2 *Foundations*

In this chapter, we review some important background material regarding key concepts from probability theory, information theory, and graph theory. This material is included in a separate introductory chapter, since it forms the basis for much of the development in the remainder of the book. Other background material — such as discrete and continuous optimization, algorithmic complexity analysis, and basic algorithmic concepts — is more localized to particular topics in the book. Many of these concepts are presented in the appendix; others are presented in concept boxes in the appropriate places in the text. All of this material is intended to focus only on the minimal subset of ideas required to understand most of the discussion in the remainder of the book, rather than to provide a comprehensive overview of the field it surveys. We encourage the reader to explore additional sources for more details about these areas.

2.1 Probability Theory

The main focus of this book is on complex probability distributions. In this section we briefly review basic concepts from probability theory.

2.1.1 Probability Distributions

When we use the word "probability" in day-to-day life, we refer to a degree of confidence that an event of an uncertain nature will occur. For example, the weather report might say "there is a low probability of light rain in the afternoon." Probability theory deals with the formal foundations for discussing such estimates and the rules they should obey.

Before we discuss the representation of probability, we need to define what the events are to which we want to assign a probability. These events might be different outcomes of throwing a die, the outcome of a horse race, the weather configurations in California, or the possible failures of a piece of machinery.

2.1.1.1 Event Spaces

event
outcome space

Formally, we define *events* by assuming that there is an agreed upon *space* of possible outcomes, which we denote by Ω. For example, if we consider dice, we might set $\Omega = \{1, 2, 3, 4, 5, 6\}$. In the case of a horse race, the space might be all possible orders of arrivals at the finish line, a much larger space.

measurable event

In addition, we assume that there is a set of *measurable events* $\mathcal{S}$ to which we are willing to assign probabilities. Formally, each event $\alpha \in \mathcal{S}$ is a subset of Ω. In our die example, the event $\{6\}$ represents the case where the die shows 6, and the event $\{1, 3, 5\}$ represents the case of an odd outcome. In the horse-race example, we might consider the event "Lucky Strike wins," which contains all the outcomes in which the horse Lucky Strike is first.

Probability theory requires that the event space satisfy three basic properties:

- It contains the *empty event* $\emptyset$, and the *trivial event* Ω.
- It is closed under union. That is, if $\alpha, \beta \in \mathcal{S}$, then so is $\alpha \cup \beta$.
- It is closed under complementation. That is, if $\alpha \in \mathcal{S}$, then so is $\Omega - \alpha$.

The requirement that the event space is closed under union and complementation implies that it is also closed under other Boolean operations, such as intersection and set difference.

2.1.1.2 Probability Distributions

Definition 2.1 probability distribution

A probability distribution P *over* $(\Omega, \mathcal{S})$ *is a mapping from events in* $\mathcal{S}$ *to real values that satisfies the following conditions:*

- $P(\alpha) \geq 0$ *for all* $\alpha \in \mathcal{S}$.
- $P(\Omega) = 1$.
- *If* $\alpha, \beta \in \mathcal{S}$ *and* $\alpha \cap \beta = \emptyset$, *then* $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$. ■

The first condition states that probabilities are not negative. The second states that the "trivial event," which allows all possible outcomes, has the maximal possible probability of 1. The third condition states that the probability that one of two mutually disjoint events will occur is the sum of the probabilities of each event. These two conditions imply many other conditions. Of particular interest are $P(\emptyset) = 0$, and $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$.

2.1.1.3 Interpretations of Probability

Before we continue to discuss probability distributions, we need to consider the interpretations that we might assign to them. Intuitively, the probability $P(\alpha)$ of an event α quantifies the degree of confidence that α will occur. If $P(\alpha) = 1$, we are certain that one of the outcomes in α occurs, and if $P(\alpha) = 0$, we consider all of them impossible. Other probability values represent options that lie between these two extremes.

This description, however, does not provide an answer to what the numbers mean. There are two common interpretations for probabilities.

frequentist interpretation

The *frequentist* interpretation views probabilities as frequencies of events. More precisely, the probability of an event is the fraction of times the event occurs if we repeat the experiment indefinitely. For example, suppose we consider the outcome of a particular die roll. In this case, the statement $P(\alpha) = 0.3$, for $\alpha = \{1, 3, 5\}$, states that if we repeatedly roll this die and record the outcome, then the fraction of times the outcomes in α will occur is 0.3. More precisely, the limit of the sequence of fractions of outcomes in α in the first roll, the first two rolls, the first three rolls, ..., the first n rolls, ... is 0.3.

The frequentist interpretation gives probabilities a tangible semantics. When we discuss concrete physical systems (for example, dice, coin flips, and card games) we can envision how these frequencies are defined. It is also relatively straightforward to check that frequencies must satisfy the requirements of proper distributions.

The frequentist interpretation fails, however, when we consider events such as "It will rain tomorrow afternoon." Although the time span of "Tomorrow afternoon" is somewhat ill defined, we expect it to occur exactly once. It is not clear how we define the frequencies of such events. Several attempts have been made to define the probability for such an event by finding a *reference class* of similar events for which frequencies are well defined; however, none of them has proved entirely satisfactory. Thus, the frequentist approach does not provide a satisfactory interpretation for a statement such as "the probability of rain tomorrow afternoon is 0.3."

reference class

An alternative interpretation views probabilities as *subjective* degrees of belief. Under this interpretation, the statement $P(\alpha) = 0.3$ represents a subjective statement about one's own degree of belief that the event α will come about. Thus, the statement "the probability of rain tomorrow afternoon is 50 percent" tells us that in the opinion of the speaker, the chances of rain and no rain tomorrow afternoon are the same. Although tomorrow afternoon will occur only once, we can still have uncertainty about its outcome, and represent it using numbers (that is, probabilities).

subjective interpretation

This description still does not resolve what exactly it means to hold a particular degree of belief. What stops a person from stating that the probability that Bush will win the election is 0.6 and the probability that he will lose is 0.8? The source of the problem is that we need to explain how subjective degrees of beliefs (something that is internal to each one of us) are reflected in our actions.

This issue is a major concern in subjective probabilities. One possible way of attributing degrees of beliefs is by a betting game. Suppose you believe that $P(\alpha) = 0.8$. Then you would be willing to place a bet of \$1 against \$3. To see this, note that with probability 0.8 you gain a dollar, and with probability 0.2 you lose \$3, so on average this bet is a good deal with expected gain of 20 cents. In fact, you might be even tempted to place a bet of \$1 against \$4. Under this bet the average gain is 0, so you should not mind. However, you would not consider it worthwhile to place a bet \$1 against \$4 and 10 cents, since that would have negative expected gain. Thus, by finding which bets you are willing to place, we can assess your degrees of beliefs.

The key point of this mental game is the following. If you hold degrees of belief that do not satisfy the rule of probability, then by a clever construction we can find a series of bets that would result in a sure negative outcome for you. Thus, the argument goes, a rational person must hold degrees of belief that satisfy the rules of probability.[1]

In the remainder of the book we discuss probabilities, but we usually do not explicitly state their interpretation. Since both interpretations lead to the same mathematical rules, the technical definitions hold for both interpretations.

1. As stated, this argument assumes as that people's preferences are directly proportional to their expected earnings. For small amounts of money, this assumption is quite reasonable. We return to this topic in chapter 22.

2.1.2 Basic Concepts in Probability

2.1.2.1 Conditional Probability

To use a concrete example, suppose we consider a distribution over a population of students taking a certain course. The space of outcomes is simply the set of all students in the population. Now, suppose that we want to reason about the students' intelligence and their final grade. We can define the event α to denote "all students with grade A," and the event β to denote "all students with high intelligence." Using our distribution, we can consider the probability of these events, as well as the probability of $\alpha \cap \beta$ (the set of intelligent students who got grade A). This, however, does not directly tell us how to update our beliefs given new evidence. Suppose we learn that a student has received the grade A; what does that tell us about her intelligence?

This kind of question arises every time we want to use distributions to reason about the real world. More precisely, after learning that an event α is true, how do we change our probability about β occurring? The answer is via the notion of *conditional probability*. Formally, the conditional probability of β given α is defined as

conditional probability

$$P(\beta \mid \alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)} \tag{2.1}$$

That is, the probability that β is true given that we know α is the relative proportion of outcomes satisfying β among these that satisfy α. (Note that the conditional probability is not defined when $P(\alpha) = 0$.)

The conditional probability given an event (say α) satisfies the properties of definition 2.1 (see exercise 2.4), and thus it is a probability distribution by its own right. Hence, we can think of the conditioning operation as taking one distribution and returning another over the same probability space.

2.1.2.2 Chain Rule and Bayes Rule

From the definition of the conditional distribution, we immediately see that

$$P(\alpha \cap \beta) = P(\alpha)P(\beta \mid \alpha). \tag{2.2}$$

chain rule

This equality is known as the *chain rule* of conditional probabilities. More generally, if $\alpha_1, \ldots, \alpha_k$ are events, then we can write

$$P(\alpha_1 \cap \ldots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 \mid \alpha_1) \cdots P(\alpha_k \mid \alpha_1 \cap \ldots \cap \alpha_{k-1}). \tag{2.3}$$

In other words, we can express the probability of a combination of several events in terms of the probability of the first, the probability of the second given the first, and so on. It is important to notice that we can expand this expression using any order of events — the result will remain the same.

