As before, let $\overline{X}$ be the sample mean of $X_1, \ldots, X_n$. As we calculated in

Example 6.22, the $X_i$ have a common expected value $p$ and variance $pq$. Therefore, by (6.17), for any $\epsilon > 0$, we have

$$\mathsf{P}[|\overline{X} - p| \geq \epsilon] \leq \frac{pq}{n\epsilon^2}. \tag{6.19}$$

The bound on the right-hand side of (6.19) decreases linearly in $n$. If one makes the stronger assumption that the $X_i$ are *mutually independent* (so that $X := X_1 + \cdots + X_n$ has a binomial distribution), one can obtain a much better bound that decreases *exponentially* in $n$:

**Theorem 6.13 (Chernoff bound).** *Let $X_1, \ldots, X_n$ be mutually independent 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 $\overline{X}$ be the sample mean of $X_1, \ldots, X_n$. Then for any $\epsilon > 0$, we have:*

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

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

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

*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 $\overline{X} - p \geq \epsilon$ if and only if $Z \geq e^{\alpha n \epsilon}$. By Markov's inequality, it follows that

$$\mathsf{P}[\overline{X} - p \geq \epsilon] = \mathsf{P}[Z \geq e^{\alpha n \epsilon}] \leq \mathsf{E}[Z]e^{-\alpha n \epsilon}. \tag{6.20}$$

So our goal is to bound $E[Z]$ from above.

For $i = 1, \ldots, n$, define the random variable $Z_i := e^{\alpha(X_i - p)}$. Note that $Z = \prod_{i=1}^n Z_i$, that the $Z_i$ are mutually independent random variables (see Theorem 6.5), and that

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

It follows that

$$\mathsf{E}[Z] = \mathsf{E}[\prod_i Z_i] = \prod_i \mathsf{E}[Z_i] = (pe^{\alpha q} + qe^{-\alpha p})^n.$$

We will prove below that

$$pe^{\alpha q} + qe^{-\alpha p} \leq e^{\alpha^2 q/2}. \tag{6.21}$$

From this, it follows that

$$\mathsf{E}[Z] \le e^{\alpha^2 qn/2}. \tag{6.22}$$

Combining (6.22) with (6.20), we obtain

$$\mathsf{P}[\overline{X} - p \ge \epsilon] \le e^{\alpha^2 qn/2 - \alpha n\epsilon}. \tag{6.23}$$

Now we choose the parameter $\alpha$ so as to minimize the quantity $\alpha^2 qn/2 - \alpha n\epsilon$. The optimal value of $\alpha$ is easily seen to be $\alpha = \epsilon/q$, and substituting this value of $\alpha$ into (6.23) yields (i).

To finish the proof of the theorem, it remains to prove the inequality (6.21). Let

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

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

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

This establishes (6.21) 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* 6.24.** Suppose we toss 10,000 coins. The expected number of heads is 5,000. What is an upper bound on the probability $\alpha$ that we get 6,000 or more heads? Using Markov's inequality, we get $\alpha \le 5/6$. Using Chebyshev's inequality, and in particular, the inequality (6.19), we get

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

Finally, using the Chernoff bound, we obtain

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

EXERCISE 6.24. You are given a biased coin. You know that if tossed, it will come up heads with probability at least 51%, or it will come up tails with probability at least 51%. Design an experiment that attempts to determine the direction of the bias (towards heads or towards tails). The experiment should work by flipping the coin some number $t$ times, and it should correctly determine the direction of the bias with probability at least 99%. Try to make $t$ as small as possible.

## 6.6 The birthday paradox

This section discusses a number of problems related to the following question: how many people must be in a room before there is a good chance that two of them were born on the same day of the year? The answer is surprisingly few, whence the "paradox."

To answer this question, we index the people in the room with integers $1, \ldots, k$, where $k$ is the number of people in the room. We abstract the problem a bit, and assume that all years have the same number of days, say $n$—setting $n = 365$ corresponds to the original problem, except that leap years are not handled correctly, but we shall ignore this detail. For $i = 1, \ldots, k$, let $X_i$ denote the day of the year on which $i$'s birthday falls. Let us assume that birthdays are uniformly distributed over $\{0, \ldots, n-1\}$; this assumption is actually not entirely realistic, as it is well known that people are somewhat more likely to be born in some months than in others.

So for any $i = 1, \ldots, k$ and $x = 0, \ldots, n-1$, we have $\mathsf{P}[X_i = x] = 1/n$.

Let $\alpha$ be the probability that no two persons share the same birthday, so that $1 - \alpha$ is the probability that there is a pair of matching birthdays. We would like to know how big $k$ must be relative to $n$ so that $\alpha$ is not too large, say, at most $1/2$.

We can compute $\alpha$ as follows, assuming the $X_i$ are *mutually independent.*

There are a total of $n^k$ sequences of integers $(x_1, \ldots, x_k)$, with each $x_i \in \{0, \ldots, n-1\}$. Among these, there are a total of $n(n-1)\cdots(n-k+1)$ that contain no repetitions: there are $n$ choices for $x_1$, and for any fixed value of $x_1$, there are $n-1$ choices for $x_2$, and so on. Therefore

$$\alpha = n(n-1)\cdots(n-k+1)/n^k = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right). \quad (6.24)$$

Using the part (i) of §A1, we obtain

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

So if $k(k-1) \geq (2\log 2)n$, we have $\alpha \leq 1/2$. Thus, when $k$ is at least a small constant times $n^{1/2}$, we have $\alpha \leq 1/2$, so the probability that two people share the same birthday is at least $1/2$. For $n = 365$, $k \geq 23$ suffices. Indeed, one can simply calculate $\alpha$ in this case numerically from equation (6.24), obtaining $\alpha \approx 0.493$. Thus, if there are 23 people in the room, there is about a 50-50 chance that two people have the same birthday.

The above analysis assumed the $X_i$ are mutually independent. However, we can still obtain useful upper bounds for $\alpha$ under much weaker independence assumptions.

For $i = 1, \ldots, k$ and $j = i+1, \ldots, k$, let us define the indicator variable

$$W_{ij} := \begin{cases} 1 & \text{if } X_i = X_j, \\ 0 & \text{if } X_i \neq X_j. \end{cases}$$

If we assume that the $X_i$ are pairwise independent, then

$$\begin{aligned} \mathsf{P}[W_{ij} = 1] = \mathsf{P}[X_i = X_j] &= \sum_{x=0}^{n-1} \mathsf{P}[X_i = x \wedge X_j = x] \\ &= \sum_{x=0}^{n-1} \mathsf{P}[X_i = x]\mathsf{P}[X_j = x] = \sum_{x=0}^{n-1} 1/n^2 = 1/n. \end{aligned}$$

We can compute the expectation and variance (see Example 6.22):

$$\mathsf{E}[W_{ij}] = \frac{1}{n}, \qquad \mathsf{Var}[W_{ij}] = \frac{1}{n}(1 - \frac{1}{n}).$$

Now consider the random variable

$$W := \sum_{i=1}^{k} \sum_{j=i+1}^{k} W_{ij},$$

which represents the number of distinct pairs of people with the same birthday. There are $k(k-1)/2$ terms in this sum, so by the linearity of expectation, we have

$$\mathsf{E}[W] = \frac{k(k-1)}{2n}.$$

Thus, for $k(k-1) \geq 2n$, we "expect" there to be at least one pair of matching birthdays. However, this does not guarantee that the probability of a matching pair of birthdays is very high, assuming just pairwise independence of the $X_i$. For example, suppose that $n$ is prime and the $X_i$ are a subset of the family of pairwise independent random variables defined in Example 6.17. That is, each $X_i$ is of the form $a_i X + Y$, where $X$ and $Y$ are uniformly and independently distributed modulo $n$. Then in fact, either all the $X_i$ are distinct, or they are all equal, where the latter event occurs exactly when $X = [0]_n$, and so with probability $1/n$— "when it rains, it pours."

