1.3 CONDITIONAL PROBABILITY

Conditional probability provides us with a way to reason about the outcome of an experiment, based on **partial information**. Here are some examples of situations we have in mind:

(a) In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?

(b) In a word guessing game, the first letter of the word is a "t". What is the likelihood that the second letter is an "h"?

(c) How likely is it that a person has a certain disease given that a medical test was negative?

(d) A spot shows up on a radar screen. How likely is it to correspond to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event $B$. We wish to quantify the likelihood that the outcome also belongs to some other given event $A$. We thus seek to construct a new probability law that takes into account the available knowledge: a probability law that for any event $A$ specifies the **conditional probability of $A$ given $B$**, denoted by $P(A \mid B)$.

We would like the conditional probabilities $P(A \mid B)$ of different events $A$ to constitute a legitimate probability law, which satisfies the probability axioms. The conditional probabilities should also be consistent with our intuition in important special cases, e.g., when all possible outcomes of the experiment are equally likely. For example, suppose that all six possible outcomes of a fair die roll are equally likely. If we are told that the outcome is even, we are left with only three possible outcomes, namely, 2, 4, and 6. These three outcomes were equally likely to start with, and so they should remain equally likely given the additional knowledge that the outcome was even. Thus, it is reasonable to let

$$P(\text{the outcome is 6} \mid \text{the outcome is even}) = \frac{1}{3}.$$ 

This argument suggests that an appropriate definition of conditional probability when all outcomes are equally likely, is given by

$$P(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$ 

Generalizing the argument, we introduce the following definition of conditional probability:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$
where we assume that $P(B) > 0$; the conditional probability is undefined if the conditioning event has zero probability. In words, out of the total probability of the elements of $B$, $P(A \mid B)$ is the fraction that is assigned to possible outcomes that also belong to $A$.

**Conditional Probabilities Specify a Probability Law**

For a fixed event $B$, it can be verified that the conditional probabilities $P(A \mid B)$ form a legitimate probability law that satisfies the three axioms. Indeed, non-negativity is clear. Furthermore,

$$P(\Omega \mid B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1,$$

and the normalization axiom is also satisfied. To verify the additivity axiom, we write for any two disjoint events $A_1$ and $A_2$,

$$P(A_1 \cup A_2 \mid B) = \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)} = \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} = \frac{P(A_1)}{P(B)} + \frac{P(A_2)}{P(B)} = P(A_1 \mid B) + P(A_2 \mid B),$$

where for the third equality, we used the fact that $A_1 \cap B$ and $A_2 \cap B$ are disjoint sets, and the additivity axiom for the (unconditional) probability law. The argument for a countable collection of disjoint sets is similar.

Since conditional probabilities constitute a legitimate probability law, all general properties of probability laws remain valid. For example, a fact such as $P(A \cup C) \leq P(A) + P(C)$ translates to the new fact

$$P(A \cup C \mid B) \leq P(A \mid B) + P(C \mid B).$$

Let us also note that since we have $P(B \mid B) = P(B)/P(B) = 1$, all of the conditional probability is concentrated on $B$. Thus, we might as well discard all possible outcomes outside $B$ and treat the conditional probabilities as a probability law defined on the new universe $B$.

Let us summarize the conclusions reached so far.
Properties of Conditional Probability

- The conditional probability of an event $A$, given an event $B$ with $P(B) > 0$, is defined by

\[ P(A | B) = \frac{P(A \cap B)}{P(B)} , \]

and specifies a new (conditional) probability law on the same sample space $\Omega$. In particular, all properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe $B$, because all of the conditional probability is concentrated on $B$.

- If the possible outcomes are finitely many and equally likely, then

\[ P(A | B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B} . \]

Example 1.6. We toss a fair coin three successive times. We wish to find the conditional probability $P(A | B)$ when $A$ and $B$ are the events

$A = \{\text{more heads than tails come up}\}, \quad B = \{\text{1st toss is a head}\}.$

The sample space consists of eight sequences.

