1

Prerequisites on Probability Theory

In this chapter we review some standard results and definitions from probability theory. The reader is assumed to have had some contact with probability theory before, and the purpose of this section is simply to brush up on some of the basic concepts and to introduce some of the notation used in the later chapters. Sections 1.1–1.3 are prerequisites for Section 2.3 and thereafter, Section 1.4 is a prerequisite for Chapter 4, and Section 1.5 is a prerequisite for Chapter 6 and Chapter 7.

1.1 Two Perspectives on Probability Theory

In many domains, the probability of seeing a certain outcome of an experiment can be interpreted as the *relative frequency* of seeing this particular outcome in all of the experiments performed. For instance, if you throw a six-sided die, then you would say that the probability of obtaining a three is 1/6, because if we throw this die a large number of times we would expect to see a three in approximately 1/6 of the throws. Along the same line of reasoning, we would also say that if we randomly draw a card from a deck consisting of 52 cards, then the probability that it will be a spade is 13/52. This interpretation of probability rests on the assumption that there is some stochastic process that can be repeated several times and from which the relative frequencies can be counted. On the other hand, we often talk about the probability of seeing a certain event although we cannot specify a frequency for it. For example, I may estimate that the probability that the Danish soccer team will win the World Cup in 2010 is p. This probability is my own personal judgment of how likely it is that the Danish team will actually win, and it is based on my belief, experience, and current state of information. However, another person may specify another probability for the same event, and it has no meaning to look for ways of determining which of us is right, if either. These probabilities are referred to as *subjective probabilities*. One way to interpret my subjective probability of Denmark winning the world cup in 2010 is to imagine the following two wagers:

1. If the Danish soccer team wins the world cup in 2010, I will receive \$100.
2. I will draw a ball from an urn containing 100 balls out of which n are white and $100 - n$ are black. If the ball drawn is white then I will receive \$100 in 2010.

If all the balls are white then I will prefer the second wager, and if all the balls are black then I will prefer the first. However, for a certain n between 0 and 100 I will be indifferent about the two wagers, and for this n, $n/100$ will be my subjective probability that the Danish soccer team will win the World Cup.

1.2 Fundamentals of Probability Theory

For both views on probability described above, we will refer to the set of possible outcomes of an experiment as the *sample space* of the experiment. Here we use the somewhat abstract term "experiment" to refer to any type of process for which the outcome is uncertain, e.g., the throw of a die and the winner of the World Cup. We shall also assume that the sample space of an experiment contains all possible outcomes of the experiment, and that each pair of outcomes are mutually exclusive. These assumptions ensure that the experiment is guaranteed to end up in exactly one of the specified outcomes in the sample space. For instance, for the die example above, the sample space would be $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, and for the soccer example the sample space would be $\mathcal{S} = \{\text{yes}, \text{no}\}$, assuming that I am interested only in whether the Danish team will win; both of the sample spaces satisfy the assumptions above. A subset of a sample space is called an *event*. For example, the event that we will get a value of three or higher with a six-sided die corresponds to the subset $\{3, 4, 5, 6\} \subseteq \{1, 2, 3, 4, 5, 6\}$, and the event will occur if the outcome of the throw is an element in the set. In general, we say that an event $\mathcal{A}$ is *true* for an experiment if the outcome of the experiment is an element of $\mathcal{A}$. When an event contains only one element, we will also refer to the event as an outcome.

To measure our degree of uncertainty about an experiment we assign a probability $P(\mathcal{A})$ to each event $\mathcal{A} \subseteq S$. These probabilities must obey the following three axioms:

The event $\mathcal{S}$ that we will get an outcome in the sample space is certain to occur and is therefore assigned the probability 1.

Axiom 1 $P(\mathcal{S}) = 1$.

Any event $\mathcal{A}$ must have a nonnegative probability.

Axiom 2 *For all* $\mathcal{A} \subseteq \mathcal{S}$ *it holds that* $P(\mathcal{A}) \geq 0$.

If two events $\mathcal{A}$ and $\mathcal{B}$ are disjoint (see Figure 1.1(a)), then the probability of the combined event is the sum of the probabilities for the two individual events:

Axiom 3 *If $\mathcal{A} \subseteq \mathcal{S}$, $\mathcal{B} \subseteq \mathcal{S}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$, then $P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B})$.*

For example, the event that a die will turn up 3, $\mathcal{B} = \{3\}$, and the event that the die will have an even number, $\mathcal{A} = \{2, 4, 6\}$, are two disjoint events, and the probability that one of these two events will occur is therefore

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) = \frac{1}{6} + \frac{3}{6} = \frac{4}{6}.$$

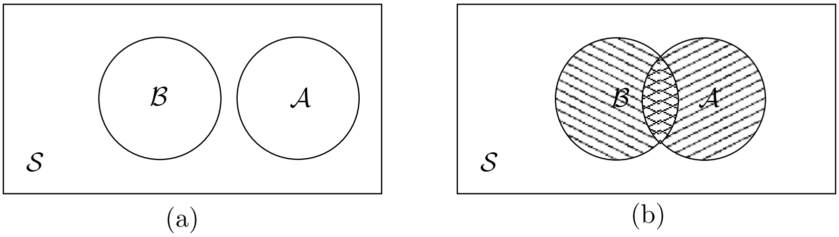

Fig. 1.1. In figure (a) the two events $\mathcal{A}$ and $\mathcal{B}$ are disjoint, whereas in figure (b), $\mathcal{A} \cap \mathcal{B} \neq \emptyset$.

