4

Multiple Discrete Variables

Overview

In this chapter we explain how to build models that capture the interactions between multiple uncertain discrete quantities. Section 4.1 shows that such quantities can be represented as random variables within the same probability space. Section 4.2 describes how to characterize the distribution of individual quantities in models with multiple variables. Section 4.3 defines the conditional distribution of a random variable, which describes its behavior when the value of other random variables is revealed. Sections 4.4 and 4.5 define independence and conditional independence for random variables. In Section 4.6 we discuss under what circumstances probabilistic models can be interpreted in terms of causal relationships between the quantities of interest. Section 4.7 is our first encounter with the notorious curse of dimensionality, which is the reason why independence assumptions are needed to make probabilistic models tractable. Sections 4.8 and 4.9 describe two models based on such assumptions: naive Bayes and Markov chains.

4.1 Multivariate Discrete Random Variables

In this section we explain how to model the joint behavior of multiple uncertain discrete quantities. In Section 4.1.1 we show that such quantities can be represented as random variables within the same probability space. Section 4.1.2 introduces the joint probability mass function (pmf), which allows us to describe the joint behavior of multiple random variables. In Section 4.1.3 we explain how to estimate the joint pmf from data.

When jointly modeling multiple random variables, we often group them as entries of a *random vector*, which can be used to represent uncertain multidimensional quantities:

$$\tilde{x} := \begin{bmatrix} \tilde{x}[1] \\ \tilde{x}[2] \\ \cdots \\ \tilde{x}[d] \end{bmatrix}. \tag{4.1}$$

Here $\tilde{x}[i]$, $1 \leq i \leq d$, denotes the random variable that corresponds to the ith entry of the d-dimensional random vector $\tilde{x}$.

4.1.1 Mathematical Definition

In Section 2.1.1 we define discrete random variables as functions from a sample space to a discrete set. This mathematical framework enables us to characterize the joint behavior of several random variables very easily: we just define them *on the same probability space*. The outcome in the sample space then simultaneously determines the value of all the random variables.

Example 4.1 (Rolling a die twice)**.** In Example 2.1, we define a probability space representing two rolls of a six-sided die. Each outcome is encoded as a two-dimensional vector,

$$\omega := \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix}, \quad \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}, \tag{4.2}$$

where ω_1 is the result of the first roll, and ω_2 the result of the second roll. The random variables

$$\tilde{a}(\omega) := \omega_1, \tag{4.3}$$

$$\tilde{b}(\omega) := \omega_2, \tag{4.4}$$

$$\tilde{c}(\omega) := \omega_1 + \omega_2, \tag{4.5}$$

represent the value of the first roll, the second roll and of the sum of the two rolls, respectively. If $\omega = \left[\begin{smallmatrix} 3 \\ 1 \end{smallmatrix}\right]$, then $\tilde{a}(\omega) = 3$, $\tilde{b}(\omega) = 1$, and $\tilde{c}(\omega) = 4$. If $\omega = \left[\begin{smallmatrix} 2 \\ 5 \end{smallmatrix}\right]$, then $\tilde{a}(\omega) = 2$, $\tilde{b}(\omega) = 5$, and $\tilde{c}(\omega) = 7$. We can therefore reason about the *joint distribution* of the random variables by defining events such as

$$\{\tilde{a} = 3\} \cap \{\tilde{c} < 6\} := \{\omega : \tilde{a}(\omega) = 3 \text{ and } \tilde{c}(\omega) < 6\} \tag{4.6}$$

$$= \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \end{bmatrix} \right\}. \tag{4.7}$$

As we discuss in Section 2.1.2, the probability space used to define random variables is a mathematical abstraction. We only build it explicitly for simple pedagogical examples like Example 4.1. Instead, we describe the random variables using probabilities. When modeling multiple random variables, we are interested in the probability that they take certain values *at the same time*.

Consider two random variables $\tilde{a}$ and $\tilde{b}$ defined on the same probability space. By Definition 2.3 the probability measure of the probability space must assign a probability to the events $\tilde{a} = a$ and $\tilde{b} = b$, for every possible value of a and b. By the properties of probability spaces (see Definitions 1.7 and 1.15), it must consequently also assign a probability to the intersection of any of these pairs of

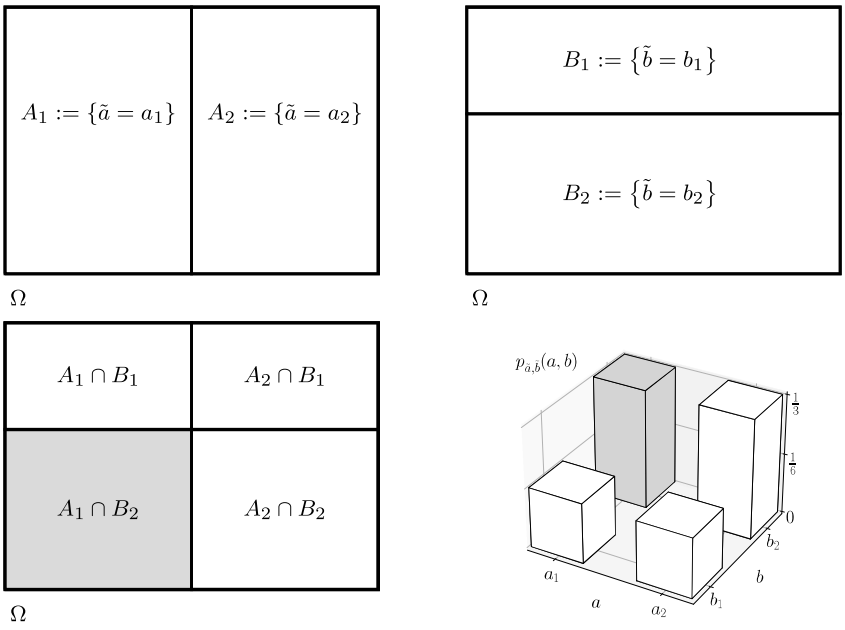

Figure 4.1 Joint probability mass function of a discrete random variable. The discrete random variables are defined on the same probability space. The top two Venn diagrams show the two partitions of the sample space Ω corresponding to $\tilde{a}$ (left) and $\tilde{b}$ (right). For $i = 1, 2$, A_i and B_i denote the events containing the outcomes mapping to a_i and b_i, respectively. The bottom left Venn diagram shows the different intersections $A_i \cap B_j$, for $i, j = 1, 2$. These events contain outcomes ω such that $\tilde{a}(\omega) = a_i$ and $\tilde{b}(\omega) = b_j$. The probability of this event is equal to the value of the joint pmf $p_{\tilde{a},\tilde{b}}$ evaluated at (a_i, b_j). The bottom right plot shows the joint pmf. The entry $p_{\tilde{a},\tilde{b}}(a_1, b_2)$ (shaded in gray) is equal to $\mathrm{P}(A_1 \cap B_2)$, represented by the area of the corresponding event $A_1 \cap B_2$ in the Venn diagram (also shaded in gray).

events

$$\left\{\omega \in \Omega : \tilde{a}(\omega) = a,\ \tilde{b}(\omega) = b\right\}, \tag{4.8}$$

which we usually just denote by $\{\tilde{a} = a, \tilde{b} = b\}$. The intersection is the event that the random variable $\tilde{a}$ equals a and simultaneously the random variable $\tilde{b}$ equals b (see Figure 4.1 for a simple example). In practice, we describe and manipulate random variables using these probabilities, encoded by the joint probability mass function, as explained in the following section.

4.1.2 The Joint Probability Mass Function

The joint probability mass function (pmf) of two random variables encodes the probability that the random variables equal any pair of values in their respective ranges.

Definition 4.2 (Joint probability mass function of two random variables)**.** *Let $\tilde{a} : \Omega \to A$ and $\tilde{b} : \Omega \to B$ be discrete random variables with discrete ranges A and B defined on the same probability space $(\Omega, \mathcal{C}, \mathrm{P})$. The joint pmf of $\tilde{a}$ and $\tilde{b}$ is*

$$p_{\tilde{a},\tilde{b}}\,(a,b) := \mathrm{P}(\tilde{a} = a, \tilde{b} = b), \quad a \in A, b \in B. \tag{4.9}$$

In words, $p_{\tilde{a},\tilde{b}}\,(a,b)$ is the probability of $\tilde{a}$ and $\tilde{b}$ being equal to a and b at the same time. Figure 4.1 shows a simple example of a joint pmf, and illustrates its connection to the underlying probability space.

We can generalize the definition of joint pmf to more than two random variables, or equivalently to the entries of a random vector.

Definition 4.3 (Joint probability mass function of a random vector)**.** *Let $\tilde{x}$ be a vector with entries equal to d random variables $\tilde{x}[1] : \Omega \to R_1$, $\tilde{x}[2] : \Omega \to R_2$, ..., $\tilde{x}[d] : \Omega \to R_d$ defined on a probability space $(\Omega, \mathcal{C}, \mathrm{P})$, where R_i is the discrete range of $x[i]$ for $1 \leq i \leq d$. The joint pmf of $\tilde{x}$ is*

$$p_{\tilde{x}}\,(x) := \mathrm{P}\Big(\tilde{x}[1] = x[1], \tilde{x}[2] = x[2], \ldots, \tilde{x}[d] = x[d]\Big). \tag{4.10}$$

The joint pmf allows us to compute the probability that random variables or random vectors belong to any subset of their ranges. This means that we can fully characterize their behavior through the joint pmf, without having to refer to the underlying probability space.

Lemma 4.4. *Let $\tilde{a}$ and $\tilde{b}$ be discrete random variables with joint pmf $p_{\tilde{a},\tilde{b}}$. For any $S \subseteq A \times B$, where A and B denote the ranges of $\tilde{a}$ and $\tilde{b}$,*

$$\mathrm{P}\Big((\tilde{a}, \tilde{b}) \in S\Big) = \sum_{(a,b)\in S} p_{\tilde{a},\tilde{b}}\,(a,b)\,. \tag{4.11}$$

Let $\tilde{x}$ be a d-dimensional random vector with joint pmf $p_{\tilde{x}}$. For any $S \subseteq R_1 \times R_2 \times \cdots \times R_d$, where R_i is the range of $\tilde{x}[i]$ for $1 \leq i \leq d$,

$$\mathrm{P}\left(\tilde{x} \in S\right) = \sum_{x \in S} p_{\tilde{x}}\,(x)\,. \tag{4.12}$$

Proof We prove the result for two variables, the general case for $d > 2$ variables follows by the same argument. The events $\{\tilde{a} = a, \tilde{b} = b\}$ for different values of a

or b are all disjoint, so

$$\mathrm{P}\left((\tilde{a}, \tilde{b}) \in S\right) = \mathrm{P}\left(\cup_{(a,b)\in S} \left\{\tilde{a} = a, \tilde{b} = b\right\}\right) \tag{4.13}$$

$$= \sum_{(a,b)\in S} \mathrm{P}\left(\tilde{a} = a, \tilde{b} = b\right) \tag{4.14}$$

$$= \sum_{(a,b)\in S} p_{\tilde{a},\tilde{b}}\left(a, b\right). \tag{4.15}$$

■

It follows from the definition that every joint pmf is nonnegative, and must sum to one. Conversely, any function that satisfies these two conditions can be interpreted as a joint pdf.

Theorem 4.5 (Properties of the joint pmf)**.** *Let $\tilde{a}$ and $\tilde{b}$ be discrete random variables with joint pmf $p_{\tilde{a},\tilde{b}}$. The joint pmf is nonnegative and satisfies*

$$\sum_{a\in A} \sum_{b\in B} p_{\tilde{a},\tilde{b}}\left(a, b\right) = 1, \tag{4.16}$$

where A and B denote the ranges of $\tilde{a}$ and $\tilde{b}$, respectively.

Let $\tilde{x}$ be a d-dimensional random vector with joint pmf $p_{\tilde{x}}$. The joint pmf is nonnegative and satisfies

$$\sum_{x[1]\in R_1} \sum_{x[2]\in R_2} \cdots \sum_{x[d]\in R_d} p_{\tilde{x}}\left(x\right) = 1, \tag{4.17}$$

where R_i denotes the range of $\tilde{x}[i]$ for $1 \leq i \leq d$.

Let R_1, R_2, ..., R_d be d discrete sets. Any nonnegative function $p : R_1 \times R_2 \times \cdots \times R_d \to [0, 1]$ satisfying

$$\sum_{x[1]\in R_1} \sum_{x[2]\in R_2} \cdots \sum_{x[d]\in R_d} p_{\tilde{x}}\left(x\right) = 1 \tag{4.18}$$

can be interpreted as the joint pmf of a random vector $\tilde{x}$ with range $R_1 \times R_2 \times \cdots \times R_d$.

Proof Joint pmfs are nonnegative because they represent probabilities. Since every outcome in the underlying space must be mapped to some pair $(a, b) \in A \times B$, by Lemma 4.4,

$$\sum_{a\in A} \sum_{b\in B} p_{\tilde{a},\tilde{b}}\left(a, b\right) = \mathrm{P}\Big(\left\{\omega \in \Omega : \left\{\tilde{a}(\omega) \in A\right\} \cup \left\{\tilde{b}(\omega) \in B\right\}\right\}\Big) \tag{4.19}$$

$$= \mathrm{P}(\Omega) = 1, \tag{4.20}$$

where Ω denotes the sample space of the probability space in which the random variables are defined. In words, the sum is equal to the probability of the whole underlying sample space. The same argument establishes (4.17).

To prove that any nonnegative function $p : R_1 \times R_2 \times \cdots \times R_d \to [0, 1]$ that sums to one is a valid joint pmf, we define a probability space where the sample

space is $R_1 \times R_2 \times \cdots \times R_d$, the collection is the power set of the sample space, and any event S is assigned the probability

$$\mathrm{P}(S) = \sum_{x \in S} p\left(x\right), \tag{4.21}$$

which yields a valid probability measure. ■

Example 4.6 (Simple example)**.** We consider two random variables $\tilde{a}$ and $\tilde{b}$ with joint pmf,

$$p_{\tilde{a},\tilde{b}}(1,1) = 0.05, \qquad p_{\tilde{a},\tilde{b}}(1,2) = 0.2, \qquad p_{\tilde{a},\tilde{b}}(1,3) = 0.1, \tag{4.22}$$

$$p_{\tilde{a},\tilde{b}}(2,1) = 0.1, \qquad p_{\tilde{a},\tilde{b}}(2,2) = 0.05, \qquad p_{\tilde{a},\tilde{b}}(2,3) = 0.2, \tag{4.23}$$

$$p_{\tilde{a},\tilde{b}}(3,1) = 0.1, \qquad p_{\tilde{a},\tilde{b}}(3,2) = 0.1, \qquad p_{\tilde{a},\tilde{b}}(3,3) = 0.1. \tag{4.24}$$

You can check that it sums up to one. Figure 4.5 shows a 3D bar plot of the joint pmf on the upper left. To compute the probability of $(\tilde{a}, \tilde{b})$ belonging to any subset of their joint range, we apply Lemma 4.4 and add the joint pmf over the elements of the subset. For example,

$$\mathrm{P}(\{\tilde{a} < 2, \tilde{b} > 1\}) = p_{\tilde{a},\tilde{b}}(1,2) + p_{\tilde{a},\tilde{b}}(1,3) \tag{4.25}$$

$$= 0.3. \tag{4.26}$$

4.1.3 The Empirical Joint Probability Mass Function

The joint pmf of multiple random variables encodes the probability that the random variables take different values. We can therefore estimate it using empirical probabilities. This is a straightforward generalization of the empirical pmf defined in Section 2.2 to multiple variables.

Definition 4.7 (Empirical joint pmf)**.** *Let $X := \{x_1, x_2, \ldots, x_n\}$ denote a dataset of d-dimensional vectors, where the jth entry takes values in a discrete set R_j for $1 \leq j \leq d$. The empirical joint pmf $p_X : R_1 \times R_2 \times \cdots \times R_d \to [0,1]$ maps each d-dimensional vector $v \in R_1 \times R_2 \times \cdots \times R_d$ to the fraction of data points that equal v,*

$$p_X(v) := \frac{\sum_{i=1}^{n} 1\left(x_i = v\right)}{n}, \tag{4.27}$$

where $1\left(x_i = v\right)$ is an indicator function that is equal to one if $x_i = v$, and to zero otherwise.

By definition, the empirical joint pmf is nonnegative and sums to one, so we can interpret it as a valid joint pmf by Theorem 4.5. As usual, we should bear in mind that empirical probabilities may be noisy if the data are limited.

Table 4.1 ***Empirical joint pmf of movie ratings.*** *The left table shows the number of users corresponding to all possible pairs of ratings for the movies Mission Impossible and Independence Day. These counts are normalized to yield the empirical joint pmf of the data on the right. The joint pmf is expressed in %.*

Independence Day

Mission Impossible

	1	**2**	**3**	**4**	**5**
1	2	3	5	1	0
2	3	12	18	11	5
3	5	14	37	41	17
4	6	15	20	47	19
5	0	0	4	12	17

Counts

Independence Day

Mission Impossible

	1	**2**	**3**	**4**	**5**
1	0.6	1	1.6	0.3	0
2	1	3.8	5.7	3.5	1.6
3	1.6	4.5	11.8	13.1	5.4
4	1.9	4.8	6.4	15	6.1
5	0	0	1.3	3.8	5.4

Empirical joint pmf (%)

Example 4.8 (Movie ratings)**.** Dataset 8 contains ratings given by a group of users to popular movies. The ratings are integers between 1 and 5. We consider users that have rated Independence Day and Mission Impossible, interpreting each pair of ratings as a joint realization of two random variables representing the movies. Table 4.1 shows the empirical joint pmf of these random variables, computed following Definition 4.7. We can approximate different probabilities related to the ratings from the joint pmf estimate via Lemma 4.4. For example,

$$\mathrm{P}(\text{both ratings} \geq 4) = \sum_{i=4}^{5} \sum_{j=4}^{5} p_X(i,j) \tag{4.28}$$

$$= 0.303. \tag{4.29}$$

Example 4.9 (Precipitation in Oregon)**.** We consider a dataset, extracted from Dataset 9, which contains hourly precipitation measurements from three Oregon weather stations in 2015: Coos Bay, Corvallis and John Day. We define a random vector of dimension three, where each entry is a Bernoulli random variable indicating whether there is any precipitation in Coos Bay ($\tilde{x}[1] = 1$), Corvallis ($\tilde{x}[2] = 1$) or John Day ($\tilde{x}[3] = 1$). The empirical joint pmf of the data is shown in Figure 4.2. We see that 82.8% of the time there is no precipitation in any of the stations.