$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

which we assume to be equally likely. The event $B$ consists of the four elements $HHH, HHT, HTH, HTT$, so its probability is

$P(B) = \frac{4}{8}.$

The event $A \cap B$ consists of the three elements $HHH, HHT, HTH$, so its probability is

$P(A \cap B) = \frac{3}{8}.$

Thus, the conditional probability $P(A | B)$ is

$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$

Because all possible outcomes are equally likely here, we can also compute $P(A | B)$ using a shortcut. We can bypass the calculation of $P(B)$ and $P(A \cap B)$, and simply
divides the number of elements shared by $A$ and $B$ (which is 3) with the number of elements of $B$ (which is 4): the same result $3/4$.

**Example 1.7.** A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let $X$ and $Y$ be the result of the 1st and the 2nd roll, respectively. We wish to determine the conditional probability $P(A|B)$, where

$$A = \{ \max(X, Y) = m \}. \quad B = \{ \min(X, Y) = 2 \}. $$

and $m$ takes each of the values 1, 2, 3, 4.

As in the preceding example, we can first determine the probabilities $P(A \cap B)$ and $P(B)$ by dividing by 16. Alternatively, we can directly divide the number of elements of $A \cap B$ with the number of elements of $B$: see Fig. 1.8.

![Sample space of an experiment involving two rolls of a 4-sided die.](image)

Figure 1.8: Sample space of an experiment involving two rolls of a 4-sided die. (cf. Example 1.7). The conditioning event $B = \{ \min(X, Y) = 2 \}$ consists of the 2-element shaded set. The set $A = \{ \max(X, Y) = m \}$ shares with $B$ two elements if $m = 3$ or $m = 4$, one element if $m = 2$, and no element if $m = 1$. Thus, we have

$$P(\{ \max(X, Y) = m \} | B) = \begin{cases} 
2/5, & \text{if } m = 3 \text{ or } m = 4. \\
1/5, & \text{if } m = 2. \\
0, & \text{if } m = 1.
\end{cases}$$

**Example 1.8.** A conservative design team, call it $C$, and an innovative design team, call it $N$, are asked to separately design a new product within a month. From past experience we know that:

(a) The probability that team $C$ is successful is 2/3.

(b) The probability that team $N$ is successful is 1/2.

(c) The probability that at least one team is successful is 3/4.
Assuming that exactly one successful design is produced, what is the probability that it was designed by team N?

There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams:

\[ SS: \text{ both succeed,} \quad FF: \text{ both fail,} \]
\[ SF: C \text{ succeeds, } N \text{ fails,} \quad FS: C \text{ fails, } N \text{ succeeds.} \]

We were given that the probabilities of these outcomes satisfy

\[ P(SS) + P(SF) = \frac{2}{3}, \quad P(SS) + P(FS) = \frac{1}{2}, \quad P(SS) + P(SF) + P(FS) = \frac{3}{4}. \]

From these relations, together with the normalization equation

\[ P(SS) + P(SF) + P(FS) + P(FF) = 1, \]

we can obtain the probabilities of individual outcomes:

\[ P(SS) = \frac{5}{12}, \quad P(SF) = \frac{1}{4}, \quad P(FS) = \frac{1}{12}, \quad P(FF) = \frac{1}{4}. \]

The desired conditional probability is

\[ P(FS \mid \{SF, FS\}) = \frac{1}{\frac{1}{4} + \frac{1}{12}} = \frac{4}{6} = \frac{1}{4}. \]

Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities. The rule \[ P(A \cap B) = P(B)P(A \mid B), \] which is a restatement of the definition of conditional probability, is often helpful in this process.

**Example 1.9. Radar Detection.** If an aircraft is present in a certain area, a radar detects it and generates an alarm signal with probability 0.99. If an aircraft is not present, the radar generates a (false) alarm, with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of no aircraft presence and a false alarm? What is the probability of aircraft presence and no detection?

A sequential representation of the experiment is appropriate here, as shown in Fig. 1.9. Let \( A \) and \( B \) be the events

\[ A = \{ \text{an aircraft is present}\}, \]
\[ B = \{ \text{the radar generates an alarm}\}, \]
and consider also their complements

\[ A^c = \{ \text{an aircraft is not present} \} \]
\[ B^c = \{ \text{the radar does not generate an alarm} \} . \]

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.9. Each possible outcome corresponds to a leaf of the tree, and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities are

\[
P(\text{not present, false alarm}) = P(A^c \cap B) = P(A^c)P(B \mid A^c) = 0.95 \cdot 0.10 = 0.095,
\]
\[
P(\text{present, no detection}) = P(A \cap B^c) = P(A)P(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005.
\]

![Tree Diagram](tree-diagram.png)

**Figure 1.9:** Sequential description of the experiment for the radar detection problem in Example 1.9.

Extending the preceding example, we have a general rule for calculating various probabilities in conjunction with a tree-based sequential description of an experiment. In particular:

(a) We set up the tree so that an event of interest is associated with a leaf. We view the occurrence of the event as a sequence of steps, namely, the traversals of the branches along the path from the root to the leaf.