On the other hand, if $\mathcal{A}$ and $\mathcal{B}$ are not disjoint (see Figure 1.1(b)), then it can easily be shown that

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}),$$

where $\mathcal{A} \cap \mathcal{B}$ is the intersection between $\mathcal{A}$ and $\mathcal{B}$ and it represents the event that *both* $\mathcal{A}$ and $\mathcal{B}$ will occur. Consider again a deck with 52 cards. The event $\mathcal{A}$ that I will draw a spade and the event $\mathcal{B}$ that I will draw a king are clearly not disjoint events; their intersection specifies the event that I will draw the king of spades, $\mathcal{A} \cap \mathcal{B} = \{\text{king of spades}\}$. Thus, the probability that I will draw either a king or a spade is

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}.$$

Notation: Sometimes we will emphasize that a probability is based on a frequency (rather than being a subjective probability), in which case we will use the notation $P^{\#}$. If the event $\mathcal{A}$ contains only one outcome a, we write $P(a)$ rather than $P(\{a\})$.

1.2.1 Conditional Probabilities

Whenever a statement about the probability $P(\mathcal{A})$ of an event $\mathcal{A}$ is given, then it is implicitly given conditioned on other known factors. For example, a statement such as "the probability of the die turning up 6 is $\frac{1}{6}$" usually has the unsaid prerequisite that it is a fair die, or rather, as long as I know nothing further, I assume it to be a fair die. This means that the statement should be "given that it is a fair die, the probability ...." In this way, any statement on probabilities is a statement conditioned on what else is known. These types of probabilities are called *conditional probabilities* and are generally statements of the following kind:

"Given the event $\mathcal{B}$, the probability of the event $\mathcal{A}$ is p."

The notation for the preceding statement is $P(\mathcal{A}|\mathcal{B}) = p$. It should be stressed that $P(\mathcal{A}|\mathcal{B}) = p$ does not mean that whenever $\mathcal{B}$ is true, then the probability of $\mathcal{A}$ is p. It means that if $\mathcal{B}$ is true, and *everything else is irrelevant for* $\mathcal{A}$, then the probability of $\mathcal{A}$ is p.

Assume that we have assigned probabilities to all subsets of the sample space $\mathcal{S}$, and let $\mathcal{A}$ and $\mathcal{B}$ be subsets of $\mathcal{S}$ (Figure 1.1(b)). The question is whether the probability assignment for $\mathcal{S}$ can be used to calculate $P(\mathcal{A}|\mathcal{B})$. If we know the event $\mathcal{B}$, then all possible outcomes are elements of $\mathcal{B}$, and the outcomes for which $\mathcal{A}$ can be true are $\mathcal{A} \cap \mathcal{B}$. So, we look for the probability assignment for $\mathcal{A} \cap \mathcal{B}$ given that we know $\mathcal{B}$. Knowing $\mathcal{B}$ does not change the proportion between the probabilities of $\mathcal{A} \cap \mathcal{B}$ and another set $\mathcal{C} \cap \mathcal{B}$ (if, for example, I will bet twice as much on $\mathcal{A} \cap \mathcal{B}$ as on $\mathcal{C} \cap \mathcal{B}$, then after knowing $\mathcal{B}$, I will still bet twice as much on $\mathcal{A} \cap \mathcal{B}$ as on $\mathcal{C} \cap \mathcal{B}$). We can conclude that the proportions $P(\mathcal{A} \cap \mathcal{B})/P(\mathcal{C} \cap \mathcal{B})$ and $P(\mathcal{A}|\mathcal{B})/P(\mathcal{C}|\mathcal{B})$ must be the same. Setting $\mathcal{C} = \mathcal{B}$, and since we know from Axiom 1 that $P(\mathcal{B}|\mathcal{B}) = 1$, we have justified the following property, which should be considered an axiom.

Property 1.1 (Conditional probability). For two events $\mathcal{A}$ and $\mathcal{B}$, with $P(\mathcal{B}) > 0$, the conditional probability for $\mathcal{A}$ given $\mathcal{B}$ is

$$P(\mathcal{A} \,|\, \mathcal{B}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}.$$

For example, the conditional probability that a die will come up 4 given that we get an even number is $P(\mathcal{A} = \{4\} \,|\, \mathcal{B} = \{2,4,6\}) = P(\{4\})/P(\{2,4,6\})$, and by assuming that the die is fair we get $\frac{1/6}{3/6} = \frac{1}{3}$.

Obviously, when working with conditional probabilities we can also condition on more than one event, in which case the definition of a conditional probability generalizes as

$$P(\mathcal{A} \,|\, \mathcal{B} \cap \mathcal{C}) = \frac{P(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C})}{P(\mathcal{B} \cap \mathcal{C})}.$$

1.2.2 Probability Calculus

The expression in Property 1.1 can be rewritten so that we obtain the so-called *fundamental rule* for probability calculus:

Theorem 1.1 (The fundamental rule).

$$P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B}). \tag{1.1}$$

That is, the fundamental rule tells us how to calculate the probability of seeing both $\mathcal{A}$ and $\mathcal{B}$ when we know the probability of $\mathcal{A}$ given $\mathcal{B}$ and the probability of $\mathcal{B}$.

By conditioning on another event $\mathcal{C}$, the fundamental rule can also be written as

$$P(\mathcal{A} \,|\, \mathcal{B} \cap \mathcal{C})P(\mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}).$$

Since $P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{B} \cap \mathcal{A})$ (and also $P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{B} \cap \mathcal{A} \,|\, \mathcal{C})$), we get that $P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A})$ from the fundamental rule. This yields the well-known *Bayes' rule*:

Theorem 1.2 (Bayes' rule).

$$P(\mathcal{A} \,|\, \mathcal{B}) = \frac{P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A})}{P(\mathcal{B})}.$$

Bayes' rule provides us with a method for updating our beliefs about an event $\mathcal{A}$ given that we get information about another event $\mathcal{B}$. For this reason $P(\mathcal{A})$ is usually called the *prior* probability of $\mathcal{A}$, whereas $P(\mathcal{A} \,|\, \mathcal{B})$ is called the *posterior* probability of $\mathcal{A}$ given $\mathcal{B}$; the probability $P(\mathcal{B} \,|\, \mathcal{A})$ is called the *likelihood* of $\mathcal{A}$ given $\mathcal{B}$. For an explanation of this strange use of the term, see Example 1.1.