Bayes' rule

Another immediate consequence of the definition of conditional probability is *Bayes' rule*

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}. \tag{2.4}$$

A more general conditional version of Bayes' rule, where all our probabilities are conditioned on some background event γ, also holds:

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}.$$

Bayes' rule is important in that it allows us to compute the conditional probability $P(\alpha \mid \beta)$ from the "inverse" conditional probability $P(\beta \mid \alpha)$.

Example 2.1

Consider the student population, and let Smart denote smart students and GradeA denote students who got grade A. Assume we believe (perhaps based on estimates from past statistics) that $P(\textit{GradeA} \mid \textit{Smart}) = 0.6$, *and now we learn that a particular student received grade A. Can we estimate the probability that the student is smart? According to Bayes' rule, this depends on our* prior *probability for students being smart (before we learn anything about them) and the prior probability of students receiving high grades. For example, suppose that* $P(\textit{Smart}) = 0.3$ *and* $P(\textit{GradeA}) = 0.2$, *then we have that* $P(\textit{Smart} \mid \textit{GradeA}) = 0.6 * 0.3/0.2 = 0.9$. *That is, an A grade strongly suggests that the student is smart. On the other hand, if the test was easier and high grades were more common, say,* $P(\textit{GradeA}) = 0.4$ *then we would get that* $P(\textit{Smart} \mid \textit{GradeA}) = 0.6 * 0.3/0.4 = 0.45$, *which is much less conclusive about the student.* ■

prior

Another classic example that shows the importance of this reasoning is in disease screening. To see this, consider the following hypothetical example (none of the mentioned figures are related to real statistics).

Example 2.2

Suppose that a tuberculosis (TB) skin test is 95 percent accurate. That is, if the patient is TB-infected, then the test will be positive with probability 0.95, *and if the patient is not infected, then the test will be negative with probability* 0.95. *Now suppose that a person gets a positive test result. What is the probability that he is infected? Naive reasoning suggests that if the test result is wrong 5 percent of the time, then the probability that the subject is infected is* 0.95. *That is, 95 percent of subjects with positive results have TB.*

If we consider the problem by applying Bayes' rule, we see that we need to consider the prior probability of TB infection, and the probability of getting positive test result. Suppose that 1 in 1000 of the subjects who get tested is infected. That is, $P(\textit{TB}) = 0.001$. *What is the probability of getting a positive test result? From our description, we see that* $0.001 \cdot 0.95$ *infected subjects get a positive result, and* $0.999 \cdot 0.05$ *uninfected subjects get a positive result. Thus,* $P(\textit{Positive}) = 0.0509$. *Applying Bayes' rule, we get that* $P(\textit{TB} \mid \textit{Positive}) = 0.001 \cdot 0.95/0.0509 \approx 0.0187$. *Thus, although a subject with a positive test is much more probable to be TB-infected than is a random subject, fewer than 2 percent of these subjects are TB-infected.* ■

2.1.3 Random Variables and Joint Distributions

2.1.3.1 Motivation

Our discussion of probability distributions deals with events. Formally, we can consider any event from the set of measurable events. The description of events is in terms of sets of outcomes. In many cases, however, it would be more natural to consider *attributes* of the outcome. For example, if we consider a patient, we might consider attributes such as "age,"

"gender," and "smoking history" that are relevant for assigning probability over possible diseases and symptoms. We would like then consider events such as "age > 55, heavy smoking history, and suffers from repeated cough."

To use a concrete example, consider again a distribution over a population of students in a course. Suppose that we want to reason about the intelligence of students, their final grades, and so forth. We can use an event such as *GradeA* to denote the subset of students that received the grade A and use it in our formulation. However, this discussion becomes rather cumbersome if we also want to consider students with grade B, students with grade C, and so on. Instead, we would like to consider a way of directly referring to a student's grade in a clean, mathematical way.

random variable

The formal machinery for discussing attributes and their values in different outcomes are *random variables*. A random variable is a way of reporting an attribute of the outcome. For example, suppose we have a random variable *Grade* that reports the final grade of a student, then the statement $P(\mathit{Grade} = A)$ is another notation for $P(\mathit{GradeA})$.

2.1.3.2 What Is a Random Variable?

Formally, a random variable, such as *Grade*, is defined by a function that associates with each outcome in Ω a value. For example, *Grade* is defined by a function f_{Grade} that maps each person in Ω to his or her grade (say, one of A, B, or C). The event $\mathit{Grade} = A$ is a shorthand for the event $\{\omega \in \Omega : f_{\mathit{Grade}}(\omega) = A\}$. In our example, we might also have a random variable *Intelligence* that (for simplicity) takes as values either "high" or "low." In this case, the event "*Intelligence* = *high*" refers, as can be expected, to the set of smart (high intelligence) students.

Random variables can take different sets of values. We can think of *categorical* (or *discrete*) random variables that take one of a few values, as in our two preceding examples. We can also talk about random variables that can take infinitely many values (for example, integer or real values), such as *Height* that denotes a student's height. We use $\mathit{Val}(X)$ to denote the set of values that a random variable X can take.

In most of the discussion in this book we examine either categorical random variables or random variables that take real values. We will usually use uppercase roman letters X, Y, Z to denote random variables. In discussing generic random variables, we often use a lowercase letter to refer to a value of a random variable. Thus, we use x to refer to a generic value of X. For example, in statements such as "$P(X = x) \geq 0$ for all $x \in \mathit{Val}(X)$." When we discuss categorical random variables, we use the notation $x^1, \ldots, x^k$, for $k = |\mathit{Val}(X)|$ (the number of elements in $\mathit{Val}(X)$), when we need to enumerate the specific values of X, for example, in statements such as

$$\sum_{i=1}^{k} P(X = x^i) = 1.$$

multinomial distribution

The distribution over such a variable is called a *multinomial*. In the case of a binary-valued random variable X, where $\mathit{Val}(X) = \{\mathit{false}, \mathit{true}\}$, we often use x^1 to denote the value *true* for X, and x^0 to denote the value *false*. The distribution of such a random variable is called a *Bernoulli distribution*.

Bernoulli distribution

We also use boldface type to denote sets of random variables. Thus, $\boldsymbol{X}$, $\boldsymbol{Y}$, or $\boldsymbol{Z}$ are typically used to denote a set of random variables, while $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ denote assignments of values to the

variables in these sets. We extend the definition of $Val(\boldsymbol{X})$ to refer to sets of variables in the obvious way. Thus, $\boldsymbol{x}$ is always a member of $Val(\boldsymbol{X})$. For $\boldsymbol{Y} \subseteq \boldsymbol{X}$, we use $\boldsymbol{x}\langle\boldsymbol{Y}\rangle$ to refer to the assignment within $\boldsymbol{x}$ to the variables in $\boldsymbol{Y}$. For two assignments $\boldsymbol{x}$ (to $\boldsymbol{X}$) and $\boldsymbol{y}$ (to $\boldsymbol{Y}$), we say that $\boldsymbol{x} \sim \boldsymbol{y}$ if they agree on the variables in their intersection, that is, $\boldsymbol{x}\langle\boldsymbol{X} \cap \boldsymbol{Y}\rangle = \boldsymbol{y}\langle\boldsymbol{X} \cap \boldsymbol{Y}\rangle$.

In many cases, the notation $P(X = x)$ is redundant, since the fact that x is a value of X is already reported by our choice of letter. Thus, in many texts on probability, the identity of a random variable is not explicitly mentioned, but can be inferred through the notation used for its value. Thus, we use $P(x)$ as a shorthand for $P(X = x)$ when the identity of the random variable is clear from the context. Another shorthand notation is that $\sum_x$ refers to a sum over all possible values that X can take. Thus, the preceding statement will often appear as $\sum_x P(x) = 1$. Finally, another standard notation has to do with conjunction. Rather than write $P((X = x) \cap (Y = y))$, we write $P(X = x, Y = y)$, or just $P(x, y)$.

2.1.3.3 Marginal and Joint Distributions

marginal distribution

Once we define a random variable X, we can consider the distribution over events that can be described using X. This distribution is often referred to as the *marginal distribution* over the random variable X. We denote this distribution by $P(X)$.

Returning to our population example, consider the random variable *Intelligence*. The marginal distribution over *Intelligence* assigns probability to specific events such as $P(\textit{Intelligence} = \textit{high})$ and $P(\textit{Intelligence} = \textit{low})$, as well as to the trivial event $P(\textit{Intelligence} \in \{\textit{high}, \textit{low}\})$. Note that these probabilities are defined by the probability distribution over the original space. For concreteness, suppose that $P(\textit{Intelligence} = \textit{high}) = 0.3$, $P(\textit{Intelligence} = \textit{low}) = 0.7$.