(b) We record the conditional probabilities associated with the branches of the tree.

(c) We obtain the probability of a leaf by multiplying the probabilities recorded along the corresponding path of the tree.
and by using the definition of conditional probability to rewrite the right-hand side above as

\[ P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid \cap_{i=1}^{n-1} A_i). \]

For the case of just two events, \( A_1 \) and \( A_2 \), the multiplication rule is simply the definition of conditional probability.

**Example 1.10.** Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn. A cumbersome approach, which we will not use, is to count the number of all card triplets that do not include a heart, and divide it with the number of all possible card triplets. Instead, we use a sequential description of the experiment in conjunction with the multiplication rule (cf. Fig. 1.11).

Define the events

\[ A_i = \{ \text{the } i\text{th card is not a heart} \}. \quad i = 1, 2, 3. \]

We will calculate \( P(A_1 \cap A_2 \cap A_3) \), the probability that none of the three cards is a heart, using the multiplication rule

\[ P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2). \]

We have

\[ P(A_1) = \frac{39}{52}, \]

since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards, 38 of which are not hearts, and

\[ P(A_2 \mid A_1) = \frac{38}{51}. \]

Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck, and

\[ P(A_3 \mid A_1 \cap A_2) = \frac{37}{50}. \]

These probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.11. The desired probability is now obtained by multiplying the probabilities recorded along the corresponding path of the tree:

\[ P(A_1 \cap A_2 \cap A_3) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}. \]
Note that once the probabilities are recorded along the tree, the probability of several other events can be similarly calculated. For example,

\[
\begin{align*}
P(1\text{st is not a heart and 2nd is a heart}) &= \frac{39}{52} \cdot \frac{13}{51}, \\
P(1\text{st and 2nd are not hearts, and 3rd is a heart}) &= \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.
\end{align*}
\]

**Example 1.11.** A class consisting of 4 graduate and 12 undergraduate students is randomly divided into 4 groups of 4. What is the probability that each group includes a graduate student? We interpret "randomly" to mean that given the assignment of some students to certain slots, any of the remaining students is equally likely to be assigned to any of the remaining slots. We then calculate the desired probability using the multiplication rule, based on the sequential description shown in Fig. 1.12. Let us denote the four graduate students by 1, 2, 3, 4, and consider the events

\[
\begin{align*}
A_1 &= \{\text{students 1 and 2 are in different groups}\}, \\
A_2 &= \{\text{students 1, 2, and 3 are in different groups}\}, \\
A_3 &= \{\text{students 1, 2, 3, and 4 are in different groups}\}.
\end{align*}
\]

We will calculate \(P(A_3)\) using the multiplication rule:

\[
P(A_3) = P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2).
\]

We have

\[
P(A_1) = \frac{12}{15},
\]

since there are 12 student slots in groups other than the one of student 1, and there are 15 student slots overall, excluding student 1. Similarly,

\[
P(A_2 \mid A_1) = \frac{8}{14},
\]
since there are 8 student slots in groups other than those of students 1 and 2, and there are 14 student slots, excluding students 1 and 2. Also,

\[ P(A_3 \mid A_1 \cap A_2) = \frac{4}{13}, \]

since there are 4 student slots in groups other than those of students 1, 2, and 3, and there are 13 student slots, excluding students 1, 2, and 3. Thus, the desired probability is

\[ \frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13}, \]

and is obtained by multiplying the conditional probabilities along the corresponding path of the tree in Fig. 1.12.

Example 1.12. The Monty Hall Problem. This is a much discussed puzzle, based on an old American game show. You are told that a prize is equally likely to be found behind any one of three closed doors in front of you. You point to one of the doors. A friend opens for you one of the remaining two doors, after making sure that the prize is not behind it. At this point, you can stick to your initial choice, or switch to the other unopened door. You win the prize if it lies behind your final choice of a door. Consider the following strategies:

(a) Stick to your initial choice.
(b) Switch to the other unopened door.
(c) You first point to door 1. If door 2 is opened, you do not switch. If door 3 is opened, you switch.