Finally, as for the fundamental rule, we can also state Bayes' rule in a context $\mathcal{C}$:

$$P(\mathcal{A} \,|\, \mathcal{B}, \mathcal{C}) = \frac{P(\mathcal{B} \,|\, \mathcal{A}, \mathcal{C})P(\mathcal{A} \,|\, \mathcal{C})}{P(\mathcal{B} \,|\, \mathcal{C})}.$$

Example 1.1. We have two diseases a_1 and a_2, both of which can cause the symptom b. Let $P(b \,|\, a_1) = 0.9$ and $P(b \,|\, a_2) = 0.3$. Assume that the prior probabilities for a_1 and a_2 are the same ($P(a_1) = P(a_2)$). Now, if b occurs, Bayes' rule gives

$$P(a_1 \,|\, b) = \frac{P(b \,|\, a_1)P(a_1)}{P(b)} = 0.9 \cdot \frac{P(a_1)}{P(b)};$$
$$P(a_2 \,|\, b) = \frac{P(b \,|\, a_2)P(a_2)}{P(b)} = 0.3 \cdot \frac{P(a_2)}{P(b)}.$$

Even though we cannot calculate the posterior probabilities, we can conclude that a_1 is three times as likely as a_2 given the symptom b.

If we furthermore know that a_1 and a_2 are the only possible causes of b, we can go even further (assuming that the probability of having both diseases is 0). Then $P(a_1 \,|\, b) + P(a_2 \,|\, b) = 1$, and we get

$$\frac{P(a_1)}{P(b)} = \frac{P(a_2)}{P(b)} = \frac{1}{0.9 + 0.3} = \frac{1}{1.2},$$

$P(a_1 \,|\, b) = 0.9/1.2 = 0.75$, and $P(a_2 \,|\, b) = 0.3/1.2 = 0.25$.

1.2.3 Conditional Independence

Sometimes information on one event $\mathcal{B}$ does not change our belief about the occurrence of another event $\mathcal{A}$, and in this case we say that $\mathcal{A}$ and $\mathcal{B}$ are *independent.*

Definition 1.1 (Independence). *The events $\mathcal{A}$ and $\mathcal{B}$ are* independent *if*

$$P(\mathcal{A} \,|\, \mathcal{B}) = P(\mathcal{A}).$$

For example, if we throw two fair dice, then seeing that the first die turns up 2 will not change our beliefs about the outcome of the second die.

This notion of independence is symmetric, so that if $\mathcal{A}$ is independent of $\mathcal{B}$, then $\mathcal{B}$ is independent of $\mathcal{A}$:

$$P(\mathcal{B} \,|\, \mathcal{A}) = \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{A})} = \frac{P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B})}{P(\mathcal{A})} = \frac{P(\mathcal{A})P(\mathcal{B})}{P(\mathcal{A})} = P(\mathcal{B}).$$

The proof requires that $P(\mathcal{A}) > 0$, so if $P(\mathcal{A}) = 0$, the calculations are not valid. However, for our considerations it does not matter; if $\mathcal{A}$ is impossible why bother considering it?

When two events are independent, then the fundamental rule can be rewritten as

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{A}) \cdot P(\mathcal{B}).$$

That is, we can calculate the probability that both events will occur by multiplying the probabilities for the individual events.

The concept of independence also appears when we are conditioning on several events. Specifically, if information about the event $\mathcal{B}$ does not change our belief about the event $\mathcal{A}$ when we already know the event $\mathcal{C}$, then we say that $\mathcal{A}$ and $\mathcal{B}$ are *conditionally independent* given $\mathcal{C}$.

Definition 1.2 (Conditional independence). *The events $\mathcal{A}$ and $\mathcal{B}$ are* conditionally independent *given the event $\mathcal{C}$ if*

$$P(\mathcal{A} \,|\, \mathcal{B} \cap \mathcal{C}) = P(\mathcal{A} \,|\, \mathcal{C}).$$

Similar to the situation above, the conditional independence statement is symmetric. If $\mathcal{A}$ is conditionally independent of $\mathcal{B}$ given $\mathcal{C}$, then $\mathcal{B}$ is conditionally independent of $\mathcal{A}$ given $\mathcal{C}$:

$$\begin{aligned} P(\mathcal{B}\,|\,\mathcal{A}\cap\mathcal{C}) &= \frac{P(\mathcal{A}\cap\mathcal{B}\,|\,C)P(\mathcal{C})}{P(\mathcal{A}\,|\,\mathcal{C})P(\mathcal{C})} = \frac{P(\mathcal{A}\,|\,\mathcal{B}\cap\mathcal{C})P(\mathcal{B}\,|\,\mathcal{C})}{P(\mathcal{A}\,|\,\mathcal{C})} = \frac{P(\mathcal{A}\,|\,\mathcal{C})P(\mathcal{B}\,|\,\mathcal{C})}{P(\mathcal{A}\,|\,\mathcal{C})} \\ &= P(\mathcal{B}\,|\,\mathcal{C}). \end{aligned}$$

Furthermore, when two events are conditionally independent, then we can use a multiplication rule similar to the one above when calculating the probability that both of the events will occur:

$$P(\mathcal{A}\cap\mathcal{B}\,|\,\mathcal{C}) = P(\mathcal{A}\,|\,\mathcal{C})\cdot P(\mathcal{B}\,|\,\mathcal{C}).$$

Note that when two events are independent it is actually a special case of conditional independence but with $\mathcal{C}=\emptyset$.