If we consider the random variable *Grade*, we can also define a marginal distribution. This is a distribution over all events that can be described in terms of the *Grade* variable. In our example, we have that $P(\textit{Grade} = A) = 0.25$, $P(\textit{Grade} = B) = 0.37$, and $P(\textit{Grade} = C) = 0.38$.

It should be fairly obvious that the marginal distribution is a probability distribution satisfying the properties of definition 2.1. In fact, the only change is that we restrict our attention to the subsets of $\mathcal{S}$ that can be described with the random variable X.

joint distribution

In many situations, we are interested in questions that involve the values of several random variables. For example, we might be interested in the event "*Intelligence* = *high* and *Grade* = *A*." To discuss such events, we need to consider the *joint distribution* over these two random variables. In general, the joint distribution over a set $\mathcal{X} = \{X_1, \ldots, X_n\}$ of random variables is denoted by $P(X_1, \ldots, X_n)$ and is the distribution that assigns probabilities to events that are specified in terms of these random variables. We use ξ to refer to a full assignment to the variables in $\mathcal{X}$, that is, $\xi \in Val(\mathcal{X})$.

The joint distribution of two random variables has to be consistent with the marginal distribution, in that $P(x) = \sum_y P(x, y)$. This relationship is shown in figure 2.1, where we compute the marginal distribution over *Grade* by summing the probabilities along each row. Similarly, we find the marginal distribution over *Intelligence* by summing out along each column. The resulting sums are typically written in the row or column margins, whence the term "marginal distribution."

Suppose we have a joint distribution over the variables $\mathcal{X} = \{X_1, \ldots, X_n\}$. The most fine-grained events we can discuss using these variables are ones of the form "$X_1 = x_1$ and $X_2 = x_2$, ..., and $X_n = x_n$" for a choice of values $x_1, \ldots, x_n$ for all the variables. Moreover,

		Intelligence		
		low	*high*	
Grade	*A*	0.07	0.18	0.25
	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

Figure 2.1 Example of a joint distribution $P(\textbf{\textit{Intelligence}}, \textbf{\textit{Grade}})$: Values of *Intelligence* (columns) and *Grade* (rows) with the associated marginal distribution on each variable.

any two such events must be either identical or disjoint, since they both assign values to all the variables in $\mathcal{X}$. In addition, any event defined using variables in $\mathcal{X}$ must be a union of a set of such events. Thus, we are effectively working in a *canonical outcome space*: a space where each outcome corresponds to a joint assignment to $X_1, \ldots, X_n$. More precisely, all our probability computations remain the same whether we consider the original outcome space (for example, all students), or the canonical space (for example, all combinations of intelligence and grade). We use ξ to denote these *atomic outcomes*: those assigning a value to each variable in $\mathcal{X}$. For example, if we let $\mathcal{X} = \{\textit{Intelligence}, \textit{Grade}\}$, there are six atomic outcomes, shown in figure 2.1. The figure also shows one possible joint distribution over these six outcomes.

canonical outcome space

atomic outcome

Based on this discussion, from now on we will not explicitly specify the set of outcomes and measurable events, and instead implicitly assume the canonical outcome space.

2.1.3.4 Conditional Probability

The notion of conditional probability extends to induced distributions over random variables. For example, we use the notation $P(\textit{Intelligence} \mid \textit{Grade} = a)$ to denote the *conditional distribution* over the events describable by *Intelligence* given the knowledge that the student's grade is A. Note that the conditional distribution over a random variable given an observation of the value of another one is not the same as the marginal distribution. In our example, $P(\textit{Intelligence} = \textit{high}) = 0.3$, and $P(\textit{Intelligence} = \textit{high} \mid \textit{Grade} = a) = 0.18/0.25 = 0.72$. Thus, clearly $P(\textit{Intelligence} \mid \textit{Grade} = a)$ is different from the marginal distribution $P(\textit{Intelligence})$. The latter distribution represents our *prior* knowledge about students before learning anything else about a particular student, while the conditional distribution represents our more informed distribution after learning her grade.

conditional distribution

We will often use the notation $P(X \mid Y)$ to represent a set of conditional probability distributions. Intuitively, for each value of Y, this object assigns a probability over values of X using the conditional probability. This notation allows us to write the shorthand version of the chain rule: $P(X, Y) = P(X)P(Y \mid X)$, which can be extended to multiple variables as

$$P(X_1, \ldots, X_k) = P(X_1)P(X_2 \mid X_1) \cdots P(X_k \mid X_1, \ldots, X_{k-1}). \tag{2.5}$$

Similarly, we can state Bayes' rule in terms of conditional probability distributions:

$$P(X \mid Y) = \frac{P(X)P(Y \mid X)}{P(Y)}. \tag{2.6}$$

2.1.4 Independence and Conditional Independence

2.1.4.1 Independence

As we mentioned, we usually expect $P(\alpha \mid \beta)$ to be different from $P(\alpha)$. That is, learning that β is true changes our probability over α. However, in some situations equality can occur, so that $P(\alpha \mid \beta) = P(\alpha)$. That is, learning that β occurs did not change our probability of α.

Definition 2.2 independent events

We say that an event α is independent *of event β in P, denoted $P \models (\alpha \perp \beta)$, if $P(\alpha \mid \beta) = P(\alpha)$ or if $P(\beta) = 0$.* ■

We can also provide an alternative definition for the concept of independence:

Proposition 2.1

A distribution P satisfies $(\alpha \perp \beta)$ if and only if $P(\alpha \cap \beta) = P(\alpha)P(\beta)$.

PROOF Consider first the case where $P(\beta) = 0$; here, we also have $P(\alpha \cap \beta) = 0$, and so the equivalence immediately holds. When $P(\beta) \neq 0$, we can use the chain rule; we write $P(\alpha \cap \beta) = P(\alpha \mid \beta)P(\beta)$. Since α is independent of β, we have that $P(\alpha \mid \beta) = P(\alpha)$. Thus, $P(\alpha \cap \beta) = P(\alpha)P(\beta)$. Conversely, suppose that $P(\alpha \cap \beta) = P(\alpha)P(\beta)$. Then, by definition, we have that

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)} = \frac{P(\alpha)P(\beta)}{P(\beta)} = P(\alpha).$$

■

As an immediate consequence of this alternative definition, we see that independence is a symmetric notion. That is, $(\alpha \perp \beta)$ implies $(\beta \perp \alpha)$.

Example 2.3

For example, suppose that we toss two coins, and let α be the event "the first toss results in a head" and β the event "the second toss results in a head." It is not hard to convince ourselves that we expect that these two events to be independent. Learning that β is true would not change our probability of α. In this case, we see two different physical processes (that is, coin tosses) leading to the events, which makes it intuitive that the probabilities of the two are independent. In certain cases, the same process can lead to independent events. For example, consider the event α denoting "the die outcome is even" and the event β denoting "the die outcome is 1 or 2." It is easy to check that if the die is fair (each of the six possible outcomes has probability $\frac{1}{6}$), then these two events are independent. ■

2.1.4.2 Conditional Independence

☞ **While independence is a useful property, it is not often that we encounter two independent events. A more common situation is when two events are independent given an additional event.** For example, suppose we want to reason about the chance that our student is accepted to graduate studies at Stanford or MIT. Denote by *Stanford* the event "admitted to Stanford" and by *MIT* the event "admitted to MIT." In most reasonable distributions, these two events are not independent. If we learn that a student was admitted to Stanford, then our estimate of her probability of being accepted at MIT is now higher, since it is a sign that she is a promising student.

Now, suppose that both universities base their decisions only on the student's grade point average (GPA), and we know that our student has a GPA of A. In this case, we might argue that learning that the student was admitted to Stanford should not change the probability that she will be admitted to MIT: Her GPA already tells us the information relevant to her chances of admission to MIT, and finding out about her admission to Stanford does not change that. Formally, the statement is

$$P(\mathit{MIT} \mid \mathit{Stanford}, \mathit{GradeA}) = P(\mathit{MIT} \mid \mathit{GradeA}).$$

In this case, we say that *MIT* is *conditionally independent* of *Stanford given GradeA*.

Definition 2.3 conditional independence

We say that an event α is conditionally independent *of event β given event γ in P, denoted $P \models (\alpha \perp \beta \mid \gamma)$, if $P(\alpha \mid \beta \cap \gamma) = P(\alpha \mid \gamma)$ or if $P(\beta \cap \gamma) = 0$.* ■

It is easy to extend the arguments we have seen in the case of (unconditional) independencies to give an alternative definition.

Proposition 2.2

P satisfies $(\alpha \perp \beta \mid \gamma)$ if and only if $P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma)P(\beta \mid \gamma)$.

2.1.4.3 Independence of Random Variables

Until now, we have focused on independence between events. Thus, we can say that two events, such as one toss landing heads and a second also landing heads, are independent. However, we would like to say that any pair of outcomes of the coin tosses is independent. To capture such statements, we can examine the generalization of independence to sets of random variables.