Which is the best strategy? To answer the question, let us calculate the probability of winning under each of the three strategies.

Under the strategy of no switching, your initial choice will determine whether you win or not, and the probability of winning is 1/3. This is because the prize is equally likely to be behind each door.

Under the strategy of switching, if the prize is behind the initially chosen door (probability 1/3), you do not win. If it is not (probability 2/3), and given that
another door without a prize has been opened for you, you will get to the winning door once you switch. Thus, the probability of winning is now 2/3, so (b) is a better strategy than (a).

Consider now strategy (c). Under this strategy, there is insufficient information for determining the probability of winning. The answer depends on the way that your friend chooses which door to open. Let us consider two possibilities.

Suppose that if the prize is behind door 1, your friend always chooses to open door 2. (If the prize is behind door 2 or 3, your friend has no choice.) If the prize is behind door 1, your friend opens door 2, you do not switch, and you win. If the prize is behind door 2, your friend opens door 3, you switch, and you win. If the prize is behind door 3, your friend opens door 2, you do not switch, and you lose. Thus, the probability of winning is 2/3. so strategy (c) in this case is as good as strategy (b).

Suppose now that if the prize is behind door 1, your friend is equally likely to open either door 2 or 3. If the prize is behind door 1 (probability 1/3), and if your friend opens door 2 (probability 1/2), you do not switch and you win (probability 1/6). But if your friend opens door 3, you switch and you lose. If the prize is behind door 2, your friend opens door 3, you switch, and you win (probability 1/3). If the prize is behind door 3, your friend opens door 2, you do not switch and you lose. Thus, the probability of winning is 1/6 + 1/3 = 1/2, so strategy (c) in this case is inferior to strategy (b).

1.4 TOTAL PROBABILITY THEOREM AND BAYES’ RULE

In this section, we explore some applications of conditional probability. We start with the following theorem, which is often useful for computing the probabilities of various events, using a “divide-and-conquer” approach.

**Total Probability Theorem**

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, \ldots, A_n$) and assume that $P(A_i) > 0$, for all $i$. Then, for any event $B$, we have

$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B)$$

$$= P(A_1)P(B | A_1) + \cdots + P(A_n)P(B | A_n).$$

The theorem is visualized and proved in Fig. 1.13. Intuitively, we are partitioning the sample space into a number of scenarios (events) $A_i$. Then, the probability that $B$ occurs is a weighted average of its conditional probability under each scenario, where each scenario is weighted according to its (unconditional) probability. One of the uses of the theorem is to compute the probability of various events $B$ for which the conditional probabilities $P(B | A_i)$ are known or
easy to derive. The key is to choose appropriately the partition $A_1 \ldots A_n$, and this choice is often suggested by the problem structure. Here are some examples.

Figure 1.13: Visualization and verification of the total probability theorem. The events $A_1, \ldots, A_n$ form a partition of the sample space, so the event $B$ can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets $A_i$, i.e.,

$$B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B).$$

Using the additivity axiom, it follows that

$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B).$$

Since, by the definition of conditional probability, we have

$$P(A_i \cap B) = P(A_i)P(B \mid A_i),$$

the preceding equality yields

$$P(B) = P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n).$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability of the leaf $A_i \cap B$ is the product $P(A_i)P(B \mid A_i)$ of the probabilities along the path leading to that leaf. The event $B$ consists of the three highlighted leaves and $P(B)$ is obtained by adding their probabilities.

Example 1.13. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

Let $A_i$ be the event of playing with an opponent of type $i$. We have

$$P(A_1) = 0.5, \quad P(A_2) = 0.25, \quad P(A_3) = 0.25.$$

Also, let $B$ be the event of winning. We have

$$P(B \mid A_1) = 0.3, \quad P(B \mid A_2) = 0.4, \quad P(B \mid A_3) = 0.5.$$
Thus, by the total probability theorem, the probability of winning is

\[
P(B) = P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + P(A_3)P(B | A_3) \\
= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 \\
= 0.375.
\]

**Example 1.14.** You roll a fair four-sided die. If the result is 1 or 2, you roll once more but otherwise, you stop. What is the probability that the sum total of your rolls is at least 4?

