

QUANTIFYING UNCERTAINTY

In which we see how to tame uncertainty with numeric degrees of belief.

12.1 Acting under Uncertainty

Agents in the real world need to handle **uncertainty**, whether due to partial observability, nondeterminism, or adversaries. An agent may never know for sure what state it is in now or where it will end up after a sequence of actions.

Uncertainty

We have seen problem-solving and logical agents handle uncertainty by keeping track of a **belief state**—a representation of the set of all possible world states that it might be in—and generating a contingency plan that handles every possible eventuality that its sensors may report during execution. This approach works on simple problems, but it has drawbacks:

- The agent must consider *every possible* explanation for its sensor observations, no matter how unlikely. This leads to a large belief-state full of unlikely possibilities.
- A correct contingent plan that handles every eventuality can grow arbitrarily large and must consider arbitrarily unlikely contingencies.
- Sometimes there is no plan that is guaranteed to achieve the goal—yet the agent must act. It must have some way to compare the merits of plans that are not guaranteed.

Suppose, for example, that an automated taxi has the goal of delivering a passenger to the airport on time. The taxi forms a plan, A_{90} , that involves leaving home 90 minutes before the flight departs and driving at a reasonable speed. Even though the airport is only 5 miles away, a logical agent will not be able to conclude with absolute certainty that “Plan A_{90} will get us to the airport in time.” Instead, it reaches the weaker conclusion “Plan A_{90} will get us to the airport in time, as long as the car doesn’t break down, and I don’t get into an accident, and the road isn’t closed, and no meteorite hits the car, and” None of these conditions can be deduced for sure, so we can’t infer that the plan succeeds. This is the logical **qualification problem** (page 259), for which we so far have seen no real solution.

Nonetheless, in some sense A_{90} *is* in fact the right thing to do. What do we mean by this? As we discussed in Chapter 2, we mean that out of all the plans that could be executed, A_{90} is expected to maximize the agent’s performance measure (where the expectation is relative to the agent’s knowledge about the environment). The performance measure includes getting to the airport in time for the flight, avoiding a long, unproductive wait at the airport, and avoiding speeding tickets along the way. The agent’s knowledge cannot guarantee any of these outcomes for A_{90} , but it can provide some degree of belief that they will be achieved. Other plans, such as A_{180} , might increase the agent’s belief that it will get to the airport on time, but also increase the likelihood of a long, boring wait. *The right thing to do—the rational decision—therefore depends on both the relative importance of various goals and*



the likelihood that, and degree to which, they will be achieved. The remainder of this section hones these ideas, in preparation for the development of the general theories of uncertain reasoning and rational decisions that we present in this and subsequent chapters.

12.1.1 Summarizing uncertainty

Let's consider an example of uncertain reasoning: diagnosing a dental patient's toothache. Diagnosis—whether for medicine, automobile repair, or whatever—almost always involves uncertainty. Let us try to write rules for dental diagnosis using propositional logic, so that we can see how the logical approach breaks down. Consider the following simple rule:

$$\text{Toothache} \Rightarrow \text{Cavity}.$$

The problem is that this rule is wrong. Not all patients with toothaches have cavities; some of them have gum disease, an abscess, or one of several other problems:

$$\text{Toothache} \Rightarrow \text{Cavity} \vee \text{GumProblem} \vee \text{Abscess} \dots$$

Unfortunately, in order to make the rule true, we have to add an almost unlimited list of possible problems. We could try turning the rule into a causal rule:

$$\text{Cavity} \Rightarrow \text{Toothache}.$$

But this rule is not right either; not all cavities cause pain. The only way to fix the rule is to make it logically exhaustive: to augment the left-hand side with all the qualifications required for a cavity to cause a toothache. Trying to use logic to cope with a domain like medical diagnosis thus fails for three main reasons:

- **Laziness:** It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule and too hard to use such rules.
- **Theoretical ignorance:** Medical science has no complete theory for the domain.
- **Practical ignorance:** Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.

The connection between toothaches and cavities is not a strict logical consequence in either direction. This is typical of the medical domain, as well as most other judgmental domains: law, business, design, automobile repair, gardening, dating, and so on. The agent's knowledge can at best provide only a **degree of belief** in the relevant sentences. Our main tool for dealing with degrees of belief is **probability theory**. In the terminology of Section 8.1, the **ontological commitments** of logic and probability theory are the same—that the world is composed of facts that do or do not hold in any particular case—but the **epistemological commitments** are different: a logical agent believes each sentence to be true or false or has no opinion, whereas a probabilistic agent may have a numerical degree of belief between 0 (for sentences that are certainly false) and 1 (certainly true).

The theory of *probability provides a way of summarizing the uncertainty that comes from our laziness and ignorance*, thereby solving the qualification problem. We might not know for sure what afflicts a particular patient, but we believe that there is, say, an 80% chance—that is, a probability of 0.8—that the patient who has a toothache has a cavity. That is, we expect that out of all the situations that are indistinguishable from the current situation as far as our knowledge goes, the patient will have a cavity in 80% of them. This belief could be derived from statistical data—80% of the toothache patients seen so far have had cavities—or from some general dental knowledge, or from a combination of evidence sources.

Laziness

Theoretical
ignorance

Practical ignorance

Degree of belief

Probability theory

One confusing point is that at the time of our diagnosis, there is no uncertainty in the actual world: the patient either has a cavity or doesn't. So what does it mean to say the probability of a cavity is 0.8? Shouldn't it be either 0 or 1? The answer is that probability statements are made with respect to a knowledge state, not with respect to the real world. We say "The probability that the patient has a cavity, *given that she has a toothache*, is 0.8." If we later learn that the patient has a history of gum disease, we can make a different statement: "The probability that the patient has a cavity, *given that she has a toothache and a history of gum disease*, is 0.4." If we gather further conclusive evidence against a cavity, we can say "The probability that the patient has a cavity, *given all we now know*, is almost 0." Note that these statements do not contradict each other; each is a separate assertion about a different knowledge state.

12.1.2 Uncertainty and rational decisions

Consider again the A_{90} plan for getting to the airport. Suppose it gives us a 97% chance of catching our flight. Does this mean it is a rational choice? Not necessarily: there might be other plans, such as A_{180} , with higher probabilities. If it is *vital* not to miss the flight, then it is worth risking the longer wait at the airport. What about A_{1440} , a plan that involves leaving home 24 hours in advance? In most circumstances, this is not a good choice, because although it almost guarantees getting there on time, it involves an intolerable wait—not to mention a possibly unpleasant diet of airport food.

To make such choices, an agent must first have **preferences** among the different possible **outcomes** of the various plans. An outcome is a completely specified state, including such factors as whether the agent arrives on time and the length of the wait at the airport. We use **utility theory** to represent preferences and reason quantitatively with them. (The term **utility** is used here in the sense of "the quality of being useful," not in the sense of the electric company or water works.) Utility theory says that every state (or state sequence) has a degree of usefulness, or utility, to an agent and that the agent will prefer states with higher utility.

The utility of a state is relative to an agent. For example, the utility of a state in which White has checkmated Black in a game of chess is obviously high for the agent playing White, but low for the agent playing Black. But we can't go strictly by the scores of 1, 1/2, and 0 that are dictated by the rules of tournament chess—some players (including the authors) might be thrilled with a draw against the world champion, whereas other players (including the former world champion) might not. There is no accounting for taste or preferences: you might think that an agent who prefers jalapeño bubble-gum ice cream to chocolate chip is odd, but you could not say the agent is irrational. A utility function can account for any set of preferences—quirky or typical, noble or perverse. Note that utilities can account for altruism, simply by including the welfare of others as one of the factors.

Preferences, as expressed by utilities, are combined with probabilities in the general theory of rational decisions called **decision theory**:

Decision theory = probability theory + utility theory.

The fundamental idea of decision theory is that *an agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action*. This is called the principle of **maximum expected utility (MEU)**. Here, "expected" means the "average," or "statistical mean" of the outcome utilities, weighted by the probability of the outcome. We saw this principle in action in Chapter 6 when we touched

Preference

Outcome

Utility theory

Decision theory



Maximum expected utility (MEU)

```

function DT-AGENT(percept) returns an action
  persistent: belief_state, probabilistic beliefs about the current state of the world
               action, the agent's action

  update belief_state based on action and percept
  calculate outcome probabilities for actions,
    given action descriptions and current belief_state
  select action with highest expected utility
    given probabilities of outcomes and utility information
  return action

```

Figure 12.1 A decision-theoretic agent that selects rational actions.

briefly on optimal decisions in backgammon; it is in fact a completely general principle for single-agent decision making.

Figure 12.1 sketches the structure of an agent that uses decision theory to select actions. The agent is identical, at an abstract level, to the agents described in Chapters 4 and 7 that maintain a belief state reflecting the history of percepts to date. The primary difference is that the decision-theoretic agent's belief state represents not just the *possibilities* for world states but also their *probabilities*. Given the belief state and some knowledge of the effects of actions, the agent can make probabilistic predictions of action outcomes and hence select the action with the highest expected utility.