Definition 2.4 conditional independence; observed variable; marginal independence

Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be sets of random variables. We say that $\boldsymbol{X}$ is conditionally independent *of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in a distribution P if P satisfies $(\boldsymbol{X} = \boldsymbol{x} \perp \boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{Z} = \boldsymbol{z})$ for all values $\boldsymbol{x} \in Val(\boldsymbol{X})$, $\boldsymbol{y} \in Val(\boldsymbol{Y})$, and $\boldsymbol{z} \in Val(\boldsymbol{Z})$. The variables in the set $\boldsymbol{Z}$ are often said to be* observed. *If the set $\boldsymbol{Z}$ is empty, then instead of writing $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \emptyset)$, we write $(\boldsymbol{X} \perp \boldsymbol{Y})$ and say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are* marginally independent. ■

Thus, an independence statement over random variables is a universal quantification over all possible values of the random variables.

The alternative characterization of conditional independence follows immediately:

Proposition 2.3

The distribution P satisfies $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ if and only if $P(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = P(\boldsymbol{X} \mid \boldsymbol{Z})P(\boldsymbol{Y} \mid \boldsymbol{Z})$.

Suppose we learn about a conditional independence. Can we conclude other independence properties that must hold in the distribution? We have already seen one such example:

symmetry

- **Symmetry**:

$$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{Y} \perp \boldsymbol{X} \mid \boldsymbol{Z}). \tag{2.7}$$

There are several other properties that hold for conditional independence, and that often provide a very clean method for proving important properties about distributions. Some key properties are:

decomposition

- **Decomposition**:

$$(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}). \tag{2.8}$$

weak union

- **Weak union**:

$$(\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W}). \tag{2.9}$$

contraction

- **Contraction**:

$$(\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}, \boldsymbol{Y}) \,\&\, (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}). \tag{2.10}$$

An additional important property does not hold in general, but it does hold in an important subclass of distributions.

Definition 2.5 positive distribution

A distribution P is said to be positive *if for all events $\alpha \in \mathcal{S}$ such that $\alpha \neq \emptyset$, we have that $P(\alpha) > 0$.* ■

For positive distributions, we also have the following property:

intersection

- **Intersection**: For positive distributions, and for mutually disjoint sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}$:

$$(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W}) \,\&\, (\boldsymbol{X} \perp \boldsymbol{W} \mid \boldsymbol{Z}, \boldsymbol{Y}) \Longrightarrow (\boldsymbol{X} \perp \boldsymbol{Y}, \boldsymbol{W} \mid \boldsymbol{Z}). \tag{2.11}$$

The proof of these properties is not difficult. For example, to prove Decomposition, assume that $(X \perp Y, W \mid Z)$ holds. Then, from the definition of conditional independence, we have that $P(X, Y, W \mid Z) = P(X \mid Z)P(Y, W \mid Z)$. Now, using basic rules of probability and arithmetic, we can show

$$\begin{aligned} P(X, Y \mid Z) &= \sum_{w} P(X, Y, w \mid Z) \\ &= \sum_{w} P(X \mid Z)P(Y, w \mid Z) \\ &= P(X \mid Z) \sum_{w} P(Y, w \mid Z) \\ &= P(X \mid Z)P(Y \mid Z). \end{aligned}$$

The only property we used here is called "reasoning by cases" (see exercise 2.6). We conclude that $(X \perp Y \mid Z)$.

2.1.5 Querying a Distribution

Our focus throughout this book is on using a joint probability distribution over multiple random variables to answer queries of interest.

2.1.5.1 Probability Queries

probability query

Perhaps the most common query type is the *probability query*. Such a query consists of two parts:

evidence

- The *evidence*: a subset $\boldsymbol{E}$ of random variables in the model, and an instantiation $\boldsymbol{e}$ to these variables;

query variables

- the *query variables*: a subset $\boldsymbol{Y}$ of random variables in the network.

Our task is to compute

$$P(\boldsymbol{Y} \mid \boldsymbol{E} = \boldsymbol{e}),$$

posterior distribution

that is, the *posterior probability distribution* over the values $\boldsymbol{y}$ of $\boldsymbol{Y}$, conditioned on the fact that $\boldsymbol{E} = \boldsymbol{e}$. This expression can also be viewed as the marginal over $\boldsymbol{Y}$, in the distribution we obtain by conditioning on $\boldsymbol{e}$.

2.1.5.2 MAP Queries

A second important type of task is that of finding a high-probability joint assignment to some subset of variables. The simplest variant of this type of task is the *MAP* query (also called *most probable explanation (MPE)*), whose aim is to find the *MAP assignment* — the most likely assignment to all of the (non-evidence) variables. More precisely, if we let $\boldsymbol{W} = \mathcal{X} - \boldsymbol{E}$, our task is to find the most likely assignment to the variables in $\boldsymbol{W}$ given the evidence $\boldsymbol{E} = \boldsymbol{e}$:

MAP assignment

$$\text{MAP}(\boldsymbol{W} \mid \boldsymbol{e}) = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}, \boldsymbol{e}), \tag{2.12}$$

where, in general, $\arg\max_x f(x)$ represents the value of x for which $f(x)$ is maximal. Note that there might be more than one assignment that has the highest posterior probability. In this case, we can either decide that the MAP task is to return the set of possible assignments, or to return an arbitrary member of that set.

It is important to understand the difference between MAP queries and probability queries. In a MAP query, we are finding the most likely *joint* assignment to $\boldsymbol{W}$. To find the most likely assignment to a single variable A, we could simply compute $P(A \mid \boldsymbol{e})$ and then pick the most likely value. **However, the assignment where each variable individually picks its most likely value can be quite different from the most likely joint assignment to all variables simultaneously.** This phenomenon can occur even in the simplest case, where we have no evidence.

Example 2.4

Consider a two node chain $A \rightarrow B$ where A and B are both binary-valued. Assume that:

a^0	a^1
0.4	0.6

A	b^0	b^1
a^0	0.1	0.9
a^1	0.5	0.5

(2.13)

We can see that $P(a^1) > P(a^0)$, so that $\text{MAP}(A) = a^1$. However, $\text{MAP}(A, B) = (a^0, b^1)$: Both values of B have the same probability given a^1. Thus, the most likely assignment containing a^1 has probability $0.6 \times 0.5 = 0.3$. On the other hand, the distribution over values of B is more skewed given a^0, and the most likely assignment (a^0, b^1) has the probability $0.4 \times 0.9 = 0.36$. Thus, we have that $\arg\max_{a,b} P(a, b) \neq (\arg\max_a P(a), \arg\max_b P(b))$. ■

2.1.5.3 Marginal MAP Queries

To motivate our second query type, let us return to the phenomenon demonstrated in example 2.4. Now, consider a medical diagnosis problem, where the most likely disease has multiple possible symptoms, each of which occurs with some probability, but not an overwhelming probability. On the other hand, a somewhat rarer disease might have only a few symptoms, each of which is very likely given the disease. As in our simple example, the MAP assignment to the data and the symptoms might be higher for the second disease than for the first one. The solution here is to look for the most likely assignment to the disease variable(s) only, rather than the most likely assignment to both the disease and symptom variables. This approach suggests the use of a more general query type. In the *marginal MAP* query, we have a subset of variables $\boldsymbol{Y}$ that forms our query. The task is to find the most likely assignment to the variables in $\boldsymbol{Y}$ given the evidence $\boldsymbol{E} = \boldsymbol{e}$:

marginal MAP

$$\text{MAP}(\boldsymbol{Y} \mid \boldsymbol{e}) = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y} \mid \boldsymbol{e}).$$

If we let $\boldsymbol{Z} = \mathcal{X} - \boldsymbol{Y} - \boldsymbol{E}$, the marginal MAP task is to compute:

$$\text{MAP}(\boldsymbol{Y} \mid \boldsymbol{e}) = \arg\max_{\boldsymbol{Y}} \sum_{\boldsymbol{Z}} P(\boldsymbol{Y}, \boldsymbol{Z} \mid \boldsymbol{e}).$$

Thus, marginal MAP queries contain both summations and maximizations; in a way, it contains elements of both a conditional probability query and a MAP query.

Note that example 2.4 shows that marginal MAP assignments are not monotonic: the most likely assignment $\text{MAP}(Y_1 \mid \boldsymbol{e})$ might be completely different from the assignment to Y_1 in $\text{MAP}(\{Y_1, Y_2\} \mid \boldsymbol{e})$. Thus, in particular, we cannot use a MAP query to give us the correct answer to a marginal MAP query.

2.1.6 Continuous Spaces

In the previous section, we focused on random variables that have a finite set of possible values. In many situations, we also want to reason about continuous quantities such as weight, height, duration, or cost that take real numbers in $\mathbb{R}$.

When dealing with probabilities over continuous random variables, we have to deal with some technical issues. For example, suppose that we want to reason about a random variable X that can take values in the range between 0 and 1. That is, $Val(X)$ is the interval $[0, 1]$. Moreover, assume that we want to assign each number in this range equal probability. What would be the probability of a number x? Clearly, since each x has the same probability, and there are infinite number of values, we must have that $P(X = x) = 0$. This problem appears even if we do not require uniform probability.