Let \( A_i \) be the event that the result of first roll is \( i \), and note that \( P(A_i) = 1/4 \) for each \( i \). Let \( B \) be the event that the sum total is at least 4. Given the event \( A_1 \), the sum total will be at least 4 if the second roll results in 3 or 4, which happens with probability \( 1/2 \). Similarly, given the event \( A_2 \), the sum total will be at least 4 if the second roll results in 2, 3, or 4, which happens with probability \( 3/4 \). Also, given the event \( A_3 \), you stop and the sum total remains below 4. Therefore,

\[
P(B | A_1) = \frac{1}{2}, \quad P(B | A_2) = \frac{3}{4}, \quad P(B | A_3) = 0, \quad P(B | A_4) = 1.
\]

By the total probability theorem,

\[
P(B) = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{9}{16}.
\]

The total probability theorem can be applied repeatedly to calculate probabilities in experiments that have a sequential character, as shown in the following example.

**Example 1.15.** Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.4 (or 0.6, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Let \( U_i \) and \( B_i \) be the events that Alice is up-to-date or behind, respectively, after \( i \) weeks. According to the total probability theorem, the desired probability \( P(U_3) \) is given by

\[
P(U_3) = P(U_2)P(U_3 | U_2) + P(B_2)P(U_3 | B_2) = P(U_2) \cdot 0.8 + P(B_2) \cdot 0.4.
\]

The probabilities \( P(U_2) \) and \( P(B_2) \) can also be calculated using the total probability theorem:

\[
P(U_2) = P(U_1)P(U_2 | U_1) + P(B_1)P(U_2 | B_1) = P(U_1) \cdot 0.8 + P(B_1) \cdot 0.4,
\]

\[
P(B_2) = P(U_1)P(B_2 | U_1) + P(B_1)P(B_2 | B_1) = P(U_1) \cdot 0.2 + P(B_1) \cdot 0.6.
\]
Finally, since Alice starts her class up-to-date, we have
\[ P(U_1) = 0.8, \quad P(B_1) = 0.2. \]
We can now combine the preceding three equations to obtain
\[ P(U_2) = 0.8 \cdot 0.8 + 0.2 \cdot 0.4 = 0.72, \]
\[ P(B_2) = 0.8 \cdot 0.2 + 0.2 \cdot 0.6 = 0.28, \]
and by using the above probabilities in the formula for \( P(U_3) \):
\[ P(U_3) = 0.72 \cdot 0.8 + 0.28 \cdot 0.4 = 0.688. \]

Note that we could have calculated the desired probability \( P(U_3) \) by constructing a tree description of the experiment, then calculating the probability of every element of \( U_3 \) using the multiplication rule on the tree, and adding. However, there are cases where the calculation based on the total probability theorem is more convenient. For example, suppose we are interested in the probability \( P(U_{20}) \) that Alice is up-to-date after 20 weeks. Calculating this probability using the multiplication rule is very cumbersome, because the tree representing the experiment is 20 stages deep and has \( 2^{20} \) leaves. On the other hand, with a computer, a sequential calculation using the total probability formulas
\[ P(U_{i+1}) = P(U_i) \cdot 0.8 + P(B_i) \cdot 0.4. \]
\[ P(B_{i+1}) = P(U_i) \cdot 0.2 + P(B_i) \cdot 0.6. \]
and the initial conditions \( P(U_1) = 0.8, \ P(B_1) = 0.2 \) is very simple.

**Inference and Bayes' Rule**

The total probability theorem is often used in conjunction with the following celebrated theorem, which relates conditional probabilities of the form \( P(A \mid B) \) with conditional probabilities of the form \( P(B \mid A) \), in which the order of the conditioning is reversed.