1.3 Probability Calculus for Variables

So far we have talked about probabilities of simple events and outcomes with respect to a certain sample space. In this book, however, we will be working with a collection of sample spaces, also called *variables*, and we will now extend the concepts above to probabilities over variables. A variable can be considered an experiment, and for each outcome of the experiment the variable has a corresponding *state*. The set of states associated with a variable A is denoted by $\mathrm{sp}(A) = (a_1, a_2, \ldots, a_n)$, and similar to the sample space these states should be *mutually exclusive* and *exhaustive*. The last assumption ensures that the variable is in one of its states (although we may not know which one), and the first assumption ensures that the variable is in only one state. For example, if we let D be a variable representing the outcome of rolling a die, then its state space would be $\mathrm{sp}(D) = (1,2,3,4,5,6)$. We will use uppercase letters for variables and lowercase letters for states, and unless otherwise stated, a variable has a finite number of states.

For a variable A with states $a_1, \ldots, a_n$, we express our uncertainty about its state through a probability distribution $P(A)$ over these states:

$$P(A) = (x_1, \ldots, x_n); \qquad x_i \geq 0; \qquad \sum_{i=1}^{n} x_i = x_1 + \cdots + x_n = 1,$$

where x_i is the probability of A being in state a_i. A distribution is called *uniform* (or *even*) if all probabilities are equal.

Notation: In general, the probability of A being in state a_i is denoted by $P(A = a_i)$, and denoted by $P(a_i)$ if the variable is obvious from the context.

As we talked about conditional probabilities for events, we can also talk about *conditional probabilities* for variables: If the variable B has states $b_1, \ldots, b_m$, then $P(A \mid B)$ contains $n \cdot m$ conditional probabilities $P(a_i \mid b_j)$ that specify the probability of seeing a_i given b_j. That is, the conditional probability for a variable given another variable is a set of probabilities (usually organized in an $n \times m$ table) with one probability for each configuration of the states of the variables involved (see Table 1.1 for an example). Moreover, since $P(A \mid B)$ specifies a probability distribution for each event $B = b_j$, we know from Axiom 1 that the probabilities over A should sum to 1 for each state of B:

$$\sum_{i=1}^{n} P(A = a_i \mid B = b_j) = 1 \text{ for each } b_j.$$

	b_1	b_2	b_3
a_1	0.4	0.3	0.6
a_2	0.6	0.7	0.4

Table 1.1. An example of a conditional probability table $P(A \mid B)$ for the binary variable A given the ternary variable B. Note that for each state of B the probabilities of A sum up to 1.

The probability of seeing joint outcomes for different experiments can be expressed by the *joint probability* for two or more variables: For each configuration (a_i, b_j) of the variables A and B, $P(A, B)$ specifies the probability of seeing both $A = a_i$ *and* $B = b_j$. Hence, $P(A, B)$ consists of $n \cdot m$ numbers, and, similar to $P(A \mid B)$, $P(A, B)$ is usually represented in an $n \times m$ table (see Table 1.2 for an example). Note that since the state spaces of both A and B are mutually exclusive and exhaustive, it follows that all combinations of their states (the Cartesian product) are also mutually exclusive and exhaustive, and they can therefore be considered a sample space. Hence, by Axiom 1,

$$P(A, B) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(A = a_i, B = b_j) = 1.$$

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

Table 1.2. An example of a joint probability table $P(A, B)$ for the binary variable A and the ternary variable B. Note that the sum of all entries is 1.

When the fundamental rule (equation (1.1)) is used on variables A and B, the procedure is to apply the rule to each of the $n \cdot m$ configurations (a_i, b_j) of the two variables:

$$P(a_i \,|\, b_j)P(b_j) = P(a_i, b_j).$$

This means that in the table $P(A \,|\, B)$, each probability in $P(A \,|\, b_j)$ is multiplied by $P(b_j)$ to obtain the table $P(A, b_j)$, and by doing this for each b_j we get $P(A, B)$. If $P(B) = (0.4, 0.4, 0.2)$, then Table 1.2 is the result of using the fundamental rule on Table 1.1 (see also Table 1.3).

$$P(A,B) = \begin{array}{|c|ccc|}\hline & b_1 & b_2 & b_3 \\ \hline a_1 & 0.4 \cdot 0.4 & 0.3 \cdot 0.4 & 0.6 \cdot 0.2 \\ a_2 & 0.6 \cdot 0.4 & 0.7 \cdot 0.4 & 0.4 \cdot 0.2 \\ \hline \end{array} = \begin{array}{|c|ccc|}\hline & b_1 & b_2 & b_3 \\ \hline a_1 & 0.16 & 0.12 & 0.12 \\ a_2 & 0.24 & 0.28 & 0.08 \\ \hline \end{array}$$

Table 1.3. The joint probability table $P(A, B)$ in Table 1.2 can be found by multiplying $P(B) = (0.4, 0.4, 0.2)$ by $P(A \,|\, B)$ in Table 1.1.

When applied to variables, the fundamental rule is expressed as follows:

Theorem 1.3 (The fundamental rule for variables).

$$P(A, B) = P(A \,|\, B)P(B),$$

and conditioned on another variable C we have

$$P(A, B \,|\, C) = P(A \,|\, B, C)P(B \,|\, C).$$

From a joint probability table $P(A, B)$, the probability distribution $P(A)$ can be calculated by considering the outcomes of B that can occur together with each state a_i of A. There are exactly m different outcomes for which A is in state a_i, namely the mutually exclusive outcomes $(a_i, b_1), \ldots, (a_i, b_m)$. Therefore, by Axiom 3,

$$P(a_i) = \sum_{j=1}^{m} P(a_i, b_j).$$

This calculation is called *marginalization*, and we say that the variable B is marginalized out of $P(A, B)$ (resulting in $P(A)$). The notation is

$$P(A) = \sum_{B} P(A, B).$$

By marginalizing B out of Table 1.2, we get

$$P(A) = (0.16 + 0.12 + 0.12, 0.24 + 0.28 + 0.08) = (0.4, 0.6),$$

and by marginalizing out A we get

$$P(B) = (0.16 + 0.24, 0.12 + 0.28, 0.12 + 0.08) = (0.4, 0.4, 0.2).$$

That is, the marginalization operation allows us to remove variables from a joint probability distribution.

Bayes' rule for events (Theorem 1.2) can also be extended to variables, by treating the division in the same way as we treated multiplication above.