2.1.6.1 Probability Density Functions

How do we define probability over a continuous random variable? We say that a function $p : \mathbb{R} \mapsto \mathbb{R}$ is a *probability density function* or (*PDF*) for X if it is a nonnegative integrable

density function

function such that

$$\int_{Val(X)} p(x)dx = 1.$$

That is, the integral over the set of possible values of X is 1. The PDF defines a distribution for X as follows: for any x in our event space:

$$P(X \leq a) = \int_{-\infty}^{a} p(x)dx.$$

cumulative distribution

The function P is the *cumulative distribution* for X. We can easily employ the rules of probability to see that by using the density function we can evaluate the probability of other events. For example,

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx.$$

Intuitively, the value of a PDF $p(x)$ at a point x is the incremental amount that x adds to the cumulative distribution in the integration process. The higher the value of p at and around x, the more mass is added to the cumulative distribution as it passes x.

The simplest PDF is the uniform distribution.

Definition 2.6

uniform distribution

A variable X has a uniform distribution *over $[a, b]$, denoted $X \sim \mathrm{Unif[a, b]}$ if it has the PDF*

$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise.} \end{cases}$$

■

Thus, the probability of any subinterval of $[a, b]$ is proportional its size relative to the size of $[a, b]$. Note that, if $b - a < 1$, then the density can be greater than 1. Although this looks unintuitive, this situation can occur even in a legal PDF, if the interval over which the value is greater than 1 is not too large. We have only to satisfy the constraint that the total area under the PDF is 1.

As a more complex example, consider the Gaussian distribution.

Definition 2.7

Gaussian distribution

A random variable X has a Gaussian distribution *with* mean μ *and* variance σ^2*, denoted $X \sim \mathcal{N}(\mu; \sigma^2)$, if it has the PDF*

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

standard Gaussian

A standard Gaussian *is one with mean* 0 *and variance* 1. ■

A Gaussian distribution has a bell-like curve, where the mean parameter μ controls the location of the peak, that is, the value for which the Gaussian gets its maximum value. The variance parameter σ^2 determines how peaked the Gaussian is: the smaller the variance, the

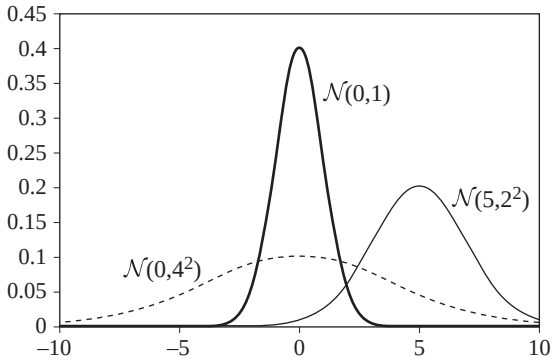

Figure 2.2 Example PDF of three Gaussian distributions

more peaked the Gaussian. Figure 2.2 shows the probability density function of a few different Gaussian distributions.

More technically, the probability density function is specified as an exponential, where the expression in the exponent corresponds to the square of the number of standard deviations σ that x is away from the mean μ. The probability of x decreases exponentially with the square of its deviation from the mean, as measured in units of its standard deviation.

2.1.6.2 Joint Density Functions

The discussion of density functions for a single variable naturally extends for joint distributions of continuous random variables.

Definition 2.8 joint density

Let P be a joint distribution over continuous random variables $X_1, \ldots, X_n$. A function $p(x_1, \ldots, x_n)$ is a joint density function *of $X_1, \ldots, X_n$ if*

- $p(x_1, \ldots, x_n) \geq 0$ *for all values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$.*
- *p is an integrable function.*
- *For any choice of $a_1, \ldots, a_n$, and $b_1, \ldots, b_n$,*

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

■

Thus, a joint density specifies the probability of any joint event over the variables of interest.

Both the uniform distribution and the Gaussian distribution have natural extensions to the multivariate case. The definition of a multivariate uniform distribution is straightforward. We defer the definition of the multivariate Gaussian to section 7.1.

From the joint density we can derive the marginal density of any random variable by integrating out the other variables. Thus, for example, if $p(x, y)$ is the joint density of X and Y,

then

$$p(x) = \int_{-\infty}^{\infty} p(x, y)dy.$$

To see why this equality holds, note that the event $a \leq X \leq b$ is, by definition, equal to the event "$a \leq X \leq b$ and $-\infty \leq Y \leq \infty$." This rule is the direct analogue of marginalization for discrete variables. Note that, as with discrete probability distributions, we abuse notation a bit and use p to denote both the joint density of X and Y and the marginal density of X. In cases where the distinction is not clear, we use subscripts, so that p_X will be the marginal density, of X, and $p_{X,Y}$ the joint density.

2.1.6.3 Conditional Density Functions

As with discrete random variables, we want to be able to describe conditional distributions of continuous variables. Suppose, for example, we want to define $P(Y \mid X = x)$. Applying the definition of conditional distribution (equation (2.1)), we run into a problem, since $P(X = x) = 0$. Thus, the ratio of $P(Y, X = x)$ and $P(X = x)$ is undefined.

To avoid this problem, we might consider conditioning on the event $x - \epsilon \leq X \leq x + \epsilon$, which can have a positive probability. Now, the conditional probability is well defined. Thus, we might consider the limit of this quantity when $\epsilon \to 0$. We define

$$P(Y \mid x) = \lim_{\epsilon \to 0} P(Y \mid x - \epsilon \leq X \leq x + \epsilon).$$

When does this limit exist? If there is a continuous joint density function $p(x, y)$, then we can derive the form for this term. To do so, consider some event on Y, say $a \leq Y \leq b$. Recall that

$$\begin{aligned} P(a \leq Y \leq b \mid x - \epsilon \leq X \leq x + \epsilon) &= \frac{P(a \leq Y \leq b, x - \epsilon \leq X \leq x + \epsilon)}{P(x - \epsilon \leq X \leq x + \epsilon)} \\ &= \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} p(x', y)dydx'}{\int_{x-\epsilon}^{x+\epsilon} p(x')dx'}. \end{aligned}$$

When ϵ is sufficiently small, we can approximate

$$\int_{x-\epsilon}^{x+\epsilon} p(x')dx' \approx 2\epsilon p(x).$$

Using a similar approximation for $p(x', y)$, we get

$$\begin{aligned} P(a \leq Y \leq b \mid x - \epsilon \leq X \leq x + \epsilon) &\approx \frac{\int_a^b 2\epsilon p(x, y)dy}{2\epsilon p(x)} \\ &= \int_a^b \frac{p(x, y)}{p(x)} dy. \end{aligned}$$

We conclude that $\frac{p(x,y)}{p(x)}$ is the density of $P(Y \mid X = x)$.

Definition 2.9 conditional density function

Let $p(x, y)$ be the joint density of X and Y. The conditional density function *of Y given X is defined as*

$$p(y \mid x) = \frac{p(x, y)}{p(x)}$$

When $p(x) = 0$, the conditional density is undefined. ■

The conditional density $p(y \mid x)$ characterizes the conditional distribution $P(Y \mid X = x)$ we defined earlier.

The properties of joint distributions and conditional distributions carry over to joint and conditional density functions. In particular, we have the chain rule

$$p(x, y) = p(x)p(y \mid x) \tag{2.14}$$

and Bayes' rule

$$p(x \mid y) = \frac{p(x)p(y \mid x)}{p(y)}. \tag{2.15}$$

As a general statement, whenever we discuss joint distributions of continuous random variables, we discuss properties with respect to the joint density function instead of the joint distribution, as we do in the case of discrete variables. Of particular interest is the notion of (conditional) independence of continuous random variables.

Definition 2.10 conditional independence

Let $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ be sets of continuous random variables with joint density $p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. We say that $\boldsymbol{X}$ is conditionally independent *of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ if*

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{z}) \text{ for all } \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \text{ such that } p(\boldsymbol{z}) > 0.$$ ■

2.1.7 Expectation and Variance

2.1.7.1 Expectation

expectation

Let X be a discrete random variable that takes numerical values; then the *expectation* of X under the distribution P is

$$\boldsymbol{E}_P[X] = \sum_x x \cdot P(x).$$

If X is a continuous variable, then we use the density function

$$\boldsymbol{E}_P[X] = \int x \cdot p(x)dx.$$

For example, if we consider X to be the outcome of rolling a fair die with probability $1/6$ for each outcome, then $\boldsymbol{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$. On the other hand, if we consider a biased die where $P(X = 6) = 0.5$ and $P(X = x) = 0.1$ for $x < 6$, then $\boldsymbol{E}[X] = 1 \cdot 0.1 + \cdots + 5 \cdot 0.1 + \cdots + 6 \cdot 0.5 = 4.5$.

Often we are interested in expectations of a function of a random variable (or several random variables). Thus, we might consider extending the definition to consider the expectation of a functional term such as $X^2 + 0.5X$. Note, however, that any function g of a set of random variables $X_1, \ldots, X_k$ is essentially defining a new random variable Y: For any outcome $\omega \in \Omega$, we define the value of Y as $g(f_{X_1}(\omega), \ldots, f_{X_k}(\omega))$.