**Bayes' Rule**

Let \( A_1, A_2, \ldots, A_n \) be disjoint events that form a partition of the sample space, and assume that \( P(A_i) > 0 \), for all \( i \). Then, for any event \( B \) such that \( P(B) > 0 \), we have
\[
P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)}.
\]
Figure 1.14: An example of the inference context that is implicit in Bayes’ rule. We observe a shade in a person’s X-ray (this is event $B$, the “effect”) and we want to estimate the likelihood of three mutually exclusive and collectively exhaustive potential causes: cause 1 (event $A_1$) is that there is a malignant tumor, cause 2 (event $A_2$) is that there is a nonmalignant tumor, and cause 3 (event $A_3$) corresponds to reasons other than a tumor. We assume that we know the probabilities $P(A_i)$ and $P(B|A_i)$, $i = 1, 2, 3$. Given that we see a shade (event $B$ occurs), Bayes’ rule gives the posterior probabilities of the various causes as

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + P(A_3)P(B | A_3)}, \quad i = 1, 2, 3.$$ 

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability $P(A_1 | B)$ of a malignant tumor is the probability of the first highlighted leaf, which is $P(A_1 \cap B)$, divided by the total probability of the highlighted leaves, which is $P(B)$.

To verify Bayes’ rule, note that by the definition of conditional probability, we have

$$P(A_i \cap B) = P(A_i)P(B | A_i) = P(A_i | B)P(B).$$

This yields the first equality. The second equality follows from the first by using the total probability theorem to rewrite $P(B)$.

Bayes’ rule is often used for inference. There are a number of “causes” that may result in a certain “effect.” We observe the effect, and we wish to infer the cause. The events $A_1, \ldots, A_n$ are associated with the causes and the event $B$ represents the effect. The probability $P(B | A_i)$ that the effect will be observed when the cause $A_i$ is present amounts to a probabilistic model of the cause-effect relation (cf. Fig. 1.14). Given that the effect $B$ has been observed, we wish to evaluate the probability $P(A_i | B)$ that the cause $A_i$ is present. We refer to $P(A_i | B)$ as the posterior probability of event $A_i$ given the information, to be distinguished from $P(A_i)$, which we call the prior probability.
Example 1.16. Let us return to the radar detection problem of Example 1.9 and Fig. 1.9. Let

\[ A = \{ \text{an aircraft is present} \}, \]
\[ B = \{ \text{the radar generates an alarm} \}. \]

We are given that

\[ P(A) = 0.05, \quad P(B \mid A) = 0.99, \quad P(B \mid A^c) = 0.1. \]

Applying Bayes' rule, with \( A_1 = A \) and \( A_2 = A^c \), we obtain

\[
P(\text{aircraft present} \mid \text{alarm}) = P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)} = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)}
\]

\[= \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.95 \cdot 0.1} \approx 0.3426. \]

Example 1.17. Let us return to the chess problem of Example 1.13. Here, \( A_i \) is the event of getting an opponent of type \( i \), and

\[ P(A_1) = 0.5, \quad P(A_2) = 0.25, \quad P(A_3) = 0.25. \]

Also, \( B \) is the event of winning, and

\[ P(B \mid A_1) = 0.3, \quad P(B \mid A_2) = 0.4, \quad P(B \mid A_3) = 0.5. \]

Suppose that you win. What is the probability \( P(A_1 \mid B) \) that you had an opponent of type 1?

Using Bayes' rule, we have

\[
P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + P(A_3)P(B \mid A_3)}
\]

\[= \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5} = 0.4. \]

Example 1.18. The False-Positive Puzzle. A test for a certain rare disease is assumed to be correct 95% of the time: if a person has the disease, the test results are positive with probability 0.95, and if the person does not have the disease, the test results are negative with probability 0.95. A random person drawn from
a certain population has probability 0.001 of having the disease. Given that the person just tested positive, what is the probability of having the disease?

If \( A \) is the event that the person has the disease, and \( B \) is the event that the test results are positive, the desired probability. \( P(A \mid B) \), is

\[
P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)}
\]

\[
= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05}
\]

\[
= 0.0187.
\]

Note that even though the test was assumed to be fairly accurate, a person who has tested positive is still very unlikely (less than 2%) to have the disease. According to *The Economist* (February 20th, 1999), 80% of those questioned at a leading American hospital substantially missed the correct answer to a question of this type; most of them thought that the probability that the person has the disease is 0.95!