This chapter and the next concentrate on the task of representing and computing with probabilistic information in general. Chapter 14 deals with methods for the specific tasks of representing and updating the belief state over time and predicting outcomes. Chapter 18 looks at ways of combining probability theory with expressive formal languages such as first-order logic and general-purpose programming languages. Chapter 15 covers utility theory in more depth, and Chapter 16 develops algorithms for planning sequences of actions in stochastic environments. Chapter 17 covers the extension of these ideas to multiagent environments.

12.2 Basic Probability Notation

For our agent to represent and use probabilistic information, we need a formal language. The language of probability theory has traditionally been informal, written by human mathematicians for other human mathematicians. Appendix A includes a standard introduction to elementary probability theory; here, we take an approach more suited to the needs of AI and connect it with the concepts of formal logic.

12.2.1 What probabilities are about

Like logical assertions, probabilistic assertions are about possible worlds. Whereas logical assertions say which possible worlds are strictly ruled out (all those in which the assertion is false), probabilistic assertions talk about how probable the various worlds are. In probability theory, the set of all possible worlds is called the **sample space**. The possible worlds are *mutually exclusive* and *exhaustive*—two possible worlds cannot both be the case, and one

possible world must be the case. For example, if we are about to roll two (distinguishable) dice, there are 36 possible worlds to consider: (1,1), (1,2), ..., (6,6). The Greek letter Ω (uppercase omega) is used to refer to the sample space, and ω (lowercase omega) refers to elements of the space, that is, particular possible worlds.

A fully specified **probability model** associates a numerical probability $P(\omega)$ with each possible world.¹ The basic axioms of probability theory say that every possible world has a probability between 0 and 1 and that the total probability of the set of possible worlds is 1:

Probability model

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1. \quad (12.1)$$

For example, if we assume that each die is fair and the rolls don't interfere with each other, then each of the possible worlds (1,1), (1,2), ..., (6,6) has probability 1/36. If the dice are loaded then some worlds will have higher probabilities and some lower, but they will all still sum to 1.

Probabilistic assertions and queries are not usually about particular possible worlds, but about sets of them. For example, we might ask for the probability that the two dice add up to 11, the probability that doubles are rolled, and so on. In probability theory, these sets are called **events**—a term already used extensively in Chapter 10 for a different concept. In logic, a set of worlds corresponds to a **proposition** in a formal language; specifically, for each proposition, the corresponding set contains just those possible worlds in which the proposition holds. (Hence, “event” and “proposition” mean roughly the same thing in this context, except that a proposition is expressed in a formal language.) The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

Event

$$\text{For any proposition } \phi, P(\phi) = \sum_{\omega \in \phi} P(\omega). \quad (12.2)$$

For example, when rolling fair dice, we have $P(\text{Total}=11) = P((5,6)) + P((6,5)) = 1/36 + 1/36 = 1/18$. Note that probability theory does not require complete knowledge of the probabilities of each possible world. For example, if we believe the dice conspire to produce the same number, we might *assert* that $P(\text{doubles}) = 1/4$ without knowing whether the dice prefer double 6 to double 2. Just as with logical assertions, this assertion *constrains* the underlying probability model without fully determining it.

Probabilities such as $P(\text{Total}=11)$ and $P(\text{doubles})$ are called **unconditional** or **prior probabilities** (and sometimes just “priors” for short); they refer to degrees of belief in propositions *in the absence of any other information*. Most of the time, however, we have *some* information, usually called **evidence**, that has already been revealed. For example, the first die may already be showing a 5 and we are waiting with bated breath for the other one to stop spinning. In that case, we are interested not in the unconditional probability of rolling doubles, but the **conditional** or **posterior** probability (or just “posterior” for short) of rolling doubles *given that the first die is a 5*. This probability is written $P(\text{doubles} | \text{Die}_1 = 5)$, where the “|” is pronounced “given.”²

Unconditional probability
Prior probability

Evidence

Conditional probability
Posterior probability

Similarly, if I am going to the dentist for a regularly scheduled checkup, then the prior probability $P(\text{cavity}) = 0.2$ might be of interest; but if I go to the dentist because I have a toothache, it's the conditional probability $P(\text{cavity} | \text{toothache}) = 0.6$ that matters.

¹ For now, we assume a discrete, countable set of worlds. The proper treatment of the continuous case brings in certain complications that are less relevant for most purposes in AI.

² Note that the precedence of “|” is such that any expression of the form $P(\dots | \dots)$ always means $P((\dots) | (\dots))$.

It is important to understand that $P(\text{cavity}) = 0.2$ is still *valid* after *toothache* is observed; it just isn't especially useful. When making decisions, an agent needs to condition on *all* the evidence it has observed. It is also important to understand the difference between conditioning and logical implication. The assertion that $P(\text{cavity} | \text{toothache}) = 0.6$ does not mean “Whenever *toothache* is true, conclude that *cavity* is true with probability 0.6” rather it means “Whenever *toothache* is true *and we have no further information*, conclude that *cavity* is true with probability 0.6.” The extra condition is important; for example, if we had the further information that the dentist found no cavities, we definitely would not want to conclude that *cavity* is true with probability 0.6; instead we need to use $P(\text{cavity} | \text{toothache} \wedge \neg \text{cavity}) = 0$.

Mathematically speaking, conditional probabilities are defined in terms of unconditional probabilities as follows: for any propositions a and b , we have

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}, \quad (12.3)$$

which holds whenever $P(b) > 0$. For example,

$$P(\text{doubles} | \text{Die}_1 = 5) = \frac{P(\text{doubles} \wedge \text{Die}_1 = 5)}{P(\text{Die}_1 = 5)}.$$

The definition makes sense if you remember that observing b rules out all those possible worlds where b is false, leaving a set whose total probability is just $P(b)$. Within that set, the worlds where a is true must satisfy $a \wedge b$ and constitute a fraction $P(a \wedge b)/P(b)$.

The definition of conditional probability, Equation (12.3), can be written in a different form called the **product rule**:

$$P(a \wedge b) = P(a|b)P(b). \quad (12.4)$$

The product rule is perhaps easier to remember: it comes from the fact that for a and b to be true, we need b to be true, and we also need a to be true given b .

12.2.2 The language of propositions in probability assertions

In this chapter and the next, propositions describing sets of possible worlds are usually written in a notation that combines elements of propositional logic and constraint satisfaction notation. In the terminology of Section 2.4.7, it is a **factored representation**, in which a possible world is represented by a set of variable/value pairs. A more expressive **structured representation** is also possible, as shown in Chapter 18.

Variables in probability theory are called **random variables**, and their names begin with an uppercase letter. Thus, in the dice example, *Total* and *Die*₁ are random variables. Every random variable is a function that maps from the domain of possible worlds Ω to some **range**—the set of possible values it can take on. The range of *Total* for two dice is the set $\{2, \dots, 12\}$ and the range of *Die*₁ is $\{1, \dots, 6\}$. Names for values are always lowercase, so we might write $\sum_x P(X=x)$ to sum over the values of X . A Boolean random variable has the range $\{\text{true}, \text{false}\}$. For example, the proposition that doubles are rolled can be written as *Doubles* = *true*. (An alternative range for Boolean variables is the set $\{0, 1\}$, in which case the variable is said to have a **Bernoulli** distribution.) By convention, propositions of the form $A = \text{true}$ are abbreviated simply as a , while $A = \text{false}$ is abbreviated as $\neg a$. (The uses of *doubles*, *cavity*, and *toothache* in the preceding section are abbreviations of this kind.)

Ranges can be sets of arbitrary tokens. We might choose the range of *Age* to be the set $\{\text{juvenile}, \text{teen}, \text{adult}\}$ and the range of *Weather* might be $\{\text{sun}, \text{rain}, \text{cloud}, \text{snow}\}$. When no

Product rule

Random variable

Range

Bernoulli

ambiguity is possible, it is common to use a value by itself to stand for the proposition that a particular variable has that value; thus, *sun* can stand for $Weather = sun$.³

The preceding examples all have finite ranges. Variables can have infinite ranges, too—either discrete (like the integers) or continuous (like the reals). For any variable with an ordered range, inequalities are also allowed, such as $NumberOfAtomsInUniverse \geq 10^{70}$.

Finally, we can combine these sorts of elementary propositions (including the abbreviated forms for Boolean variables) by using the connectives of propositional logic. For example, we can express “The probability that the patient has a cavity, given that she is a teenager with no toothache, is 0.1” as follows:

$$P(cavity | \neg toothache \wedge teen) = 0.1.$$

In probability notation, it is also common to use a comma for conjunction, so we could write $P(cavity | \neg toothache, teen)$.