Theorem 1.4 (Bayes' rule for variables).

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{P(A, B)}{\sum_B P(A, B)},$$

and conditioned on another variable C we have

$$P(B \mid A, C) = \frac{P(A \mid B, C)P(B \mid C)}{P(A \mid C)} = \frac{P(A, B \mid C)}{\sum_B P(A, B \mid C)}.$$

Note that the two equalities in the equations follow from (1) the fundamental rule and (2) the marginalization operator described above.

By applying Bayes' rule using $P(A)$, $P(B)$, and $P(A \mid B)$ as specified above, we get $P(B \mid A)$ shown in Table 1.4.

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \begin{array}{|c|cc|}\hline & a_1 & a_2 \\ \hline b_1 & \frac{0.4\cdot 0.4}{0.4} & \frac{0.6\cdot 0.4}{0.6} \\ b_2 & \frac{0.3\cdot 0.4}{0.4} & \frac{0.7\cdot 0.4}{0.6} \\ b_3 & \frac{0.6\cdot 0.2}{0.4} & \frac{0.4\cdot 0.2}{0.6} \\ \hline \end{array} = \begin{array}{|c|cc|}\hline & a_1 & a_2 \\ \hline b_1 & 0.4 & 0.4 \\ b_2 & 0.3 & 0.47 \\ b_3 & 0.3 & 0.13 \\ \hline \end{array}$$

Table 1.4. The conditional probability $P(B \mid A)$ obtained by applying Bayes' rule to $P(A \mid B)$ in Table 1.1, $P(A) = (0.4, 0.6)$, and $P(B) = (0.4, 0.4, 0.2)$. Note that the probabilities over B sum to 1 for each state of A.

The concept of (conditional) independence is also defined for variables.

Definition 1.3 (Conditional independence for variables). *Two variables A and C are said to be* conditionally independent *given the variable B if*

$$P(a_i \mid c_k, b_j) = P(a_i \mid b_j)$$

for each $a_i \in \mathrm{sp}(A)$, $b_j \in \mathrm{sp}(B)$, and $c_k \in \mathrm{sp}(C)$.

As a shorthand notation we will sometimes write $P(A \mid C, B) = P(A \mid B)$.

This means that when the state of B is known, then no knowledge of C will alter the probability of A. Observe that we require the independence statement to hold for each state of B; if the conditioning set is empty then we say that A and C are *marginally independent* or just independent (written as $P(A \mid C) = P(A)$).

When two variables A and C are conditionally independent given B, then the fundamental rule (Theorem 1.3) can be simplified:

$$P(A, C \mid B) = P(A \mid B, C)P(C \mid B) = P(A \mid B)P(C \mid B).$$

In the expression above, we multiply two conditional probability tables over different domains. Fortunately, the method for doing this multiplication is a straightforward extension of what we have done so far:

$$P(a_i, c_k \mid b_j) = P(a_i \mid b_j)P(c_k \mid b_j).$$

For example, by multiplying $P(A \mid B)$ and $P(C \mid B)$ (specified in Table 1.1 and Table 1.5, respectively) we get the joint probability $P(A, C \mid B)$ in Table 1.6.

	b_1	b_2	b_3
c_1	0.2	0.9	0.3
c_2	0.05	0.05	0.2
c_3	0.75	0.05	0.5

Table 1.5. The conditional probability table $P(C \mid B)$ for the ternary variable C given the ternary variable B.

$P(A, C \mid B) = P(A \mid B)P(C \mid B)$

=

	b_1	b_2	b_3
c_1	$(0.2 \cdot 0.4, 0.2 \cdot 0.6)$	$(0.9 \cdot 0.3, 0.9 \cdot 0.7)$	$(0.3 \cdot 0.6, 0.3 \cdot 0.4)$
c_2	$(0.05 \cdot 0.4, 0.05 \cdot 0.6)$	$(0.05 \cdot 0.3, 0.05 \cdot 0.7)$	$(0.2 \cdot 0.6, 0.2 \cdot 0.4)$
c_3	$(0.75 \cdot 0.4, 0.75 \cdot 0.6)$	$(0.05 \cdot 0.3, 0.05 \cdot 0.7)$	$(0.5 \cdot 0.6, 0.5 \cdot 0.4)$

=

	b_1	b_2	b_3
c_1	$(0.08, 0.12)$	$(0.27, 0.63)$	$(0.18, 0.12)$
c_2	$(0.02, 0.03)$	$(0.015, 0.035)$	$(0.12, 0.08)$
c_3	$(0.3, 0.45)$	$(0.015, 0.035)$	$(0.3, 0.2)$

Table 1.6. If A and C are conditionally independent given B, then $P(A, C \mid B)$ can be found by multiplying $P(A \mid B)$ and $P(C \mid B)$ as specified in Table 1.1 and Table 1.5, respectively.

1.3.1 Calculations with Probability Tables: An Example

To illustrate the theorems above, assume that we have three variables, A, B, and C, with the probabilities as in Table 1.7. We receive evidence $A = a_2$ and

$C = c_1$ and we would now like to calculate the conditional probability table $P(B \mid a_2, c_1)$.

	b_1	b_2	b_3
a_1	(0, 0.05, 0.05)	(0.05, 0.05, 0)	(0.05, 0.05, 0.05)
a_2	(0.1, 0.1, 0)	(0.1, 0, 0.1)	(0.2, 0, 0.05)

Table 1.7. A joint probability table for the variables A, B, and C. The three numbers in each entry correspond to the states c_1, c_2, and c_3.