### 1.5 Independence

We have introduced the conditional probability \( P(A \mid B) \) to capture the partial information that event \( B \) provides about event \( A \). An interesting and important special case arises when the occurrence of \( B \) provides no such information and does not alter the probability that \( A \) has occurred, i.e.,

\[
P(A \mid B) = P(A).
\]

When the above equality holds, we say that \( A \) is independent of \( B \). Note that by the definition \( P(A \mid B) = P(A \cap B) / P(B) \), this is equivalent to

\[
P(A \cap B) = P(A)P(B).
\]

We adopt this latter relation as the definition of independence because it can be used even when \( P(B) = 0 \), in which case \( P(A \mid B) \) is undefined. The symmetry of this relation also implies that independence is a symmetric property; that is, if \( A \) is independent of \( B \), then \( B \) is independent of \( A \), and we can unambiguously say that \( A \) and \( B \) are independent events.

Independence is often easy to grasp intuitively. For example, if the occurrence of two events is governed by distinct and noninteracting physical processes, such events will turn out to be independent. On the other hand, independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events \( A \) and \( B \) with \( P(A) > 0 \) and \( P(B) > 0 \) are never independent, since their intersection \( A \cap B \) is empty and has probability 0.
For example, an event \( A \) and its complement \( A^c \) are not independent \([\text{unless } P(A) = 0 \text{ or } P(A) = 1]\), since knowledge that \( A \) has occurred provides precise information about whether \( A^c \) has occurred.

**Example 1.19.** Consider an experiment involving two successive rolls of a 4-sided die in which all 16 possible outcomes are equally likely and have probability 1/16.

(a) Are the events

\[ A_i = \{1\text{st roll results in } i\}, \quad B_j = \{2\text{nd roll results in } j\}, \]

independent? We have

\[
P(A_i \cap B_j) = P(\text{the outcome of the two rolls is } (i, j)) = \frac{1}{16},
\]

\[
P(A_i) = \frac{\text{number of elements of } A_i}{\text{total number of possible outcomes}} = \frac{4}{16},
\]

\[
P(B_j) = \frac{\text{number of elements of } B_j}{\text{total number of possible outcomes}} = \frac{4}{16}.
\]

We observe that \( P(A_i \cap B_j) = P(A_i)P(B_j) \), and the independence of \( A_i \) and \( B_j \) is verified. Thus, our choice of the discrete uniform probability law implies the independence of the two rolls.

(b) Are the events

\[ A = \{1\text{st roll is a } 1\}, \quad B = \{\text{sum of the two rolls is a } 5\}, \]

independent? The answer here is not quite obvious. We have

\[
P(A \cap B) = P(\text{the result of the two rolls is } (1, 4)) = \frac{1}{16},
\]

and also

\[
P(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{4}{16}.
\]

The event \( B \) consists of the outcomes (1, 4), (2, 3), (3, 2), and (4, 1). and

\[
P(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{4}{16}.
\]

Thus, we see that \( P(A \cap B) = P(A)P(B) \), and the events \( A \) and \( B \) are independent.

(c) Are the events

\[ A = \{\text{maximum of the two rolls is } 2\}, \quad B = \{\text{minimum of the two rolls is } 2\}, \]
independent? Intuitively, the answer is "no" because the minimum of the two rolls conveys some information about the maximum. For example, if the minimum is 2, the maximum cannot be 1. More precisely, to verify that $A$ and $B$ are not independent, we calculate

$$P(A \cap B) = P(\text{the result of the two rolls is } (2,2)) = \frac{1}{16},$$

and also

$$P(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{3}{16},$$

$$P(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{5}{16}.$$

We have $P(A)P(B) = 15/(16)^2$, so that $P(A \cap B) \neq P(A)P(B)$, and $A$ and $B$ are not independent.

We finally note that, as mentioned earlier, if $A$ and $B$ are independent, the occurrence of $B$ does not provide any new information on the probability of $A$ occurring. It is then intuitive that the non-occurrence of $B$ should also provide no information on the probability of $A$. Indeed, it can be verified that if $A$ and $B$ are independent, the same holds true for $A$ and $B^c$ (see the end-of-chapter problems).

**Conditional Independence**

We noted earlier that the conditional probabilities of events, conditioned on a particular event, form a legitimate probability law. We can thus talk about independence of various events with respect to this conditional law. In particular, given an event $C$, the events $A$ and $B$ are called **conditionally independent** if

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

To derive an alternative characterization of conditional independence, we use the definition of the conditional probability and the multiplication rule, to write

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

We now compare the preceding two expressions, and after eliminating the common factor $P(B \mid C)$, assumed nonzero, we see that conditional independence is the same as the condition

$$P(A \mid B \cap C) = P(A \mid C).$$
In words, this relation states that if \( C \) is known to have occurred, the additional knowledge that \( B \) also occurred does not change the probability of \( A \).