4.2 Marginal Distributions

In probabilistic models with multiple variables, it is often of interest to isolate the individual behavior of a single variable. This can be achieved by computing the pmf of the random variable, *marginalizing* out the rest of variables by summing

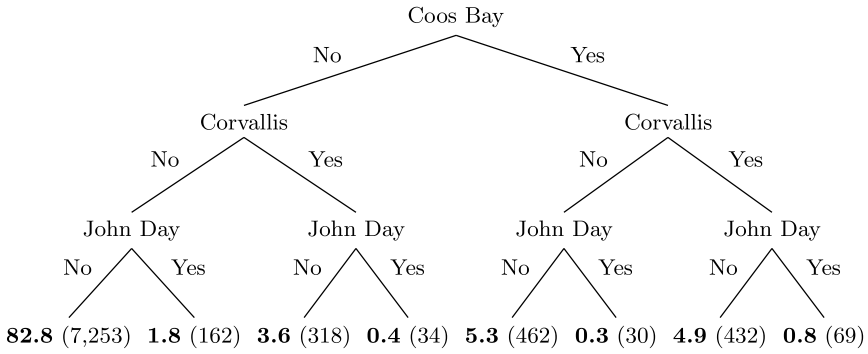

Figure 4.2 Empirical joint pmf of precipitation. The tree diagram shows the joint distribution of precipitation in the Oregon stations of Coos Bay, Corvallis and John Day in 2015, based on hourly measurements. The counts in parentheses are normalized to yield the empirical joint pmf (in bold). The joint pmf is expressed in %.

over them. In this context, the resulting pmf is called the *marginal* pmf of the random variable.

Theorem 4.10 (Marginal pmf)**.** *Let $\tilde{a}$ and $\tilde{b}$ be discrete random variables with ranges A and B, respectively, and joint pmf $p_{\tilde{a},\tilde{b}}$. The marginal pmf of $\tilde{a}$ is obtained by summing the joint pmf over all the possible values of $\tilde{b}$,*

$$p_{\tilde{a}}(a) = \sum_{b \in B} p_{\tilde{a},\tilde{b}}(a,b). \tag{4.30}$$

Let $\tilde{x}$ be a random vector with joint pmf $p_{\tilde{x}}$. The marginal pmf of the ith entry $\tilde{x}[i]$ is obtained by summing the joint pmf over all the possible values of the other entries,

$$p_{\tilde{x}[i]}(a) = \sum_{b_1 \in X_1} \cdots \sum_{b_{i-1} \in X_{i-1}} \sum_{b_{i+1} \in X_{i+1}} \cdots \sum_{b_d \in X_d} p_{\tilde{x}}(b_1, ..., b_{i-1}, a, b_{i+1}, ..., b_d), \tag{4.31}$$

where X_i denotes the range of the ith entry.

Proof We prove the bivariate case; the vector case follows by the same argument. By Lemma 4.4,

$$\begin{aligned} p_{\tilde{a}}(a) &= \mathrm{P}(\tilde{a} = a) && (4.32)\\ &= \mathrm{P}(\cup_{b \in B} \{\tilde{a} = a, \tilde{b} = b\}) && (4.33)\\ &= \sum_{b \in B} \mathrm{P}(\tilde{a} = a, \tilde{b} = b) && (4.34)\\ &= \sum_{b \in B} p_{\tilde{a},\tilde{b}}(a,b). && (4.35) \end{aligned}$$

■

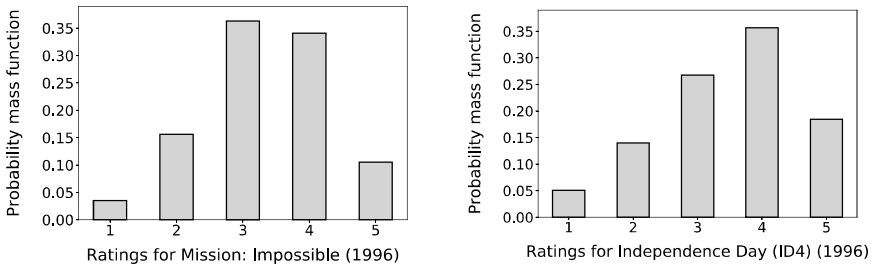

Figure 4.3 Marginal pmfs of movie ratings. Marginal pmf of the ratings of the movies Mission Impossible (left) and Independence Day (right), computed from the empirical joint pmf in Table 4.1 (see Example 4.8).

If we are interested in computing the joint pmf of several entries in a random vector, instead of just one, the marginalization process is essentially the same. Notation gets a bit complicated, so let us just consider an example with four entries. To compute the marginal pmf of the first and fourth entry, we marginalize out the second and third entries,

$$p_{\tilde{x}[1],\tilde{x}[4]}(a,d) = \mathrm{P}\left(\cup_{b\in X_2, c\in X_3} \{\tilde{x}[1]=a, \tilde{x}[2]=b, \tilde{x}[3]=c, \tilde{x}[4]=d\}\right) \tag{4.36}$$

$$= \sum_{b\in X_2} \sum_{c\in X_3} p_{\tilde{x}}\left(a,b,c,d\right). \tag{4.37}$$

Example 4.11 (Simple example: Marginal distribution)**.** Consider the random variables $\tilde{a}$ and $\tilde{b}$ in Example 4.6. We compute the marginal pmf of $\tilde{a}$ by summing over b,

$$p_{\tilde{a}}(1) = 0.35, \qquad p_{\tilde{a}}(2) = 0.35, \qquad p_{\tilde{a}}(3) = 0.3. \tag{4.38}$$

The marginal pmf is displayed on the upper right of Figure 4.5. Conversely, the marginal pmf of $\tilde{b}$ is obtained by summing over a,

$$p_{\tilde{b}}(1) = 0.25, \qquad p_{\tilde{b}}(2) = 0.35, \qquad p_{\tilde{b}}(3) = 0.4. \tag{4.39}$$

Figure 4.3 shows the marginal pmfs of the movie ratings in Example 4.8, obtained by applying Theorem 4.10 to the joint pmf. It turns out that this group of users enjoyed Independence Day more than Mission Impossible.

Figure 4.4 shows the marginal distributions corresponding to the precipitation data in Example 4.9. By Theorem 4.10, to obtain the marginal joint pmfs of each pair of stations, we can sum the joint pmf over the remaining entry of the random vector. For example,

$$p_{\tilde{x}[1],\tilde{x}[2]}\left(a,b\right) = p_{\tilde{x}}\left(a,b,0\right) + p_{\tilde{x}}\left(a,b,1\right), \quad a,b\in\{0,1\}. \tag{4.40}$$

Similarly, to obtain the marginal pmf of each station, we sum over the other two.

Corvallis

Coos Bay	No	Yes
No	84.70	4.02
Yes	5.62	5.72

John Day

Coos Bay	No	Yes
No	86.43	2.24
Yes	10.21	1.13

John Day

Corvallis	No	Yes
No	88.07	2.19
Yes	8.56	1.18

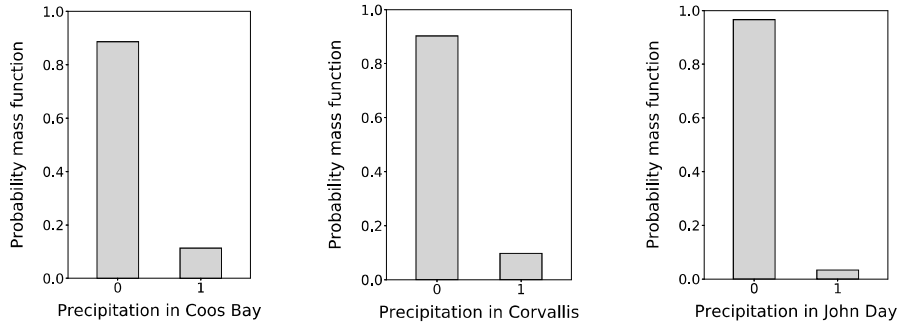

Figure 4.4 Marginal pmfs for precipitation data. The top row shows the marginal joint pmf of each pair of weather stations in Example 4.9, obtained from the empirical joint pmf in Figure 4.2. The graphs below show the marginal pmf corresponding to each individual station.

For example,

$$p_{\tilde{x}[2]}(b) = \sum_{a=0}^{1} \sum_{c=0}^{1} p_{\tilde{x}}(a, b, c), \quad b \in \{0, 1\}. \tag{4.41}$$

From the data, we see that the weather in John Day is substantially drier than in Coos Bay or Corvallis.

In practice, we usually estimate marginal pmfs by directly applying Definition 2.11 to the data associated with the relevant variable, instead of first estimating the empirical joint pmf and then marginalizing. The following lemma shows that this is equivalent.

Lemma 4.12 (Marginal empirical pmf)**.** *Let* (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) *be* n *pairs of real-valued data. We denote by* $X := \{x_1, x_2, \ldots, x_n\}$ *and* $Y := \{y_1, y_2, \ldots, y_n\}$ *the bags of first and second entries, respectively. By Definition 4.7, the joint empirical pmf of* X *and* Y *is*

$$p_{X,Y}(a, b) := \frac{\sum_{i=1}^{n} 1(x_i = a, y_i = b)}{n}, \tag{4.42}$$

where $1(x_i = a, y_i = b)$ *is an indicator function that equals one if both* $x_i = a$ *and* $y_i = b$*, and zero otherwise. By Definition 2.11, the empirical pmf of* X *is*

$$p_X(a) := \frac{\sum_{i=1}^{n} 1(x_i = a)}{n}, \tag{4.43}$$

where $1(x_i = a)$ *is an indicator function that equals one if* $x_i = a$ *and zero otherwise. For any* $a \in A$ *we have*

$$p_X(a) = \sum_{b \in B} p_{X,Y}(a, b), \tag{4.44}$$

where A *and* B *denote the sets of distinct values in* X *and* Y*, respectively.*

Proof Note that $\sum_{b \in B} 1(x_i = a, y_i = b) = 1(x_i = a)$ because y_i must equal one of the values in B, so the expression is one if and only if $x_i = a$. As a result,

$$\sum_{b \in B} p_{X,Y}(a, b) = \frac{\sum_{i=1}^{n} \sum_{b \in B} 1(x_i = a, y_i = b)}{n} \tag{4.45}$$

$$= \frac{\sum_{i=1}^{n} 1(x_i = a)}{n} \tag{4.46}$$

$$= p_X(a). \tag{4.47}$$

■

4.3 Conditional Distributions

Conditional probabilities allow us to update our uncertainty about a certain variable when new information is revealed. The conditional distribution of a random variable describes its behavior under the assumption that other random variables take fixed values. For discrete random variables, the distribution is typically specified using the conditional pmf.

Definition 4.13 (Conditional probability mass function)**.** *Let* $\tilde{a}$ *and* $\tilde{b}$ *be discrete random variables with ranges* A *and* B *respectively. The conditional probability mass function of* $\tilde{b}$ *given* $\tilde{a}$ *is*

$$p_{\tilde{b} \,|\, \tilde{a}}(b \,|\, a) := \mathrm{P}(\tilde{b} = b \,|\, \tilde{a} = a) \tag{4.48}$$

$$= \frac{\mathrm{P}(\tilde{a} = a, \tilde{b} = b)}{\mathrm{P}(\tilde{a} = a)} \tag{4.49}$$

$$= \frac{p_{\tilde{a},\tilde{b}}(a, b)}{p_{\tilde{a}}(a)} \tag{4.50}$$

if $p_{\tilde{a}}(a) > 0$ *and is undefined otherwise.*

Let $\tilde{x}$ *be a random vector. The conditional pmf of* $\tilde{x}[i]$ *given* $\tilde{x}[j] = a_j$ *for* $j \neq i$ *is*

$$p_{\tilde{x}[i] \,|\, \tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}(b \,|\, a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d) \tag{4.51}$$

$$= \frac{p_{\tilde{x}}(a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_d)}{p_{\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d)}, \tag{4.52}$$

if $p_{\tilde{x}[1],\ldots,\tilde{x}[i-1],\tilde{x}[i+1],\ldots,\tilde{x}[d]}(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_d) > 0$.

The conditional pmf $p_{\tilde{b}\,|\,\tilde{a}}\,(\cdot\,|\,a)$ characterizes our uncertainty about $\tilde{b}$ conditioned on the event $\{\tilde{a} = a\}$. This object is a valid pmf, since it is nonnegative and

$$\sum_{b\in B} p_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a) = \frac{\sum_{b\in B} p_{\tilde{a},\tilde{b}}\,(a,b)}{p_{\tilde{a}}\,(a)} \tag{4.53}$$

$$= \frac{p_{\tilde{a}}\,(a)}{p_{\tilde{a}}\,(a)} = 1, \tag{4.54}$$

for any a such that $p_{\tilde{a}}\,(a) \neq 0$. However, the conditional pmf is *not* a pmf of the variable we condition on (in this case $\tilde{a}$). Therefore there is no reason for $\sum_{a\in A} p_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a)$ to add up to one!

It is also possible to compute the conditional joint pmf of a group of variables given another group of variables, or equivalently, of a subset of entries of a random vector given another subset. The notation quickly becomes cumbersome, so we just consider an example where $d = 4$. The conditional joint pmf of the second and third entries of a random vector $\tilde{x}$ given the first and fourth entry is

$$p_{\tilde{x}[2],\tilde{x}[3]\,|\,\tilde{x}[1],\tilde{x}[4]}(b,c\,|\,a,d) = \mathrm{P}\left(\tilde{x}[2] = b, \tilde{x}[3] = c\,|\,\tilde{x}[1] = a, \tilde{x}[4] = d\right) \tag{4.55}$$

$$= \frac{p_{\tilde{x}}(a,b,c,d)}{p_{\tilde{x}[1],\tilde{x}[4]}(a,d)}. \tag{4.56}$$

From the definition of the conditional pmf, we derive a chain rule for discrete random variables and vectors.

Theorem 4.14 (Chain rule for discrete random variables and vectors)**.** *For any discrete random variables $\tilde{a}$ and $\tilde{b}$,*

$$p_{\tilde{a},\tilde{b}}\,(a,b) = p_{\tilde{a}}\,(a)\,p_{\tilde{b}\,|\,\tilde{a}}\,(b\,|\,a) \tag{4.57}$$

$$= p_{\tilde{b}}\,(b)\,p_{\tilde{a}\,|\,\tilde{b}}\,(a\,|\,b)\,. \tag{4.58}$$

For any d-dimensional random vector $\tilde{x}$

$$p_{\tilde{x}}\,(x) = p_{\tilde{x}[1]}\,(x[1]) \prod_{i=2}^{n} p_{\tilde{x}[i]\,|\,\tilde{x}[1],\ldots,\tilde{x}[i-1]}\,(x[i]\,|\,x[1],\ldots,x[i-1])\,. \tag{4.59}$$

The order of indices in the random vector is completely arbitrary (any order works).

Example 4.15 (Simple example: Conditional distribution)**.** Consider the random variables $\tilde{a}$ and $\tilde{b}$ in Example 4.6. To condition on $\tilde{a} = 2$, we could be tempted to just fix that value in the joint pmf, which yields

$$p_{\tilde{a},\tilde{b}}(2,1) = 0.1, \qquad p_{\tilde{a},\tilde{b}}(2,2) = 0.05, \qquad p_{\tilde{a},\tilde{b}}(2,3) = 0.2. \tag{4.60}$$

However, this is not a valid pmf! It does not add up to one. We can normalize dividing by the sum of the entries. Since this probability is equal to $p_{\tilde{a}}(2) = 0.35$ (the marginal pmf evaluated at 2), this is equivalent to dividing by the probability

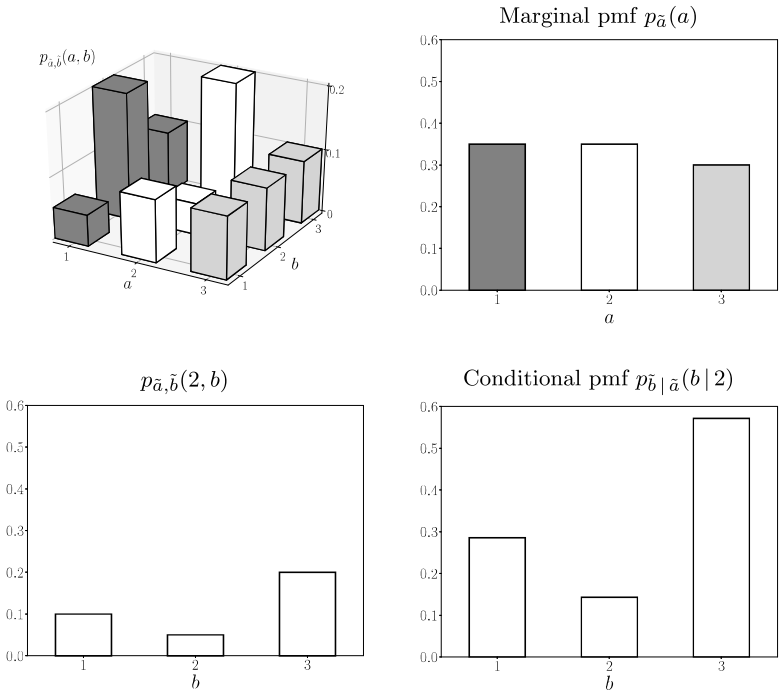

Figure 4.5 Joint, marginal and conditional pmfs. The joint pmf of the random variables in Example 4.6 is represented as a 3D bar plot in the upper left. The upper-right bar plot shows the marginal pmf of $\tilde{a}$. Each bar is obtained by summing over the bars in the 3D plot with the same color. The lower-left bar plot shows the joint pmf for $\tilde{a} = 2$ (in white in the 3D bar plot). The bar plot of the corresponding conditional pmf is shown in the lower right. It is obtained by normalizing these values with their sum, which equals the value of the marginal pmf at 2.

of the event we are conditioning on ($\tilde{a} = 2$), to obtain the conditional probabilities encoded in the conditional pmf:

$$p_{\tilde{b}\,|\,\tilde{a}}(1\,|\,2) = \frac{2}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(2\,|\,2) = \frac{1}{7}, \qquad p_{\tilde{b}\,|\,\tilde{a}}(3\,|\,2) = \frac{4}{7}. \tag{4.61}$$

The plots at the bottom of Figure 4.5 show the unnormalized and normalized values.