First, we focus on the part of the table corresponding to $A = a_2$ and $C = c_1$, and we get

$$P(a_2, B, c_1) = (0.1, 0.1, 0.2). \tag{1.2}$$

To calculate $P(B \mid a_2, c_1)$, we can use Theorem 1.4:

$$P(B \mid a_2, c_1) = \frac{P(a_2, B, c_1)}{P(a_2, c_1)} = \frac{P(a_2, B, c_1)}{\sum_B P(a_2, B, c_1)}. \tag{1.3}$$

By marginalizing B out of equation (1.2) we get

$$P(a_2, c_1) = 0.1 + 0.1 + 0.2 = 0.4.$$

Finally, by performing the division in equation (1.3) we get

$$P(B \mid a_2, c_1) = \left(\frac{0.1}{0.4}, \frac{0.1}{0.4}, \frac{0.2}{0.4}\right) = (0.25, 0.25, 0.5).$$

Another way of doing the same is to say that we wish to transform $P(a_2, B, c_1)$ into a probability distribution. Because the numbers do not add up to one, we *normalize* the distribution by dividing each number by the sum of all the numbers.

Suppose now that we were given only the evidence $A = a_2$, and we want to calculate $P(B \mid a_2, C)$. The calculation of this probability table follows the same steps as above, except that we now work with tables during the calculations. As before, we start by focusing on the part of $P(A, B, C)$ corresponding to $A = a_2$ and we get the result in Table 1.8.

To calculate $P(B \mid a_2, C)$ we use

$$P(B \mid a_2, C) = \frac{P(a_2, B, C)}{P(a_2, C)} = \frac{P(a_2, B, C)}{\sum_B P(a_2, B, C)}. \tag{1.4}$$

The probability $P(a_2, C)$ is found by marginalizing B out of Table 1.8:

$$P(a_2, C) = (0.1 + 0.1 + 0.2, 0.1 + 0 + 0, 0 + 0.1 + 0.05) = (0.4, 0.1, 0.15), \tag{1.5}$$

and by inserting this in equation (1.4) we get the result shown in Table 1.2.

	b_1	b_2	b_3
c_1	0.1	0.1	0.2
c_2	0.1	0	0
c_2	0	0.1	0.05

Table 1.8. The probability table $P(a_2, B, C)$ that corresponds to the part of the probability table in Table 1.8 restricted to $A = a_2$.

$$P(B \mid a_2, C) = \begin{array}{|c|ccc|} \hline & b_1 & b_2 & b_3 \\ \hline c_1 & \frac{0.1}{0.4} & \frac{0.1}{0.4} & \frac{0.2}{0.4} \\ c_2 & \frac{0.1}{0.1} & \frac{0}{0.1} & \frac{0}{0.1} \\ c_2 & \frac{0}{0.15} & \frac{0.1}{0.15} & \frac{0.05}{0.15} \\ \hline \end{array} = \begin{array}{|c|ccc|} \hline & b_1 & b_2 & b_3 \\ \hline c_1 & 0.25 & 0.25 & 0.5 \\ c_2 & 1 & 0 & 0 \\ c_2 & 0 & 2/3 & 1/3 \\ \hline \end{array}$$

Table 1.9. The calculation of $P(B \mid a_2, C)$ using $P(a_2, B, C)$ (Table 1.1) and $P(a_2, C)$ (equation (1.5)).

1.4 An Algebra of Potentials

Below we list some properties of the algebra of multiplication and marginalization of tables. The tables need not be (conditional) probabilities, and they are generally called *potentials.*

A potential ϕ is a real-valued function over a *domain* of finite variables $\mathcal{X}$:

$$\phi : \mathrm{sp}(\mathcal{X}) \to \mathbb{R}$$

The domain of a potential is denoted by $\mathrm{dom}\,(\phi)$. For example, the domain of the potential $P(A, B \mid C)$ is $\mathrm{dom}\,(P(A, B \mid C)) = \{A, B, C\}$.

Two potentials can be *multiplied*, denoted by an (often suppressed) dot. Multiplication has the following properties:

1. $\mathrm{dom}\,(\phi_1\phi_2) = \mathrm{dom}\,(\phi_1) \cup \mathrm{dom}\,(\phi_2)$.
2. **The commutative law**: $\phi_1\phi_2 = \phi_2\phi_1$.
3. **The associative law**: $(\phi_1\phi_2)\phi_3 = \phi_1(\phi_2\phi_3)$.
4. **Existence of unit**: The unit potential $\mathbf{1}$ is a potential that contains only 1's and is defined over any domain such that $\mathbf{1} \cdot \phi = \phi$, for all potentials ϕ.

The marginalization operator defined in Section 1.3 can be generalized to potentials so that $\sum_A \phi$ is a potential over $\mathrm{dom}(\phi) \setminus \{A\}$. Furthermore, marginalization is *commutative*:

$$\sum_A \sum_B \phi = \sum_B \sum_A \phi.$$

For potentials of the form $P(A \mid \mathcal{V})$, where $\mathcal{V}$ is a set of variables, we have

5. **The unit potential property**: $\sum_A P(A \mid \mathcal{V}) = \mathbf{1}$.

For marginalization of a product, the following holds

6. **The distributive law**: If $A \notin \text{dom}(\phi_1)$, then $\sum_A \phi_1\phi_2 = \phi_1 \sum_A \phi_2$.

The distributive law is usually known as $ab + ac = a(b + c)$, and the preceding formula is actually the same law applied to tables. To verify it, consider the calculations in Tables 1.10–1.14. Here we see that Table 1.12 and Table 1.14 are equal and correspond to the left-hand and right-hand sides of the distributive law.

$B \setminus A$	a_1	a_2
b_1	x_1	x_2
b_2	x_3	x_4

$B \setminus C$	c_1	c_2
b_1	y_1	y_2
b_2	y_3	y_4

Table 1.10. $\phi_1(A, B)$ and $\phi_2(C, B)$.

$B \setminus A$	a_1	a_2
b_1	(x_1y_1, x_1y_2)	(x_2y_1, x_2y_2)
b_2	(x_3y_3, x_3y_4)	(x_4y_3, x_4y_4)

Table 1.11. $\phi_1(A, B) \cdot \phi_2(C, B)$. The two numbers in each entry correspond to the states c_1 and c_2.