Sometimes we will want to talk about the probabilities of *all* the possible values of a random variable. We could write:

$$\begin{aligned} P(Weather = sun) &= 0.6 \\ P(Weather = rain) &= 0.1 \\ P(Weather = cloud) &= 0.29 \\ P(Weather = snow) &= 0.01, \end{aligned}$$

but as an abbreviation we will allow

$$\mathbf{P}(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle,$$

where the bold **P** indicates that the result is a vector of numbers, and where we assume a predefined ordering $\langle sun, rain, cloud, snow \rangle$ on the range of *Weather*. We say that the **P** statement defines a **probability distribution** for the random variable *Weather*—that is, an assignment of a probability for each possible value of the random variable. (In this case, with a finite, discrete range, the distribution is called a **categorical distribution**.) The **P** notation is also used for conditional distributions: $\mathbf{P}(X|Y)$ gives the values of $P(X=x_i | Y=y_j)$ for each possible i, j pair.

Probability
distribution

Categorical
distribution

For continuous variables, it is not possible to write out the entire distribution as a vector, because there are infinitely many values. Instead, we can define the probability that a random variable takes on some value x as a parameterized function of x , usually called a **probability density function**. For example, the sentence

$$P(NoonTemp = x) = Uniform(x; 18C, 26C)$$

Probability density
function

expresses the belief that the temperature at noon is distributed uniformly between 18 and 26 degrees Celsius.

Probability density functions (sometimes called **pdfs**) differ in meaning from discrete distributions. Saying that the probability density is uniform from 18C to 26C means that there is a 100% chance that the temperature will fall somewhere in that 8C-wide region and a 50% chance that it will fall in any 4C-wide sub-region, and so on. We write the probability density for a continuous random variable X at value x as $P(X=x)$ or just $P(x)$; the intuitive

³ These conventions taken together lead to a potential ambiguity in notation when summing over values of a Boolean variable: $P(a)$ is the probability that A is *true*, whereas in the expression $\sum_a P(a)$ it just refers to the probability of one of the values a of A .

definition of $P(x)$ is the probability that X falls within an arbitrarily small region beginning at x , divided by the width of the region:

$$P(x) = \lim_{dx \rightarrow 0} P(x \leq X \leq x + dx) / dx.$$

For *NoonTemp* we have

$$P(\text{NoonTemp} = x) = \text{Uniform}(x; 18C, 26C) = \begin{cases} \frac{1}{8C} & \text{if } 18C \leq x \leq 26C \\ 0 & \text{otherwise} \end{cases},$$

where C stands for centigrade (not for a constant). In $P(\text{NoonTemp} = 20.18C) = \frac{1}{8C}$, note that $\frac{1}{8C}$ is not a probability, it is a probability density. The probability that *NoonTemp* is *exactly* 20.18C is zero, because 20.18C is a region of width 0. Some authors use different symbols for discrete probabilities and probability densities; we use P for specific probability values and \mathbf{P} for vectors of values in both cases, since confusion seldom arises and the equations are usually identical. Note that probabilities are unitless numbers, whereas density functions are measured with a unit, in this case reciprocal degrees centigrade. If the same temperature interval were to be expressed in degrees Fahrenheit, it would have a width of 14.4 degrees, and the density would be $1/14.4F$.

In addition to distributions on single variables, we need notation for distributions on multiple variables. Commas are used for this. For example, $\mathbf{P}(\text{Weather}, \text{Cavity})$ denotes the probabilities of all combinations of the values of *Weather* and *Cavity*. This is a 4×2 table of probabilities called the **joint probability distribution** of *Weather* and *Cavity*. We can also mix variables and specific values; $\mathbf{P}(\text{sun}, \text{Cavity})$ would be a two-element vector giving the probabilities of a cavity with a sunny day and no cavity with a sunny day.

The \mathbf{P} notation makes certain expressions much more concise than they might otherwise be. For example, the product rules (see Equation (12.4)) for all possible values of *Weather* and *Cavity* can be written as a single equation:

$$\mathbf{P}(\text{Weather}, \text{Cavity}) = \mathbf{P}(\text{Weather} | \text{Cavity}) \mathbf{P}(\text{Cavity}),$$

instead of as these $4 \times 2 = 8$ equations (using abbreviations W and C):

$$\begin{aligned} P(W = \text{sun} \wedge C = \text{true}) &= P(W = \text{sun} | C = \text{true}) P(C = \text{true}) \\ P(W = \text{rain} \wedge C = \text{true}) &= P(W = \text{rain} | C = \text{true}) P(C = \text{true}) \\ P(W = \text{cloud} \wedge C = \text{true}) &= P(W = \text{cloud} | C = \text{true}) P(C = \text{true}) \\ P(W = \text{snow} \wedge C = \text{true}) &= P(W = \text{snow} | C = \text{true}) P(C = \text{true}) \\ P(W = \text{sun} \wedge C = \text{false}) &= P(W = \text{sun} | C = \text{false}) P(C = \text{false}) \\ P(W = \text{rain} \wedge C = \text{false}) &= P(W = \text{rain} | C = \text{false}) P(C = \text{false}) \\ P(W = \text{cloud} \wedge C = \text{false}) &= P(W = \text{cloud} | C = \text{false}) P(C = \text{false}) \\ P(W = \text{snow} \wedge C = \text{false}) &= P(W = \text{snow} | C = \text{false}) P(C = \text{false}). \end{aligned}$$

As a degenerate case, $\mathbf{P}(\text{sun}, \text{cavity})$ has no variables and thus is a zero-dimensional vector, which we can think of as a scalar value.

Now we have defined a syntax for propositions and probability assertions and we have given part of the semantics: Equation (12.2) defines the probability of a proposition as the sum of the probabilities of worlds in which it holds. To complete the semantics, we need to say what the worlds are and how to determine whether a proposition holds in a world. We borrow this part directly from the semantics of propositional logic, as follows. *A possible world is defined to be an assignment of values to all of the random variables under consideration.*

It is easy to see that this definition satisfies the basic requirement that possible worlds be mutually exclusive and exhaustive (Exercise 12.EXEX). For example, if the random variables

are *Cavity*, *Toothache*, and *Weather*, then there are $2 \times 2 \times 4 = 16$ possible worlds. Furthermore, the truth of any given proposition can be determined easily in such worlds by the same recursive truth calculation we used for propositional logic (see page 236).

Note that some random variables may be redundant, in that their values can be obtained in all cases from the values of other variables. For example, the *Doubles* variable in the two-dice world is true exactly when $Die_1 = Die_2$. Including *Doubles* as one of the random variables, in addition to *Die*₁ and *Die*₂, seems to increase the number of possible worlds from 36 to 72, but of course exactly half of the 72 will be logically impossible and will have probability 0.

From the preceding definition of possible worlds, it follows that a probability model is completely determined by the joint distribution for all of the random variables—the so-called **full joint probability distribution**. For example, given *Cavity*, *Toothache*, and *Weather*, the full joint distribution is $\mathbf{P}(\text{Cavity}, \text{Toothache}, \text{Weather})$. This joint distribution can be represented as a $2 \times 2 \times 4$ table with 16 entries. Because every proposition's probability is a sum over possible worlds, a full joint distribution suffices, in principle, for calculating the probability of any proposition. We will see examples of how to do this in Section 12.3.

Full joint probability
distribution

12.2.3 Probability axioms and their reasonableness

The basic axioms of probability (Equations (12.1) and (12.2)) imply certain relationships among the degrees of belief that can be accorded to logically related propositions. For example, we can derive the familiar relationship between the probability of a proposition and the probability of its negation:

$$\begin{aligned}
 P(\neg a) &= \sum_{\omega \in \neg a} P(\omega) && \text{by Equation (12.2)} \\
 &= \sum_{\omega \in \neg a} P(\omega) + \sum_{\omega \in a} P(\omega) - \sum_{\omega \in a} P(\omega) \\
 &= \sum_{\omega \in \Omega} P(\omega) - \sum_{\omega \in a} P(\omega) && \text{grouping the first two terms} \\
 &= 1 - P(a) && \text{by (12.1) and (12.2).}
 \end{aligned}$$

We can also derive the well-known formula for the probability of a disjunction, sometimes called the **inclusion–exclusion principle**:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b). \quad (12.5)$$

Inclusion–exclusion
principle

This rule is easily remembered by noting that the cases where *a* holds, together with the cases where *b* holds, certainly cover all the cases where $a \vee b$ holds; but summing the two sets of cases counts their intersection twice, so we need to subtract $P(a \wedge b)$.

Equations (12.1) and (12.5) are often called **Kolmogorov's axioms** in honor of the mathematician Andrei Kolmogorov, who showed how to build up the rest of probability theory from this simple foundation and how to handle the difficulties caused by continuous variables.⁴ While Equation (12.2) has a definitional flavor, Equation (12.5) reveals that the axioms really do constrain the degrees of belief an agent can have concerning logically related propositions. This is analogous to the fact that a logical agent cannot simultaneously believe *A*, *B*, and $\neg(A \wedge B)$, because there is no possible world in which all three are true. With probabilities, however, statements refer not to the world directly, but to the agent's own state of knowledge. Why, then, can an agent not hold the following set of beliefs (even though they violate Kolmogorov's axioms)?