Based on this discussion, we often define new random variables by a functional term. For example $Y = X^2$, or $Y = e^X$. We can also consider functions that map values of one or more categorical random variables to numerical values. One such function that we use quite often is the *indicator function*, which we denote $\mathbf{1}\{\boldsymbol{X} = \boldsymbol{x}\}$. This function takes value 1 when $\boldsymbol{X} = \boldsymbol{x}$, and 0 otherwise.

indicator function

In addition, we often consider expectations of functions of random variables without bothering to name the random variables they define. For example $\boldsymbol{E}_P[X + Y]$. Nonetheless, we should keep in mind that such a term does refer to an expectation of a random variable.

We now turn to examine properties of the expectation of a random variable.

First, as can be easily seen, the expectation of a random variable is a linear function in that random variable. Thus,

$$\boldsymbol{E}[a \cdot X + b] = a\boldsymbol{E}[X] + b.$$

A more complex situation is when we consider the expectation of a function of several random variables that have some joint behavior. An important property of expectation is that the expectation of a sum of two random variables is the sum of the expectations.

Proposition 2.4

$$\boldsymbol{E}[X + Y] = \boldsymbol{E}[X] + \boldsymbol{E}[Y].$$

linearity of expectation

This property is called *linearity of expectation*. It is important to stress that this identity is true even when the variables are not independent. As we will see, this property is key in simplifying many seemingly complex problems.

Finally, what can we say about the expectation of a product of two random variables? In general, very little:

Example 2.5

Consider two random variables X and Y, each of which takes the value $+1$ with probability $1/2$, and the value -1 with probability $1/2$. If X and Y are independent, then $\boldsymbol{E}[X \cdot Y] = 0$. On the other hand, if X and Y are correlated in that they always take the same value, then $\boldsymbol{E}[X \cdot Y] = 1$. ■

However, when X and Y are independent, then, as in our example, we can compute the expectation simply as a product of their individual expectations:

Proposition 2.5

If X and Y are independent, then

$$\boldsymbol{E}[X \cdot Y] = \boldsymbol{E}[X] \cdot \boldsymbol{E}[Y].$$

conditional expectation

We often also use the expectation given some evidence. The *conditional expectation* of X given $\boldsymbol{y}$ is

$$\boldsymbol{E}_P[X \mid \boldsymbol{y}] = \sum_x x \cdot P(x \mid \boldsymbol{y}).$$

2.1.7.2 Variance

variance

The expectation of X tells us the mean value of X. However, It does not indicate how far X deviates from this value. A measure of this deviation is the *variance* of X.

$$\mathbf{Var}_P[X] = \mathbf{E}_P\left[(X - \mathbf{E}_P[X])^2\right].$$

Thus, the variance is the expectation of the squared difference between X and its expected value. It gives us an indication of the spread of values of X around the expected value.

An alternative formulation of the variance is

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 . \tag{2.16}$$

(see exercise 2.11).

Similar to the expectation, we can consider the expectation of a functions of random variables.

Proposition 2.6 *If X and Y are independent, then*

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y].$$

It is straightforward to show that the variance scales as a quadratic function of X. In particular, we have:

$$\mathbf{Var}[a \cdot X + b] = a^2 \mathbf{Var}[X].$$

standard deviation

For this reason, we are often interested in the square root of the variance, which is called the *standard deviation* of the random variable. We define

$$\sigma_X = \sqrt{\mathbf{Var}[X]}.$$

The intuition is that it is improbable to encounter values of X that are farther than several standard deviations from the expected value of X. Thus, σ_X is a normalized measure of "distance" from the expected value of X.

As an example consider the Gaussian distribution of definition 2.7.

Proposition 2.7 *Let X be a random variable with Gaussian distribution $N(\mu, \sigma^2)$, then $\mathbf{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$.*

Thus, the parameters of the Gaussian distribution specify the expectation and the variance of the distribution. As we can see from the form of the distribution, the density of values of X drops exponentially fast in the distance $\frac{x-\mu}{\sigma}$.

Not all distributions show such a rapid decline in the probability of outcomes that are distant from the expectation. However, even for arbitrary distributions, one can show that there is a decline.

Theorem 2.1 **(Chebyshev inequality)**:

Chebyshev's inequality

$$P(|X - \mathbf{E}_P[X]| \geq t) \leq \frac{\mathbf{Var}_P[X]}{t^2}.$$

We can restate this inequality in terms of standard deviations: We write $t = k\sigma_X$ to get

$$P(|X - \boldsymbol{E}_P[X]| \geq k\sigma_X) \leq \frac{1}{k^2}.$$

Thus, for example, the probability of X being more than two standard deviations away from $\boldsymbol{E}[X]$ is less than $1/4$.

2.2 Graphs

Perhaps the most pervasive concept in this book is the representation of a probability distribution using a graph as a data structure. In this section, we survey some of the basic concepts in graph theory used in the book.

2.2.1 Nodes and Edges

A graph is a data structure $\mathcal{K}$ consisting of a set of nodes and a set of edges. Throughout most this book, we will assume that the set of nodes is $\mathcal{X} = \{X_1, \ldots, X_n\}$. A pair of nodes X_i, X_j can be connected by a *directed edge* $X_i \to X_j$ or an *undirected edge* X_i—X_j. Thus, the set of edges $\mathcal{E}$ is a set of pairs, where each pair is one of $X_i \to X_j$, $X_j \to X_i$, or X_i—X_j, for $X_i, X_j \in \mathcal{X}$, $i < j$. We assume throughout the book that, for each pair of nodes X_i, X_j, at most one type of edge exists; thus, we cannot have both $X_i \to X_j$ and $X_j \to X_i$, nor can we have $X_i \to X_j$ and X_i—X_j.[2] The notation $X_i \leftarrow X_j$ is equivalent to $X_j \to X_i$, and the notation X_j—X_i is equivalent to X_i—X_j. We use $X_i \rightleftharpoons X_j$ to represent the case where X_i and X_j are connected via some edge, whether directed (in any direction) or undirected.

directed edge

undirected edge

In many cases, we want to restrict attention to graphs that contain only edges of one kind or another. We say that a graph is *directed* if all edges are either $X_i \to X_j$ or $X_j \to X_i$. We usually denote directed graphs as $\mathcal{G}$. We say that a graph is *undirected* if all edges are X_i—X_j. We denote undirected graphs as $\mathcal{H}$. We sometimes convert a general graph to an undirected graph by ignoring the directions on the edges.

directed graph

undirected graph

Definition 2.11 *Given a graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, its* undirected version *is a graph $\mathcal{H} = (\mathcal{X}, \mathcal{E}')$ where $\mathcal{E}' = \{X$—$Y : X \rightleftharpoons Y \in \mathcal{E}\}$.* ■

graph's undirected version

Whenever we have that $X_i \to X_j \in \mathcal{E}$, we say that X_j is the *child* of X_i in $\mathcal{K}$, and that X_i is the *parent* of X_j in $\mathcal{K}$. When we have X_i—$X_j \in \mathcal{E}$, we say that X_i is a *neighbor* of X_j in $\mathcal{K}$ (and vice versa). We say that X and Y are adjacent whenever $X \rightleftharpoons Y \in \mathcal{E}$. We use Pa_X to denote the parents of X, Ch_X to denote its children, and Nb_X to denote its neighbors. We define the *boundary* of X, denoted $\mathrm{Boundary}_X$, to be $\mathrm{Pa}_X \cup \mathrm{Nb}_X$; for DAGs, this set is simply X's parents, and for undirected graphs X's neighbors.[3] Figure 2.3 shows an example of a graph $\mathcal{K}$. There, we have that A is the only parent of C, and F, I are the children of C. The only neighbor of C is D, but its adjacent nodes are A, D, F, I. The *degree* of a node X is the number of edges in which it participates. Its *indegree* is the number of directed edges $Y \to X$. The *degree* of a graph is the maximal degree of a node in the graph.

child

parent

neighbor

boundary

degree

indegree

2. Note that our definition is somewhat restricted, in that it disallows cycles of length two, where $X_i \to X_j \to X_i$, and allows self-loops where $X_i \to X_i$.