In Example 1.25 we compute conditional probabilities to analyze voting data from the United States House of Representatives (Dataset 1). Let us use the same example as a simple illustration of joint, marginal and conditional distributions.

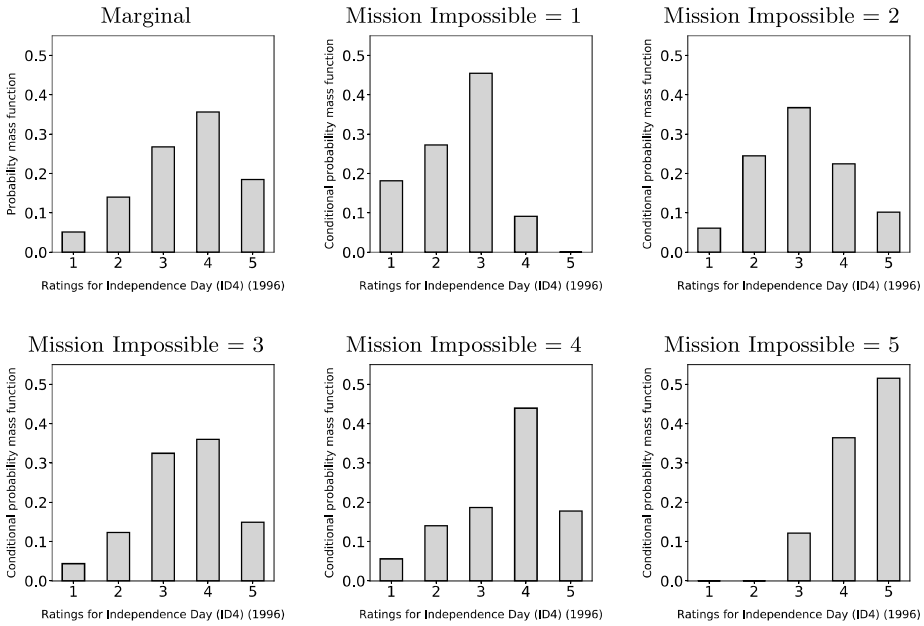

Figure 4.6 Conditional pmfs of movie ratings. The marginal pmf of the ratings for Independence Day is shown on the upper left. The remaining graphs show the conditional pmf of the ratings of Independence Day given different ratings for Mission Impossible (indicated above each graph). Users who like Mission Impossible tend to rate Independence Day higher than users who don't.

We define a Bernoulli random variable to represent each event:

$$\tilde{b} = \begin{cases} 1 & \text{Yes on budget,} \\ 0 & \text{No on budget,} \end{cases} \tag{4.62}$$

$$\tilde{d} = \begin{cases} 1 & \text{Yes on duty-free exports,} \\ 0 & \text{No on duty-free exports.} \end{cases} \tag{4.63}$$

We compute the empirical joint pmf of the data from Table 1.2. Then we use the joint pmf to compute the marginal pmfs of the individual random variables, and their conditional pmfs. Table 4.2 shows the results.

Figure 4.6 shows the conditional pmf of the ratings for Independence Day in Example 4.8 given the different ratings for Mission Impossible, and compares it to the marginal pmf. The rating for Mission Impossible provides a lot of information about the rating for Independence Day. Users who like Mission Impossible tend to give Independence Day substantially higher ratings than users who don't like Mission Impossible.

Table 4.3 provides an example of conditional joint pmfs of two random variables given a third one. It shows the conditional joint pmfs of precipitation at each pair

Table 4.2 ***Joint, marginal and conditional pmfs.*** *Joint, marginal and conditional pmfs of the random variables $\tilde{b}$ and $\tilde{d}$ in Example 1.25 estimated from the corresponding empirical probabilities.*

		d				
	$p_{\tilde{b},\tilde{d}}$	0	1	$p_{\tilde{b}}$	$p_{\tilde{b}\mid\tilde{d}}\left(\cdot\mid 0\right)$	$p_{\tilde{b}\mid\tilde{d}}\left(\cdot\mid 1\right)$
b	0	0.35	0.05	0.40	0.61	0.12
	1	0.22	0.38	0.60	0.39	0.88

$p_{\tilde{d}}$	0.57	0.43

$p_{\tilde{d}\mid\tilde{b}}\left(\cdot\mid 0\right)$	0.87	0.13

$p_{\tilde{d}\mid\tilde{b}}\left(\cdot\mid 1\right)$	0.37	0.63

of weather stations in Example 4.9, given precipitation at the remaining station, and compares them to the marginal joint pmfs. These conditional joint pmfs are obtained from the joint pmf of the three values by applying Definition 4.13. For example, the conditional joint pmf of $\tilde{x}[1]$ and $\tilde{x}[2]$ given $\tilde{x}[3] = 1$ is

$$p_{\tilde{x}[1],\tilde{x}[2]\,|\,\tilde{x}[3]}\left(a, b \,|\, 1\right) = \frac{p_{\tilde{x}}\left(a, b, 1\right)}{p_{\tilde{x}[3]}(1)}, \quad a, b \in \{0, 1\}\,. \tag{4.64}$$

Figure 4.7 compares the marginal pmf of precipitation at Coos Bay (first row) with the conditional pmf given a single other station (second row). It also shows the conditional pmf given two other stations (third row). To condition on two stations, we again apply Definition 4.13. For example, the conditional joint pmf of $\tilde{x}[1]$ (Coos Bay) given $\tilde{x}[2] = 0$ and $\tilde{x}[3] = 1$ is

$$p_{\tilde{x}[1]\,|\,\tilde{x}[2],\tilde{x}[3]}\left(a \,|\, 0, 1\right) = \frac{p_{\tilde{x}}\left(a, 0, 1\right)}{p_{\tilde{x}[2],\tilde{x}[3]}(0, 1)}, \quad a \in \{0, 1\}\,. \tag{4.65}$$

The conditional pmfs indicate that the precipitation in Corvallis provides more information about precipitation in Coos Bay, than the precipitation in John Day. When there is precipitation in Corvallis, the probability of precipitation in Coos Bay jumps from 11.3% to 58.7%. When there isn't it falls to 6.2%. In contrast, when there is precipitation in John Day, the probability of precipitation in Coos Bay only increases to 33.6%. When there isn't, it only decreases to 10.6%.

Table 4.3 ***Conditional joint pmfs for precipitation data.*** *The left column shows the marginal joint pmfs of precipitation at each pair of weather stations in Example 4.9. The center and right columns show the conditional joint pmfs of each pair given precipitation at the remaining station. Conditioning has a substantial effect on the distribution, especially if there is precipitation in the station we condition on. All pmfs are computed from the empirical joint pmf in Figure 4.2.*

Coos Bay \ Corvallis	**0**	**1**
0	84.7	4.0
1	5.6	5.7

Marginal

Coos Bay \ Corvallis	**0**	**1**
0	85.7	3.8
1	5.5	5.1

John Day = 0

Coos Bay \ Corvallis	**0**	**1**
0	54.9	11.5
1	10.2	23.4

John Day = 1

Coos Bay \ John Day	**0**	**1**
0	86.4	2.2
1	10.2	1.1

Marginal

Coos Bay \ John Day	**0**	**1**
0	91.7	2.0
1	5.8	0.4

Corvallis = 0

Coos Bay \ John Day	**0**	**1**
0	37.3	4.0
1	50.6	8.1

Corvallis = 1

Corvallis \ John Day	**0**	**1**
0	88.1	2.2
1	8.6	1.2

Marginal

Corvallis \ John Day	**0**	**1**
0	93.4	2.1
1	4.1	0.4

Coos Bay = 0

Corvallis \ John Day	**0**	**1**
0	46.5	3.0
1	43.5	7.0

Coos Bay = 1

4.4 Independence

When knowledge about a random variable $\tilde{a}$ does not affect our uncertainty about another random variable $\tilde{b}$, we say that $\tilde{a}$ and $\tilde{b}$ are *independent*. This can be expressed in terms of the conditional distributions of the random variables. If $\tilde{a}$ and $\tilde{b}$ are discrete, then for any a and b, we should have

$$p_{\tilde{a}\,|\,\tilde{b}}(a\,|\,b) = \mathrm{P}(\tilde{a} = a\,|\,\tilde{b} = b) \tag{4.66}$$
$$= \mathrm{P}(\tilde{a} = a) \tag{4.67}$$
$$= p_{\tilde{a}}(a), \tag{4.68}$$

assuming the conditional pmf is well defined (the pmf of $\tilde{b}$ is nonzero at b). Equivalently,

$$p_{\tilde{a},\tilde{b}}(a, b) = p_{\tilde{a}}(a) p_{\tilde{b}}(b). \tag{4.69}$$

This is consistent with the definition of independence of events (see Section 1.5). Here, the events $\tilde{a} = a$ and $\tilde{b} = b$ are independent.

Definition 4.16 (Independent discrete random variables)**.** *Two discrete random*

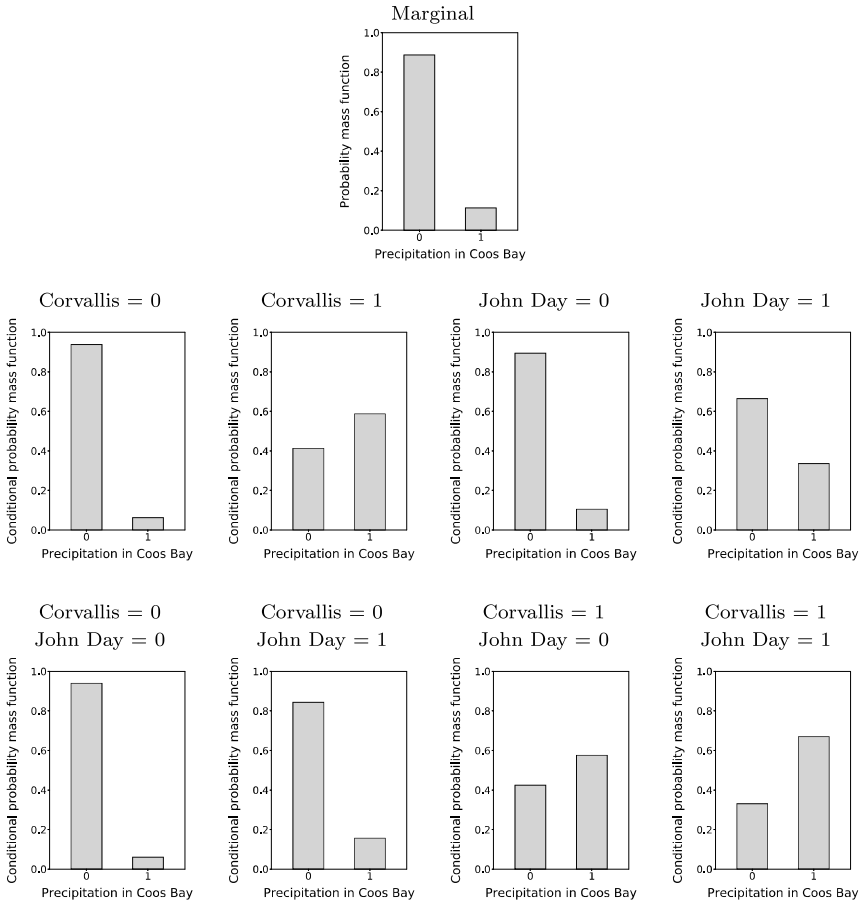

Figure 4.7 Conditional pmfs for precipitation data. The graph on the top shows the marginal pmf of precipitation in Coos Bay in Example 4.9. The second row shows the conditional pmf of precipitation in Coos Bay given different values of the random variable representing precipitation in Corvallis (first and second column), and also given different values of the random variable representing precipitation in John Day (third and fourth column). The third row shows the pmf conditioned on both random variables representing Corvallis and John Day. All pmfs are computed from the empirical joint pmf in Figure 4.2.

variables $\tilde{a}$ *and* $\tilde{b}$ *with respective ranges* A *and* B *defined on the same probability space are independent if and only if*

$$p_{\tilde{a},\tilde{b}}\left(a,b\right) = p_{\tilde{a}}\left(a\right) p_{\tilde{b}}\left(b\right), \qquad \textit{for all } a \in A, b \in B. \tag{4.70}$$

The definition can be extended to multiple random variables and, equivalently, to the entries of a random vector. In order to ensure that all possible conditional

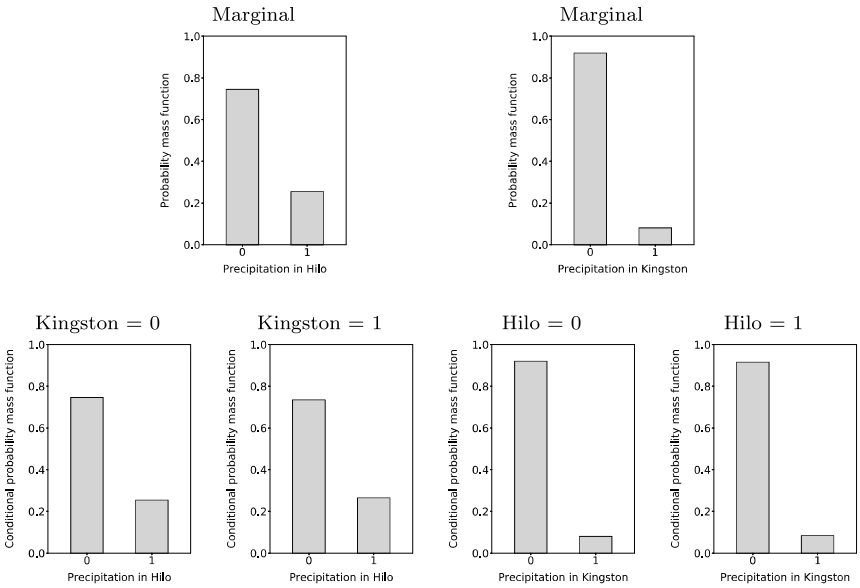

Figure 4.8 Precipitation in Hilo and Kingston is approximately independent. The graph on the top shows the marginal pmf of precipitation in Hilo (left) and Kingston (right) in Example 4.18. The second row shows the conditional pmf of precipitation in Hilo given the precipitation in Kingston (first and second column) and vice versa (third and fourth column). There is barely any change in the distributions when we condition, indicating that precipitation in both stations is approximately independent.

probabilities are the same as the marginal probabilities, we require that the joint pmf factorizes completely.

Definition 4.17 (Random vector with independent entries)**.** *The d entries $\tilde{x}[1]$, $\tilde{x}[2]$, $\ldots$, $\tilde{x}[d]$ in a discrete random vector $\tilde{x}$ are independent if and only if*

$$p_{\tilde{x}}\left(x\right) = \prod_{i=1}^{d} p_{\tilde{x}[i]}\left(x[i]\right), \tag{4.71}$$

for all possible values of the entries.

Example 4.18 (Precipitation in Hawaii and Rhode Island)**.** We consider hourly precipitation data measured at weather stations in Hilo, Hawaii and Kingston, Rhode Island on 2015, extracted from Dataset 9. As in Example 4.9, we define a random vector of dimension two, where each entry is a Bernoulli random variable indicating whether there is precipitation in Hilo ($\tilde{x}[1] = 1$), and Kingston ($\tilde{x}[2] = 1$). We compute the empirical joint pmf of the random vector, and use it to obtain the marginal and conditional pmfs of its entries. Figure 4.8 shows the pmfs. The probability of precipitation at Hilo increases from 25.5% to 26.5% if there is

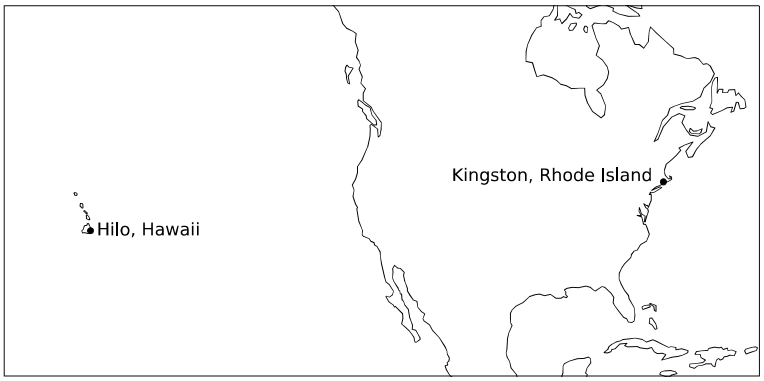

Figure 4.9 Weather stations in Kingston and Hilo. Location of the weather stations of Kingston (Rhode Island) and Hilo (Hawaii).

precipitation in Kingston, and decreases very slightly to 25.4% if there isn't. The probability of precipitation at Kingston increases from 8.1% to 8.4% if there is precipitation in Hilo, and decreases again very slightly to 8.0% if there isn't. Strictly speaking, the marginal and conditional pmfs are not exactly the same, but the changes are very slight. Practically speaking, precipitation in both stations is approximately independent. This is not surprising given their geographical location, shown in Figure 4.9.