To get a useful upper bound on the probability $\alpha$ that there are no matching birthdays, it suffices to assume that the $X_i$ are *4-wise independent*. In this case, it is easy to verify that the variables $W_{ij}$ are *pairwise* independent, since any two of the $W_{ij}$ are determined by at most four of the $X_i$. Therefore, in this case, the variance of the sum is equal to the sum of the variances, and so

$$\mathsf{Var}[W] = \frac{k(k-1)}{2n}(1 - \frac{1}{n}) \leq \mathsf{E}[W].$$

Furthermore, by Chebyshev's inequality,

$$\begin{aligned}\alpha = \mathsf{P}[W = 0] &\leq \mathsf{P}[|W - \mathsf{E}[W]| \geq \mathsf{E}[W]] \\ &\leq \mathsf{Var}[W]/\mathsf{E}[W]^2 \leq 1/\mathsf{E}[W] = \frac{2n}{k(k-1)}.\end{aligned}$$

Thus, if $k(k-1) \geq 4n$, then $\alpha \leq 1/2$.

In many practical applications, it is more important to bound $\alpha$ from *below*, rather than from above; that is, to bound from above the probability $1 - \alpha$ that there are any collisions. For this, pairwise independence of the $X_i$ suffices, since than we have $\mathsf{P}[W_{ij} = 1] = 1/n$, and by (6.5), we have

$$1 - \alpha \leq \sum_{i=1}^{k} \sum_{j=i+1}^{k} \mathsf{P}[W_{ij} = 1] = \frac{k(k-1)}{2n},$$

which is at most $1/2$ provided $k(k-1) \leq n$.

Exercise 6.25. Let $\alpha_1, \ldots, \alpha_n$ be real numbers with $\sum_{i=1}^{n} \alpha_i = 1$. Show that

$$0 \leq \sum_{i=1}^{n} (\alpha_i - 1/n)^2 = \sum_{i=1}^{n} \alpha_i^2 - 1/n,$$

and in particular,

$$\sum_{i=1}^{n} \alpha_i^2 \geq 1/n.$$

Exercise 6.26. Let $\mathcal{X}$ be a set of size $n \geq 1$, and let $X$ and $X'$ be independent random variables, taking values in $\mathcal{X}$, and with the same distribution. Show that

$$\mathsf{P}[X = X'] = \sum_{x \in \mathcal{X}} \mathsf{P}[X = x]^2 \geq \frac{1}{n}.$$

Exercise 6.27. Let $\mathcal{X}$ be a set of size $n \geq 1$, and let $x_0$ be an arbitrary, fixed element of $\mathcal{X}$. Consider a random experiment in which a function $F$ is chosen uniformly from among all $n^n$ functions from $\mathcal{X}$ into $\mathcal{X}$. Let us define random variables $X_i$, for $i = 0, 1, 2, \ldots$, as follows:

$$X_0 := x_0, \quad X_{i+1} := F(X_i) \;\; (i = 0, 1, 2, \ldots).$$

Thus, the value of $X_i$ is obtained by applying the function $F$ a total of $i$ times to the starting value $x_0$. Since $\mathcal{X}$ has size $n$, the sequence $\{X_i\}$ must repeat at some point; that is, there exists a positive integer $k$ (with $k \le n$) such that $X_k = X_i$ for some $i = 0, \ldots, k-1$. Define the random variable $K$ to be the smallest such value $k$.

(a) Show that for any $i \ge 0$ and any fixed values of $x_1, \ldots, x_i \in \mathcal{X}$ such that $x_0, x_1, \ldots, x_i$ are distinct, the conditional distribution of $X_{i+1}$ given that $X_1 = x_1, \ldots, X_i = x_i$ is uniform over $\mathcal{X}$.

(b) Show that for any integer $k \ge 1$, we have $K \ge k$ if and only if $X_0, X_1, \ldots, X_{k-1}$ take on distinct values.

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

$$\mathsf{P}[K \ge k \mid K \ge k-1] = 1 - (k-1)/n,$$

and conclude that

$$\mathsf{P}[K \ge k] = \prod_{i=1}^{k-1} (1 - i/n) \le e^{-k(k-1)/2n}.$$

(d) Show that

$$\sum_{k=1}^{\infty} e^{-k(k-1)/2n} = O(n^{1/2})$$

and then conclude from part (c) that

$$\mathsf{E}[K] = \sum_{k=1}^{n} \mathsf{P}[K \ge k] \le \sum_{k=1}^{\infty} e^{-k(k-1)/2n} = O(n^{1/2}).$$

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

EXERCISE 6.28. The setup for this exercise is identical to that of the previous exercise, except that now, the function $F$ is chosen uniformly from among all $n!$ *permutations* of $\mathcal{X}$.

(a) Show that if $K = k$, then $X_k = X_0$.

(b) Show that for any $i \ge 0$ and any fixed values of $x_1, \ldots, x_i \in \mathcal{X}$ such that $x_0, x_1, \ldots, x_i$ are distinct, the conditional distribution of $X_{i+1}$ given that $X_1 = x_1, \ldots, X_i = x_i$ is uniform over $\mathcal{X} \setminus \{x_1, \ldots, x_i\}$.

(c) Show that for any $k = 2, \ldots, n$, we have

$$\mathsf{P}[K \ge k \mid K \ge k-1] = 1 - \frac{1}{n-k+2},$$

and conclude that for all $k = 1, \dots, n$, we have

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

(d) From part (c), show that $K$ is uniformly distributed over $\{1, \dots, n\}$, and in particular,

$$\mathsf{E}[K] = \frac{n+1}{2}.$$

## 6.7 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.

The scenario is as follows. We have finite, non-empty sets $\mathcal{A}$ and $\mathcal{Z}$, with $|\mathcal{A}| = k$ and $|\mathcal{Z}| = n$, and a finite, non-empty set $\mathcal{H}$ of **hash functions**, each of which map elements of $\mathcal{A}$ into $\mathcal{Z}$. More precisely, each element $h \in \mathcal{H}$ defines a function that maps $a \in \mathcal{A}$ to an element $z \in \mathcal{Z}$, and we write $z = h(a)$; the value $z$ is called the **hash code of** $a$ **(under** $h$**)**, and we say that $a$ **hashes to** $z$ **(under** $h$**)**. Note that two distinct elements of $\mathcal{H}$ may happen to define the same function. We call $\mathcal{H}$ a **family of hash functions (from** $\mathcal{A}$ **to** $\mathcal{Z}$**)**.

Let $H$ be a random variable whose distribution is uniform on $\mathcal{H}$. For any $a \in \mathcal{A}$, $H(a)$ denotes the random variable whose value is $z = h(a)$ when $H = h$. For any $\ell = 1, \dots, k$, we say that $\mathcal{H}$ is an **$\ell$-wise independent** family of hash functions if each $H(a)$ is uniformly distributed over $\mathcal{Z}$, and the collection of all $H(a)$ is $\ell$-wise independent; in case $\ell = 2$, we say that $\mathcal{H}$ is a **pairwise independent** family of hash functions. Pairwise independence is equivalent to saying that for all $a, a' \in \mathcal{A}$, with $a \neq a'$, and all $z, z' \in \mathcal{Z}$,

$$\mathsf{P}[H(a) = z \wedge H(a') = z'] = \frac{1}{n^2}.$$

***Example* 6.25.** Examples 6.17 and 6.18 provide explicit constructions for pairwise independent families of hash functions. In particular, from the discussion in Example 6.17, if $n$ is prime, and we take $\mathcal{A} := \mathbb{Z}_n$, $\mathcal{Z} := \mathbb{Z}_n$, and $\mathcal{H} := \{h_{x,y} : x, y \in \mathbb{Z}_n\}$, where for $h_{x,y} \in \mathcal{H}$ and $a \in \mathcal{A}$ we define $h_{x,y}(a) := ax + y$, then $\mathcal{H}$ is a pairwise independent family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$.

Similarly, Example 6.18 yields a pairwise independent family of hash functions from $\mathcal{A} := \mathbb{Z}_n^{\times t}$ to $\mathcal{Z} := \mathbb{Z}_n$, with $\mathcal{H} := \{h_{x_1,\dots,x_t,y} : x_1, \dots, x_t, y \in \mathbb{Z}_n\}$, where for $h_{x_1,\ldots,x_t,y} \in \mathcal{H}$ and $(a_1, \ldots, a_t) \in \mathcal{A}$, we define

$$h_{x_1,\ldots,x_t,y}(a_1, \ldots, a_t) := a_1x_1 + \cdots + a_tx_t + y.$$

In practice, the inputs to such a hash function may be long bit strings, which we chop into small pieces so that each piece can be viewed as an element of $\mathbb{Z}_n$. □

### *6.7.1 Hash tables*

Pairwise independent families of hash functions may be used to implement a data structure known as a **hash table**, which in turn may be used to implement a **dictionary**.