Kolmogorov's
axioms

$$P(a) = 0.4 \quad P(b) = 0.3 \quad P(a \wedge b) = 0.0 \quad P(a \vee b) = 0.8. \quad (12.6)$$

⁴ The difficulties include the **Vitali set**, a well-defined subset of the interval $[0, 1]$ with no well-defined size.

Proposition	Agent 1's belief	Agent 2 bets	Agent 1 bets	Agent 1 payoffs for each outcome			
				a, b	$a, \neg b$	$\neg a, b$	$\neg a, \neg b$
a	0.4	\$4 on a	\$6 on $\neg a$	-\$6	-\$6	\$4	\$4
b	0.3	\$3 on b	\$7 on $\neg b$	-\$7	\$3	-\$7	\$3
$a \vee b$	0.8	\$2 on $\neg(a \vee b)$	\$8 on $a \vee b$	\$2	\$2	\$2	-\$8
				-\$11	-\$1	-\$1	-\$1

Figure 12.2 Because Agent 1 has inconsistent beliefs, Agent 2 is able to devise a set of three bets that guarantees a loss for Agent 1, no matter what the outcome of a and b .

This kind of question has been the subject of decades of intense debate between those who advocate the use of probabilities as the only legitimate form for degrees of belief and those who advocate alternative approaches.

One argument for the axioms of probability, first stated in 1931 by Bruno de Finetti (see de Finetti, 1993, for an English translation), is as follows: If an agent has some degree of belief in a proposition a , then the agent should be able to state odds at which it is indifferent to a bet for or against a .⁵ Think of it as a game between two agents: Agent 1 states, “my degree of belief in event a is 0.4.” Agent 2 is then free to choose whether to wager for or against a at stakes that are consistent with the stated degree of belief. That is, Agent 2 could choose to accept Agent 1’s bet that a will occur, offering \$6 against Agent 1’s \$4. Or Agent 2 could accept Agent 1’s bet that $\neg a$ will occur, offering \$4 against Agent 1’s \$6. Then we observe the outcome of a , and whoever is right collects the money. If one’s degrees of belief do not accurately reflect the world, then one would expect to lose money over the long run to an opposing agent whose beliefs more accurately reflect the state of the world.

► De Finetti’s theorem is not concerned with choosing the right values for individual probabilities, but with choosing values for the probabilities of logically related propositions: *If Agent 1 expresses a set of degrees of belief that violate the axioms of probability theory then there is a combination of bets by Agent 2 that guarantees that Agent 1 will lose money every time.* For example, suppose that Agent 1 has the set of degrees of belief from Equation (12.6). Figure 12.2 shows that if Agent 2 chooses to bet \$4 on a , \$3 on b , and \$2 on $\neg(a \vee b)$, then Agent 1 always loses money, regardless of the outcomes for a and b . De Finetti’s theorem implies that no rational agent can have beliefs that violate the axioms of probability.

One common objection to de Finetti’s theorem is that this betting game is rather contrived. For example, what if one refuses to bet? Does that end the argument? The answer is that the betting game is an abstract model for the decision-making situation in which every agent is *unavoidably* involved at every moment. Every action (including inaction) is a kind of bet, and every outcome can be seen as a payoff of the bet. Refusing to bet is like refusing to allow time to pass.

Other strong philosophical arguments have been put forward for the use of probabilities, most notably those of Cox (1946), Carnap (1950), and Jaynes (2003). They each construct a

⁵ One might argue that the agent’s preferences for different bank balances are such that the possibility of losing \$1 is not counterbalanced by an equal possibility of winning \$1. One possible response is to make the bet amounts small enough to avoid this problem. Savage’s analysis (1954) circumvents the issue altogether.

set of axioms for reasoning with degrees of beliefs: no contradictions, correspondence with ordinary logic (for example, if belief in A goes up, then belief in $\neg A$ must go down), and so on. The only controversial axiom is that degrees of belief must be numbers, or at least act like numbers in that they must be transitive (if belief in A is greater than belief in B , which is greater than belief in C , then belief in A must be greater than C) and comparable (the belief in A must be one of equal to, greater than, or less than belief in B). It can then be proved that probability is the only approach that satisfies these axioms.

The world being the way it is, however, practical demonstrations sometimes speak louder than proofs. The success of reasoning systems based on probability theory has been much more effective than philosophical arguments in making converts. We now look at how the axioms can be deployed to make inferences.

12.3 Inference Using Full Joint Distributions

In this section we describe a simple method for **probabilistic inference**—that is, the computation of posterior probabilities for **query** propositions given observed evidence. We use the full joint distribution as the “knowledge base” from which answers to all questions may be derived. Along the way we also introduce several useful techniques for manipulating equations involving probabilities.

Probabilistic
inference
Query

We begin with a simple example: a domain consisting of just the three Boolean variables *Toothache*, *Cavity*, and *Catch* (the dentist’s nasty steel probe catches in my tooth). The full joint distribution is a $2 \times 2 \times 2$ table as shown in Figure 12.3.

	<i>toothache</i>		\neg <i>toothache</i>	
	<i>catch</i>	\neg <i>catch</i>	<i>catch</i>	\neg <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
\neg <i>cavity</i>	0.016	0.064	0.144	0.576

Figure 12.3 A full joint distribution for the *Toothache*, *Cavity*, *Catch* world.

Notice that the probabilities in the joint distribution sum to 1, as required by the axioms of probability. Notice also that Equation (12.2) gives us a direct way to calculate the probability of any proposition, simple or complex: simply identify those possible worlds in which the proposition is true and add up their probabilities. For example, there are six possible worlds in which $cavity \vee toothache$ holds:

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28.$$

One particularly common task is to extract the distribution over some subset of variables or a single variable. For example, adding the entries in the first row gives the unconditional or **marginal probability**⁶ of *cavity*:

Marginal probability

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2.$$

⁶ So called because of a common practice among actuaries of writing the sums of observed frequencies in the margins of insurance tables.

Marginalization

This process is called **marginalization**, or **summing out**—because we sum up the probabilities for each possible value of the other variables, thereby taking them out of the equation. We can write the following general marginalization rule for any sets of variables \mathbf{Y} and \mathbf{Z} :

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{Z} = \mathbf{z}), \quad (12.7)$$

where $\sum_{\mathbf{z}}$ sums over all the possible combinations of values of the set of variables \mathbf{Z} . As usual we can abbreviate $\mathbf{P}(\mathbf{Y}, \mathbf{Z} = \mathbf{z})$ in this equation by $\mathbf{P}(\mathbf{Y}, \mathbf{z})$. For the *Cavity* example, Equation (12.7) corresponds to the following equation:

$$\begin{aligned} \mathbf{P}(\text{Cavity}) &= \mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) + \mathbf{P}(\text{Cavity}, \text{toothache}, \neg \text{catch}) \\ &\quad + \mathbf{P}(\text{Cavity}, \neg \text{toothache}, \text{catch}) + \mathbf{P}(\text{Cavity}, \neg \text{toothache}, \neg \text{catch}) \\ &= \langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle + \langle 0.072, 0.144 \rangle + \langle 0.008, 0.576 \rangle \\ &= \langle 0.2, 0.8 \rangle. \end{aligned}$$

Conditioning

Using the product rule (Equation (12.4)), we can replace $\mathbf{P}(\mathbf{Y}, \mathbf{z})$ in Equation (12.7) by $\mathbf{P}(\mathbf{Y} | \mathbf{z})P(\mathbf{z})$, obtaining a rule called **conditioning**:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y} | \mathbf{z})P(\mathbf{z}). \quad (12.8)$$

Marginalization and conditioning turn out to be useful rules for all kinds of derivations involving probability expressions.