$B \setminus A$	a_1	a_2
b_1	$x_1y_1 + x_1y_2$	$x_2y_1 + x_2y_2$
b_2	$x_3y_3 + x_3y_4$	$x_4y_3 + x_4y_4$

Table 1.12. $\sum_C \phi_1(A, B) \cdot \phi_2(C, B)$.

B	
b_1	$y_1 + y_2$
b_2	$y_3 + y_4$

Table 1.13. $\sum_C \phi_2(C, B)$.

We also use the term *projection* for marginalization. For example, if A and B are marginalized out of $\phi(A, B, C)$, we may say that ϕ is *projected* down to C, and we use the notation $\phi^{\downarrow C}$. With this notation, the properties of marginalization look as follows ($\mathcal{V}$ and $\mathcal{W}$ denote sets of variables):

$B \setminus A$	a_1	a_2
b_1	$x_1(y_1+y_2)$	$x_2(y_1+y_2)$
b_2	$x_3(y_3+y_4)$	$x_4(y_3+y_4)$

Table 1.14. $\phi_1(A,B)\sum_C \phi_2(C,B)$.

7. **The commutative law**: $(\phi^{\downarrow\mathcal{V}})^{\downarrow\mathcal{W}} = (\phi^{\downarrow\mathcal{W}})^{\downarrow\mathcal{V}}$.
8. **The distributive law**: If $\mathrm{dom}(\phi_1) \subseteq \mathcal{V}$, then $(\phi_1\phi_2)^{\downarrow\mathcal{V}} = \phi_1(\phi_2^{\downarrow\mathcal{V}})$.

1.5 Random Variables

Let $\mathcal{S}$ be a sample space. A *random variable* is a real-valued function on $\mathcal{S}$; $V : \mathcal{S} \to \mathbb{R}$. If, for example, you throw a die, and you win \$1 if you get 4 or above, and you lose \$1 if you get 3 or below, then the corresponding random variable is a function with value -1 on $\{1,2,3\}$ and 1 on $\{4,5,6\}$.

The *mean value* of a random variable V on $\mathcal{S}$ is defined as

$$\mu(V) = \sum_{s\in\mathcal{S}} V(s)P(s). \tag{1.6}$$

For the example above, the mean value is $-1\frac{1}{6}+-1\frac{1}{6}+-1\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=0$ (provided that the die is fair). The mean value is also called the *expected value*.

A measure of how much a random variable varies between its values is the *variance*, σ^2. It is defined as the mean of the square of the difference between value and mean:

$$\sigma^2(V) = \sum_{s\in\mathcal{S}} (V(s)-\mu(V))^2 P(s). \tag{1.7}$$

For the example above we have

$$\sigma^2 = 3(-1-0)^2\frac{1}{6} + 3(1-0)^2\frac{1}{6} = 1.$$

1.5.1 Continuous Distributions

Consider an experiment, where an arrow is thrown at the $[0,1]\times[0,1]$ square. The possible outcomes are the points (x,y) in the unit square. Since the probability is zero for any particular outcome, the probability distribution is assigned to subsets of the unit square. We may think of this assignment as a process of distributing a probability mass of 1 over the sample space. We may, for example, assign a probability for landing in the small square $[x, x+\epsilon]\times[y, y+\epsilon]$. To be more systematic, let n be a natural number, then the unit square can be partitioned into small squares of the type $[\frac{i}{n}, \frac{i+1}{n}]\times[\frac{j}{n}, \frac{j+1}{n}]$, and we can assign probabilities $P([\frac{i}{n}, \frac{i+1}{n}]\times[\frac{j}{n}, \frac{j+1}{n}])$ to these squares with area

$\frac{1}{n^2}$. Now, if $P([\frac{i}{n}, \frac{i+1}{n}] \times [\frac{j}{n}, \frac{j+1}{n}]) = x$, then you can say that the probability mass x is distributed over the small square with an average density of n^2x, and we define the *density function* (also called the *frequency function*) $f(x, y)$ as

$$f(x,y) = \lim_{n\to\infty} n^2 P\left(\left[x, x+\frac{1}{n}\right] \times \left[y, y+\frac{1}{n}\right]\right).$$

In general, if $\mathcal{S}$ is a continuous sample space, the density function is a nonnegative real-valued function f on $\mathcal{S}$, for which it holds that for any subset $\mathcal{A}$ of $\mathcal{S}$,

$$\int_{\mathcal{A}} f(s)ds = P(\mathcal{A}).$$

In particular,

$$\int_{\mathcal{S}} f(s)ds = 1.$$

When $\mathcal{S}$ is an interval $[a, b]$ (possibly infinite), the outcomes are real numbers (such as height or weight), and you may be interested in the mean (height or weight). It is defined as

$$\mu = \int_a^b x f(x)dx,$$

and the variance is given by

$$\sigma^2 = \int_a^b (\mu - x)^2 f(x)dx.$$

Mathematically, the mean and variance are the mean and variance of the identity function $I(x) = x$, but we use the term "mean and variance of the *distribution*."

1.6 Exercises

Exercise 1.1. Given Axioms 1 to 3, prove that

$$P(\mathcal{A} \cup \mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) .$$

Exercise 1.2. Consider the experiment of rolling a red and a blue fair six-sided die. Give an example of a sample space for the experiment along with probabilities for each outcome. Suppose then that we are interested only in the sum of the dice (that is, the experiment consists in rolling the dice and adding up the numbers). Give another example of a sample space for this experiment and probabilities for the outcomes.

Exercise 1.3. Consider the experiment of flipping a fair coin, and if it lands heads, rolling a fair four-sided die, and if it lands tails, rolling a fair six-sided die. Suppose that we are interested only in the number rolled by the die, and a sample space $\mathcal{S}_A$ for the experiment could thus be the numbers from 1 to 6. Another sample space could be $\mathcal{S}_B = \{t1, \ldots, t6, h1, \ldots, h4\}$, with for example $t2$ meaning "tails and a roll of 2" and $h4$ meaning "heads and a roll of 4." Choose either $\mathcal{S}_A$ or $\mathcal{S}_B$ and associate probabilities with it. According to your sample space and probability distribution, what is the probability of rolling either 3 or 5.