Assume that $\mathcal{H}$ is a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $|\mathcal{A}| = k$ and $|\mathcal{Z}| = n$. A hash function is chosen at random from $\mathcal{H}$; an element $a \in \mathcal{A}$ is inserted into the hash table by storing the value of $a$ into a **bin** indexed by the hash code of $a$; likewise, to see if a particular value $a \in \mathcal{A}$ is stored in the hash table, one must search in the bin indexed by the hash code of $a$.

So as to facilitate fast storage and retrieval, one typically wants the elements stored in the hash table to be distributed in roughly equal proportions among all the bins.

Assuming that $\mathcal{H}$ is a pairwise independent family of hash functions, one can easily derive some useful results, such as the following:

- If the hash table holds $q$ values, then for any value $a \in \mathcal{A}$, the expected number of other values that are in the bin indexed by $a$'s hash code is at most $q/n$. This result bounds the expected amount of "work" we have to do to search for a value in its corresponding bin, which is essentially equal to the size of the bin. In particular, if $q = O(n)$, then the expected amount of work is constant. See Exercise 6.32 below.
- If the table holds $q$ values, with $q(q-1) \leq n$, then with probability at least $1/2$, each value lies in a distinct bin. This result is useful if one wants to find a "perfect" hash function that hashes $q$ fixed values to distinct bins: if $n$ is sufficiently large, we can just choose hash functions at random until we find one that works. See Exercise 6.33 below.
- If the table holds $n$ values, then the expected value of the maximum number of values in any bin is $O(n^{1/2})$. See Exercise 6.34 below.

Results such as these, and others, can be obtained using a broader notion of hashing called **universal hashing**. We call $\mathcal{H}$ a **universal** family of hash functions if for all $a, a' \in \mathcal{A}$, with $a \neq a'$, we have

$$\mathsf{P}[H(a) = H(a')] \leq \frac{1}{n}.$$

Note that the pairwise independence property implies the universal property (see Exercise 6.29 below). There are even weaker notions that are relevant in practice; for example, in some applications, it is sufficient to require that $\mathsf{P}[H(a) = H(a')] \leq c/n$ for some constant $c$.

Exercise 6.29. Show that any pairwise independent family of hash functions is also a universal family of hash functions.

Exercise 6.30. Let $\mathcal{A} := \mathbb{Z}_n^{\times(t+1)}$ and $\mathcal{Z} := \mathbb{Z}_n$, where $n$ is prime. Let $\mathcal{H} := \{h_{x_1,\ldots,x_t} : x_1, \ldots, x_t \in \mathbb{Z}_n\}$ be a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where for $h_{x_1,\ldots,x_t} \in \mathcal{H}$, and for $(a_0, a_1, \ldots, a_t) \in \mathcal{A}$, we define

$$h_{x_1,\ldots,x_t}(a_0, a_1, \ldots, a_t) := a_0 + a_1 x_1 + \cdots + a_t x_t.$$

Show that $\mathcal{H}$ is universal, but not pairwise independent.

Exercise 6.31. Let $k$ be a prime and let $n$ be any positive integer. Let $\mathcal{A} := \{0, \ldots, k-1\}$ and $\mathcal{Z} := \{0, \ldots, n-1\}$. Let

$$\mathcal{H} := \{h_{x,y} : x = 1, \ldots, k-1,\ y = 0, \ldots, k-1\},$$

be a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where for $h_{x,y} \in \mathcal{H}$ and for $a \in \mathcal{A}$, we define

$$h_{x,y}(a) := ((ax + y) \bmod k) \bmod n.$$

Show that $\mathcal{H}$ is universal. Hint: first show that for any $a, a' \in \mathcal{A}$ with $a \neq a'$, the number of $h \in \mathcal{H}$ such that $h(a) = h(a')$ is equal to the number of pairs of integers $(r, s)$ such that

$$0 \leq r < k,\ 0 \leq s < k,\ r \neq s,\ \text{and } r \equiv s \pmod{n}.$$

In the following three exercises, assume that $\mathcal{H}$ is a universal family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $|\mathcal{A}| = k$ and $|\mathcal{Z}| = n$, and that $H$ is a random variable uniformly distributed over $\mathcal{H}$.

Exercise 6.32. Let $a_1, \ldots, a_q$ be distinct elements of $\mathcal{A}$, and let $a \in \mathcal{A}$. Define $L$ to be the number of indices $i = 1, \ldots, q$ such that $H(a_i) = H(a)$. Show that

$$\mathsf{E}[L] \leq \begin{cases} 1 + (q-1)/n & \text{if } a \in \{a_1, \ldots, a_q\}; \\ q/n & \text{otherwise.} \end{cases}$$

Exercise 6.33. Let $a_1, \ldots, a_q$ be distinct elements of $\mathcal{A}$, and assume that $q(q-1) \le n$. Show that the probability that $H(a_i) = H(a_j)$ for some $i, j$ with $i \ne j$, is at most 1/2.

Exercise 6.34. Assume $k \ge n$, and let $a_1, \ldots, a_n$ be distinct elements of $\mathcal{A}$. For $z \in \mathcal{Z}$, define the random variable $B_z := \{a_i : H(a_i) = z\}$. Define the random variable $M := \max\{|B_z| : z \in \mathcal{Z}\}$. Show that $\mathsf{E}[M] = O(n^{1/2})$.

Exercise 6.35. A family $\mathcal{H}$ of hash functions from $\mathcal{A}$ to $\mathcal{Z}$ is called $\epsilon$-**universal** if for $H$ uniformly distributed over $\mathcal{H}$, and for all $a, a' \in \mathcal{A}$ with $a \ne a'$, we have $\mathsf{P}[H(a) = H(a')] \le \epsilon$. Show that if $\mathcal{H}$ is $\epsilon$-universal, then we must have

$$\epsilon \ge \frac{1}{|\mathcal{Z}|} - \frac{1}{|\mathcal{A}|}.$$

Hint: using Exercise 6.26, first show that if $H, A, A'$ are mutually independent random variables, with $H$ uniformly distributed over $\mathcal{H}$, and $A$ and $A'$ uniformly distributed over $\mathcal{A}$, then $\mathsf{P}[A \ne A' \wedge H(A) = H(A')] \ge 1/|\mathcal{Z}| - 1/|\mathcal{A}|$.

### 6.7.2 Message authentication

Pairwise independent families of hash functions may be used to implement a **message authentication scheme**, which is a mechanism to detect if a message has been tampered with in transit between two parties. Unlike an error correcting code (such as the one discussed in §4.5.1), a message authentication scheme should be effective against *arbitrary* tampering.

As above, assume that $\mathcal{H}$ is a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $|\mathcal{A}| = k$ and $|\mathcal{Z}| = n$. Suppose that Alice and Bob somehow agree upon a hash function chosen at random from $\mathcal{H}$. At some later time, Alice transmits a message $a \in \mathcal{A}$ to Bob over an insecure network. In addition to sending $a$, Alice also sends the hash code $z$ of $a$. Upon receiving a pair $(a, z)$, Bob checks that the hash code of $a$ is indeed equal to $z$: if so, he accepts the message as authentic (i.e., originating from Alice); otherwise, he rejects the message.

Now suppose that an adversary is trying to trick Bob into accepting an inauthentic message (i.e., one not originating from Alice). Assuming that $\mathcal{H}$ is a pairwise independent family of hash functions, it is not too hard to see that the adversary can succeed with probability no better than $1/n$, regardless of the strategy or computing power of the adversary. Indeed, on the one hand, suppose the adversary gives Bob a pair $(a', z')$ at some time before Alice sends her message. In this case, the adversary knows nothing about the hash function, and so the correct value of the hash code of $a'$ is completely unpredictable: it is equally likely to be any element of $\mathcal{Z}$. Therefore, no matter how clever the adversary is in choosing $a'$ and $z'$, Bob will accept $(a', z')$ as authentic with probability only $1/n$. On the other hand, suppose the adversary waits until Alice sends her message, intercepting the message/hash code pair $(a, z)$ sent by Alice, and gives Bob a pair $(a', z')$, where $a' \neq a$, instead of the pair $(a, z)$. Again, since the adversary does not know anything about the hash function other than the fact that the hash code of $a$ is equal to $z$, the correct hash code of $a'$ is completely unpredictable, and again, Bob will accept $(a', z')$ as authentic with probability only $1/n$.