Interestingly, independence of two events \( A \) and \( B \) with respect to the unconditional probability law does not imply conditional independence, and vice versa, as illustrated by the next two examples.

**Example 1.20.** Consider two independent fair coin tosses, in which all four possible outcomes are equally likely. Let

\[
H_1 = \{1\text{st toss is a head}\}, \\
H_2 = \{2\text{nd toss is a head}\}, \\
D = \{\text{the two tosses have different results}\}.
\]

The events \( H_1 \) and \( H_2 \) are (unconditionally) independent. But

\[
P(H_1 \mid D) = \frac{1}{2}, \quad P(H_2 \mid D) = \frac{1}{2}, \quad P(H_1 \cap H_2 \mid D) = 0,
\]

so that \( P(H_1 \cap H_2 \mid D) \neq P(H_1 \mid D)P(H_2 \mid D) \), and \( H_1, H_2 \) are not conditionally independent.

This example can be generalized. For any probabilistic model, let \( A \) and \( B \) be independent events, and let \( C \) be an event such that \( P(C) > 0 \), \( P(A \mid C) > 0 \), and \( P(B \mid C) > 0 \), while \( A \cap B \cap C \) is empty. Then, \( A \) and \( B \) cannot be conditionally independent (given \( C \)) since \( P(A \cap B \mid C) = 0 \) while \( P(A \mid C)P(B \mid C) > 0 \).

**Example 1.21.** There are two coins, a blue and a red one. We choose one of the two at random, each being chosen with probability 1/2, and proceed with two independent tosses. The coins are biased: with the blue coin, the probability of heads in any given toss is 0.99, whereas for the red coin it is 0.01.

Let \( B \) be the event that the blue coin was selected. Let also \( H_i \) be the event that the \( i \)th toss resulted in heads. Given the choice of a coin, the events \( H_1 \) and \( H_2 \) are independent because of our assumption of independent tosses. Thus,

\[
P(H_1 \cap H_2 \mid B) = P(H_1 \mid B)P(H_2 \mid B) = 0.99 \cdot 0.99.
\]

On the other hand, the events \( H_1 \) and \( H_2 \) are not independent. Intuitively, if we are told that the first toss resulted in heads, this leads us to suspect that the blue coin was selected, in which case, we expect the second toss to also result in heads. Mathematically, we use the total probability theorem to obtain

\[
P(H_1) = P(B)P(H_1 \mid B) + P(B^c)P(H_1 \mid B^c) = \frac{1}{2} \cdot 0.99 + \frac{1}{2} \cdot 0.01 = \frac{1}{2},
\]

as should be expected from symmetry considerations. Similarly, we have \( P(H_2) = 1/2 \). Now notice that

\[
P(H_1 \cap H_2) = P(B)P(H_1 \cap H_2 \mid B) + P(B^c)P(H_1 \cap H_2 \mid B^c)
\]

\[
= \frac{1}{2} \cdot 0.99 \cdot 0.99 + \frac{1}{2} \cdot 0.01 \cdot 0.01 \approx \frac{1}{2}.
\]
Thus, $\mathbf{P}(H_1 \cap H_2) \neq \mathbf{P}(H_1)\mathbf{P}(H_2)$, and the events $H_1$ and $H_2$ are dependent, even though they are conditionally independent given $B$.

We now summarize.

**Independence**

- Two events $A$ and $B$ are said to be **independent** if
  \[
  \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).
  \]
  If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition
  \[
  \mathbf{P}(A \mid B) = \mathbf{P}(A).
  \]

- If $A$ and $B$ are independent, so are $A$ and $B^c$.

- Two events $A$ and $B$ are said to be **conditionally independent**, given another event $C$ with $\mathbf{P}(C) > 0$, if
  \[
  \mathbf{P}(A \cap B \mid C) = \mathbf{P}(A \mid C)\mathbf{P}(B \mid C).
  \]
  If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition
  \[
  \mathbf{P}(A \mid B \cap C) = \mathbf{P}(A \mid C).
  \]

- Independence does not imply conditional independence, and vice versa.