Exercise 1.4. Draw a Venn diagram (like that in Figure 1.1) over $\mathcal{S}_B$ defined in Exercise 1.3. The diagram should show the events corresponding to "rolling a 3," "flipping tails," and "flipping tails and rolling a 3."

Exercise 1.5. Let $\mathcal{S}_B$ be defined as in Exercise 1.3, but with a loaded coin and loaded dice. A probability distribution is given in Table 1.15. What is the probability that the loaded coin lands "tails"? What is the conditional probability of rolling a 4, given that the coin lands tails? Which of the loaded dice has the highest chance of rolling 4 or more?

$t1$	$\frac{5}{18}$	$t6$	$\frac{1}{18}$
$t2$	$\frac{1}{9}$	$h1$	$\frac{1}{24}$
$t3$	$\frac{1}{9}$	$h2$	$\frac{1}{24}$
$t4$	$\frac{1}{18}$	$h3$	$\frac{1}{8}$
$t5$	$\frac{1}{18}$	$h4$	$\frac{1}{8}$

Table 1.15. Probabilities for $\mathcal{S}_B$ in Exercise 1.5.

Exercise 1.6. Prove that

$$P(\mathcal{A} \,|\, \mathcal{B} \cup \mathcal{C}) P(\mathcal{B} \,|\, \mathcal{C}) = P(\mathcal{A} \cap \mathcal{B} \,|\, \mathcal{C}) \ .$$

Exercise 1.7. A farmer has a cow, which he suspects is pregnant. He administers a test to the urine of the cow to determine whether it is pregnant. There are four outcomes in this experiment:

1. The cow is pregnant and the test is positive.
2. The cow is pregnant, but the test is negative.
3. The cow is not pregnant, but the test is positive.
4. The cow is not pregnant, and the test is negative.

The prior probability of the event that the cow is pregnant is 0.05, the probability of the event that the test is positive, when the cow indeed is pregnant, is 0.98 and the probability that the test is negative, when the cow is not pregnant, is 0.999. The test turns out to be positive. What is the posterior probability of the cow being pregnant?

Exercise 1.8. Consider the following two experiments: One consists in throwing a red six-sided die, and one consists in throwing a blue six-sided die. We let R be a variable representing the roll of the red die, having a set of states $\{r1, r2, r3, r4, r5, r6\}$, and B be a variable representing the roll of the blue die (states $\{b1, b2, b3, b4, b5, b6\}$). Assume that the red die is fair so that $P(R = r1) = \cdots = P(R = r6) = \frac{1}{6}$, and that the variable for the blue die has probabilities $P(B = b1) = P(B = b2) = P(B = b3) = \frac{1}{12}$ and $P(B = b4) = P(B = b5) = P(B = b6) = \frac{1}{4}$. Give an example of a sample space for an experiment consisting of throwing both the red and the blue die. Using $P(R)$ and $P(B)$, what is the probability distribution for your sample space?

Exercise 1.9. Consider the sample space $\mathcal{S}_B$ from Exercise 1.3, with probability distribution as defined in Table 1.15. Recast the sample space as variables. What is the probability distribution for each variable?

Exercise 1.10. Prove the fundamental rule for variables:

$$P(A, B) = P(A \mid B)P(B) .$$

Exercise 1.11. Calculate $P(A)$, $P(B)$, $P(A \mid B)$, and $P(B \mid A)$ from the table for $P(A, B)$ (Table 1.16).

	b_1	b_2	b_3
a_1	0.05	0.10	0.05
a_2	0.15	0.00	0.25
a_3	0.10	0.20	0.10

Table 1.16. $P(A, B)$ for Exercise 1.11.

Exercise 1.12. Table 1.17 describes a test T for an event A. The number 0.01 is the frequency of *false negatives*, and the number 0.001 is the frequency of *false positives*.

(*i*) The police can order a blood test on drivers under the suspicion of having consumed too much alcohol. The test has the above characteristics. Experience says that 20% of the drivers under suspicion do in fact drive with too much alcohol in their blood. A suspicious driver has a positive blood test. What is the probability that the driver is guilty of driving under the influence of alcohol?

(*ii*) The police block a road, take blood samples of all drivers, and use the same test. It is estimated that one out of 1,000 drivers have too much alcohol in their blood. A driver has a positive test result. What is the probability that the driver is guilty of driving under the influence of alcohol?

	$A = yes$	$A = no$
$T = yes$	0.99	0.001
$T = no$	0.01	0.999

Table 1.17. Table for Exercise 1.12. Conditional probabilities $P(T \mid A)$ characterizing test T for A.

Exercise 1.13. In Table 1.18, a joint probability table for the binary variables A, B, and C is given.

- Calculate $P(B, C)$ and $P(B)$.
- Are A and C independent given B?

	b_1	b_2
a_1	$(0.006, 0.054)$	$(0.048, 0.432)$
a_2	$(0.014, 0.126)$	$(0.032, 0.288)$

Table 1.18. $P(A, B, C)$ for Exercise 1.13.

Exercise 1.14. Write a short algorithm that given an $n \times m$ potential $\phi(A, B)$ calculates $\sum_A \phi$. Use your algorithm on the joint probability table $P(A, B)$ in Table 1.2 and on the conditional probability table $P(A|B)$ in Table 1.1.

Exercise 1.15. Prove that the associative, commutative, and distributive laws hold for potentials.

Exercise 1.16. Let $\phi(x) = ax$ be a distribution on $[0, 1]$. Determine a. What are the mean and the variance of ϕ?

Exercise 1.17. Let $\phi(x) = a \sin(x)$ be a distribution on $[0, \pi]$. Determine a and the mean of ϕ.

Part I

Probabilistic Graphical Models