One can easily make $n$ large enough so that the probability that an adversary succeeds is so small that for all practical purposes it is impossible to trick Bob (e.g., $n \approx 2^{100}$).

More formally, and more generally, one can define an **$\epsilon$-forgeable message authentication scheme** to be a family $\mathcal{H}$ of hash functions from $\mathcal{A}$ to $\mathcal{Z}$ with the following property: if $H$ is uniformly distributed over $\mathcal{H}$, then

(i) for all $a \in \mathcal{A}$ and $z \in \mathcal{Z}$, we have $\mathsf{P}[H(a) = z] \leq \epsilon$, and

(ii) for all $a \in \mathcal{A}$ and all functions $f : \mathcal{Z} \to \mathcal{A}$ and $g : \mathcal{Z} \to \mathcal{Z}$, we have

$$\mathsf{P}[A' \neq a \wedge H(A') = Z'] \leq \epsilon,$$

where $Z := H(a)$, $A' := f(Z)$, and $Z' := g(Z)$.

Intuitively, part (i) of this definition says that it is impossible to guess the hash code of any message with probability better than $\epsilon$; further, part (ii) of this definition says that even after seeing the hash code of one message, it is impossible to guess the hash code of a different message with probability better than $\epsilon$, regardless the choice of the first message (i.e., the value $a$) and regardless of the strategy used to pick the second message and its putative hash code, given the hash code of the first message (i.e., the functions $f$ and $g$).

Exercise 6.36. Suppose that a family $\mathcal{H}$ of hash functions from $\mathcal{A}$ to $\mathcal{Z}$ is an $\epsilon$-forgeable message authentication scheme. Show that $\epsilon \geq 1/|\mathcal{Z}|$.

Exercise 6.37. Suppose that $\mathcal{H}$ is a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$ and that $|\mathcal{A}| > 1$. Show that if $\mathcal{H}$ satisfies part (ii) of the definition of an $\epsilon$-forgeable message authentication scheme, then it also satisfies part (i) of the definition.

Exercise 6.38. Let $\mathcal{H}$ be a family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$. For $\epsilon \geq 0$, we call $\mathcal{H}$ **pairwise $\epsilon$-predictable** if the following holds: for $H$ uniformly distributed over $\mathcal{H}$, for all $a, a' \in \mathcal{A}$, and for all $z, z' \in \mathcal{Z}$, we have $\mathsf{P}[H(a) = z] \leq \epsilon$ and

$$\mathsf{P}[H(a) = z] > 0 \text{ and } a' \neq a \ \text{ implies } \ \mathsf{P}[H(a') = z' \mid H(a) = z] \leq \epsilon.$$

(a) Show that if $\mathcal{H}$ is pairwise $\epsilon$-predictable, then it is an $\epsilon$-forgeable message authentication scheme.

(b) Show that if $\mathcal{H}$ is pairwise independent, then it is pairwise $1/|\mathcal{Z}|$-predictable. Combining this with part (a), we see that if $\mathcal{H}$ is pairwise independent, then it is a $1/|\mathcal{Z}|$-forgeable message authentication scheme (which makes rigorous the intuitive argument given above).

(c) Give an example of a family of hash functions that is an $\epsilon$-forgeable message authentication scheme for some $\epsilon < 1$, but is *not* pairwise $\epsilon$-predictable for any $\epsilon < 1$.

Exercise 6.39. Give an example of an $\epsilon$-forgeable message authentication scheme, where $\epsilon$ is very small, but where if Alice authenticates *two* distinct messages using the same hash function, an adversary can easily forge the hash code of any message he likes (after seeing Alice's two messages and their hash codes). This shows that, as we have defined a message authentication scheme, Alice should only authenticate a *single* message per hash function ($t$ messages may be authenticated using $t$ hash functions).

Exercise 6.40. Let $\mathcal{H}$ be an $\epsilon$-universal family of hash functions from $\mathcal{A}$ to $\mathcal{Y}$ (see Exercise 6.35), and let $\mathcal{H}'$ be a pairwise independent family of hash functions from $\mathcal{Y}$ to $\mathcal{Z}$. Define the composed family $\mathcal{H}' \circ \mathcal{H}$ of hash functions from $\mathcal{A}$ to $\mathcal{Z}$ as $\mathcal{H}' \circ \mathcal{H} := \{\phi_{h',h} : h' \in \mathcal{H}',\ h \in \mathcal{H}\}$, where $\phi_{h',h}(a) := h'(h(a))$ for $\phi_{h',h} \in \mathcal{H}' \circ \mathcal{H}$ and for $a \in \mathcal{A}$. Show that $\mathcal{H}' \circ \mathcal{H}$ is an $(\epsilon + 1/|\mathcal{Z}|)$-forgeable message authentication scheme.

## 6.8 Statistical distance

This section discusses a useful measure "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 on a finite set

$\mathcal{V}$. We define the **statistical distance between** $X$ **and** $Y$ as

$$\Delta[X;Y] := \frac{1}{2}\sum_{v\in\mathcal{V}} |\mathsf{P}[X=v] - \mathsf{P}[Y=v]|.$$

**Theorem 6.14.** *For random variables $X, Y, Z$, we have*

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

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

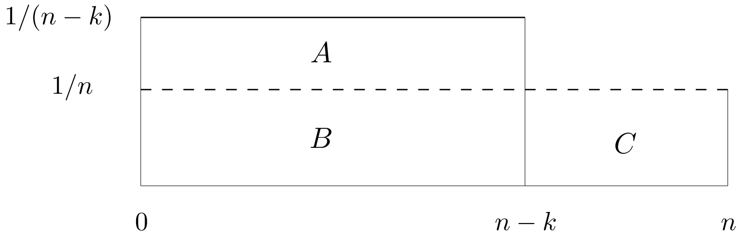
*(iii)* $\Delta[X;Y] = \Delta[Y;X]$*, and*

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

*Proof.* Exercise. □

Note that $\Delta[X;Y]$ depends only on the individual distributions of $X$ and $Y$, and not on the joint distribution of $X$ and $Y$. As such, one may speak of the statistical distance between two distributions, rather than between two random variables.

***Example* 6.26.** Suppose $X$ has the uniform distribution on $\{1, \ldots, n\}$, and $Y$ has the uniform distribution on $\{1, \ldots, n-k\}$, where $0 \le k \le n-1$. 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,

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

The following characterization of statistical distance is quite useful:

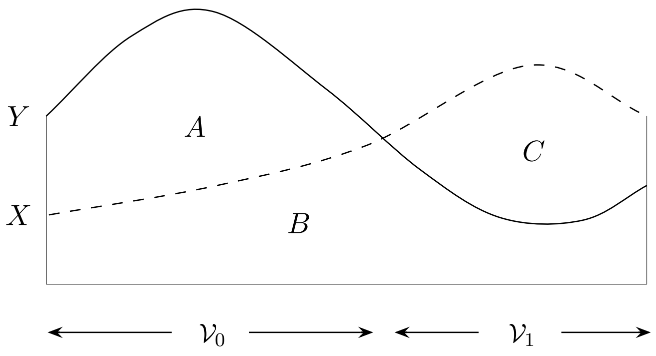
**Theorem 6.15.** *Let $X$ and $Y$ be random variables taking values on a set*

*$\mathcal{V}$. For any $\mathcal{W} \subseteq \mathcal{V}$, we have*

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

*and equality holds if $\mathcal{W}$ is either the set of all $v \in \mathcal{V}$ such that $\mathsf{P}[X = v] < \mathsf{P}[Y = v]$, or the complement of this set.*

*Proof.* Suppose we partition the set $\mathcal{V}$ into two sets: the set $\mathcal{V}_0$ consisting of those $v \in \mathcal{V}$ such that $\mathsf{P}[X = v] < \mathsf{P}[Y = v]$, and the set $\mathcal{V}_1$ consisting of those $v \in \mathcal{V}$ such that $\mathsf{P}[X = v] \geq \mathsf{P}[Y = v]$. Consider the following rough graph of the distributions of $X$ and $Y$, where the elements of $\mathcal{V}_0$ are placed to the left of the elements of $\mathcal{V}_1$:



Now, as in Example 6.26,

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

Further, consider any subset $\mathcal{W}$ of $\mathcal{V}$. The quantity $|\mathsf{P}[X \in \mathcal{W}] - \mathsf{P}[Y \in \mathcal{W}]|$ is equal to the absolute value of the difference of the area of the subregion of $A$ that lies above $\mathcal{W}$ and the area of the subregion of $C$ that lies above $\mathcal{W}$. This quantity is maximized when $\mathcal{W} = \mathcal{V}_0$ or $\mathcal{W} = \mathcal{V}_1$, in which case it is equal to $\Delta[X;Y]$. □

We can restate Theorem 6.15 as follows:

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

This implies that when $\Delta[X;Y]$ is very small, then for *any* predicate $\phi$, 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 6.16.** *Let $X, Y$ be random variables taking values on a set $\mathcal{V}$, and let $f$ be a function from $\mathcal{V}$ into a set $\mathcal{W}$. Then $\Delta[f(X); f(Y)] \le \Delta[X; Y]$.*

*Proof.* By Theorem 6.15, for any subset $\mathcal{W}'$ of $\mathcal{W}$, we have

$$|\mathsf{P}[f(X) \in \mathcal{W}'] - \mathsf{P}[f(Y) \in \mathcal{W}']| = \\ |\mathsf{P}[X \in f^{-1}(\mathcal{W}')] - \mathsf{P}[Y \in f^{-1}(\mathcal{W}')]| \le \Delta[X; Y].$$

In particular, again by Theorem 6.15,

$$\Delta[f(X); f(Y)] = |\mathsf{P}[f(X) \in \mathcal{W}'] - \mathsf{P}[f(Y) \in \mathcal{W}']|$$

for some $\mathcal{W}'$. □

***Example* 6.27.** Let $X$ be uniformly distributed on the set $\{0, \ldots, n-1\}$, and let $Y$ be uniformly distributed on the set $\{0, \ldots, m-1\}$, for $m \ge n$. Let $f(y) := y \bmod n$. We want to compute an upper bound on the statistical distance between $X$ and $f(Y)$. We can do this as follows. Let $m = qn - r$, where $0 \le r < n$, so that $q = \lceil m/n \rceil$. Also, let $Z$ be uniformly distributed over $\{0, \ldots, qn-1\}$. Then $f(Z)$ is uniformly distributed over $\{0, \ldots, n-1\}$, since every element of $\{0, \ldots, n-1\}$ has the same number (namely, $q$) of pre-images under $f$ which lie in the set $\{0, \ldots, qn-1\}$. Therefore, by the previous theorem,

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

and as we saw in Example 6.26,

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

Therefore,

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

We close this section with two useful theorems.

**Theorem 6.17.** *Let $X$ and $Y$ be random variables taking values on a set $\mathcal{V}$, and let $W$ be a random variable taking values on a set $\mathcal{W}$. Further, suppose that $X$ and $W$ are independent, and that $Y$ and $W$ are independent. Then the statistical distance between $(X, W)$ and $(Y, W)$ is equal to the statistical distance between $X$ and $Y$; that is,*

$$\Delta[X, W; Y, W] = \Delta[X, Y].$$

*Proof.* From the definition of statistical distance,

$$\begin{aligned} 2\Delta[X, W; Y, W] &= \sum_{v,w} |\mathsf{P}[X = v \wedge W = w] - \mathsf{P}[Y = v \wedge W = w]| \\ &= \sum_{v,w} |\mathsf{P}[X = v]\mathsf{P}[W = w] - \mathsf{P}[Y = v]\mathsf{P}[W = w]| \\ &\quad \text{(by independence)} \\ &= \sum_{v,w} \mathsf{P}[W = w]|\mathsf{P}[X = v] - \mathsf{P}[Y = v]| \\ &= (\sum_{w} \mathsf{P}[W = w])(\sum_{v} |\mathsf{P}[X = v] - \mathsf{P}[Y = v]|) \\ &= 1 \cdot 2\Delta[X; Y]. \quad \square \end{aligned}$$

**Theorem 6.18.** *Let $U_1, \ldots, U_\ell, V_1, \ldots, V_\ell$ be mutually independent random variables. We have*

$$\Delta[U_1, \ldots, U_\ell; V_1, \ldots, V_\ell] \le \sum_{i=1}^{\ell} \Delta[U_i; V_i].$$

*Proof.* We introduce random variables $W_0, \ldots, W_\ell$, defined as follows:

$$\begin{aligned} W_0 &:= (U_1, \ldots, U_\ell), \\ W_i &:= (V_1, \ldots, V_i, U_{i+1}, \ldots, U_\ell) \quad \text{for } i = 1, \ldots, \ell - 1, \text{ and} \\ W_\ell &:= (V_1, \ldots, V_\ell). \end{aligned}$$

By definition,

$$\Delta[U_1, \ldots, U_\ell; V_1, \ldots, V_\ell] = \Delta[W_0; W_\ell].$$

Moreover, by part (iv) of Theorem 6.14, we have

$$\Delta[W_0; W_\ell] \le \sum_{i=1}^{\ell} \Delta[W_{i-1}; W_i].$$

Now consider any fixed index $i = 1, \ldots, \ell$. By Theorem 6.17, we have

$$\begin{aligned} \Delta[W_{i-1}; W_i] &= \Delta[\, U_i,\ (V_1, \ldots, V_{i-1}, U_{i+1}, \ldots, U_\ell); \\ &\qquad\quad V_i,\ (V_1, \ldots, V_{i-1}, U_{i+1}, \ldots, U_\ell)] \\ &= \Delta[U_i; V_i]. \end{aligned}$$

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" variables $W_0, W_1, \ldots, W_\ell$, and shows that the distance between each consecutive pair of variables is small.

EXERCISE 6.41. 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 6.42. Let $n$ be a large integer that is the product of two distinct primes of roughly 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 $\Delta[X; Y] = O(n^{-1/2})$.

EXERCISE 6.43. Let $\mathcal{V}$ be a finite set, and consider any function $\phi : \mathcal{V} \to \{0, 1\}$. 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 $\mathcal{V}$, and assume that $X_b$ and $B$ are independent. Show that

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

EXERCISE 6.44. Let $X, Y$ be random variables on a probability distribution, and let $\mathcal{B}_1, \ldots, \mathcal{B}_n$ be events that partition of the underlying sample space, where each $\mathcal{B}_i$ occurs with non-zero probability. For $i = 1, \ldots, n$, let $X_i$ and $Y_i$ denote the random variables $X$ and $Y$ in the conditional probability distribution given $\mathcal{B}_i$; that is, $\mathsf{P}[X_i = v] = \mathsf{P}[X = v \mid \mathcal{B}_i]$, and $\mathsf{P}[Y_i = v] = \mathsf{P}[Y = v \mid \mathcal{B}_i]$. Show that

$$\Delta[X; Y] \le \sum_{i=1}^{n} \Delta[X_i; Y_i]\mathsf{P}[\mathcal{B}_i].$$

EXERCISE 6.45. Let $X$ and $Y$ be random variables that take the same value unless a certain event $\mathcal{F}$ occurs. Show that $\Delta[X; Y] \le \mathsf{P}[\mathcal{F}]$.

EXERCISE 6.46. Let $M$ be a large integer. Consider three random experiments. In the first, we generate a random integer $n$ between 1 and $M$, and then a random integer $w$ between 1 and $n$. In the second, we generate a random integer $n$ between 2 and $M$, and then generate a random integer $w$ between 1 and $n$. In the third, we generate a random integer $n$ between 2 and $M$, and then a random integer $w$ between 2 and $n$. For $i = 1, 2, 3$, let $X_i$ denote the outcome $(n, w)$ of the $i$th experiment. Show that $\Delta[X_1; X_2] = O(1/M)$ and $\Delta[X_2; X_3] = O(\log M/M)$, and conclude that $\Delta[X_1; X_3] = O(\log M/M)$.

EXERCISE 6.47. Show that Theorem 6.17 is not true if we drop the independence assumptions.

EXERCISE 6.48. Show that the hypothesis of Theorem 6.18 can be weakened: all one needs to assume is that $X_1, \ldots, X_\ell$ are mutually independent, and that $Y_1, \ldots, Y_\ell$ are mutually independent.