In most cases, we are interested in computing *conditional* probabilities of some variables, given evidence about others. Conditional probabilities can be found by first using Equation (12.3) to obtain an expression in terms of unconditional probabilities and then evaluating the expression from the full joint distribution. For example, we can compute the probability of a cavity, given evidence of a toothache, as follows:

$$\begin{aligned} P(\text{cavity} | \text{toothache}) &= \frac{P(\text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6. \end{aligned}$$

Just to check, we can also compute the probability that there is no cavity, given a toothache:

$$\begin{aligned} P(\neg \text{cavity} | \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4. \end{aligned}$$

The two values sum to 1.0, as they should. Notice that the term $P(\text{toothache})$ is in the denominator for both of these calculations. If the variable *Cavity* had more than two values, it would be in the denominator for all of them. In fact, it can be viewed as a **normalization** constant for the distribution $\mathbf{P}(\text{Cavity} | \text{toothache})$, ensuring that it adds up to 1. Throughout the chapters dealing with probability, we use α to denote such constants. With this notation, we can write the two preceding equations in one:

$$\begin{aligned} \mathbf{P}(\text{Cavity} | \text{toothache}) &= \alpha \mathbf{P}(\text{Cavity}, \text{toothache}) \\ &= \alpha [\mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) + \mathbf{P}(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\ &= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] = \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle. \end{aligned}$$

In other words, we can calculate $\mathbf{P}(\text{Cavity}|\text{toothache})$ even if we don't know the value of $P(\text{toothache})$! We temporarily forget about the factor $1/P(\text{toothache})$ and add up the values for *cavity* and $\neg\text{cavity}$, getting 0.12 and 0.08. Those are the correct relative proportions, but they don't sum to 1, so we normalize them by dividing each one by $0.12 + 0.08$, getting the true probabilities of 0.6 and 0.4. Normalization turns out to be a useful shortcut in many probability calculations, both to make the computation easier and to allow us to proceed when some probability assessment (such as $P(\text{toothache})$) is not available.

From the example, we can extract a general inference procedure. We begin with the case in which the query involves a single variable, X (*Cavity* in the example). Let \mathbf{E} be the list of evidence variables (just *Toothache* in the example), let \mathbf{e} be the list of observed values for them, and let \mathbf{Y} be the remaining unobserved variables (just *Catch* in the example). The query is $\mathbf{P}(X|\mathbf{e})$ and can be evaluated as

$$\mathbf{P}(X|\mathbf{e}) = \alpha \mathbf{P}(X, \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y}), \quad (12.9)$$

where the summation is over all possible \mathbf{y} s (i.e., all possible combinations of values of the unobserved variables \mathbf{Y}). Notice that together the variables X , \mathbf{E} , and \mathbf{Y} constitute the complete set of variables for the domain, so $\mathbf{P}(X, \mathbf{e}, \mathbf{y})$ is simply a subset of probabilities from the full joint distribution.

Given the full joint distribution to work with, Equation (12.9) can answer probabilistic queries for discrete variables. It does not scale well, however: for a domain described by n Boolean variables, it requires an input table of size $O(2^n)$ and takes $O(2^n)$ time to process the table. In a realistic problem we could easily have $n = 100$, making $O(2^n)$ impractical—a table with $2^{100} \approx 10^{30}$ entries! The problem is not just memory and computation: the real issue is that if each of the 10^{30} probabilities has to be estimated separately from examples, the number of examples required will be astronomical.

For these reasons, the full joint distribution in tabular form is seldom a practical tool for building reasoning systems. Instead, it should be viewed as the theoretical foundation on which more effective approaches may be built, just as truth tables formed a theoretical foundation for more practical algorithms like DPLL in Chapter 7. The remainder of this chapter introduces some of the basic ideas required in preparation for the development of realistic systems in Chapter 13.

12.4 Independence

Let us expand the full joint distribution in Figure 12.3 by adding a fourth variable, *Weather*. The full joint distribution then becomes $\mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather})$, which has $2 \times 2 \times 2 \times 4 = 32$ entries. It contains four “editions” of the table shown in Figure 12.3, one for each kind of weather. What relationship do these editions have to each other and to the original three-variable table? How is the value of $P(\text{toothache}, \text{catch}, \text{cavity}, \text{cloud})$ related to the value of $P(\text{toothache}, \text{catch}, \text{cavity})$? We can use the product rule (Equation (12.4)):

$$\begin{aligned} P(\text{toothache}, \text{catch}, \text{cavity}, \text{cloud}) \\ = P(\text{cloud}|\text{toothache}, \text{catch}, \text{cavity})P(\text{toothache}, \text{catch}, \text{cavity}). \end{aligned}$$

Now, unless one is in the deity business, one should not imagine that one's dental problems influence the weather. And for indoor dentistry, at least, it seems safe to say that the weather

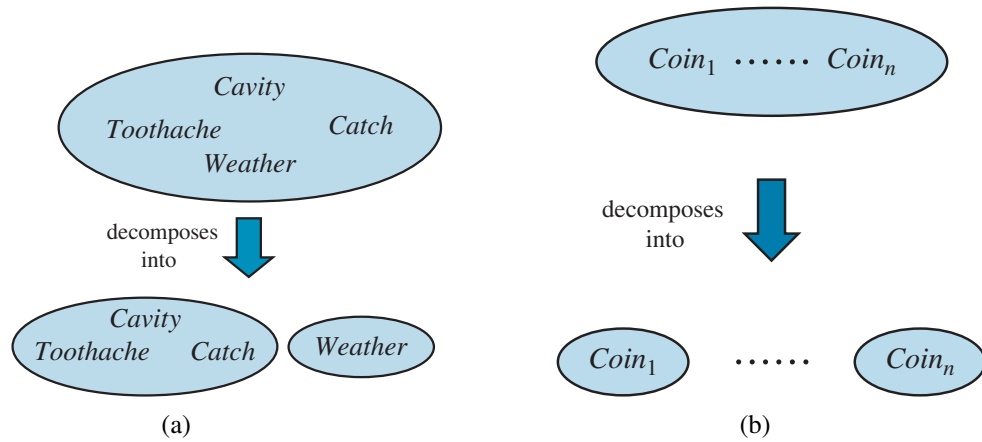


Figure 12.4 Two examples of factoring a large joint distribution into smaller distributions, using absolute independence. (a) Weather and dental problems are independent. (b) Coin flips are independent.

does not influence the dental variables. Therefore, the following assertion seems reasonable:

$$P(\text{cloud} | \text{toothache}, \text{catch}, \text{cavity}) = P(\text{cloud}). \quad (12.10)$$

From this, we can deduce

$$P(\text{toothache}, \text{catch}, \text{cavity}, \text{cloud}) = P(\text{cloud})P(\text{toothache}, \text{catch}, \text{cavity}).$$

A similar equation exists for *every* entry in $\mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather})$. In fact, we can write the general equation

$$\mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather}) = \mathbf{P}(\text{Toothache}, \text{Catch}, \text{Cavity})\mathbf{P}(\text{Weather}).$$

Thus, the 32-element table for four variables can be constructed from one 8-element table and one 4-element table. This decomposition is illustrated schematically in Figure 12.4(a).

The property we used in Equation (12.10) is called **independence** (also **marginal independence** and **absolute independence**). In particular, the weather is independent of one's dental problems. Independence between propositions a and b can be written as

$$P(a|b) = P(a) \quad \text{or} \quad P(b|a) = P(b) \quad \text{or} \quad P(a \wedge b) = P(a)P(b). \quad (12.11)$$

All these forms are equivalent (Exercise 12.INDI). Independence between variables X and Y can be written as follows (again, these are all equivalent):

$$\mathbf{P}(X|Y) = \mathbf{P}(X) \quad \text{or} \quad \mathbf{P}(Y|X) = \mathbf{P}(Y) \quad \text{or} \quad \mathbf{P}(X, Y) = \mathbf{P}(X)\mathbf{P}(Y).$$

Independence assertions are usually based on knowledge of the domain. As the toothache-weather example illustrates, they can dramatically reduce the amount of information necessary to specify the full joint distribution. If the complete set of variables can be divided into independent subsets, then the full joint distribution can be *factored* into separate joint distributions on those subsets. For example, the full joint distribution on the outcome of n independent coin flips, $\mathbf{P}(C_1, \dots, C_n)$, has 2^n entries, but it can be represented as the product of n single-variable distributions $\mathbf{P}(C_i)$. In a more practical vein, the independence of dentistry and meteorology is a good thing, because otherwise the practice of dentistry might require intimate knowledge of meteorology, and vice versa.

When they are available, then, independence assertions can help in reducing the size of the domain representation and the complexity of the inference problem. Unfortunately, clean separation of entire sets of variables by independence is quite rare. Whenever a connection, however indirect, exists between two variables, independence will fail to hold. Moreover, even independent subsets can be quite large—for example, dentistry might involve dozens of diseases and hundreds of symptoms, all of which are interrelated. To handle such problems, we need more subtle methods than the straightforward concept of independence.

12.5 Bayes' Rule and Its Use

On page 408, we defined the **product rule** (Equation (12.4)). It can actually be written in two forms:

$$P(a \wedge b) = P(a|b)P(b) \quad \text{and} \quad P(a \wedge b) = P(b|a)P(a).$$

Equating the two right-hand sides and dividing by $P(a)$, we get

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}. \quad (12.12)$$

This equation is known as **Bayes' rule** (also Bayes' law or Bayes' theorem). This simple Bayes' rule equation underlies most modern AI systems for probabilistic inference.