Independence is a crucial modeling assumption when we estimate probabilities from data. Both approaches to estimate pmfs described in Chapter 2, the nonparametric empirical pmf estimator from Section 2.2 and the parametric maximum-likelihood estimator from Section 2.4, rely on the premise that the data are independent and identically distributed (i.i.d.). We already defined this concept using independence of events in Definition 2.23. The following definition is equivalent.

Definition 4.19 (Independent identically distributed random variables)**.** *Let $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ be discrete random variables belonging to the same probability space. The random variables are identically distributed if their marginal pmfs are the same*

$$p_{\tilde{a}_1} = p_{\tilde{a}_2} = \cdots = p_{\tilde{a}_n} = p_{\tilde{a}} \tag{4.72}$$

for some pmf $p_{\tilde{a}}$. They are independent and identically distributed (i.i.d.) if their joint pmf is equal to the product of the marginal pmfs

$$p_{\tilde{a}_1,\tilde{a}_2,\ldots,\tilde{a}_n}(a_1, a_2, \ldots, a_n) = \prod_{i=1}^{d} p_{\tilde{a}}(a_i), \tag{4.73}$$

for all possible values of a_1, a_2, ..., a_n.

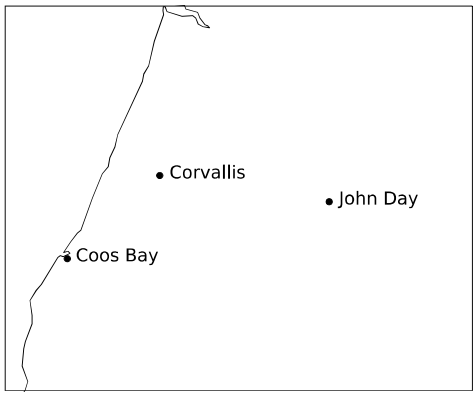

Figure 4.10 Weather stations in Oregon. Location of the weather stations of Coos Bay, Corvallis, and John Day in Oregon. John Day is in the interior, which explains why it is much drier. Coos Bay and Corvallis are quite close, so their precipitation patterns are very dependent.

4.5 Conditional Independence

In order to motivate the definition of conditional independence, consider the conditional pmfs in Figure 4.7. Coos Bay is clearly dependent on John Day. When there is precipitation in John Day, the probability of precipitation in Coos Bay increases from 11.3% to 33.6%, and when there isn't it decreases to 10.6%.

Now, let us assume we know there is precipitation in Corvallis. In that case, knowing that there is also precipitation in John Day is much less informative. Conditioned just on precipitation in Corvallis, the probability of precipitation in Coos Bay is 58.7%. Conditioned on precipitation in both Corvallis and John Day, the probability is 67.0%. Conditioned on precipitation in Corvallis and no precipitation in John Day, it is 57.6%. Similarly, once we know that there is no precipitation in Corvallis, knowing whether there is precipitation or not in John Day does not affect the conditional probability of precipitation in Coos Bay as much (it increases from 6.2% to 15.6% if there is, and decreases to 6.0% if there isn't).

In summary, the dependence between Coos Bay and John Day is *greatly reduced when we condition on Corvallis*. This makes sense given their geographical locations; Corvallis is situated between them and closer to Coos Bay (see the map in Figure 4.10). Conditional independence occurs when this effect is taken to the extreme, and the dependence between two random variables is completely *explained away* if the value of a third random variable is known.

Two random variables $\tilde{a}$ and $\tilde{b}$ are conditionally independent given a third random variable $\tilde{c}$, if our uncertainty about $\tilde{a}$ is not modified when $\tilde{b}$ is revealed, *as long as the value of* $\tilde{c}$ *is known.*

Definition 4.20 (Conditionally independent random variables)**.** *Two discrete*

random variables $\tilde{a}$ and $\tilde{b}$ with respective ranges A and B defined on the same probability space are conditionally independent given a random variable $\tilde{c}$ with range C if and only if

$$p_{\tilde{a},\tilde{b}\,|\,\tilde{c}}\,(a,b\,|\,c) = p_{\tilde{a}\,|\,\tilde{c}}\,(a\,|\,c)\,p_{\tilde{b}\,|\,\tilde{c}}\,(b\,|\,c)\,, \qquad \textit{for all } a \in A, b \in B, c \in C. \tag{4.74}$$

The definition can be extended to multiple random variables or random vectors, conditioned on multiple random variables. Notation can get a bit complicated, but the main idea is the same as independence: all conditional distributions must equal the (conditional) marginals, which implies that the joint pmfs factorize into the product of these marginals.

Definition 4.21 (Conditionally independent random vectors)**.** *The d entries $\tilde{x}[1]$, $\tilde{x}[2]$, ..., $\tilde{x}[d]$ in a discrete random vector $\tilde{x}$ are conditionally independent given a random variable $\tilde{a}$ if and only if*

$$p_{\tilde{x}\,|\,\tilde{a}}\,(x\,|\,a) = \prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{a}}\,(x[i]\,|\,a)\,, \tag{4.75}$$

for all possible values of x and a.

As discussed in Section 1.6, independence does *not* imply conditional independence or vice versa.

4.6 Causal Inference

Causal inference aims to identify causal effects from data, which is a key task in many data-science applications. In Section 4.6.1 we introduce the framework of potential outcomes, which allows us to define causal relationships formally using random variables. Section 4.6.2 illustrates the difficulty of estimating such relationships in the presence of confounding factors through Simpson's paradox. Section 4.6.3 describes how to neutralize confounding factors via randomization. Finally, in Section 4.6.4 we explain how to control for confounding factors when randomization is not possible.

4.6.1 Potential Outcomes

Consider the problem of evaluating the efficacy of a new drug. Let $\tilde{t}$ be a Bernoulli random variable indicating whether a patient received the drug ($\tilde{t} = 1$) or not ($\tilde{t} = 0$), and let $\tilde{y}$ be another Bernoulli random variable indicating whether the patient recovered ($\tilde{y} = 1$) or not ($\tilde{y} = 0$). If $\mathrm{P}(\tilde{y} = 1\,|\,\tilde{t} = 1)$ is larger than $\mathrm{P}(\tilde{y} = 1\,|\,\tilde{t} = 0)$, then treated patients recover more often than untreated patients. Can we conclude that the treatment is causing the recovery? No! To explain why, we introduce the concept of potential outcomes.∗

∗Notice that here outcome does not refer to an element in a sample space, as in Section 1.2. Instead, it is a random variable representing a quantity of interest that may (or may not) be affected by the treatment.

Table 4.4 ***Potential outcomes.*** *In order to characterize causal effects, we model both potential outcomes* $\widetilde{\text{po}}_0$ *and* $\widetilde{\text{po}}_1$ *of a treatment* $\tilde{t}$ *probabilistically. However, in practice only one of the potential outcomes is observed. If* $\tilde{t} = 0$ *(indicated by a cross) we observe* $\tilde{y} := \widetilde{\text{po}}_0$*. If* $\tilde{t} = 1$ *(indicated by a tick) we observe* $\tilde{y} := \widetilde{\text{po}}_1$*. In this example, the outcomes are either positive (smiley face) or negative (frownie face).*

Treatment $\tilde{t}$	Observed outcome $\tilde{y}$	Outcome if not treated $\widetilde{\text{po}}_0$	Outcome if treated $\widetilde{\text{po}}_1$
✗	☹	☹	?
✗	☺	☺	?
✓	☺	?	☺
✓	☹	?	☹
✓	☺	?	☺

In order to characterize the causal effect of a treatment, we model what happens in a hypothetical scenario where we treat all patients, and also in another hypothetical scenario where we treat no patients at all. The potential outcomes are random variables that represent these two alternatives. The potential outcome $\widetilde{\text{po}}_0$ indicates whether a patient would recover ($\widetilde{\text{po}}_0 = 1$) or not ($\widetilde{\text{po}}_0 = 0$) in the absence of treatment, *regardless of whether they are actually treated or not*. Similarly, the potential outcome $\widetilde{\text{po}}_1$ indicates whether a patient would recover ($\widetilde{\text{po}}_1 = 1$) or not ($\widetilde{\text{po}}_1 = 0$) if they were treated, and is also defined for patients that are *not treated*. In reality, both potential outcomes cannot be observed simultaneously: we cannot treat and not treat a patient at the same time! As illustrated in Table 4.4, the observed outcome $\tilde{y}$ is equal to either $\widetilde{\text{po}}_0$ or $\widetilde{\text{po}}_1$ depending on the treatment $\tilde{t}$,

$$\tilde{y} := \begin{cases} \widetilde{\text{po}}_0 & \text{if} \quad \tilde{t} = 0, \\ \widetilde{\text{po}}_1 & \text{if} \quad \tilde{t} = 1. \end{cases} \tag{4.76}$$

If $\tilde{t} = 0$, we observe $\widetilde{\text{po}}_0$, but not $\widetilde{\text{po}}_1$. In that case we call $\widetilde{\text{po}}_1$ a *counterfactual*, because it captures what occurs in a situation that is counter to factual reality. Conversely, if $\tilde{t} = 1$, we observe $\widetilde{\text{po}}_1$, and $\widetilde{\text{po}}_0$ is an unobserved counterfactual.

By incorporating both potential outcomes into a model, we can determine whether the treatment $\tilde{t}$ makes a difference or not. We consider that $\tilde{t}$ has a causal effect on $\tilde{y}$ if

$$\text{P}\left(\widetilde{\text{po}}_0 = 1\right) \neq \text{P}\left(\widetilde{\text{po}}_1 = 1\right), \tag{4.77}$$

because this implies that the probability of recovery depends on whether the patient takes the drug or not. The key challenge of causal inference is that it is not always possible to compute these probabilities from the available data, especially when the data are from an observational study, where we have no control over what patients receive the treatment.

Example 4.22 (Drug efficacy: Observational study)**.** We consider a fictitious observational study to evaluate a new drug for a certain disease. Treatment with the drug is represented by the random variable $\tilde{t}$. The outcome of interest is recovery from the disease, represented by a random variable $\tilde{y}$ defined as in (4.76), where $\widetilde{\mathrm{po}}_1 = 1$ and $\widetilde{\mathrm{po}}_0 = 1$ indicate recovery with and without the drug, respectively.

It turns out that the efficacy of the drug depends on the age of the patients. The random variable $\tilde{a}$ indicates whether a patient is old ($\tilde{a} = \text{old}$) or young ($\tilde{a} = \text{young}$). Half of the patients are young and half are old, so $p_{\tilde{a}}(\text{old}) = p_{\tilde{a}}(\text{young}) = 0.5$. For young patients, the recovery rate is relatively high (80%), as they are generally healthier, but the drug is completely useless

$$p_{\widetilde{\mathrm{po}}_0 \mid \tilde{a}}(1 \mid \text{young}) = p_{\widetilde{\mathrm{po}}_1 \mid \tilde{a}}(1 \mid \text{young}) := 0.8. \tag{4.78}$$

For old patients, the recovery rate is lower and the drug is quite effective,

$$p_{\widetilde{\mathrm{po}}_0 \mid \tilde{a}}(1 \mid \text{old}) := 0.2, \qquad p_{\widetilde{\mathrm{po}}_1 \mid \tilde{a}}(1 \mid \text{old}) := 0.4. \tag{4.79}$$

Old patients are distrustful of the new drug. As a result, they are treated at a lower rate than young patients:

$$p_{\tilde{t} \mid \tilde{a}}(1 \mid \text{young}) := 0.7, \qquad p_{\tilde{t} \mid \tilde{a}}(1 \mid \text{old}) := 0.3. \tag{4.80}$$

Our goal is to determine the efficacy of the drug, which can be quantified by comparing $p_{\widetilde{\mathrm{po}}_0}(1)$ and $p_{\widetilde{\mathrm{po}}_1}(1)$:

$$\begin{aligned} p_{\widetilde{\mathrm{po}}_0}(1) &= \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\mathrm{po}}_0 \mid \tilde{a}}(1 \mid a) && (4.81)\\ &= 0.5 \cdot 0.8 + 0.5 \cdot 0.2 = 0.5, && (4.82)\\ p_{\widetilde{\mathrm{po}}_1}(1) &= \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\mathrm{po}}_1 \mid \tilde{a}}(1 \mid a) && (4.83)\\ &= 0.5 \cdot 0.8 + 0.5 \cdot 0.4 = 0.6. && (4.84) \end{aligned}$$

Notice that the efficacy associated to each potential outcome does *not* depend on what patients were actually treated: it corresponds to the recovery rate if everyone were treated ($p_{\widetilde{\mathrm{po}}_0}(1)$) or if nobody were treated ($p_{\widetilde{\mathrm{po}}_1}(1)$). The drug is quite effective, it increases the chance of recovery by 10%.

Now, let us consider the data we observe in our observational study. The observed probabilities of recovery for untreated and treated patients, respectively, are

$$\begin{aligned} p_{\tilde{y} \mid \tilde{t}}(1 \mid 0) &= p_{\widetilde{\mathrm{po}}_0 \mid \tilde{t}}(1 \mid 0), && (4.85)\\ p_{\tilde{y} \mid \tilde{t}}(1 \mid 1) &= p_{\widetilde{\mathrm{po}}_1 \mid \tilde{t}}(1 \mid 1). && (4.86) \end{aligned}$$

Unfortunately, these probabilities can be very different from $p_{\widetilde{\mathrm{po}}_0}(1)$ and $p_{\widetilde{\mathrm{po}}_1}(1)$,

due to the influence of age. We have

$$p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,0) = p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{t}}(1\,|\,0) = \sum_{a\in\{\text{young,old}\}} p_{\widetilde{\mathrm{po}}_0,\tilde{a}\,|\,\tilde{t}}(1,a\,|\,0) \tag{4.87}$$

$$= \sum_{a\in\{\text{young,old}\}} p_{\tilde{a}\,|\,\tilde{t}}(a\,|\,0)p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a},\tilde{t}}(1\,|\,a,0). \tag{4.88}$$

Let us assume that $\widetilde{\mathrm{po}}_0$ and $\tilde{t}$ are conditionally independent given $\tilde{a}$, and so are $\widetilde{\mathrm{po}}_1$ and $\tilde{t}$, so that

$$p_{\tilde{a},\widetilde{\mathrm{po}}_0,\tilde{t}} = p_{\tilde{a}} \cdot p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a}} \cdot p_{\tilde{t}\,|\,\tilde{a}}, \qquad p_{\tilde{a},\widetilde{\mathrm{po}}_1,\tilde{t}} = p_{\tilde{a}} \cdot p_{\widetilde{\mathrm{po}}_1\,|\,\tilde{a}} \cdot p_{\tilde{t}\,|\,\tilde{a}}. \tag{4.89}$$

As we discuss in Example 4.26, this conditional-independence assumption means that within each age group there is no systematic difference between the treated and untreated patients. By (4.88), under this assumption,

$$p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,0) = \sum_{a\in\{\text{young,old}\}} p_{\tilde{a}\,|\,\tilde{t}}(a\,|\,0)p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a}}(1\,|\,a). \tag{4.90}$$

This is almost equal to

$$p_{\widetilde{\mathrm{po}}_0}(1) = \sum_{a\in\{\text{young,old}\}} p_{\tilde{a}}(a)p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a}}(1\,|\,a), \tag{4.91}$$

but it is *not* because by Bayes' rule

$$p_{\tilde{a}\,|\,\tilde{t}}(\text{young}\,|\,0) = \frac{p_{\tilde{t}\,|\,\tilde{a}}(0\,|\,\text{young})p_{\tilde{a}}(\text{young})}{p_{\tilde{t}\,|\,\tilde{a}}(0\,|\,\text{young})p_{\tilde{a}}(\text{young}) + p_{\tilde{t}\,|\,\tilde{a}}(0\,|\,\text{old})p_{\tilde{a}}(\text{old})} \tag{4.92}$$

$$= \frac{0.3\cdot 0.5}{0.3\cdot 0.5 + 0.7\cdot 0.5} = 0.3 \neq 0.5 = p_{\tilde{a}}(\text{young}), \tag{4.93}$$

and consequently $p_{\tilde{a}\,|\,\tilde{t}}(\text{old}\,|\,0) = 0.7 \neq 0.5 = p_{\tilde{a}}(\text{old})$. The fraction of young patients in the untreated group is lower and the fraction of old patients higher than in the overall population, which distorts the observed efficacy:

$$p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,0) = \sum_{a\in\{\text{young,old}\}} p_{\tilde{a}\,|\,\tilde{t}}(a\,|\,0)p_{\widetilde{\mathrm{po}}_0\,|\,\tilde{a}}(1\,|\,a) \tag{4.94}$$

$$= \mathbf{0.3}\cdot 0.8 + \mathbf{0.7}\cdot 0.2 \tag{4.95}$$

$$= 0.38 \neq 0.5 = p_{\widetilde{\mathrm{po}}_0}(1). \tag{4.96}$$

Similarly, by Bayes' rule $p_{\tilde{a}\,|\,\tilde{t}}(\text{young}\,|\,1) = 0.7$, so by the same reasoning

$$p_{\tilde{y}\,|\,\tilde{t}}(1\,|\,1) = \sum_{a\in\{\text{young,old}\}} p_{\tilde{a}\,|\,\tilde{t}}(a\,|\,1)p_{\widetilde{\mathrm{po}}_1\,|\,\tilde{a}}(1\,|\,a) \tag{4.97}$$

$$= \mathbf{0.7}\cdot 0.8 + \mathbf{0.3}\cdot 0.4 \tag{4.98}$$

$$= 0.68 \neq 0.6 = p_{\widetilde{\mathrm{po}}_1}(1). \tag{4.99}$$

The observed efficacy of the drug is distorted by the discrepancy in treatment between old patients (0.3, highlighted in bold) and young patients (0.7, highlighted in bold). Young patients are overrepresented in the treatment group, which boosts

Table 4.5 ***Who is the better shooter?*** *The table reports the 3-point shooting percentages of Stephen Curry and Courtney Lee in the 2014/2015 NBA season. Lee has a better overall percentage, but Curry has better percentages both for long threes (taken at a distance of more than 24 feet) and for short threes. This paradox is due to Curry shooting a larger proportion of long threes, which are more difficult to make.*

	Stephen Curry	Courtney Lee
Short threes ($\leq$ 24 feet)	45/90 = **50.0%**	56/116 = 48.3%
Long threes ($>$ 24 feet)	145/366 = **39.6%**	19/55 = 34.5%
Total	190/456 = 41.7%	75/171 = **43.9%**

the apparent efficacy of the drug because those patients are more likely to recover *regardless of the treatment*. As a result, it looks like the drug increases the probability of recovery by 30% instead of 10%!