EXERCISE 6.49. Let $Y_1, \ldots, Y_\ell$ be mutually independent random variables, where each $Y_i$ is uniformly distributed on $\{0, \ldots, m-1\}$. For $i = 1, \ldots, \ell$, define $Z_i := \sum_{j=1}^{i} jY_j$. Let $n$ be a prime greater than $\ell$. Let $\mathcal{S}$ be any finite subset of $\mathbb{Z}^{\times \ell}$. Let $\mathcal{A}$ be the event that for some $(a_1, \ldots, a_\ell) \in \mathcal{S}$, we have $Z_i \equiv a_i \pmod{n}$ for $i = 1, \ldots, \ell$. Show that

$$\mathsf{P}[\mathcal{A}] \leq |\mathcal{S}|/n^\ell + \ell n/m.$$

EXERCISE 6.50. Let $\mathcal{X}$ be a set of size $n \geq 1$. Let $F$ be a random function from $\mathcal{X}$ into $\mathcal{X}$. Let $G$ be a random permutation of $\mathcal{X}$. Let $x_1, \ldots, x_\ell$ be distinct, fixed elements of $\mathcal{X}$. Show that

$$\Delta[F(x_1), \ldots, F(x_\ell); G(x_1), \ldots, G(x_\ell)] \leq \frac{\ell(\ell-1)}{2n}.$$

EXERCISE 6.51. Let $\mathcal{H}$ be a family hash functions from $\mathcal{A}$ to $\mathcal{Z}$ such that (i) each $h \in \mathcal{H}$ maps $\mathcal{A}$ injectively into $\mathcal{Z}$, and (ii) there exists $\epsilon$, with $0 \leq \epsilon \leq 1$, such that $\Delta[H(a); H(a')] \leq \epsilon$ for all $a, a' \in \mathcal{A}$, where $H$ is uniformly distributed over $\mathcal{H}$. Show that $|\mathcal{H}| \geq (1-\epsilon)|\mathcal{A}|$.

## 6.9 Measures of randomness and the leftover hash lemma (∗)

In this section, we discuss different ways to measure "how random" a probability distribution is, and relations among them. Consider a distribution defined on a finite sample space $\mathcal{V}$. In some sense, the "most random" distribution on $\mathcal{V}$ is the uniform distribution, while the least random would be a "point mass" distribution, that is, a distribution where one point $v \in \mathcal{V}$ in the sample space has probability 1, and all other points have probability 0.

We define three measures of randomness. Let $X$ be a random variable taking values on a set $\mathcal{V}$ of size $N$.

1. We say $X$ is **$\delta$-uniform on $\mathcal{V}$** if the statistical distance between $X$ and the uniform distribution on $\mathcal{V}$ is equal to $\delta$; that is,

$$\delta = \frac{1}{2}\sum_{v \in \mathcal{V}} |\mathsf{P}[X = v] - 1/N|.$$

2. The **guessing probability** $\gamma(X)$ of $X$ is defined to be

$$\gamma(X) := \max\{\mathsf{P}[X = v] : v \in \mathcal{V}\}.$$

3. The **collision probability** $\kappa(X)$ of $X$ is defined to be

$$\kappa(X) := \sum_{v \in \mathcal{V}} \mathsf{P}[X = v]^2.$$

Observe that if $X$ is uniformly distributed on $\mathcal{V}$, then it is 0-uniform on $\mathcal{V}$, and $\gamma(X) = \kappa(X) = 1/N$. Also, if $X$ has a point mass distribution, then it is $(1-1/N)$-uniform on $\mathcal{V}$, and $\gamma(X) = \kappa(X) = 1$. The quantity $\log_2(1/\gamma(X))$ is sometimes called the **min entropy** of $X$, and the quantity $\log_2(1/\kappa(X))$ is sometimes called the **Renyi entropy** of $X$. The collision probability $\kappa(X)$ has the following interpretation: if $X$ and $X'$ are identically distributed independent random variables, then $\kappa(X) = \mathsf{P}[X = X']$ (see Exercise 6.26).

We first state some easy inequalities:

**Theorem 6.19.** *Let $X$ be a random variable taking values on a set $\mathcal{V}$ of size $N$, such that $X$ is $\delta$-uniform on $\mathcal{V}$, $\gamma := \gamma(X)$, and $\kappa := \kappa(X)$. Then we have:*

*(i)* $\kappa \geq 1/N$;

*(ii)* $\gamma^2 \leq \kappa \leq \gamma \leq 1/N + \delta$.

*Proof.* Part (i) is immediate from Exercise 6.26. 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 ones intuition.

While the above theorem implies that $\gamma$ and $\kappa$ are close to $1/N$ when $\delta$ is small, the following theorem provides a converse of sorts:

**Theorem 6.20.** *If $X$ is $\delta$-uniform on $\mathcal{V}$, $\kappa := \kappa(X)$, and $N := |\mathcal{V}|$, then*

$$\kappa \geq \frac{1 + 4\delta^2}{N}.$$

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

For $v \in \mathcal{V}$, let $p_v := \mathsf{P}[X = v]$. We have $\delta = \frac{1}{2}\sum_v |p_v - 1/N|$, and hence

$1 = \sum_v q_v$, where $q_v := |p_v - 1/N|/(2\delta)$. So we have

$$\begin{aligned}\frac{1}{N} &\le \sum_v q_v^2 \quad \text{(by Exercise 6.25)}\\ &= \frac{1}{4\delta^2}\sum_v (p_v - 1/N)^2\\ &= \frac{1}{4\delta^2}(\sum_v p_v^2 - 1/N) \quad \text{(again by Exercise 6.25)}\\ &= \frac{1}{4\delta^2}(\kappa - 1/N),\end{aligned}$$

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 a universal family of hash functions (see §6.7.1).

**Theorem 6.21 (Leftover hash lemma).** *Let $\mathcal{H}$ be a universal family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $\mathcal{Z}$ is of size $n$. Let $H$ denote a random variable with the uniform distribution on $\mathcal{H}$, and let $A$ denote a random variable taking values in $\mathcal{A}$, and with $H, A$ independent. Let $\kappa := \kappa(A)$. Then $(H, H(A))$ is $\delta$-uniform on $\mathcal{H} \times \mathcal{Z}$, where*

$$\delta \le \sqrt{n\kappa}/2.$$

*Proof.* Let $Z$ denote a random variable uniformly distributed on $\mathcal{Z}$, with $H, A, Z$ mutually independent. Let $m := |\mathcal{H}|$ and $\delta := \Delta[H, H(A); H, Z]$.

Let us compute the collision probability $\kappa(H, H(A))$. Let $H'$ have the same distribution as $H$ and $A'$ have the same distribution as $A$, with $H, H', A, A'$ mutually independent. Then

$$\begin{aligned}\kappa(H, H(A)) &= \mathsf{P}[H = H' \wedge H(A) = H'(A')]\\ &= \mathsf{P}[H = H']\mathsf{P}[H(A) = H(A')]\\ &= \frac{1}{m}\Big(\mathsf{P}[H(A) = H(A') \mid A = A']\mathsf{P}[A = A'] +\\ &\qquad \mathsf{P}[H(A) = H(A') \mid A \ne A']\mathsf{P}[A \ne A']\Big)\\ &\le \frac{1}{m}(\mathsf{P}[A = A'] + \mathsf{P}[H(A) = H(A') \mid A \ne A'])\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{m}(\kappa + 1/n) \\
&= \frac{1}{mn}(n\kappa + 1).
\end{aligned}
$$

Applying Theorem 6.20 to the random variable $(H, H(A))$, which takes values on the set $\mathcal{H} \times \mathcal{Z}$ of size $N := mn$, we see that $4\delta^2 \leq n\kappa$, from which the theorem immediately follows. □

***Example* 6.28.** Suppose $A$ is uniformly distributed over a subset $\mathcal{A}'$ of $\mathcal{A}$, where $|\mathcal{A}'| \geq 2^{160}$, so that $\kappa(A) \leq 2^{-160}$. Suppose that $\mathcal{H}$ is a universal family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $|\mathcal{Z}| \leq 2^{64}$. If $H$ is uniformly distributed over $\mathcal{H}$, independently of $A$, then the leftover hash lemma says that $(H, H(A))$ is $\delta$-uniform on $\mathcal{H} \times \mathcal{Z}$, with

$$\delta \leq \sqrt{2^{64} 2^{-160}}/2 = 2^{-49}. \quad \square$$

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 $Z$ whose distribution is uniform on a set $\mathcal{Z}$ of size $n$, or at least $\delta$-uniform for a small value of $\delta$. 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 probability 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 $\mathcal{Z}$, obtaining a "high quality," nearly uniform, distribution on $\mathcal{Z}$.