The more general case of Bayes' rule for multivalued variables can be written in the **P** notation as follows:

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)}.$$

As before, this is to be taken as representing a set of equations, each dealing with specific values of the variables. We will also have occasion to use a more general version conditionalized on some background evidence **e**:

$$\mathbf{P}(Y|X, \mathbf{e}) = \frac{\mathbf{P}(X|Y, \mathbf{e})\mathbf{P}(Y|\mathbf{e})}{\mathbf{P}(X|\mathbf{e})}. \quad (12.13)$$

12.5.1 Applying Bayes' rule: The simple case

On the surface, Bayes' rule does not seem very useful. It allows us to compute the single term $P(b|a)$ in terms of three terms: $P(a|b)$, $P(b)$, and $P(a)$. That seems like two steps backwards; but Bayes' rule is useful in practice because there are many cases where we do have good probability estimates for these three numbers and need to compute the fourth. Often, we perceive as evidence the *effect* of some unknown *cause* and we would like to determine that cause. In that case, Bayes' rule becomes

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}.$$

The conditional probability $P(\text{effect}|\text{cause})$ quantifies the relationship in the **causal** Causal direction, whereas $P(\text{cause}|\text{effect})$ describes the **diagnostic** Diagnostic direction. In a task such as medical diagnosis, we often have conditional probabilities on causal relationships. The doctor knows $P(\text{symptoms}|\text{disease})$ and wants to derive a diagnosis, $P(\text{disease}|\text{symptoms})$.

For example, a doctor knows that the disease meningitis causes a patient to have a stiff neck, say, 70% of the time. The doctor also knows some unconditional facts: the prior

probability that any patient has meningitis is 1/50,000, and the prior probability that any patient has a stiff neck is 1%. Letting s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis, we have

$$\begin{aligned} P(s|m) &= 0.7 \\ P(m) &= 1/50000 \\ P(s) &= 0.01 \\ P(m|s) &= \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014. \end{aligned} \quad (12.14)$$

That is, we expect only 0.14% of patients with a stiff neck to have meningitis. Notice that even though a stiff neck is quite strongly indicated by meningitis (with probability 0.7), the probability of meningitis in patients with stiff necks remains small. This is because the prior probability of stiff necks (from any cause) is much higher than the prior for meningitis.

Section 12.3 illustrated a process by which one can avoid assessing the prior probability of the evidence (here, $P(s)$) by instead computing a posterior probability for each value of the query variable (here, m and $\neg m$) and then normalizing the results. The same process can be applied when using Bayes' rule. We have

$$\mathbf{P}(M|s) = \alpha \langle P(s|m)P(m), P(s|\neg m)P(\neg m) \rangle.$$

Thus, to use this approach we need to estimate $P(s|\neg m)$ instead of $P(s)$. There is no free lunch—sometimes this is easier, sometimes it is harder. The general form of Bayes' rule with normalization is

$$\mathbf{P}(Y|X) = \alpha \mathbf{P}(X|Y)\mathbf{P}(Y), \quad (12.15)$$

where α is the normalization constant needed to make the entries in $\mathbf{P}(Y|X)$ sum to 1.

One obvious question to ask about Bayes' rule is why one might have available the conditional probability in one direction, but not the other. In the meningitis domain, perhaps the doctor knows that a stiff neck implies meningitis in 1 out of 5000 cases; that is, the doctor has quantitative information in the **diagnostic** direction from symptoms to causes. Such a doctor has no need to use Bayes' rule.

Unfortunately, *diagnostic knowledge is often more fragile than causal knowledge*. If there is a sudden epidemic of meningitis, the unconditional probability of meningitis, $P(m)$, will go up. The doctor who derived the diagnostic probability $P(m|s)$ directly from statistical observation of patients before the epidemic will have no idea how to update the value, but the doctor who computes $P(m|s)$ from the other three values will see that $P(m|s)$ should go up proportionately with $P(m)$. Most important, the causal information $P(s|m)$ is *unaffected* by the epidemic, because it simply reflects the way meningitis works. The use of this kind of direct causal or model-based knowledge provides the crucial robustness needed to make probabilistic systems feasible in the real world.

12.5.2 Using Bayes' rule: Combining evidence

We have seen that Bayes' rule can be useful for answering probabilistic queries conditioned on one piece of evidence—for example, the stiff neck. In particular, we have argued that probabilistic information is often available in the form $P(\text{effect}|\text{cause})$. What happens when we have two or more pieces of evidence? For example, what can a dentist conclude if her

nasty steel probe catches in the aching tooth of a patient? If we know the full joint distribution (Figure 12.3), we can read off the answer:

$$\mathbf{P}(\text{Cavity} | \text{toothache} \wedge \text{catch}) = \alpha \langle 0.108, 0.016 \rangle \approx \langle 0.871, 0.129 \rangle.$$

We know, however, that such an approach does not scale up to larger numbers of variables. We can try using Bayes' rule to reformulate the problem:

$$\begin{aligned} \mathbf{P}(\text{Cavity} | \text{toothache} \wedge \text{catch}) \\ = \alpha \mathbf{P}(\text{toothache} \wedge \text{catch} | \text{Cavity}) \mathbf{P}(\text{Cavity}). \end{aligned} \quad (12.16)$$

For this reformulation to work, we need to know the conditional probabilities of the conjunction $\text{toothache} \wedge \text{catch}$ for each value of Cavity . That might be feasible for just two evidence variables, but again it does not scale up. If there are n possible evidence variables (X rays, diet, oral hygiene, etc.), then there are $O(2^n)$ possible combinations of observed values for which we would need to know conditional probabilities. This is no better than using the full joint distribution.

To make progress, we need to find some additional assertions about the domain that will enable us to simplify the expressions. The notion of **independence** in Section 12.4 provides a clue, but needs refining. It would be nice if *Toothache* and *Catch* were independent, but they are not: if the probe catches in the tooth, then it is likely that the tooth has a cavity and that the cavity causes a toothache. These variables *are* independent, however, *given the presence or the absence of a cavity*. Each is directly caused by the cavity, but neither has a direct effect on the other: toothache depends on the state of the nerves in the tooth, whereas the probe's accuracy depends primarily on the dentist's skill, to which the toothache is irrelevant.⁷ Mathematically, this property is written as

$$\mathbf{P}(\text{toothache} \wedge \text{catch} | \text{Cavity}) = \mathbf{P}(\text{toothache} | \text{Cavity}) \mathbf{P}(\text{catch} | \text{Cavity}). \quad (12.17)$$

This equation expresses the **conditional independence** of *toothache* and *catch* given *Cavity*. We can plug it into Equation (12.16) to obtain the probability of a cavity:

Conditional independence

$$\begin{aligned} \mathbf{P}(\text{Cavity} | \text{toothache} \wedge \text{catch}) \\ = \alpha \mathbf{P}(\text{toothache} | \text{Cavity}) \mathbf{P}(\text{catch} | \text{Cavity}) \mathbf{P}(\text{Cavity}). \end{aligned} \quad (12.18)$$

Now the information requirements are the same as for inference, using each piece of evidence separately: the prior probability $\mathbf{P}(\text{Cavity})$ for the query variable and the conditional probability of each effect, given its cause.

The general definition of **conditional independence** of two variables X and Y , given a third variable Z , is

$$\mathbf{P}(X, Y | Z) = \mathbf{P}(X | Z) \mathbf{P}(Y | Z).$$

In the dentist domain, for example, it seems reasonable to assert conditional independence of the variables *Toothache* and *Catch*, given *Cavity*:

$$\mathbf{P}(\text{Toothache}, \text{Catch} | \text{Cavity}) = \mathbf{P}(\text{Toothache} | \text{Cavity}) \mathbf{P}(\text{Catch} | \text{Cavity}). \quad (12.19)$$

Notice that this assertion is somewhat stronger than Equation (12.17), which asserts independence only for specific values of *Toothache* and *Catch*. As with absolute independence in Equation (12.11), the equivalent forms

$$\mathbf{P}(X | Y, Z) = \mathbf{P}(X | Z) \quad \text{and} \quad \mathbf{P}(Y | X, Z) = \mathbf{P}(Y | Z)$$

⁷ We assume that the patient and dentist are distinct individuals.

can also be used (see Exercise 12.PXYZ). Section 12.4 showed that absolute independence assertions allow a decomposition of the full joint distribution into much smaller pieces. It turns out that the same is true for conditional independence assertions. For example, given the assertion in Equation (12.19), we can derive a decomposition as follows:

$$\begin{aligned} & \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \\ &= \mathbf{P}(\textit{Toothache}, \textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \quad (\text{product rule}) \\ &= \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity}) \quad (\text{using 12.19}). \end{aligned}$$

(The reader can easily check that this equation does in fact hold in Figure 12.3.) In this way, the original large table is decomposed into three smaller tables. The original table has 7 independent numbers. (The table has $2^3 = 8$ entries, but they must sum to 1, so 7 are independent). The smaller tables contain a total of $2 + 2 + 1 = 5$ independent numbers. (For a conditional probability distribution such as $\mathbf{P}(\textit{Toothache} \mid \textit{Cavity})$ there are two rows of two numbers, and each row sums to 1, so that's two independent numbers; for a prior distribution such as $\mathbf{P}(\textit{Cavity})$ there is only one independent number.) Going from 7 to 5 might not seem like a major triumph, but the gains can be much greater with larger number of symptoms.