4.6.2 Confounders And Simpson's Paradox

In Example 4.22 the age of the patients $\tilde{a}$ occludes the underlying causal relationship between recovery and treatment. Such variables are called *confounders* or confounding factors. In this section, we show that confounders can make conditional probabilities very difficult to interpret.

While looking at data from NBA games in the 2014/2015 season (Dataset 10), I was puzzled to find that Courtney Lee from the Memphis Grizzlies had a better 3-point percentage (43.9%) than Stephen Curry (41.7%).* In case you don't follow basketball, Stephen Curry is considered the best 3-point shooter of all time. During the 2014/2015 season he led the Warriors to an NBA championship and was declared the NBA Most Valuable Player, mostly due to his 3-point shooting prowess.

Let $\tilde{y}$ be a Bernoulli random variable representing whether a 3-point shot is made ($\tilde{y} = 1$) or not ($\tilde{y} = 0$). We are interested in determining the causal effect of the shooter being Lee or Curry. We interpret the player taking the shot as a *treatment* $\tilde{t}$. As you can see, in causal inference the term treatment is often used in a figurative sense. According to the data, shown in Table 4.5,

$$0.439 = P\left(\tilde{y} = 1 \,|\, \tilde{t} = \text{Lee}\right) > P\left(\tilde{y} = 1 \,|\, \tilde{t} = \text{Curry}\right) = 0.417, \tag{4.100}$$

which suggests that Lee shoots better than Curry! With all due respect to Lee, who was a fantastic player, this is very unexpected.

To dig a bit deeper, we divide the 3-point shots in the dataset into short threes, taken at a distance 24 feet or less, and long threes, taken at a distance of more than 24 feet. In the NBA the 3-point line is 22 feet away in the corners, and 23-foot 9-inch away elsewhere. Surprisingly, when we analyze the different types

*For some reason, they do not contain all games from the season, so the percentages reported in Table 4.5 differ from the official percentages.

of shots separately, the pattern is exactly the opposite to the one we observe for the aggregated data! Curry's percentage is better for long threes and also for short threes,

$$0.500 = P(\tilde{y} = 1 \,|\, \tilde{t} = \text{Curry}, \tilde{d} = \text{short}) > P(\tilde{y} = 1 \,|\, \tilde{t} = \text{Lee}, \tilde{d} = \text{short}) = 0.483,$$
$$0.396 = P(\tilde{y} = 1 \,|\, \tilde{t} = \text{Curry}, \tilde{d} = \text{long}) > P(\tilde{y} = 1 \,|\, \tilde{t} = \text{Lee}, \tilde{d} = \text{short}) = 0.345,$$

where $\tilde{d}$ is a random variable indicating whether the shot is long or short. This is known as Simpson's paradox.

Simpson's paradox occurs due to the presence of a confounder. In our example the confounder is the shot distance $\tilde{d}$. Lee takes many more short threes than Curry:

$$0.678 = \mathrm{P}(\tilde{d} = \text{short} \,|\, \tilde{t} = \text{Lee}) > \mathrm{P}(\tilde{d} = \text{short} \,|\, \tilde{t} = \text{Curry}) = 0.197. \tag{4.101}$$

Even though he doesn't shoot short threes as well as Curry, this shot selection boosts Lee's overall percentage, because long threes are more difficult for both players. The following decomposition of the conditional probabilities of $\tilde{y}$ given $\tilde{t}$ reveals the effect of the confounder $\tilde{d}$,

$$\mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = \text{Lee}\right) \tag{4.102}$$
$$= \mathrm{P}(\tilde{y} = 1, \tilde{d} = \text{short} \,|\, \tilde{t} = \text{Lee}) + \mathrm{P}(\tilde{y} = 1, \tilde{d} = \text{long} \,|\, \tilde{t} = \text{Lee}) \tag{4.103}$$
$$= \mathrm{P}(\tilde{d} = \text{short} \,|\, \tilde{t} = \text{Lee})\mathrm{P}(\tilde{y} = 1 \,|\, \tilde{d} = \text{short}, \tilde{t} = \text{Lee}) + \mathrm{P}(\tilde{d} = \text{long} \,|\, \tilde{t} = \text{Lee})\mathrm{P}(\tilde{y} = 1 \,|\, \tilde{d} = \text{long}, \tilde{t} = \text{Lee}) \tag{4.104}$$
$$= \mathbf{0.678} \cdot 0.483 + \mathbf{0.322} \cdot 0.345 = 0.439, \tag{4.105}$$
$$\mathrm{P}\left(\tilde{y} = 1 \,|\, \tilde{t} = \text{Curry}\right) \tag{4.106}$$
$$= \mathrm{P}(\tilde{y} = 1, \tilde{d} = \text{short} \,|\, \tilde{t} = \text{Curry}) + \mathrm{P}(\tilde{y} = 1, \tilde{d} = \text{long} \,|\, \tilde{t} = \text{Curry}) \tag{4.107}$$
$$= \mathrm{P}(\tilde{d} = \text{short} \,|\, \tilde{t} = \text{Curry})\mathrm{P}(\tilde{y} = 1 \,|\, \tilde{d} = \text{short}, \tilde{t} = \text{Curry}) + \mathrm{P}(\tilde{d} = \text{long} \,|\, \tilde{t} = \text{Curry})\mathrm{P}(\tilde{y} = 1 \,|\, \tilde{d} = \text{long}, \tilde{t} = \text{Curry}) \tag{4.108}$$
$$= \mathbf{0.197} \cdot 0.500 + \mathbf{0.803} \cdot 0.396 = 0.417. \tag{4.109}$$

Even though the conditional probability of $\tilde{y} = 1$ given $\tilde{d} = \text{short}$ and $\tilde{d} = \text{long}$ are both larger for Curry, they are reweighted by the conditional probability of $\tilde{d}$ given $\tilde{t}$ (in bold) to yield a higher overall percentage for Lee.

4.6.3 Randomized Experiments

Example 4.22 and Section 4.6.2 show that confounders can make it very challenging to perform causal inference from observational data. Fortunately, their effect can be neutralized by ensuring that the treatment is independent from the potential outcomes. In that case, we can easily estimate the probability of the potential outcomes associated to the treatment from data.

Theorem 4.23 (Randomization enables causal inference)**.** *Let $\tilde{y}$ be a random*

variable representing an observed outcome corresponding to two potential outcomes $\widetilde{\mathrm{po}}_0$ *and* $\widetilde{\mathrm{po}}_1$ *associated to a treatment* $\tilde{t}$*,*

$$\tilde{y} := \begin{cases} \widetilde{\mathrm{po}}_0 & \text{if} \quad \tilde{t} = 0, \\ \widetilde{\mathrm{po}}_1 & \text{if} \quad \tilde{t} = 1. \end{cases} \tag{4.110}$$

If $\tilde{t}$ *and* $\widetilde{\mathrm{po}}_0$ *are independent, and* $\tilde{t}$ *and* $\widetilde{\mathrm{po}}_1$ *are also independent, then*

$$p_{\widetilde{\mathrm{po}}_0}(1) = p_{\tilde{y} \mid \tilde{t}}(1 \mid 0), \tag{4.111}$$

$$p_{\widetilde{\mathrm{po}}_1}(1) = p_{\tilde{y} \mid \tilde{t}}(1 \mid 1). \tag{4.112}$$

Proof By definition, $p_{\widetilde{\mathrm{po}}_0 \mid \tilde{t}}(1 \mid 0) = p_{\tilde{y} \mid \tilde{t}}(1 \mid 0)$ and $p_{\widetilde{\mathrm{po}}_1 \mid \tilde{t}}(1 \mid 1) = p_{\tilde{y} \mid \tilde{t}}(1 \mid 1)$. By the independence assumption, $p_{\widetilde{\mathrm{po}}_0}(1) = p_{\widetilde{\mathrm{po}}_0 \mid \tilde{t}}(1 \mid 0)$ and $p_{\widetilde{\mathrm{po}}_1}(1) = p_{\widetilde{\mathrm{po}}_1 \mid \tilde{t}}(1 \mid 1)$. ■

The simplest and most effective way to render the treatment independent from the potential outcomes is to perform a *randomized experiment*∗. The experiment subjects are divided into a treatment group and a control group. Each subject is assigned to one of the groups, independently from the rest, ensuring that the independence condition in Theorem 4.23 is satisfied.

In a randomized experiment, the members of the treatment and control groups are very unlikely to differ systematically, which neutralizes any possible confounders, even if we are not aware of their existence! In medicine, these experiments are known as *randomized controlled trials*. When possible, the trials are double blind, meaning that the participants and the personnel involved in the trials do not know what group each patient belongs to. This is necessary to avoid placebo effects, which could jeopardize the independence between the potential outcomes and the treatment.

Example 4.24 (Drug efficacy: Randomized controlled trial)**.** New drugs have to undergo strict randomized controlled trials in order to be officially approved. In this example, we analyze a randomized controlled trial to evaluate the drug in Example 4.22. Each participant in the trial is administered the drug independently at random with probability 1/2. As a result, the treatment is independent from the potential outcomes. This might sound confusing at first. How can the treatment be independent from the outcome?

The answer is that randomization renders the *potential* outcomes independent from the treatment, but *not the observed outcome*. In our example, if the treatment is randomized then,

$$p_{\widetilde{\mathrm{po}}_1 \mid \tilde{t}}(1 \mid 0) = p_{\widetilde{\mathrm{po}}_1 \mid \tilde{t}}(1 \mid 1) = p_{\widetilde{\mathrm{po}}_1}(1) = 0.6. \tag{4.113}$$

This means that probability that the drug works is the same (0.6) for patients in the treatment group and in the control group. Similarly,

$$p_{\widetilde{\mathrm{po}}_0 \mid \tilde{t}}(1 \mid 0) = p_{\widetilde{\mathrm{po}}_0 \mid \tilde{t}}(1 \mid 1) = p_{\widetilde{\mathrm{po}}_0}(1) = 0.5. \tag{4.114}$$

∗Here randomized experiment refers to an approach to gather data, as opposed to the virtual experiment in Section 1.1, which we use as a thought experiment to gain intuition about the properties of probability.

However, the observed outcome $\tilde{y}$ is definitely not independent from the treatment. It is equal to the potential outcome $p_{\widetilde{\text{po}}_0}$ when $\tilde{t} = 0$, and to $p_{\widetilde{\text{po}}_1}$ when $\tilde{t} = 1$, so

$$p_{\tilde{y} \mid \tilde{t}}(1 \mid 0) = p_{\widetilde{\text{po}}_0 \mid \tilde{t}}(1 \mid 0) = 0.5, \tag{4.115}$$

$$p_{\tilde{y} \mid \tilde{t}}(1 \mid 1) = p_{\widetilde{\text{po}}_1 \mid \tilde{t}}(1 \mid 1) = 0.6. \tag{4.116}$$

This confirms that thanks to randomization, the conditional probabilities of the observed outcome given the treatment provide an accurate estimate of the corresponding potential outcomes, as established in Theorem 4.23.

Notice that, in contrast to the observational study in Example 4.22, in the randomized control trial, young and old patients receive the drug *at the same rate*, because the treatment $\tilde{t}$ is independent from the random variable $\tilde{a}$ representing age,

$$p_{\tilde{t} \mid \tilde{a}}(1 \mid \text{young}) = p_{\tilde{t} \mid \tilde{a}}(1 \mid \text{old}) = p_{\tilde{t}}(1) = 0.5. \tag{4.117}$$

This shows why randomization is able to neutralize confounders: the random assignment automatically balances the treatment and control groups with respect to all possible confounders, even if they are unknown.

Randomized controlled trials are essential in the development of vaccines, as illustrated by the following real-world example.

Example 4.25 (COVID-19 vaccine)**.** Vaccines against COVID-19 played a crucial role in slowing the spread of the disease and reducing its severity. Randomized controlled trials were the main tool used to evaluate their efficacy and safety. In the case of the Pfizer vaccine, 43,448 patients were randomly divided into a treatment group of 21,720 patients, who received two doses of the vaccine, and a control group of 21,728, who received two placebo doses.

The number of COVID-19 cases with onset at least 7 days after the second dose were 8 for the treatment group (0.037%), and 162 for the control group (0.746%) (**?**). The randomized assignment guarantees that the only systematic difference between the two groups was the vaccine, which makes the results very convincing. In Examples 10.33 and Figure 10.15, we apply hypothesis testing and confidence intervals to analyze these results further.

4.6.4 Adjusting For Confounders

Section 4.6.3 describes how randomization enables us to perform causal inference without having to worry about confounders. Unfortunately, randomized experiments have several important limitations. First, they are very expensive. Randomized controlled trials for new drugs typically cost at least several million dollars. Second, they cannot always be applied due to ethical reasons. For example, we cannot administer a treatment that is known to have a negative effect, such as smoking or playing American football with a shoddy helmet. Third, in

some situations controlled experiments are not possible; all we have is observational data. This is definitely the case in our basketball example; choosing what player takes each shot is not an option.

The following example shows that it is possible to *adjust* or *control* for the effect of a confounder as long as (1) we know about it, and (2) the treatment is conditionally independent from the potential outcomes given the confounder. This conditional-independence assumption guarantees that there are no systematic differences between untreated and treated subjects associated to each possible value of the confounder.

Example 4.26 (Drug efficacy: Adjusting for age)**.** Imagine that the observational data in Example 4.22 is all we have. We cannot perform a randomized controlled trial. However, we do have access to the age of each patient. How can we use this information to improve our estimate of the drug efficacy?

We have

$$p_{\widetilde{\text{po}}_0}(1) = \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\text{po}}_0 \,|\, \tilde{a}}(1 \,|\, a). \tag{4.118}$$

If we are able to estimate the different terms of this equation from the data, we can combine them to approximate $p_{\widetilde{\text{po}}_0}(1)$. The marginal pmf $p_{\tilde{a}}$ can be determined from the age of the patients. The key question is whether we can somehow obtain $p_{\widetilde{\text{po}}_0 \,|\, \tilde{a}}(1 \,|\, a)$. The observed rate of recovery for untreated young patients is

$$p_{\tilde{y} \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 0) = p_{\widetilde{\text{po}}_0 \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 0). \tag{4.119}$$

In order for this to equal the conditional pmf of the potential outcome, $\widetilde{\text{po}}_0$ must be conditionally independent from $\tilde{t}$ given $\tilde{a} = \text{young}$, so that

$$p_{\widetilde{\text{po}}_0 \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 0) = p_{\widetilde{\text{po}}_0 \,|\, \tilde{a}}(1 \,|\, \text{young}). \tag{4.120}$$

In that case, the observed rate of recovery for young patients $p_{\tilde{y} \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 0)$ is equal to the true conditional rate $p_{\widetilde{\text{po}}_0 \,|\, \tilde{a}}(1 \,|\, \text{young}) = 0.8$! Crucially, equality only holds when patient recovery ($\widetilde{\text{po}}_0$ or $\widetilde{\text{po}}_1$) is conditionally independent from the treatment ($\tilde{t}$) given $\tilde{a} = \text{young}$. This guarantees that there are no other confounders producing systematic differences among the treated and untreated young patients.