3. When the graph is not clear from context, we often add the graph as an additional argument.

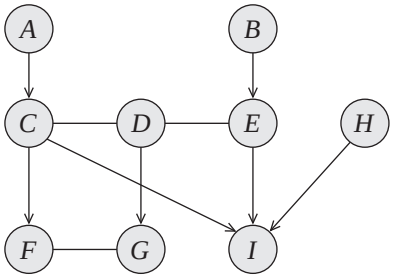

Figure 2.3 An example of a partially directed graph $\mathcal{K}$

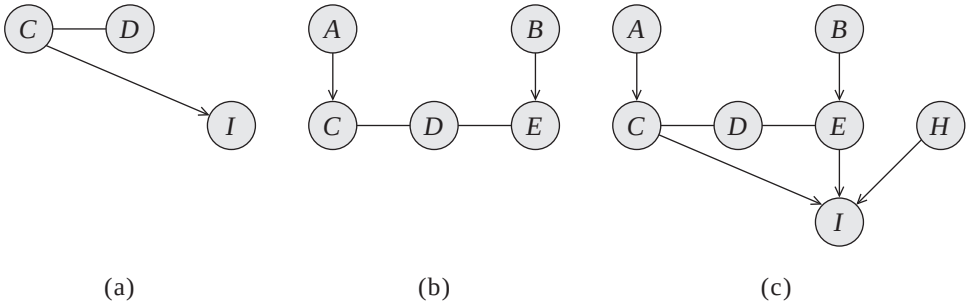

Figure 2.4 Induced graphs and their upward closure: (a) The induced subgraph $\mathcal{K}[C, D, I]$. (b) The upwardly closed subgraph $\mathcal{K}^+[C]$. (c) The upwardly closed subgraph $\mathcal{K}^+[C, D, I]$.

2.2.2 Subgraphs

In many cases, we want to consider only the part of the graph that is associated with a particular subset of the nodes.

Definition 2.12 induced subgraph

Let $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and let $\boldsymbol{X} \subset \mathcal{X}$. We define the induced subgraph $\mathcal{K}[\boldsymbol{X}]$ *to be the graph $(\boldsymbol{X}, \mathcal{E}')$ where $\mathcal{E}'$ are all the edges $X \rightleftharpoons Y \in \mathcal{E}$ such that $X, Y \in \boldsymbol{X}$.* ■

For example, figure 2.4a shows the induced subgraph $\mathcal{K}[C, D, I]$.

A type of subgraph that is often of particular interest is one that contains all possible edges.

Definition 2.13 complete subgraph, clique

A subgraph over $\boldsymbol{X}$ is complete *if every two nodes in $\boldsymbol{X}$ are connected by some edge. The set $\boldsymbol{X}$ is often called a* clique*; we say that a clique $\boldsymbol{X}$ is* maximal *if for any superset of nodes $\boldsymbol{Y} \supset \boldsymbol{X}$, $\boldsymbol{Y}$ is not a clique.* ■

Although the subset of nodes $\boldsymbol{X}$ can be arbitrary, we are often interested in sets of nodes that preserve certain aspects of the graph structure.

Definition 2.14 upward closure

We say that a subset of nodes $\boldsymbol{X} \in \mathcal{X}$ is upwardly closed *in $\mathcal{K}$ if, for any $X \in \boldsymbol{X}$, we have that* $\text{Boundary}_X \subset \boldsymbol{X}$. *We define the* upward closure *of $\boldsymbol{X}$ to be the minimal upwardly closed subset*

$\boldsymbol{Y}$ that contains $\boldsymbol{X}$. We define the upwardly closed subgraph *of $\boldsymbol{X}$, denoted $\mathcal{K}^+[\boldsymbol{X}]$, to be the induced subgraph over $\boldsymbol{Y}$, $\mathcal{K}[\boldsymbol{Y}]$.* ■

For example, the set A, B, C, D, E is the upward closure of the set $\{C\}$ in $\mathcal{K}$. The upwardly closed subgraph of $\{C\}$ is shown in figure 2.4b. The upwardly closed subgraph of $\{C, D, I\}$ is shown in figure 2.4c.

2.2.3 Paths and Trails

Using the basic notion of edges, we can define different types of longer-range connections in the graph.

Definition 2.15 path

We say that $X_1, \ldots, X_k$ form a path *in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ if, for every $i = 1, \ldots, k-1$, we have that either $X_i \rightarrow X_{i+1}$ or X_i—X_{i+1}. A path is* directed *if, for at least one i, we have $X_i \rightarrow X_{i+1}$.* ■

Definition 2.16 trail

We say that $X_1, \ldots, X_k$ form a trail *in the graph $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ if, for every $i = 1, \ldots, k-1$, we have that $X_i \rightleftharpoons X_{i+1}$.* ■

In the graph $\mathcal{K}$ of figure 2.3, A, C, D, E, I is a path, and hence also a trail. On the other hand, A, C, F, G, D is a trail, which is not a path.

Definition 2.17 connected graph

A graph is connected *if for every X_i, X_j there is a trail between X_i and X_j.* ■

We can now define longer-range relationships in the graph.

Definition 2.18 ancestor descendant

We say that X is an ancestor *of Y in $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and that Y is a* descendant *of X, if there exists a directed path $X_1, \ldots, X_k$ with $X_1 = X$ and $X_k = Y$. We use $\mathrm{Descendants}_X$ to denote X's descendants, $\mathrm{Ancestors}_X$ to denote X's ancestors, and $\mathrm{NonDescendants}_X$ to denote the set of nodes in $\mathcal{X} - \mathrm{Descendants}_X$.* ■

In our example graph $\mathcal{K}$, we have that F, G, I are descendants of C. The ancestors of C are A, via the path A, C, and B, via the path B, E, D, C.

A final useful notion is that of an ordering of the nodes in a directed graph that is consistent with the directionality its edges.

Definition 2.19 topological ordering

Let $\mathcal{G} = (\mathcal{X}, \mathcal{E})$ be a graph. An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering *relative to $\mathcal{K}$ if, whenever we have $X_i \rightarrow X_j \in \mathcal{E}$, then $i < j$.* ■

Appendix A.3.1 presents an algorithm for finding such a topological ordering.

2.2.4 Cycles and Loops

Note that, in general, we can have a cyclic path that leads from a node to itself, making that node its own descendant.

Definition 2.20

cycle

acyclic

A cycle *in $\mathcal{K}$ is a directed path $X_1, \ldots, X_k$ where $X_1 = X_k$. A graph is* acyclic *if it contains no cycles.* ■

For most of this book, we will restrict attention to graphs that do not allow such cycles, since it is quite difficult to define a coherent probabilistic model over graphs with directed cycles.

DAG

A *directed acyclic graph (DAG)* is one of the central concepts in this book, as DAGs are the basic graphical representation that underlies Bayesian networks. For some of this book, we also use acyclic graphs that are partially directed. The graph $\mathcal{K}$ of figure 2.3 is acyclic. However, if we add the undirected edge A—E to $\mathcal{K}$, we have a path A, C, D, E, A from A to itself. Clearly, adding a directed edge $E \rightarrow A$ would also lead to a cycle. Note that prohibiting cycles does not imply that there is no *trail* from a node to itself. For example, $\mathcal{K}$ contains several trails: C, D, E, I, C as well as C, D, G, F, C.

PDAG

chain component

An acyclic graph containing both directed and undirected edges is called a *partially directed acyclic graph* or *PDAG*. The acyclicity requirement on a PDAG implies that the graph can be decomposed into a directed graph of *chain components*, where the nodes within each chain component are connected to each other only with undirected edges. The acyclicity of a PDAG guarantees us that we can order the components so that all edges point from lower-numbered components to higher-numbered ones.

Definition 2.21

Let $\mathcal{K}$ be a PDAG over $\mathcal{X}$. Let $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_\ell$ be a disjoint partition of $\mathcal{X}$ such that:

- *the induced subgraph over $\boldsymbol{K}_i$ contains no directed edges;*
- *for any pair of nodes $X \in \boldsymbol{K}_i$ and $Y \in \boldsymbol{K}_j$ for $i < j$, an edge between X and Y can only be a directed edge $X \rightarrow Y$.*

chain component

Each component $\boldsymbol{K}_i$ is called a chain component. ■

chain graph

Because of its chain structure, a PDAG is also called a *chain graph*.

Example 2.6

In the PDAG of figure 2.3, we have six chain components: $\{A\}$, $\{B\}$, $\{C, D, E\}$, $\{F, G\}$, $\{H\}$, and $\{I\}$. This ordering of the chain components is one of several possible legal orderings. ■

Note that when the PDAG is an undirected graph, the entire graph forms a single chain component. Conversely, when the PDAG is a directed graph (and therefore acyclic), each node in the graph is its own chain component.

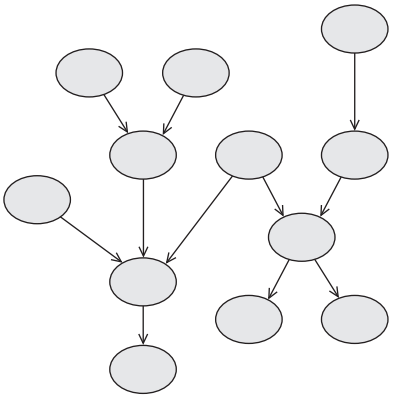

Figure 2.5 An example of a polytree

Different from a cycle is the notion of a loop:

Definition 2.22

loop

singly connected

leaf

polytree

forest

tree

A loop *in $\mathcal{K}$ is a trail $X_1, \ldots, X_k$ where $X_1 = X_k$. A graph is* singly connected *if it contains no loops. A node in a singly connected graph is called a* leaf *if it has exactly one adjacent node. A singly connected directed graph is also called a* polytree. *A singly connected undirected graph is called a* forest*; if it is also connected, it is called a* tree. ■

We can also define a notion of a forest, or of a tree, for directed graphs.