Separation

In general, for n symptoms that are all conditionally independent given *Cavity*, the size of the representation grows as $O(n)$ instead of $O(2^n)$. That means that *conditional independence assertions can allow probabilistic systems to scale up; moreover, they are much more commonly available than absolute independence assertions*. Conceptually, *Cavity* **separates** *Toothache* and *Catch* because it is a direct cause of both of them. The decomposition of large probabilistic domains into weakly connected subsets through conditional independence is one of the most important developments in the recent history of AI.

12.6 Naive Bayes Models

The dentistry example illustrates a commonly occurring pattern in which a single cause directly influences a number of effects, all of which are conditionally independent, given the cause. The full joint distribution can be written as

$$\mathbf{P}(\textit{Cause}, \textit{Effect}_1, \dots, \textit{Effect}_n) = \mathbf{P}(\textit{Cause}) \prod_i \mathbf{P}(\textit{Effect}_i \mid \textit{Cause}). \quad (12.20)$$

Naive Bayes

Such a probability distribution is called a **naive Bayes** model—“naive” because it is often used (as a simplifying assumption) in cases where the “effect” variables are *not* strictly independent given the cause variable. (The naive Bayes model is sometimes called a **Bayesian classifier**, a somewhat careless usage that has prompted true Bayesians to call it the **idiot Bayes** model.) In practice, naive Bayes systems often work very well, even when the conditional independence assumption is not strictly true.

To use a naive Bayes model, we can apply Equation (12.20) to obtain the probability of the cause given some observed effects. Call the observed effects $\mathbf{E} = \mathbf{e}$, while the remaining effect variables \mathbf{Y} are unobserved. Then the standard method for inference from the joint distribution (Equation (12.9)) can be applied:

$$\mathbf{P}(\textit{Cause} \mid \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(\textit{Cause}, \mathbf{e}, \mathbf{y}).$$

From Equation (12.20), we then obtain

$$\begin{aligned}
 \mathbf{P}(\text{Cause} | \mathbf{e}) &= \alpha \sum_{\mathbf{y}} \mathbf{P}(\text{Cause}) \mathbf{P}(\mathbf{y} | \text{Cause}) \left(\prod_j \mathbf{P}(e_j | \text{Cause}) \right) \\
 &= \alpha \mathbf{P}(\text{Cause}) \left(\prod_j \mathbf{P}(e_j | \text{Cause}) \right) \sum_{\mathbf{y}} \mathbf{P}(\mathbf{y} | \text{Cause}) \\
 &= \alpha \mathbf{P}(\text{Cause}) \prod_j \mathbf{P}(e_j | \text{Cause})
 \end{aligned} \tag{12.21}$$

where the last line follows because the summation over \mathbf{y} is 1. Reinterpreting this equation in words: for each possible cause, multiply the prior probability of the cause by the product of the conditional probabilities of the observed effects given the cause; then normalize the result. The run time of this calculation is linear in the number of observed effects and does not depend on the number of unobserved effects (which may be very large in domains such as medicine). We will see in the next chapter that this is a common phenomenon in probabilistic inference: evidence variables whose values are unobserved usually “disappear” from the computation altogether.

12.6.1 Text classification with naive Bayes

Let’s see how a naive Bayes model can be used for the task of **text classification**: given a text, decide which of a predefined set of classes or categories it belongs to. Here the “cause” is the *Category* variable, and the “effect” variables are the presence or absence of certain key words, *HasWord_i*. Consider these two example sentences, taken from newspaper articles:

Text classification

1. Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.
2. Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations.

The task is to classify each sentence into a *Category*—the major sections of the newspaper: *news*, *sports*, *business*, *weather*, or *entertainment*. The naive Bayes model consists of the prior probabilities $\mathbf{P}(\text{Category})$ and the conditional probabilities $\mathbf{P}(\text{HasWord}_i | \text{Category})$. For each category c , $P(\text{Category}=c)$ is estimated as the fraction of all previously seen documents that are of category c . For example, if 9% of articles are about weather, we set $P(\text{Category}=\text{weather})=0.09$. Similarly, $\mathbf{P}(\text{HasWord}_i | \text{Category})$ is estimated as the fraction of documents of each category that contain word i ; perhaps 37% of articles about business contain word 6, “stocks,” so $P(\text{HasWord}_6=\text{true} | \text{Category}=\text{business})$ is set to 0.37.⁸

To categorize a new document, we check which key words appear in the document and then apply Equation (12.21) to obtain the posterior probability distribution over categories. If we have to predict just one category, we take the one with the highest posterior probability. Notice that, for this task, every effect variable is observed, since we can always tell whether a given word appears in the document.

⁸ One needs to be careful not to assign probability zero to words that have not been seen previously in a given category of documents, since the zero would wipe out all the other evidence in Equation (12.21). Just because you haven’t seen a word yet doesn’t mean you will *never* see it. Instead, reserve a small portion of the probability distribution to represent “previously unseen” words. See Chapter 21 for more on this issue in general, and Section 24.1.4 for the particular case of word models.

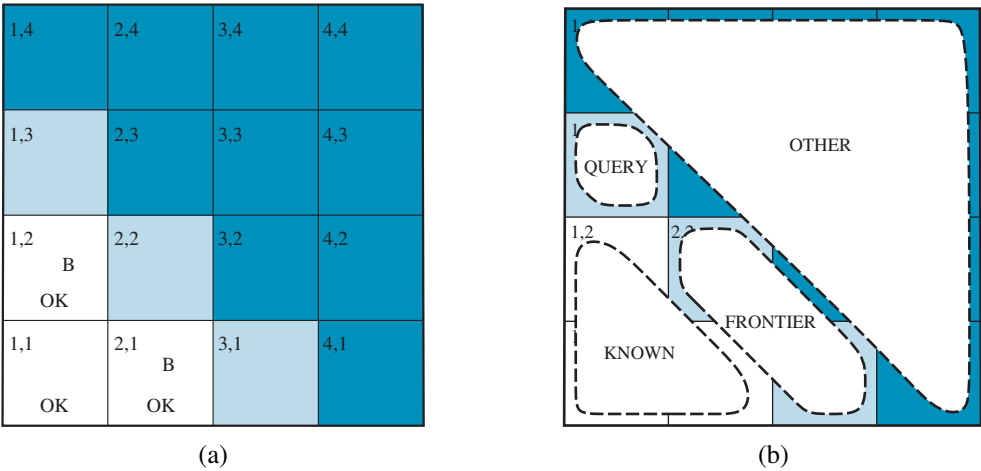


Figure 12.5 (a) After finding a breeze in both [1,2] and [2,1], the agent is stuck—there is no safe place to explore. (b) Division of the squares into *Known*, *Frontier*, and *Other*, for a query about [1,3].

The naive Bayes model assumes that words occur independently in documents, with frequencies determined by the document category. This independence assumption is clearly violated in practice. For example, the phrase “first quarter” occurs more frequently in business (or sports) articles than would be suggested by multiplying the probabilities of “first” and “quarter.” The violation of independence usually means that the final posterior probabilities will be much closer to 1 or 0 than they should be; in other words, the model is overconfident in its predictions. On the other hand, even with these errors, the *ranking* of the possible categories is often quite accurate.

Naive Bayes models are widely used for language determination, document retrieval, spam filtering, and other classification tasks. For tasks such as medical diagnosis, where the actual values of the posterior probabilities really matter—for example, in deciding whether to perform an appendectomy—one would usually prefer to use the more sophisticated models described in the next chapter.

12.7 The Wumpus World Revisited

We can combine the ideas in this chapter to solve probabilistic reasoning problems in the wumpus world. (See Chapter 7 for a complete description of the wumpus world.) Uncertainty arises in the wumpus world because the agent’s sensors give only partial information about the world. For example, Figure 12.5 shows a situation in which each of the three unvisited but reachable squares—[1,3], [2,2], and [3,1]—might contain a pit. Pure logical inference can conclude nothing about which square is most likely to be safe, so a logical agent might have to choose randomly. We will see that a probabilistic agent can do much better than the logical agent.

Our aim is to calculate the probability that each of the three squares contains a pit. (For this example we ignore the wumpus and the gold.) The relevant properties of the wumpus

world are that (1) a pit causes breezes in all neighboring squares, and (2) each square other than [1,1] contains a pit with probability 0.2. The first step is to identify the set of random variables we need:

- As in the propositional logic case, we want one Boolean variable P_{ij} for each square, which is true iff square $[i, j]$ actually contains a pit.
- We also have Boolean variables B_{ij} that are true iff square $[i, j]$ is breezy; we include these variables only for the observed squares—in this case, [1,1], [1,2], and [2,1].