If we assume that $\widetilde{\text{po}}_0$ is conditionally independent from $\tilde{t}$ given $\tilde{a}$ (and in particular, given $\tilde{a} = \text{old}$), the observed rate of recovery also reveals the conditional distribution of the potential outcome for untreated old patients,

$$p_{\tilde{y} \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{old}, 0) = p_{\widetilde{\text{po}}_0 \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{old}, 0) = p_{\widetilde{\text{po}}_0 \,|\, \tilde{a}}(1 \,|\, \text{old}) = 0.2. \tag{4.121}$$

By the same argument, if $\widetilde{\text{po}}_1$ is conditionally independent from $\tilde{t}$ given $\tilde{a}$

$$p_{\tilde{y} \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 1) = p_{\widetilde{\text{po}}_1 \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{young}, 1) = p_{\widetilde{\text{po}}_1 \,|\, \tilde{a}}(1 \,|\, \text{young}) = 0.8, \tag{4.122}$$

$$p_{\tilde{y} \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{old}, 1) = p_{\widetilde{\text{po}}_1 \,|\, \tilde{a},\tilde{t}}(1 \,|\, \text{old}, 1) = p_{\widetilde{\text{po}}_1 \,|\, \tilde{a}}(1 \,|\, \text{old}) = 0.4. \tag{4.123}$$

Aggregating these probabilities weighted by the marginal pmf of $\tilde{a}$ yields

$$p_{\widetilde{\text{po}}_0}(1) = \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\text{po}}_0 \mid \tilde{a}}(1 \mid a) \tag{4.124}$$

$$= \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\tilde{y} \mid \tilde{a}, \tilde{t}}(1 \mid a, 0) \tag{4.125}$$

$$= 0.5 \cdot 0.8 + 0.5 \cdot 0.2 = 0.5, \tag{4.126}$$

$$p_{\widetilde{\text{po}}_1}(1) = \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\widetilde{\text{po}}_1 \mid \tilde{a}}(1 \mid a) \tag{4.127}$$

$$= \sum_{a \in \{\text{young,old}\}} p_{\tilde{a}}(a) p_{\tilde{y} \mid \tilde{a}, \tilde{t}}(1 \mid a, 1) \tag{4.128}$$

$$= 0.5 \cdot 0.8 + 0.5 \cdot 0.4 = 0.6. \tag{4.129}$$

We obtain an exact estimate of the probability of the potential outcomes! Unfortunately, in practice we often cannot completely rule out the presence of additional unknown confounders, which would violate the conditional-independence assumption.

The procedure derived in Example 4.26 is a key tool in causal inference.

Theorem 4.27 (Adjusting for a confounder)**.** *Let $\tilde{y}$ be a random variable that represents the observed outcome corresponding to two potential outcomes $\widetilde{\text{po}}_0$ and $\widetilde{\text{po}}_1$ associated to a treatment $\tilde{t}$,*

$$\tilde{y} := \begin{cases} \widetilde{\text{po}}_0 & \text{if} \quad \tilde{t} = 0, \\ \widetilde{\text{po}}_1 & \text{if} \quad \tilde{t} = 1, \end{cases}$$

and let $\tilde{c}$ be a confounder, represented by a discrete random variable with range C. If the treatment $\tilde{t}$ and the potential outcomes $\widetilde{\text{po}}_0$ and $\widetilde{\text{po}}_1$ are conditionally independent given $\tilde{c}$, then

$$p_{\widetilde{\text{po}}_0}(1) = \sum_{c \in C} p_{\tilde{c}}(c) p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 0), \tag{4.130}$$

$$p_{\widetilde{\text{po}}_1}(1) = \sum_{c \in C} p_{\tilde{c}}(c) p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 1). \tag{4.131}$$

Proof By the conditional-independence assumption, we can estimate the conditional probability of each potential outcome given the confounder:

$$p_{\widetilde{\text{po}}_0 \mid \tilde{c}}(1 \mid c) = p_{\widetilde{\text{po}}_0 \mid \tilde{c}, \tilde{t}}(1 \mid c, 0) \tag{4.132}$$

$$= p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 0), \tag{4.133}$$

$$p_{\widetilde{\text{po}}_1 \mid \tilde{c}}(1 \mid c) = p_{\widetilde{\text{po}}_1 \mid \tilde{c}, \tilde{t}}(1 \mid c, 1) \tag{4.134}$$

$$= p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 1). \tag{4.135}$$

The desired result then follows by marginalization and the chain rule,

$$p_{\widetilde{\text{po}}_0}(1) = \sum_{c \in C} p_{\tilde{c}}(c) p_{\widetilde{\text{po}}_0 \mid \tilde{c}}(1 \mid c) \tag{4.136}$$

$$= \sum_{c \in C} p_{\tilde{c}}(c) p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 0), \tag{4.137}$$

$$p_{\widetilde{\text{po}}_1}(1) = \sum_{c \in C} p_{\tilde{c}}(c) p_{\widetilde{\text{po}}_1 \mid \tilde{c}}(1 \mid c) \tag{4.138}$$

$$= \sum_{c \in C} p_{\tilde{c}}(c) p_{\tilde{y} \mid \tilde{c}, \tilde{t}}(1 \mid c, 1). \tag{4.139}$$

■

Example 4.28 (3-point shooting: Adjusting for shot distance)**.** We apply Theorem 4.27 to the example in Section 4.6.2 in order to estimate the probabilities of the potential outcomes associated to Curry and Lee. This yields an adjusted 3-point percentage, which suggests that Curry is the better shooter

$$p_{\widetilde{\text{po}}_{\text{Curry}}}(1) = \sum_{d \in \{\text{short,long}\}} p_{\tilde{d}}(d) p_{\tilde{y} \mid \tilde{d}, \tilde{t}}(1 \mid d, \text{Curry}) \tag{4.140}$$

$$= 0.329 \cdot 0.5 + 0.671 \cdot 0.396 = 0.430, \tag{4.141}$$

$$p_{\widetilde{\text{po}}_{\text{Lee}}}(1) = \sum_{d \in \{\text{short,long}\}} p_{\tilde{d}}(d) p_{\tilde{y} \mid \tilde{d}, \tilde{t}}(1 \mid d, \text{Lee}) \tag{4.142}$$

$$= 0.329 \cdot 0.483 + 0.671 \cdot 0.345 = 0.391, \tag{4.143}$$

where $p_{\tilde{d}}$ is estimated from the data in Table 4.5 using the corresponding empirical pmf.

The assumption of Theorem 4.27 is unlikely to hold exactly for this setting. Even controlling for shot distance, Curry's shots are probably still more difficult than Lee's because opposing defenses are very focused on him. Nonetheless, our adjusted percentage is definitively more informative about the shooting abilities of the two players than the raw observed 3-point percentage! This often happens in applications of causal inference to observational data. In the absence of randomized experiments, it is very difficult to account for all possible confounders, but at least we should adjust for the ones we know.

4.7 The Curse Of Dimensionality

In the examples we have considered up to now, we model the joint distribution of at most three variables. In practice, probabilistic models often have many more. For instance, Dataset 9 contains measurements from 134 locations. Imagine that we want to build a model for precipitation where each location is represented by a random variable. To simplify matters, we only model the presence or absence of precipitation, as in Examples 4.9 and 4.18, so each variable has just two possible values. If we try to estimate the joint pmf of these 134 Bernoulli random

variables using empirical probabilities, we run into a problem. The joint pmf has 2^{134} entries, which is more than 10^{40}! In comparison, the number of available data (8,760 hourly measurements) is ridiculously small. In fact, as reported in Figure 4.11, 89% of the precipitation patterns in the data are observed just once, and only 1.9% of patterns are observed more than three times. As a result, the empirical probabilities that we compute are completely inaccurate: the vast majority are zero, and among the nonzero probabilities, most equal $1/n$ (where n is the number of data). This is highly problematic for any practical application of our model. For example, we cannot use it for forecasting future precipitation patterns, as we almost certainly will not have observed them previously.

Our precipitation example is not an exception, but rather the norm: estimating the joint pmf of multiple random variables is usually *intractable* unless the number of variables is very small. In other words, it is often impossible to estimate the joint pmf of high-dimensional random vectors! This phenomenon is known as the *curse of dimensionality*. In order to ensure tractability, we need to make assumptions about the observed data, and design simplified models with a number of parameters that does not explode exponentially with the number of variables. This is usually achieved via independence assumptions.

In our precipitation example, we can assume that the measurements from all stations are mutually independent. In that case we only need to estimate 134 parameters to determine the marginal distribution of each of the 134 Bernoulli random variables, which is easy to do given that we have 8,760 data points. Of course, the bad news is that we are not modeling dependencies that may be very important to our analysis. For instance, there is clearly a strong dependence between Coos Bay and Corvallis, which we may want to take into account. The key to effective probabilistic modeling is striking a balance between incorporating such dependencies, while keeping the number of variables small enough to ensure that we can actually fit the model with the available data. The following sections describe two popular models based on this philosophy: naive Bayes and Markov chains.

4.8 Classification Via Naive Bayes

In statistics and machine learning, classification is the problem of assigning a category or class to a data point. For example, we may want to identify what object is present in a picture, or whether an email is spam or not. Assume that we have gathered a dataset of n examples (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Each example consists of a d-dimensional *feature* vector x_i (e.g. a picture, or the text in an email) and its corresponding *label* $y_i \in \{1, 2, \ldots, c\}$, indicating which of the c classes is associated to the example (e.g. *car* in the case of the picture, or *spam* in the case of the email). The goal is to estimate the label of new examples from the corresponding features.

To perform classification, we model the data using a d-dimensional random vector $\tilde{x}$ and the corresponding class label as a random variable $\tilde{y}$. We then estimate the conditional pmf $p_{\tilde{y}\,|\,\tilde{x}}$ of $\tilde{y}$ given $\tilde{x}$ from the available data, and use it to clas-

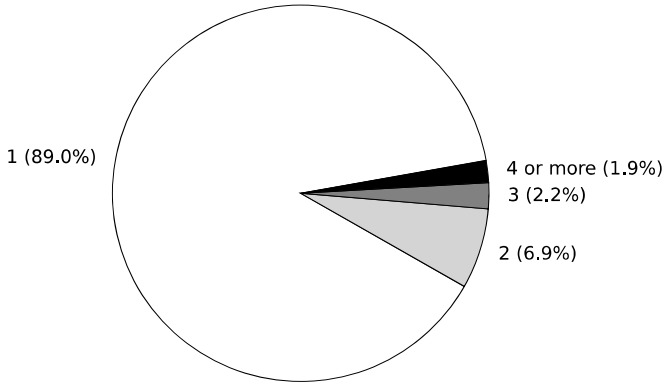

Figure 4.11 The curse of dimensionality. We consider Dataset 9, which contains 8,760 hourly precipitation measurements at 134 weather stations. The precipitation pattern for each measurement encodes in what stations there is precipitation at that time. The pie chart reports what fraction of precipitation patterns are repeated in the data. Most precipitation patterns occur just once (89.0%). Only 1.9% of patterns are observed more than three times.

sify new examples by selecting the most likely class given the observed features. This approach is known as *maximum a posteriori* (MAP) estimation, because the conditional distribution is the *posterior* distribution of the class once the features have been observed (this term is commonly used in Bayesian modeling, as explained in Section 6.7.1).

Definition 4.29 (MAP estimator)**.** *Given a discrete random variable $\tilde{y}$ with range Y and a discrete random vector $\tilde{x}$, the maximum a posteriori (MAP) estimator of $\tilde{y}$ given $\tilde{x} = x$ is*

$$\text{MAP}(x) := \arg\max_{y \in Y} p_{\tilde{y} \,|\, \tilde{x}}(y \,|\, x). \tag{4.144}$$

The MAP estimate is optimal in the sense that it minimizes the probability of error.

Theorem 4.30 (MAP estimation is optimal)**.** *Given a discrete random variable $\tilde{y}$ and a d-dimensional discrete random vector $\tilde{x}$, the maximum a posteriori (MAP) estimator of $\tilde{y}$ given $\tilde{x} = x$ minimizes the probability of error. Equivalently, the probability that the MAP estimator is correct is greater than or equal to the probability that any estimator $h : \mathbb{R}^d \to \mathbb{R}$ is correct:*

$$\text{P}\left(\text{MAP}(\tilde{x}) = \tilde{y}\right) \geq \text{P}\left(h(\tilde{x}) = \tilde{y}\right). \tag{4.145}$$

Proof To simplify notation, we denote the set of all possible values of the

random vector $\tilde{x}$ by $\mathcal{X}$. By definition of the MAP estimator, for any $x \in \mathcal{X}$, $p_{\tilde{y}\,|\,\tilde{x}}(\text{MAP}\,(x)\,|\,x) \geq p_{\tilde{y}\,|\,\tilde{x}}(h\,(x)\,|\,x)$. Therefore, by the law of total probability

$$\begin{aligned}
\text{P}\,(h(\tilde{x}) = \tilde{y}) &= \sum_{x \in \mathcal{X}} \text{P}\,(\tilde{x} = x)\,\text{P}\,(\tilde{y} = h\,(x)\,|\,\tilde{x} = x) && (4.146)\\
&= \sum_{x \in \mathcal{X}} p_{\tilde{x}}(x) p_{\tilde{y}\,|\,\tilde{x}}(h\,(x)\,|\,x) && (4.147)\\
&\leq \sum_{x \in \mathcal{X}} p_{\tilde{x}}(x) p_{\tilde{y}\,|\,\tilde{x}}(\text{MAP}\,(x)\,|\,x) && (4.148)\\
&= \sum_{x \in \mathcal{X}} \text{P}\,(\tilde{x} = x)\,\text{P}\,(\tilde{y} = \text{MAP}\,(x)\,|\,\tilde{x} = x) && (4.149)\\
&= \text{P}\,(\text{MAP}(\tilde{x}) = \tilde{y})\,. && (4.150)
\end{aligned}$$

■

Unfortunately, it is often intractable to compute the MAP estimator due the curse of dimensionality. Let us illustrate this using the voting data from Dataset 1. The data consist of votes on 16 different issues by representatives in the United States House of Representatives, who are affiliated with the Democratic or Republican party. Our goal is to guess the affiliation of a politician from their voting record.

We separate the dataset into a training set with 425 representatives and a test set with 10 representatives. We model the 16-dimensional voting record as a random vector $\tilde{x}$ with entries defined as follows,

$$\tilde{x}[i] := \begin{cases} 1 & \text{if the representative voted Yes on issue } i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.151}$$

We model the affiliation as the Bernoulli random variable:

$$\tilde{y} = \begin{cases} R & \text{if the representative is a Republican,} \\ D & \text{if the representative is a Democrat.} \end{cases} \tag{4.152}$$

In order to determine the affiliation of a representative from their voting record x, we leverage the conditional pmf of $\tilde{y}$ given $\tilde{x}$: if $p_{\tilde{y}\,|\,\tilde{x}}(R\,|\,x) > p_{\tilde{y}\,|\,\tilde{x}}(D\,|\,x)$, we classify them as a Republican; if not, we classify them as a Democrat. The problem is that it is impossible to estimate the full conditional pmf: There are $2^{16} = 65,536$ possible values of x, and we only have 425 data points! After fitting the model with these data, we are very likely to encounter politicians with voting records that have never been observed.

As mentioned in Section 4.7, we can address the curse of dimensionality by making independence assumptions that simplify our probabilistic model. Here we cannot assume independence of the data $\tilde{x}$ and the corresponding label $\tilde{y}$; this defeats the whole purpose of classification! Instead, we assume that each entry of $\tilde{x}$ is conditionally independent of the rest given $\tilde{y}$. This dramatically reduces the

number of parameters. Under the conditional-independence assumption,

$$p_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x) = \frac{p_{\tilde{x},\tilde{y}}(x,y)}{\sum_{w\in\{1,2,\ldots,c\}} p_{\tilde{x},\tilde{y}}(x,w)} \tag{4.153}$$

$$= \frac{p_{\tilde{y}}(y)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,y)}{\sum_{w=1}^{c} p_{\tilde{y}}(w)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,w)}. \tag{4.154}$$

This is an application of Bayes' rule under a *naive* assumption, since we do not expect the conditional-independence assumption to hold in reality. The classification method is therefore called *naive Bayes*.

Definition 4.31 (Naive Bayes classifier)**.** *Assume the availability of a dataset of n examples (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n), where x_i denotes a d-dimensional discrete vector and $y_i \in \{1, 2, \ldots, c\}$ its corresponding class label (c is the number of classes). To classify a new example x, we:*

1 *Use empirical probabilities to estimate the c conditional pmfs $p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,k)$ for $k \in \{1, 2, \ldots, c\}$ and $i \in \{1, 2, \ldots, d\}$. The random variables $\tilde{x}[i]$ and $\tilde{y}$ represent the ith entry of the data and the class, respectively.*

2 *Compute the empirical pmf of $\tilde{y}$ (see Definition 2.11).*

3 *Choose the MAP estimate of $\tilde{y}$ given $\tilde{x} = x$ under the assumption that all entries of $\tilde{x}$ are conditionally independent given $\tilde{y}$,*

$$\text{MAP}\,(x) := \arg\max_{y\in\{1,2,\ldots,c\}} p_{\tilde{y}\,|\,\tilde{x}}(y\,|\,x) \tag{4.155}$$

$$= \arg\max_{y\in\{1,2,\ldots,c\}} \frac{p_{\tilde{y}}(y)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,y)}{\sum_{k=1}^{c} p_{\tilde{y}}(k)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,k)}. \tag{4.156}$$

In the case of the voting example, we denote the two labels by R (Republican) or D (Democrat). In order to apply naive Bayes, we need to determine $p_{\tilde{y}}$, $p_{\tilde{x}[i]\,|\,\tilde{y}}(\cdot\,|\,R)$ for $1 \leq i \leq 16$, and $p_{\tilde{x}[i]\,|\,\tilde{y}}(\cdot\,|\,D)$ for $1 \leq i \leq 16$. This requires estimating 33 parameters (there is only one free parameter for each pmf, because they need to sum up to one), which is a dramatic reduction from the 65,535 parameters required to estimate the conditional pmf $\tilde{y}$ given $\tilde{x}$ without any assumptions. As a result, it is now tractable to estimate the required probabilities from data.