Definition 2.23

A directed graph is a forest *if each node has at most one parent. A directed forest is a* tree *if it is also connected.* ■

Note that polytrees are very different from trees. For example, figure 2.5 shows a graph that is a polytree but is not a tree, because several nodes have more than one parent. As we will discuss later in the book, loops in the graph increase the computational cost of various tasks.

We conclude this section with a final definition relating to loops in the graph. This definition will play an important role in evaluating the cost of reasoning using graph-based representations.

Definition 2.24

chordal graph

Let X_1—X_2—$\cdots$—X_k—X_1 be a loop in the graph; a chord *in the loop is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. An undirected graph $\mathcal{H}$ is said to be* chordal *if any loop X_1—X_2—$\cdots$—X_k—X_1 for $k \geq 4$ has a chord.* ■

Thus, for example, a loop A—B—C—D—A (as in figure 1.1b) is nonchordal, but adding an edge A—C would render it chordal. In other words, in a chordal graph, the longest "minimal loop" (one that has no shortcut) is a triangle. Thus, chordal graphs are often also called *triangulated*.

triangulated graph

We can extend the notion of chordal graphs to graphs that contain directed edges.

Definition 2.25

A graph $\mathcal{K}$ is said to be chordal *if its underlying undirected graph is chordal.* ■

2.3 Relevant Literature

Section 1.4 provides some history on the development of of probabilistic methods. There are many good textbooks about probability theory; see, for example, DeGroot (1989), Ross (1988) or Feller (1970). The distinction between the frequentist and subjective view of probability was a major issue during much of the late nineteenth and early twentieth centuries. Some references that touch on this discussion include Cox (2001) and Jaynes (2003) on the Bayesian side, and Feller (1970) on the frequentist side; these books also contain much useful general material about probabilistic reasoning.

Dawid (1979, 1980) was the first to propose the axiomatization of conditional independence properties, and he showed how they can help unify a variety of topics within probability and statistics. These axioms were studied in great detail by Pearl and colleagues, work that is presented in detail in Pearl (1988).

2.4 Exercises

Exercise 2.1

Prove the following properties using basic properties of definition 2.1:

a. $P(\emptyset) = 0$.
b. If $\alpha \subseteq \beta$, then $P(\alpha) \leq P(\beta)$.
c. $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$.

Exercise 2.2

a. Show that for binary random variables X, Y, the event-level independence $(x^0 \perp y^0)$ implies random-variable independence $(X \perp Y)$.
b. Show a counterexample for nonbinary variables.
c. Is it the case that, for a binary-valued variable Z, we have that $(X \perp Y \mid z^0)$ implies $(X \perp Y \mid Z)$?

Exercise 2.3

Consider two events α and β such that $P(\alpha) = p_a$ and $P(\beta) = p_b$. Given only that knowledge, what is the maximum and minimum values of the probability of the events $\alpha \cap \beta$ and $\alpha \cup \beta$. Can you characterize the situations in which each of these extreme values occurs?

Exercise 2.4⋆

Let P be a distribution over $(\Omega, \mathcal{S})$, and let $a \in \mathcal{S}$ be an event such that $P(\alpha) > 0$. The conditional probability $P(\cdot \mid \alpha)$ assigns a value to each event in $\mathcal{S}$. Show that it satisfies the properties of definition 2.1.

Exercise 2.5

Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ be three disjoint subsets of variables such that $\mathcal{X} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$. Prove that $P \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ if and only if we can write P in the form:

$$P(\mathcal{X}) = \phi_1(\boldsymbol{X}, \boldsymbol{Z})\phi_2(\boldsymbol{Y}, \boldsymbol{Z}).$$

Exercise 2.6

An often useful rule in dealing with probabilities is known as *reasoning by cases*. Let X, Y, and Z be random variables, then

$$P(X \mid Y) = \sum_z P(X, z \mid Y).$$

Prove this equality using the chain rule of probabilities and basic properties of (conditional) distributions.

Exercise 2.7⋆

In this exercise, you will prove the properties of conditional independence discussed in section 2.1.4.3.

a. Prove that the weak union and contraction properties hold for any probability distribution P.
b. Prove that the intersection property holds for any positive probability distribution P.
c. Provide a counterexample to the intersection property in cases where the distribution P is not positive.

Exercise 2.8

a. Show that for binary random variables X and Y, $(x^1 \perp y^1)$ implies $(X \perp Y)$.
b. Provide a counterexample to this property for nonbinary variables.
c. Is it the case that, for binary Z, $(X \perp Y \mid z^1)$ implies $(X \perp Y \mid Z)$? Prove or provide a counterexample.

Exercise 2.9

Show how you can use breadth-first search to determine whether a graph $\mathcal{K}$ is cyclic.

Exercise 2.10⋆

In appendix A.3.1, we describe an algorithm for finding a topological ordering for a directed graph. Extend this algorithm to one that finds a topological ordering for the chain components in a PDAG. Your algorithm should construct both the chain components of the PDAG, as well as an ordering over them that satisfies the conditions of definition 2.21. Analyze the asymptotic complexity of your algorithm.

Exercise 2.11

Use the properties of expectation to show that we can rewrite the variance of a random variable X as

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 .$$

Exercise 2.12⋆

Prove the following property of expectations

Theorem 2.2 Markov inequality

(Markov inequality) *Let X be a random variable such that $P(X \geq 0) = 1$, then for any $t \geq 0$,*

$$P(X \geq t) \leq \frac{\mathbf{E}_P[X]}{t}.$$

You may assume in your proof that X is a discrete random variable with a finite number of values.

Exercise 2.13

Prove Chebyshev's inequality using the Markov inequality shown in exercise 2.12. (Hint: define a new random variable Y, so that the application of the Markov inequality with respect to this random variable gives the required result.)

Exercise 2.14⋆

Let $X \sim \mathcal{N}\left(\mu; \sigma^2\right)$, and define a new variable $Y = a \cdot X + b$. Show that $Y \sim \mathcal{N}\left(a \cdot \mu + b; a^2\sigma^2\right)$.

Exercise 2.15⋆

concave function
convex function

A function f is *concave* if for any $0 \leq \alpha \leq 1$ and any x and y, we have that $f(\alpha x + (1-\alpha)y) \geq \alpha f(x) + (1-\alpha)f(y)$. A function is *convex* if the opposite holds, that is, $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$.

a. Prove that a continuous and differentiable function f is concave if and only if $f''(x) \leq 0$ for all x.
b. Show that $\log(x)$ is concave (over the positive real numbers).

Exercise 2.16⋆

Proposition 2.8

Jensen inequality

Jensen's inequality: Let f be a concave function and P a distribution over a random variable X. Then

$$\boldsymbol{E}_P[f(X)] \leq f(\boldsymbol{E}_P[X])$$

Use this inequality to show that:

a. $\boldsymbol{H}_P(X) \leq \log |Val(X)|$.
b. $\boldsymbol{H}_P(X) \geq 0$.
c. $\boldsymbol{D}(P \| Q) \geq 0$.

See appendix A.1 for the basic definitions.

Exercise 2.17

Show that, for any probability distribution $P(X)$, we have that

$$\boldsymbol{H}_P(X) = \log K - \boldsymbol{D}(P(X) \| P_u(X))$$

where $P_u(X)$ is the uniform distribution over $Val(X)$ and $K = |Val(X)|$.

Exercise 2.18⋆

Prove proposition A.3, and use it to show that $\boldsymbol{I}(X;Y) \geq 0$.

Exercise 2.19

conditional mutual information

As with entropies, we can define the notion of *conditional mutual information*

$$\boldsymbol{I}_P(X;Y \mid Z) = \boldsymbol{E}_P\left[\log \frac{P(X \mid Y,Z)}{P(X \mid Z)}\right].$$

Prove that:

a. $\boldsymbol{I}_P(X;Y \mid Z) = \boldsymbol{H}_P(X \mid Z) - \boldsymbol{H}_P(X,Y \mid Z)$.
b. The *chain rule of mutual information*:

chain rule of mutual information

$$\boldsymbol{I}_P(X;Y,Z) = \boldsymbol{I}_P(X;Y) + \boldsymbol{I}_P(X;Z \mid Y).$$

Exercise 2.20

Use the chain law of mutual information to prove that

$$\boldsymbol{I}_P(X;Y) \leq \boldsymbol{I}_P(X;Y,Z).$$

That is, the information that Y and Z together convey about X cannot be less than what Y alone conveys about X.

Exercise 2.21⋆

Consider a sequence of N independent samples from a binary random variable X whose distribution is $P(x^1) = p$, $P(x^0) = 1 - p$. As in appendix A.2, let S_N be the number of trials whose outcome is x^1. Show that

$$P(S_N = r) \approx \exp[-N \cdot \boldsymbol{D}((p, 1-p) \| (r/N, 1 - r/N))].$$

Your proof should use Stirling's approximation to the factorial function:

$$m! \approx \frac{1}{2\pi m} m^m e^{-m}.$$