Of course, this approach requires a random hash function, which may be just as difficult to generate as a random element of $\mathcal{Z}$. 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 6.22.** *Let $\mathcal{H}$ be a universal family of hash functions from $\mathcal{A}$ to $\mathcal{Z}$, where $\mathcal{Z}$ is of size $n$. Let $H$ denote a random variable with the uniform distribution on $\mathcal{H}$, and let $A_1, \ldots, A_\ell$ denote random variables taking values in $\mathcal{A}$, with $H, A_1, \ldots, A_\ell$ mutually independent. Let $\kappa := \max\{\kappa(\mathcal{A}_1), \ldots, \kappa(\mathcal{A}_\ell)\}$. Then $(H, H(A_1), \ldots, H(A_\ell))$ is $\delta'$-uniform on $\mathcal{H} \times \mathcal{Z}^{\times \ell}$, where*

$$\delta' \leq \ell\sqrt{n\kappa}/2.$$

*Proof.* Let $Z_1, \ldots, Z_\ell$ denote random variables with the uniform distribution on $\mathcal{Z}$, with $H, A_1, \ldots, A_\ell, Z_1, \ldots, Z_\ell$ mutually independent. We shall make a hybrid argument (as in the proof of Theorem 6.18). Define random variables $W_0, W_1, \ldots, W_\ell$ as follows:

$$\begin{aligned} W_0 &:= (H, H(A_1), \ldots, H(A_\ell)), \\ W_i &:= (H, Z_1, \ldots, Z_i, H(A_{i+1}), \ldots, H(A_\ell)) \quad \text{for } i = 1, \ldots, \ell - 1, \text{ and} \\ W_\ell &:= (H, Z_1, \ldots, Z_\ell). \end{aligned}$$

We have

$$\begin{aligned} \delta' &= \Delta[W_0; W_\ell] \\ &\leq \sum_{i=1}^{\ell} \Delta[W_{i-1}; W_i] \quad \text{(by part (iv) of Theorem 6.14)} \\ &\leq \sum_{i=1}^{\ell} \Delta[H, Z_1, \ldots, Z_{i-1}, H(A_i), A_{i+1}, \ldots, A_\ell; \\ &\qquad\qquad H, Z_1, \ldots, Z_{i-1}, \quad Z_i, \quad A_{i+1}, \ldots, A_\ell] \\ &\quad \text{(by Theorem 6.16)} \\ &= \sum_{i=1}^{\ell} \Delta[H, H(A_i); H, Z_i] \quad \text{(by Theorem 6.17)} \\ &\leq \ell\sqrt{n\kappa}/2 \quad \text{(by Theorem 6.21).} \quad \square \end{aligned}$$

Another source of "low quality" randomness arises in certain cryptographic applications, where we have a "secret" random variable $A$ that is distributed uniformly over a large subset of some set $\mathcal{A}$, but we want to derive from $A$ a "secret key" whose distribution is close to that of the uniform distribution on a specified "key space" $\mathcal{Z}$ (typically, $\mathcal{Z}$ is the set of all bit strings of some specified length). The leftover hash lemma, combined with Theorem 6.22, 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.

Exercise 6.52. Consider again the situation in Theorem 6.21. Suppose that $\mathcal{Z} = \{0, \ldots, n-1\}$, but that we would rather have an almost-uniform distribution over $\mathcal{Z}' = \{0, \ldots, t-1\}$, for some $t < n$. While it may be possible to work with a different family of hash functions, we do not have to if $n$ is large enough with respect to $t$, in which case we can just use the value $H(A) \bmod t$. If $Z'$ is uniformly distributed over $\mathcal{Z}'$, show that

$$\Delta[H, H(A) \bmod t; H, Z'] \leq \sqrt{n\kappa}/2 + t/n.$$

Exercise 6.53. Suppose $X$ and $Y$ are random variables with images $\mathcal{X}$ and $\mathcal{Y}$, respectively, and suppose that for some $\epsilon$, we have $\mathsf{P}[X = x \mid Y = y] \leq \epsilon$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Let $\mathcal{H}$ be a universal family of hash functions from $\mathcal{X}$ to $\mathcal{Z}$, where $\mathcal{Z}$ is of size $n$. Let $H$ denote a random variable with the uniform distribution on $\mathcal{H}$, and $Z$ denote a random variable with the uniform distribution on $\mathcal{Z}$, where the three variables $H$, $Z$, and $(X, Y)$ are mutually independent. Show that the statistical distance between $(Y, H, H(X))$ and $(Y, H, Z)$ is at most $\sqrt{n\epsilon}/2$.

## 6.10 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, 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 §§6.1–6.5 extends fairly easily to the countably infinite setting, and in this section, we discuss how this is done.

### *6.10.1 Basic definitions*

To say that the sample space $\mathcal{U}$ is countably infinite simply means that there is a bijection $f$ from the set of positive integers onto $\mathcal{U}$; thus, we can enumerate the elements of $\mathcal{U}$ as $u_1, u_2, u_3, \ldots$, where $u_i = f(i)$.

As in the finite case, the probability function assigns to each $u \in \mathcal{U}$ a value $\mathsf{P}[u] \in [0, 1]$. The basic requirement that the probabilities sum to one (equation (6.1)) is the requirement that the infinite series $\sum_{i=1}^{\infty} \mathsf{P}[u_i]$ converges to one. Luckily, the convergence properties of an infinite series whose terms are all non-negative is invariant under a re-ordering of terms (see §A4), so it does not matter how we enumerate the elements of $\mathcal{U}$.

***Example* 6.29.** Suppose we flip a fair coin repeatedly until it comes up "heads," and let the outcome $u$ of the experiment denote the number of coins flipped. We can model this experiment as a discrete probability distribution $\mathbf{D} = (\mathcal{U}, \mathsf{P})$, where $\mathcal{U}$ consists of the set of all positive integers, and where for $u \in \mathcal{U}$, we set $\mathsf{P}[u] = 2^{-u}$. We can check that indeed $\sum_{u=1}^{\infty} 2^{-u} = 1$, 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* 6.30.** 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 the outcome $u$ of the experiment denote the number of trials executed. Then we associate the probability $\mathsf{P}[u] = q^{u-1}p$ with each positive integer $u$, where $q := 1 - p$, since we have $u - 1$ failures before the one success. One can easily check that these probabilities sum to 1. Such a distribution is called a **geometric distribution**. □

***Example* 6.31.** The series $\sum_{i=1}^{\infty} 1/i^3$ converges to some positive number $c$. Therefore, we can define a probability distribution on the set of positive integers, where we associate with each $i \geq 1$ the probability $1/ci^3$. □

***Example* 6.32.** More generally, if $x_i, i = 1, 2, \ldots$, are non-negative numbers, and $0 < c := \sum_{i=1}^{\infty} x_i < \infty$, then we can define a probability distribution on the set of positive integers, assigning the probability $x_i/c$ to $i$. □

As in the finite case, an event is an arbitrary subset $\mathcal{A}$ of $\mathcal{U}$. The probability $\mathsf{P}[\mathcal{A}]$ of $\mathcal{A}$ is defined as the sum of the probabilities associated with the elements of $\mathcal{A}$—in the definition (6.2), the 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* 6.33.** Consider the geometric distribution discussed in Example 6.30, where $p$ is the success probability of each Bernoulli trial, and $q := 1 - p$. For 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

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

It is an easy matter to check that all the statements made in §6.1 carry over *verbatim* to the case of countably infinite sample spaces. Moreover, it also makes sense in the countably infinite case to consider events that are a union or intersection of a countably infinite number of events:

**Theorem 6.23.** *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be an infinite sequence of events.*

*(i) If $\mathcal{A}_i \subseteq \mathcal{A}_{i+1}$ for all $i \geq 1$, then $\mathsf{P}[\bigcup_{i \geq 1} \mathcal{A}_i] = \lim_{i \to \infty} \mathsf{P}[\mathcal{A}_i]$.*

*(ii) In general, we have $\mathsf{P}[\bigcup_{i \geq 1} \mathcal{A}_i] \leq \sum_{i \geq 1} \mathsf{P}[\mathcal{A}_i]$.*

*(iii) If the $\mathcal{A}_i$ are pairwise disjoint, then $\mathsf{P}[\bigcup_{i \geq 1} \mathcal{A}_i] = \sum_{i \geq 1} \mathsf{P}[\mathcal{A}_i]$.*

*(iv) If $\mathcal{A}_i \supseteq \mathcal{A}_{i+1}$ for all $i \geq 1$, then $\mathsf{P}[\bigcap_{i \geq 1} \mathcal{A}_i] = \lim_{i \to \infty} \mathsf{P}[\mathcal{A}_i]$.*

*Proof.* For (i), let $\mathcal{A} := \bigcup_{i \geq 1} \mathcal{A}_i$, and let $a_1, a_2, \ldots$ be an enumeration of the elements of $\mathcal{A}$. For any $\epsilon > 0$, there exists a value $k_0$ such that $\sum_{i=1}^{k_0} a_i > \mathsf{P}[\mathcal{A}] - \epsilon$. Also, there is some $k_1$ such that $\{a_1, \ldots, a_{k_0}\} \subseteq \mathcal{A}_{k_1}$. Therefore, for any $k \geq k_1$, we have $\mathsf{P}[\mathcal{A}] - \epsilon < \mathsf{P}[\mathcal{A}_k] \leq \mathsf{P}[\mathcal{A}]$.