There are 263 Democrats and 162 Republicans, so $p_{\tilde{y}}(R) = 0.381$ and $p_{\tilde{y}}(D) = 0.619$. For each vote, we compute the fraction of Republicans who voted Yes to estimate $p_{\tilde{x}[i]\,|\,\tilde{y}}(1\,|\,R)$ and the fraction of Democrats who voted Yes to approximate $p_{\tilde{x}[i]\,|\,\tilde{y}}(1\,|\,D)$. Table 4.6 shows the resulting probabilities in the top rows. Finally, we classify each representative in the test set by computing the MAP estimate (4.155). The voting record of one of the representatives is shown in Table 4.6 (row marked *Example D*). The representative did not vote for issues 2 and 16, so we omit the corresponding entries of $\tilde{x}$ in our calculation. The naive-Bayes

Table 4.6 ***Classification of political affiliation via naive Bayes.*** *The table shows the empirical conditional probability of voting Yes on the 16 issues in the US House of Representatives dataset for Republican (R) and Democratic (D) representatives. We also show the voting record for two Democratic representatives from the test set. The first (row marked* Example D*) has a voting pattern which is consistent with other Democrats in a majority of the issues. To show this we highlight in bold the issues for which their vote is in agreement with Democrats (i.e. they voted Yes and* $p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, D) > 0.5$*, or they voted No and* $p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, D) < 0.5$*). In contrast, the representative who is wrongly classified has a voting pattern aligned with Republicans rather than with Democrats in a majority of the issues. To show this we highlight in bold the issues for which their vote is in agreement with Republicans (i.e. they voted Yes and* $p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, R) > 0.5$*, or they voted No and* $p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, R) < 0.5$*).*

i	1	2	3	4	5	6	7	8
$p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, R)$	0.19	0.50	0.14	0.99	0.95	0.90	0.24	0.15
$p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, D)$	0.61	0.50	0.89	0.05	0.22	0.47	0.78	0.83
Example D	N	–	**Y**	**N**	**N**	Y	**Y**	**Y**
Misclassified D	Y	Y	–	**Y**	**Y**	**Y**	**N**	**N**

i	9	10	11	12	13	14	15	16
$p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, R)$	0.11	0.55	0.14	0.87	0.86	0.98	0.09	0.66
$p_{\tilde{x}[i] \,|\, \tilde{y}}(1 \,|\, D)$	0.76	0.47	0.51	0.15	0.29	0.35	0.64	0.94
Example D	N	Y	N	**N**	**N**	**N**	**Y**	–
Misclassified D	Y	**Y**	–	**Y**	**Y**	**Y**	**N**	N

estimate of the probability that the candidate is a Democrat is

$$\begin{aligned} p_{\tilde{y} \,|\, \tilde{x}}(D \,|\, x) &= \frac{p_{\tilde{y}}(D) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \,|\, \tilde{y}}(x[i] \,|\, D)}{p_{\tilde{y}}(D) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \,|\, \tilde{y}}(x[i] \,|\, D) + p_{\tilde{y}}(R) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \,|\, \tilde{y}}(x[i] \,|\, R)} \\ &= 1 - 1.410^{-8}. \end{aligned} \tag{4.157}$$

We conclude correctly that the representative is a Democrat. Naive Bayes identifies the right class for 9 of the 10 examples in the test set. The only error corresponds to a Democratic representative whose voting record, is more aligned with the Republican representatives in the training set (see row marked *Misclassified D* in Table 4.6).

4.9 Markov Chains

Markov chains are one of the most popular models for data that have temporal structure. Here we focus on time-homogeneous Markov chains, which we introduce in Section 4.9.1. Section 4.9.2 shows that these Markov chains can be represented in terms of a state vector and a transition matrix, which governs the time evolution of the model. Section 4.9.3 explains how to analyze the asymptotic behavior of Markov chains and determine whether they converge to a stationary distribution.

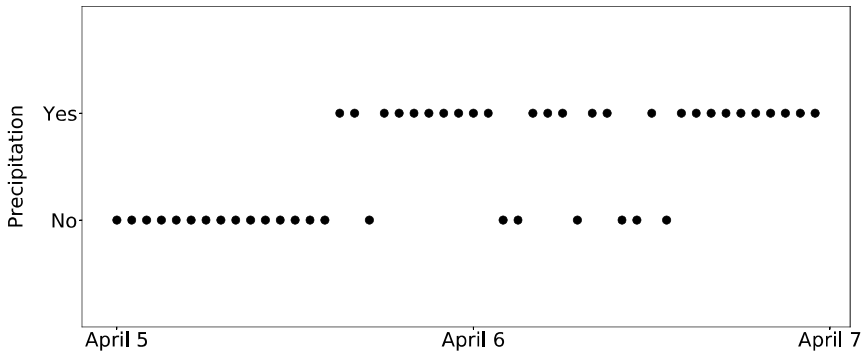

Figure 4.12 Precipitation in Coos Bay. Hourly precipitation in Coos Bay over two days from April 2015, extracted from Dataset 9. There is clear temporal dependence in the data: changes from no precipitation to precipitation, and vice versa, are relatively rare.

4.9.1 Time-Homogeneous Markov Chains

Imagine that we want to model a dataset of discrete measurements $x_1, x_2, \ldots, x_n$ representing the evolution in time of a quantity of interest. For example, x_i may indicate whether it rained or not at time i in a certain location. Up to now, we have analyzed such data by computing the corresponding pmf, either using empirical probabilities (see Section 2.2) or via maximum likelihood (see Section 2.4). Implicitly, this assumes that the data are i.i.d. samples from a single random variable. However, this analysis may neglect important structure in the data. Specifically, the relationship between adjacent data points is not captured due to the independence assumption.

Figure 4.12 shows the hourly precipitation in Coos Bay (Oregon) in 2015, extracted from Dataset 9. There is obvious temporal structure in the data: we observe long stretches with and without precipitation. In order to account for such structure, we can model the data as realizations from *a sequence of random variables* $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$, instead of just i.i.d. samples from a single random variable.

Modeling our data of interest as realizations from a sequence of random variables comes at a price: we have to deal with the curse of dimensionality, described in Section 4.7. For simplicity, let us assume that the data are a binary sequence of zeros and ones. In Figure 4.12 zero and one represent no precipitation and precipitation, respectively. If we interpret these data as realizations of a sequence of Bernoulli random variables $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$, then the joint pmf has 2^n entries, corresponding to each possible binary sequence of length n.* That is unfortunate, given that we only see a single binary sequence! We need to make some assumptions to reduce the complexity of our model. Assuming the data are i.i.d. as in

*Strictly speaking, the degrees of freedom in the joint pmf are $2^n - 1$, because the entries of the joint pmf must sum to one.

Definition 4.19 takes care of this problem, but ignores crucial temporal dependence. We need a model that is simple enough to be fit from our limited data, but also incorporates temporal dependence. Enter the Markov chain, which assumes that a sequence is completely characterized by one-step transitions.

Definition 4.32 (Markov chain)**.** *Let $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ be discrete random variables belonging to the same probability space. The random variables form a Markov chain if for any $1 < i < n$, $\tilde{a}_{i+1}$ is conditionally independent of $\tilde{a}_1, \ldots, \tilde{a}_{i-1}$ given $\tilde{a}_i$, i.e.*

$$p_{\tilde{a}_{i+1} \,|\, \tilde{a}_1, \ldots, \tilde{a}_i} \left(a_{i+1} \,|\, a_1, a_2, \ldots, a_i \right) = p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i} \left(a_{i+1} \,|\, a_i \right), \tag{4.158}$$

for any values of a_1, a_2, ..., a_n. Equivalently, by the chain rule, the joint pmf of $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ equals

$$p_{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n} \left(a_1, a_2, \ldots, a_n \right) = p_{\tilde{a}_1} \left(a_1 \right) \prod_{i=1}^{n-1} p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i} \left(a_{i+1} \,|\, a_i \right), \tag{4.159}$$

for any values of a_1, a_2, ..., a_n.

Assuming that our data are realizations from a Markov chain reduces the number of parameters enormously. In order to characterize the joint pmf in (4.159), we only need to estimate the one-step conditional probabilities $p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i}$, for $1 \leq i \leq n-1$ (as well as $p_{\tilde{a}_1}$). If the random variables are Bernoulli, this reduces the number of parameters of the joint pmf from $2^n - 1$ (remember that it must add to one), to $2n - 1$, since we need to estimate a conditional pmf $p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i}(\cdot \,|\, a_i)$ for $1 \leq i \leq n-1$ and $a_i \in \{0, 1\}$ (and also $p_{\tilde{a}_1}$). Unfortunately, this is still problematic: we only have a single data point (x_i) from which to estimate each conditional pmf $p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i}$! In order to make estimation tractable, a common assumption is that all of these conditional pmfs are the same. In that case, the Markov chain is said to be *time homogeneous*.

Definition 4.33 (Time-homogeneous Markov chain)**.** *A Markov chain $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ is time homogeneous if the conditional pmfs $p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i} \left(a_{i+1} \,|\, a_i \right)$ are all equal to the same function p_{cond} for all values of i between 1 and $n - 1$,*

$$p_{\tilde{a}_{i+1} \,|\, \tilde{a}_i} \left(a_{i+1} \,|\, a_i \right) = p_{\text{cond}} \left(a_{i+1} \,|\, a_i \right), \quad 1 \leq i \leq n-1, \tag{4.160}$$

for any values of a_1, a_2, ..., a_n. By the chain rule, the joint pmf of $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$ equals

$$p_{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n} \left(a_1, a_2, \ldots, a_n \right) = p_{\tilde{a}_1} \left(a_1 \right) \prod_{i=1}^{n-1} p_{\text{cond}} \left(a_{i+1} \,|\, a_i \right), \tag{4.161}$$

for any values of a_1, a_2, ..., a_n.

When fitting a time-homogeneous Markov chain using data, we only need to estimate the one-step transition conditional pmf p_{cond} and the initial pmf $p_{\tilde{a}_1}$ to completely characterize the joint pmf. If the random variables are Bernoulli, this requires estimating three parameters, corresponding to $p_{\tilde{a}_1}$, $p_{\text{cond}}(\cdot \,|\, 0)$ and

Table 4.7 ***Precipitation statistics at Coos Bay.*** *The tables show the statistics of hourly precipitation in Coos Bay, Oregon during 2015 represented as percentages. The table at the top left shows the fraction of hours with and without precipitation. The table at the top right shows the frequency of the different possible one-step transitions between precipitation and no precipitation. The bottom two tables shows the frequency of two-step transitions.*

Marginal probabilities

No	88.7
Yes	11.3

1-step conditional probabilities

Hour $h+1$ \ *Hour h*	**No**	**Yes**
No	96.0	31.2
Yes	4.0	68.8

2-step conditional probabilities

No precipitation at hour $h-1$

Hour $h+1$ \ *Hour h*	**No**	**Yes**
No	97.1	49.4
Yes	2.9	50.6

Precipitation at hour $h-1$

Hour $h+1$ \ *Hour h*	**No**	**Yes**
No	70.6	23.0
Yes	29.4	77.0

$p_{\text{cond}}(\cdot \,|\, 1)$. We have finally obtained a model that is tractable to fit using n data points. The following example applies it to the precipitation data from Figure 4.12.

Example 4.34 (Predicting hourly precipitation)**.** We consider the problem of predicting hourly precipitation in Coos Bay, Oregon. We use binary hourly precipitation measurements from 2015, extracted from Dataset 9, as training data (a small segment is shown in Figure 4.12). We begin by interpreting the data as i.i.d. and estimate the corresponding pmf using empirical probabilities. This is a fancy way of saying that we compute the fraction of measurements with precipitation. The estimated pmf is Bernoulli with parameter 0.113. Using this model, the best possible prediction is that it never rains. The resulting accuracy when we apply this naive prediction to test data corresponding to hourly measurements from 2016 is 83.4%, because in 2016 it rained more often than in 2015 (16.6% of the time).

In order to take into account the temporal structure of the data observed in Figure 4.12, we fit a two-state time-homogeneous Markov chain to the data. We estimate the one-step transition probabilities using the corresponding empirical conditional probabilities (see Definition 1.24). Table 4.7 shows the estimated one-step transition conditional probabilities (top right). According to the model, when there is precipitation at hour h, there is still precipitation at hour $h+1$ with probability 0.688. In contrast, when there is no precipitation at hour h, the probability of precipitation at hour $h+1$ plummets to 0.04. Based on this model, we

predict precipitation if there is precipitation the hour before, and no precipitation otherwise. This yields an accuracy of 87.3% for the 2016 test dataset.

A natural extension of the one-step Markov-chain model is to take into account additional previous measurements. The bottom of Table 4.7 shows two-step transition probabilities estimated using the corresponding empirical conditional probabilities. We immediately see that the data are not really realizations from a Markov chain, because they violate the conditional-independence assumption in Definition 4.32:

$$\mathrm{P}(\text{Yes at } h+1 \,|\, \text{Yes at } h) \tag{4.162}$$

$$= 0.688 \tag{4.163}$$

$$\neq 0.770 = \mathrm{P}(\text{Yes at } h+1 \,|\, \text{Yes at } h, \text{Yes at } h-1) \tag{4.164}$$

$$\neq 0.506 = \mathrm{P}(\text{Yes at } h+1 \,|\, \text{Yes at } h, \text{No at } h-1). \tag{4.165}$$

According to this two-step model, regardless of what happens at $h-1$, if there is precipitation at h, then precipitation is more likely at $h+1$, and if there is no precipitation at h, then no precipitation is more likely at $h+1$. Consequently, our binary prediction does not change from the prediction based on the one-step model (although the predicted probabilities do change). This example illustrates two important points. First, conditional-independence assumptions almost never hold exactly. Second, there is often diminishing returns when modeling temporal dependence. In this case, incorporating one-step transitions improves perfomance substantially, but modeling the two-step transitions does not make a difference in our binary prediction.

4.9.2 The State Vector And The Transition Matrix

In this section we explain how to define and manipulate time-homogeneous Markov chains which take values in the same finite set finite set $S := \{s_1, s_2, \ldots, s_m\}$ at every time step. We call the set of possible values the *state space* of the Markov chain. For each element $s \in S$, if $\tilde{a}_i = s$ we say that the Markov chain is in state s at time i. If we denote the Markov chain by $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$, we can represent the marginal pmf at each time step $\tilde{a}_i$ using an m-dimensional *state vector* π_i,

$$\pi_i := \begin{bmatrix} p_{\tilde{a}_i}(s_1) \\ p_{\tilde{a}_i}(s_2) \\ \cdots \\ p_{\tilde{a}_i}(s_m) \end{bmatrix}. \tag{4.166}$$

Notice that $\sum_{i=1}^{m} \pi_i = 1$ for all i because the entries of the vector form a complete pmf. In Example 4.34 there are only two states ($m = 2$): precipitation and no precipitation.

Finite-state time-homogeneous Markov chains are characterized by the transition probabilities between states, denoted by p_{cond} in Definition 4.33. There are

$m \times m$ such probabilities, which we store in the *transition matrix* of the Markov chain

$$T := \begin{bmatrix} p_{\text{cond}}(s_1 \,|\, s_1) & p_{\text{cond}}(s_1 \,|\, s_2) & \cdots & p_{\text{cond}}(s_1 \,|\, s_m) \\ p_{\text{cond}}(s_2 \,|\, s_1) & p_{\text{cond}}(s_2 \,|\, s_2) & \cdots & p_{\text{cond}}(s_2 \,|\, s_m) \\ \cdots & \cdots & \cdots & \cdots \\ p_{\text{cond}}(s_m \,|\, s_1) & p_{\text{cond}}(s_m \,|\, s_2) & \cdots & p_{\text{cond}}(s_m \,|\, s_m) \end{bmatrix}. \tag{4.167}$$

Each column of the transition matrix contains a full conditional pmf, so its entries must sum to one

$$\sum_{i=1}^{m} T_{ij} = 1, \quad 1 \leq j \leq m. \tag{4.168}$$

In Example 4.34, the transition matrix for the 1-step model equals

$$T := \begin{bmatrix} 0.960 & 0.312 \\ 0.040 & 0.688 \end{bmatrix}, \tag{4.169}$$

where the states s_1 and s_2 correspond to no precipitation and precipitation, respectively. The transition probabilities of finite-space time-homogeneous Markov chains can be visualized using a state diagram, which shows each state and the probability of every possible transition. Figure 4.13 shows the state diagram corresponding to the following example.

Example 4.35 (Car rental)**.** A car-rental company hires you to model the location of their cars. The company operates in Los Angeles, San Francisco and San Jose. Customers have the option to take a car and drop it off in different cities. Your goal is to compute how likely it is for a car to end up in each city. You decide to model the location of the car as a three-state time-homogeneous Markov chain $\tilde{a}_1$, $\tilde{a}_2$, ..., $\tilde{a}_n$. The random variable $\tilde{a}_i$ represents the location of the car after the ith customer i. The three states correspond to the three possible locations. The transition probabilities, obtained from past data, are

San Francisco	Los Angeles	San Jose	
0.6	0.1	0.3	San Francisco
0.2	0.8	0.3	Los Angeles
0.2	0.1	0.4	San Jose

To be clear, the probability that a customer moves the car from San Francisco to LA is 0.2, the probability that the car stays in San Francisco is 0.6, and so on.

The company allocates new cars evenly between the three cities. According to this information, the initial state vector and the transition matrix of the Markov

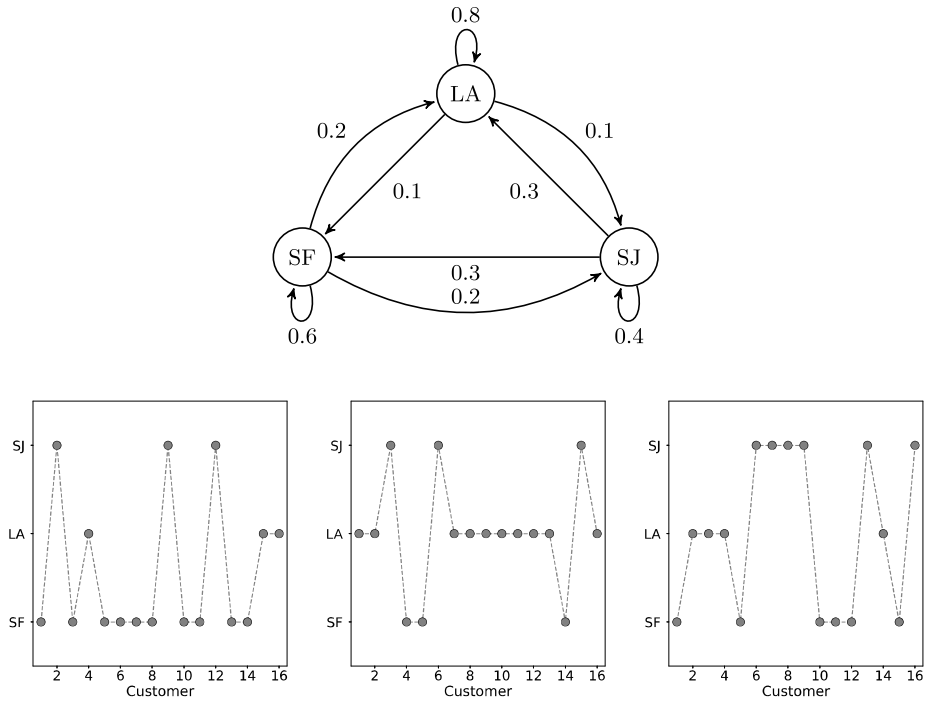

Figure 4.13 Finite-state time-homogeneous Markov chain modeling a rental car. The state diagram (top) describes the Markov chain in Example (4.35). Each arrow represents a transition between states, annotated by the corresponding transition probability. The graphs below depict three realizations of the Markov chain.

chain are

$$\pi_1 := \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \qquad T := \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.2 & 0.8 & 0.3 \\ 0.2 & 0.1 & 0.4 \end{bmatrix}. \tag{4.170}$$

State 1 is assigned to *San Francisco*, State 2 to *Los Angeles* and State 3 to *San Jose*. Figure 4.13 shows a state diagram of the Markov chain, as well as some realizations.

The initial state vector and the transition matrix completely characterize the joint pmf of the Markov chain. For example, we can compute the probability that the car starts in San Francisco and ends up in San Jose right after the second

customer:

$$p_{\tilde{a}_1,\tilde{a}_3}(1,3) = \sum_{i=1}^{3} p_{\tilde{a}_1,\tilde{a}_2,\tilde{a}_3}(1,i,3) \tag{4.171}$$

$$= \sum_{i=1}^{3} p_{\tilde{a}_1}(1)\, p_{\tilde{a}_2 \mid \tilde{a}_1}(i \mid 1)\, p_{\tilde{a}_3 \mid \tilde{a}_2}(3 \mid i) \tag{4.172}$$

$$= \pi_1[1] \sum_{i=1}^{3} T_{i1} T_{3i} \tag{4.173}$$

$$= \frac{0.6 \cdot 0.2 + 0.2 \cdot 0.1 + 0.2 \cdot 0.4}{3} \approx 7.33 \cdot 10^{-2}. \tag{4.174}$$

The following lemma shows that we can obtain the state vector at time i by multiplying the transition matrix and the state vector at time $i-1$.

Theorem 4.36 (State vector and transition matrix)**.** *For an n-dimensional finite-state time-homogeneous Markov chain with transition matrix T and initial state vector π_1,*

$$\pi_i = T\,\pi_{i-1} \tag{4.175}$$

$$= T^{i-1}\,\pi_1, \qquad 2 \le i \le n, \tag{4.176}$$

where T^{i-1} means that we apply T $i-1$ times.

Proof Let $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n$ denote the Markov chain. Equation (4.175) follows directly from the chain rule,

$$\pi_i := \begin{bmatrix} p_{\tilde{a}_i}(s_1) \\ p_{\tilde{a}_i}(s_2) \\ \cdots \\ p_{\tilde{a}_i}(s_m) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} p_{\tilde{a}_{i-1}}(s_j)\, p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_1 \mid s_j) \\ \sum_{j=1}^{m} p_{\tilde{a}_{i-1}}(s_j)\, p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_2 \mid s_j) \\ \cdots \\ \sum_{j=1}^{m} p_{\tilde{a}_{i-1}}(s_j)\, p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_m \mid s_j) \end{bmatrix} \tag{4.177}$$

$$= \begin{bmatrix} p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_1 \mid s_1) & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_1 \mid s_2) & \cdots & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_1 \mid s_m) \\ p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_2 \mid s_1) & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_2 \mid s_2) & \cdots & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_2 \mid s_m) \\ & & \cdots & \\ p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_m \mid s_1) & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_m \mid s_2) & \cdots & p_{\tilde{a}_i \mid \tilde{a}_{i-1}}(s_m \mid s_m) \end{bmatrix} \begin{bmatrix} p_{\tilde{a}_{i-1}}(s_1) \\ p_{\tilde{a}_{i-1}}(s_2) \\ \cdots \\ p_{\tilde{a}_{i-1}}(s_m) \end{bmatrix}$$

$$= T\pi_{i-1}. \tag{4.178}$$

Equation (4.176) is obtained by applying (4.175) $i-1$ times. ■

Example 4.37 (Car rental: 5th customer)**.** We want to estimate the probability that a new car ends in each of the three possible locations after the 5th customer.

Applying Theorem 4.36, we obtain

$$\pi_6 = T^5 \pi_1 \tag{4.179}$$

$$= \begin{bmatrix} 0.281 \\ 0.534 \\ 0.185 \end{bmatrix}, \tag{4.180}$$

so, for example, the probability that the car ends up in Los Angeles, is 0.534, according to the model.

4.9.3 Stationary Distribution Of A Markov Chain

In this section we study the evolution of the state vector of finite-state time-homogeneous Markov chains. The following example shows that, in order to analyze this evolution, it is very useful to leverage the eigendecomposition of the transition matrix.

Example 4.38 (Car rental: Asymptotic distribution)**.** We are interested in studying the probability that a car from the rental company in Example 4.35 eventually ends up in each of the different cities. This is given by the limit $\lim_{i\to\infty} \pi_i$, where π_i denotes the state vector at time i. By Theorem 4.36 the limit can be expressed in terms of the transition matrix and the initial state vector

$$\lim_{i\to\infty} \pi_i = \lim_{i\to\infty} T^{i-1} \pi_1. \tag{4.181}$$

In order to analyze this expression, we compute the eigendecomposition of the transition matrix T:

$$T = \underbrace{\begin{bmatrix} 0.27 & 0.37 & 0.37 \\ 0.55 & -0.50 & 0.13 \\ 0.18 & 0.13 & -0.50 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.57 & 0 \\ 0 & 0 & 0.23 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1.28 & -0.87 & 0.70 \\ 0.70 & 0.13 & -1.44 \end{bmatrix}}_{Q^{-1}}. \tag{4.182}$$

The columns of Q contain the three eigenvectors of T, and the diagonal entries of Λ contain the corresponding eigenvalues. The eigendecomposition enables us to express the $i-1$th power of the matrix T in terms of Q, Q^{-1} and the $i-1$th power of the diagonal matrix of eigenvalues Λ, which is equal to a diagonal matrix containing the $i-1$th power of each eigenvalue:

$$T^{i-1} = (Q\Lambda Q^{-1})^{i-1} = Q\Lambda Q^{-1} Q\Lambda Q^{-1} \cdots Q\Lambda Q^{-1} \tag{4.183}$$

$$= Q\Lambda^{i-1} Q^{-1}. \tag{4.184}$$

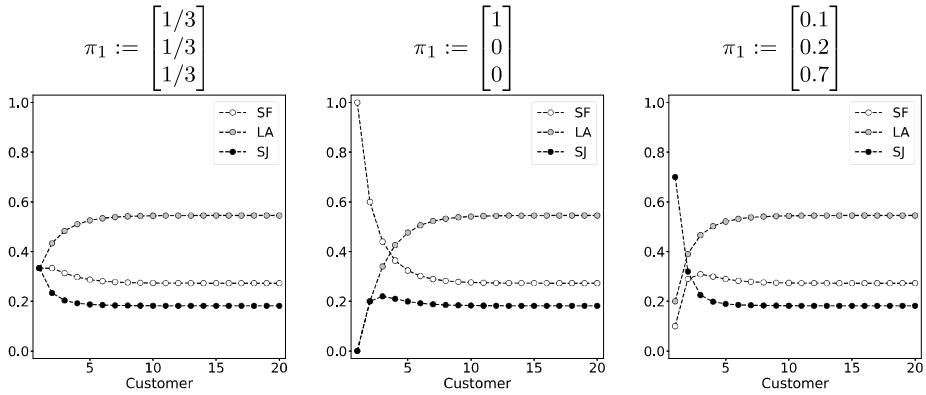

Figure 4.14 Evolution of the state vector of the Markov chain in Example (4.38)**.** The state vector quickly converges to the eigenvector with eigenvalue equal to one for the three different initializations.

Combining (4.181) and (4.184) we obtain

$$\lim_{i\to\infty} \pi_i = \lim_{i\to\infty} (Q\Lambda Q^{-1})^{i-1}\pi_1 \tag{4.185}$$

$$= \lim_{i\to\infty} Q \begin{bmatrix} 1^{i-1} & 0 & 0 \\ 0 & 0.57^{i-1} & 0 \\ 0 & 0 & 0.23^{i-1} \end{bmatrix} Q^{-1}\pi_1 \tag{4.186}$$

$$= Q \begin{bmatrix} 1 & 0 & 0 \\ 0 & \lim_{i\to\infty} 0.57^{i-1} & 0 \\ 0 & 0 & \lim_{i\to\infty} 0.23^{i-1} \end{bmatrix} Q^{-1}\pi_1 \tag{4.187}$$

$$= Q \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^{-1}\pi_1 = \begin{bmatrix} 0.27 \\ 0.55 \\ 0.18 \end{bmatrix} \sum_{j=1}^{3} \pi_1[j] = \begin{bmatrix} 0.27 \\ 0.55 \\ 0.18 \end{bmatrix}, \tag{4.188}$$

where we have used the fact that the entries of any initial state vector must add up to one. The state vector always converges to the same pmf for all possible initial vectors! Specifically, it converges to the eigenvector corresponding to the eigenvalue equal to one, because the terms that depend on the other eigenvalues are multiplied by 0.57^{i-1} and 0.23^{i-1}, and therefore decrease exponentially. The numerical simulations in Figure 4.14 show that convergence is indeed very fast. We conclude that no matter how the company allocates the new cars, eventually, 27.3% will end up in San Francisco, 54.5% in LA, and 18.2% in San Jose.

The following theorem captures the relationship between the dynamics of a finite-state time-homogeneous Markov chain and the eigendecomposition of its transition matrix.

Theorem 4.39. *We consider an m-state time-homogeneous Markov chain with*

transition matrix $T \in \mathbb{R}^{m \times m}$. *Assume that* T *has an eigendecomposition* $Q \Lambda Q^{-1}$, *where the columns of* Q *are the* m *eigenvectors* $q_1, \ldots, q_m$ *and* Λ *is a diagonal matrix containing the corresponding eigenvalues* $\lambda_1, \ldots, \lambda_m$. *We express the initial state vector* $\pi_1 \in \mathbb{R}^m$ *in terms of the eigenvectors:*

$$\pi_1 = \sum_{j=1}^{m} \alpha_j q_j, \tag{4.189}$$

where α_i *is the* i*th entry of* $\alpha := Q^{-1}\pi_1$. *For any* $i \geq 1$

$$\pi_i = \sum_{j=1}^{m} \lambda_j^{i-1} \alpha_j q_j. \tag{4.190}$$

Proof The proof follows the same argument we apply in Example 4.38. By Theorem 4.36,

$$\pi_i = T^{i-1}\pi_1 \tag{4.191}$$

$$= \left(Q \Lambda Q^{-1}\right)^{i-1} \pi_1 \tag{4.192}$$

$$= Q \Lambda^{i-1} Q^{-1} \pi_1. \tag{4.193}$$

■

An important consequence of Theorem 4.39 is that if a certain state vector π_* is an eigenvector of the transition matrix with eigenvalue equal to one, then this state vector is *stationary*. If the state vector of the Markov chain equals π_* at any point, then it will equal π_* forever afterwards.

Definition 4.40 (Stationary distribution)**.** *A marginal pmf or state vector* $\pi_* \in \mathbb{R}^m$ *is a stationary distribution of a finite-state time-homogeneous Markov chain with transition matrix* T *if* $T\pi_* = \pi_*$.

As illustrated by Example 4.38, if a finite-state time-homogeneous Markov chain has a single stationary distribution, and the remaining eigenvalues have magnitudes smaller than one, then the state vector of the Markov chain converges to the stationary distribution *for any initial state vector*. The principle is the same as in the celebrated power method to compute eigenvalues: as we apply the transition matrix over and over, the component of the state vector associated to the stationary distribution is preserved because its corresponding eigenvalue equals one, whereas the other components shrink exponentially because their eigenvalues have magnitudes smaller than one. The following example shows that this occurs for the real precipitation data in Example 4.34.

Example 4.41 (Stationary distribution of precipitation)**.** The eigendecomposition of the transition matrix in Example 4.34 is

$$T = \underbrace{\begin{bmatrix} 0.887 & -0.632 \\ 0.113 & -0.632 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 0.648 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 1 \\ -0.179 & 1.40 \end{bmatrix}}_{Q^{-1}}. \tag{4.194}$$

We denote the two eigenvalues by λ_1 and λ_2, and the two corresponding eigenvectors (columns of Q) by q_1 and q_2. For any initial state vector π_1, let $\alpha := Q^{-1}\pi_1$ denote the coefficients of π_1 in the basis of eigenvectors. By Theorem 4.39,

$$\lim_{i\to\infty} \pi_i = \lim_{i\to\infty} \sum_{j=1}^{m} \lambda_j^{i-1} \alpha_j q_j \tag{4.195}$$

$$= \begin{bmatrix} 0.887 \\ 0.113 \end{bmatrix}, \tag{4.196}$$

because $\lambda_1^{i-1} = 1$ and $\lim_{i\to\infty} \lambda_2^{i-1} = 0$. The stationary distribution to which the Markov chain converges should look familiar. It is exactly equal to the marginal pmf estimated using the empirical probabilities of each state from the data, reported in Table 4.7. This shows that the Markov chain indeed converges to its stationary distribution in this real-world dataset.

In some cases, the state vector of a finite-state time-homogeneous Markov chain may not converge to a stationary distribution, as described in the following example.

Example 4.42 (Periodic Markov chain)**.** We consider the three-state time-homogeneous Markov chain with the state diagram depicted in Figure 4.15. The transition matrix and corresponding eigendecomposition equal

$$T = \begin{bmatrix} 0 & 0.1 & 0 \\ 1 & 0 & 1 \\ 0 & 0.9 & 0 \end{bmatrix} \tag{4.197}$$

$$= \underbrace{\begin{bmatrix} 0.05 & 0.05 & 0.477 \\ 0.5 & -0.5 & 0 \\ 0.45 & 0.45 & -0.477 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1.89 & 0 & -0.21 \end{bmatrix}}_{Q^{-1}}. \tag{4.198}$$

The Markov chain has a stationary distribution corresponding to the eigenvector with eigenvalue equal to one,

$$\pi_* = \begin{bmatrix} 0.05 \\ 0.5 \\ 0.45 \end{bmatrix}. \tag{4.199}$$

If the Markov chain is initialized at this state vector, then the state vector will equal π_* indefinitely. However, let us assume an initial state vector where $\pi[2] = 0$,

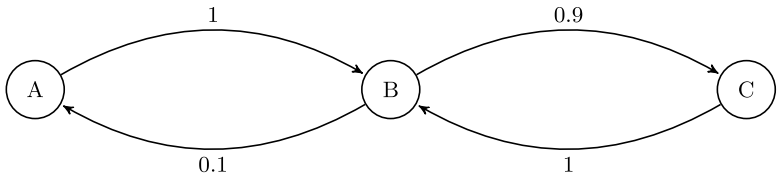

Figure 4.15 Periodic Markov chain. State diagram of a three-state time-homogeneous Markov chain with periodic dynamics. It alternates between states A or C, and state B.

which means that the initial state is either A or C. In that case, by Theorem 4.39

$$\pi_i = Q\Lambda^{i-1}Q^{-1}\pi_1 \tag{4.200}$$

$$= Q \begin{bmatrix} 1^{i-1} & 0 & 0 \\ 0 & (-1)^{i-1} & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^{-1}\pi_1 \tag{4.201}$$

$$= \left(\begin{bmatrix} 0.05 \\ 0.5 \\ 0.45 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} + (-1)^{i-1} \begin{bmatrix} 0.05 \\ -0.5 \\ 0.45 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} \right) \pi_1 \tag{4.202}$$

$$= \begin{bmatrix} 0.05 \\ 0.5 \\ 0.45 \end{bmatrix} \sum_{j=1}^{3} \pi_1[j] + (-1)^{i-1} \begin{bmatrix} 0.05 \\ -0.5 \\ 0.45 \end{bmatrix} (\pi_1[1] - \pi[2] + \pi_1[3]) \tag{4.203}$$

$$= \begin{cases} \begin{bmatrix} 0.1 \\ 0 \\ 0.9 \end{bmatrix} & \text{if } i \text{ is odd,} \\ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} & \text{if } i \text{ is even.} \end{cases} \tag{4.204}$$

The state vector oscillates between two possible values. This makes sense given the state diagram, if we begin at A or C at $i = 1$, then the Markov chain is always in state B for even i, returning to either A or C with probability 0.1 and 0.9 respectively for odd i.