(ii) and (iii) follow by applying (i) to the sequence $\{\bigcup_{j=1}^{i} \mathcal{A}_j\}_i$, and making use of (6.5) and (6.6), respectively.

(iv) follows by applying (i) to the sequence $\{\overline{\mathcal{A}_i}\}$, using (the infinite version of) DeMorgan's law. □

### 6.10.2 Conditional probability and independence

All of the definitions and results in §6.2 carry over *verbatim* to the countably infinite case. Equation (6.7) as well as Bayes' theorem (equation 6.8) and equation (6.9) extend *mutatis mutandus* to the case of an infinite partition $\mathcal{B}_1, \mathcal{B}_2, \ldots$.

### 6.10.3 Random variables

All of the definitions and results in §6.3 carry over *verbatim* to the countably infinite case (except Theorem 6.2, which of course only makes sense in the finite setting).

### *6.10.4 Expectation and variance*

We define the expected value of a real random variable $X$ exactly as before:

$$\mathsf{E}[X] := \sum_{u \in \mathcal{U}} X(u) \cdot \mathsf{P}[u],$$

where, of course, the sum is an infinite series. However, if $X$ may take negative values, then we require that the series converges *absolutely*; that is, we require that $\sum_{u \in \mathcal{U}} |X(u)| \cdot \mathsf{P}[u] < \infty$ (see §A4). Otherwise, we say the expected value of $X$ **does not exist**. Recall from calculus that a series that converges absolutely will itself converge, and will converge to the same value under a re-ordering of terms. Thus, if the expectation exists at all, its value is independent of the ordering on $\mathcal{U}$. For a non-negative random variable $X$, if its expectation does not exist, one may express this as "$\mathsf{E}[X] = \infty$."

All of the results in §6.4 carry over essentially unchanged, except that one must pay some attention to "convergence issues."

Equations (6.13) and (6.14) hold, but with the following caveats (verify):

- If $X$ is a real random variable with image $\mathcal{X}$, then its expected value $\mathsf{E}[X]$ exists if and only if the series $\sum_{x \in \mathcal{X}} x\mathsf{P}[X = x]$ converges absolutely, in which case $\mathsf{E}[X]$ is equal to the value of the latter series.
- If $X$ is a random variable with image $\mathcal{X}$ and $f$ a real-valued function on $\mathcal{X}$, then $\mathsf{E}[f(X)]$ exists if and only if the series $\sum_{x \in \mathcal{X}} f(x)\mathsf{P}[X = x]$ converges absolutely, in which case $\mathsf{E}[f(X)]$ is equal to the value of the latter series.

***Example* 6.34.** Let $X$ be a random variable whose distribution is as in Example 6.31. Since the series $\sum 1/n^2$ converges and the series $\sum 1/n$ diverges, the expectation $\mathsf{E}[X]$ exists, while $\mathsf{E}[X^2]$ does not. □

Theorems 6.6 and 6.7 hold under the additional hypothesis that $\mathsf{E}[X]$ and $\mathsf{E}[Y]$ exist.

If $X_1, X_2, \ldots$ is an infinite sequence of real random variables, then the random variable $X := \sum_{i=1}^{\infty} X_i$ is well defined provided the series $\sum_{i=1}^{\infty} X_i(u)$ converges for all $u \in \mathcal{U}$. 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 all $u$; nevertheless, it is true when the $X_i$ are non-negative:

**Theorem 6.24.** *Let $X := \sum_{i \geq 1} X_i$, where each $X_i$ takes non-negative values only. Then we have*

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

*Proof.* We have

$$\begin{aligned}\sum_{i\geq 1}\mathsf{E}[X_i] &= \sum_{i\geq 1}\sum_{u\in\mathcal{U}} X_i(u)\mathsf{P}[u] = \sum_{u\in\mathcal{U}}\sum_{i\geq 1} X_i(u)\mathsf{P}[u] \\ &= \sum_{u\in\mathcal{U}}\mathsf{P}[u]\sum_{i\geq 1} X_i(u) = \mathsf{E}[X],\end{aligned}$$

where we use the fact that we may reverse the order of summation in an infinite double summation of non-negative terms (see §A5). □

Using this theorem, one can prove the analog of Theorem 6.8 for countably infinite sample spaces, using exactly the same argument.

**Theorem 6.25.** *If $X$ is a random variable that takes non-negative integer values, then*

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

A nice picture to keep in mind with regards to Theorem 6.25 is the following. Let $p_i := \mathsf{P}[X = i]$ for $i = 0, 1, \ldots$, and let us arrange the probabilities $p_i$ in a table as follows:

$$\begin{array}{lll} p_1 & & \\ p_2 & p_2 & \\ p_3 & p_3 & p_3 \\ & \vdots & \quad \ddots \end{array}$$

Summing the $i$th row of this table, we get $i\mathsf{P}[X = i]$, and so $\mathsf{E}[X]$ is equal to the sum of all the entries in the table. However, we may compute the same sum column by column, and the sum of the entries in the $i$th column is $\mathsf{P}[X \geq i]$.

***Example* 6.35.** Suppose $X$ is a random variable with a geometric distribution, as in Example 6.30, with an associated success probability $p$ and failure probability $q := 1 - p$. As we saw in Example 6.33, for all integer $i \geq 1$, we have $\mathsf{P}[X \geq i] = q^{i-1}$. We may therefore apply Theorem 6.25 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* 6.36.** To illustrate that Theorem 6.24 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 $\mathsf{E}[X] = 1$. □

The variance $\mathsf{Var}[X]$ of $X$ exists if and only if $\mathsf{E}[X]$ and $\mathsf{E}[(X - \mathsf{E}[X])^2]$ exist, which holds if and only if $\mathsf{E}[X]$ and $\mathsf{E}[X^2]$ exist.

Theorem 6.9 holds under the additional hypothesis that $\mathsf{E}[X]$ and $\mathsf{E}[X^2]$ exist. Similarly, Theorem 6.10 holds under the additional hypothesis that $\mathsf{E}[X_i]$ and $\mathsf{E}[X_i^2]$ exist for each $i$.

The definition of conditional expectation carries over verbatim, as do equations (6.15) and (6.16). The analog of (6.16) for infinite partitions $\mathcal{B}_1, \mathcal{B}_2, \ldots$ does not hold in general, but does hold if $X$ is always non-negative.

### 6.10.5 Some useful bounds

Both Theorems 6.11 and 6.12 (Markov's and Chebyshev's inequalities) hold, under the additional hypothesis that the relevant expectations and variances exist.

Exercise 6.54. 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 6.55. 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 coin is flipped, 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; each time he plays, if he wins, he pockets his winnings and goes home, and otherwise, he doubles his bet and plays again.

(a) Show that if the gambler has an infinite amount of money (so he can keep playing no matter how many times he looses), then his expected winnings are one dollar. Hint: model the gambler's winnings as a random variable on a geometric distribution, and compute its expected value.

(b) Show that if the gambler has a finite amount of money (so that he can only afford to loose a certain number of times), then his expected winnings are zero (regardless of how much money he starts with).

Hint: in this case, you can model the gambler's winnings as a random variable on a finite probability distribution.

## 6.11 Notes

Our Chernoff bound (Theorem 6.13) 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, 99].

The leftover hash lemma (Theorem 6.21) was originally stated and proved by Impagliazzo, Levin, and Luby [46], 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 [47], who also present further applications.