The next step is to specify the full joint distribution, $\mathbf{P}(P_{1,1}, \dots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$. Applying the product rule, we have

$$\mathbf{P}(P_{1,1}, \dots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \dots, P_{4,4}) \mathbf{P}(P_{1,1}, \dots, P_{4,4}).$$

This decomposition makes it easy to see what the joint probability values should be. The first term is the conditional probability distribution of a breeze configuration, given a pit configuration; its values are 1 if all the breezy squares are adjacent to the pits and 0 otherwise. The second term is the prior probability of a pit configuration. Each square contains a pit with probability 0.2, independently of the other squares; hence,

$$\mathbf{P}(P_{1,1}, \dots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}). \quad (12.22)$$

For a particular configuration with exactly n pits, the probability is $0.2^n \times 0.8^{16-n}$.

In the situation in Figure 12.5(a), the evidence consists of the observed breeze (or its absence) in each square that is visited, combined with the fact that each such square contains no pit. We abbreviate these facts as $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$ and $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$. We are interested in answering queries such as $\mathbf{P}(P_{1,3} \mid known, b)$: how likely is it that [1,3] contains a pit, given the observations so far?

To answer this query, we can follow the standard approach of Equation (12.9), namely, summing over entries from the full joint distribution. Let *Unknown* be the set of $P_{i,j}$ variables for squares other than the known squares and the query square [1,3]. Then, by Equation (12.9), we have

$$\mathbf{P}(P_{1,3} \mid known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, known, b, unknown). \quad (12.23)$$

The full joint probabilities have already been specified, so we are done—that is, unless we care about computation. There are 12 unknown squares; hence the summation contains $2^{12} = 4096$ terms. In general, the summation grows exponentially with the number of squares.

Surely, one might ask, aren't the other squares irrelevant? How could [4,4] affect whether [1,3] has a pit? Indeed, this intuition is roughly correct, but it needs to be made more precise. What we really mean is that if we knew the values of all the pit variables adjacent to the squares we care about, then pits (or their absence) in other, more distant squares could have no further effect on our belief.

Let *Frontier* be the pit variables (other than the query variable) that are adjacent to visited squares, in this case just [2,2] and [3,1]. Also, let *Other* be the pit variables for the other unknown squares; in this case, there are 10 other squares, as shown in Figure 12.5(b). With these definitions, $Unknown = Frontier \cup Other$. The key insight given above can now be stated as follows: the observed breezes are *conditionally independent* of the other variables, given

the known, frontier, and query variables. To use this insight, we manipulate the query formula into a form in which the breezes are conditioned on all the other variables, and then we apply conditional independence:

$$\begin{aligned}
 & \mathbf{P}(P_{1,3} | \text{known}, b) \\
 &= \alpha \sum_{\text{unknown}} \mathbf{P}(P_{1,3}, \text{known}, b, \text{unknown}) \quad (\text{from Equation (12.23)}) \\
 &= \alpha \sum_{\text{unknown}} \mathbf{P}(b | P_{1,3}, \text{known}, \text{unknown}) \mathbf{P}(P_{1,3}, \text{known}, \text{unknown}) \quad (\text{product rule}) \\
 &= \alpha \sum_{\text{frontier}} \sum_{\text{other}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}, \text{other}) \mathbf{P}(P_{1,3}, \text{known}, \text{frontier}, \text{other}) \\
 &= \alpha \sum_{\text{frontier}} \sum_{\text{other}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}) \mathbf{P}(P_{1,3}, \text{known}, \text{frontier}, \text{other}),
 \end{aligned}$$

where the final step uses conditional independence: b is independent of other given known , $P_{1,3}$, and frontier . Now, the first term in this expression does not depend on the Other variables, so we can move the summation inward:

$$\begin{aligned}
 & \mathbf{P}(P_{1,3} | \text{known}, b) \\
 &= \alpha \sum_{\text{frontier}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}) \sum_{\text{other}} \mathbf{P}(P_{1,3}, \text{known}, \text{frontier}, \text{other}).
 \end{aligned}$$

By independence, as in Equation (12.22), the term on the right can be factored, and then the terms can be reordered:

$$\begin{aligned}
 & \mathbf{P}(P_{1,3} | \text{known}, b) \\
 &= \alpha \sum_{\text{frontier}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}) \sum_{\text{other}} \mathbf{P}(P_{1,3}) P(\text{known}) P(\text{frontier}) P(\text{other}) \\
 &= \alpha P(\text{known}) \mathbf{P}(P_{1,3}) \sum_{\text{frontier}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}) P(\text{frontier}) \sum_{\text{other}} P(\text{other}) \\
 &= \alpha' \mathbf{P}(P_{1,3}) \sum_{\text{frontier}} \mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier}) P(\text{frontier}),
 \end{aligned}$$

where the last step folds $P(\text{known})$ into the normalizing constant and uses the fact that $\sum_{\text{other}} P(\text{other})$ equals 1.

Now, there are just four terms in the summation over the frontier variables, $P_{2,2}$ and $P_{3,1}$. The use of independence and conditional independence has completely eliminated the other squares from consideration.

Notice that the probabilities in $\mathbf{P}(b | \text{known}, P_{1,3}, \text{frontier})$ are 1 when the breeze observations are consistent with the other variables and 0 otherwise. Thus, for each value of $P_{1,3}$, we sum over the *logical models* for the frontier variables that are consistent with the known facts. (Compare with the enumeration over models in Figure 7.5 on page 233.) The models and their associated prior probabilities— $P(\text{frontier})$ —are shown in Figure 12.6. We have

$$\mathbf{P}(P_{1,3} | \text{known}, b) = \alpha' \langle 0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16) \rangle \approx \langle 0.31, 0.69 \rangle.$$

That is, [1,3] (and [3,1] by symmetry) contains a pit with roughly 31% probability. A similar calculation, which the reader might wish to perform, shows that [2,2] contains a pit with roughly 86% probability. The wumpus agent should definitely avoid [2,2]! Note that our logical agent from Chapter 7 did not know that [2,2] was worse than the other squares. Logic

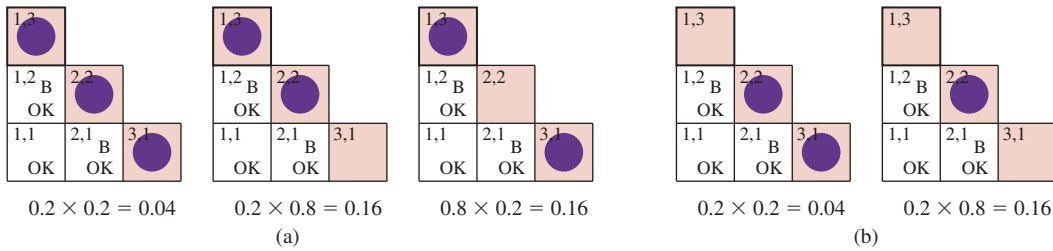


Figure 12.6 Consistent models for the frontier variables, $P_{2,2}$ and $P_{3,1}$, showing $P(\text{frontier})$ for each model: (a) three models with $P_{1,3} = \text{true}$ showing two or three pits, and (b) two models with $P_{1,3} = \text{false}$ showing one or two pits.

can tell us that it is unknown whether there is a pit in $[2, 2]$, but we need probability to tell us how likely it is.

What this section has shown is that even seemingly complicated problems can be formulated precisely in probability theory and solved with simple algorithms. To get *efficient* solutions, independence and conditional independence relationships can be used to simplify the summations required. These relationships often correspond to our natural understanding of how the problem should be decomposed. In the next chapter, we develop formal representations for such relationships as well as algorithms that operate on those representations to perform probabilistic inference efficiently.

Summary

This chapter has suggested probability theory as a suitable foundation for uncertain reasoning and provided a gentle introduction to its use.

- Uncertainty arises because of both laziness and ignorance. It is inescapable in complex, nondeterministic, or partially observable environments.
- **Probabilities** express the agent's inability to reach a definite decision regarding the truth of a sentence. Probabilities summarize the agent's beliefs relative to the evidence.
- **Decision theory** combines the agent's beliefs and desires, defining the best action as the one that maximizes expected **utility**.
- Basic probability statements include **prior** or **unconditional probabilities** and **posterior** or **conditional probabilities** over simple and complex propositions.
- The axioms of probability constrain the probabilities of logically related propositions. An agent that violates the axioms must behave irrationally in some cases.
- The **full joint probability distribution** specifies the probability of each complete assignment of values to random variables. It is usually too large to create or use in its explicit form, but when it is available it can be used to answer queries simply by adding up entries for the possible worlds corresponding to the query propositions.
- **Absolute independence** between subsets of random variables allows the full joint distribution to be factored into smaller joint distributions, greatly reducing its complexity.

- **Bayes' rule** allows unknown probabilities to be computed from known conditional probabilities, usually in the causal direction. Applying Bayes' rule with many pieces of evidence runs into the same scaling problems as does the full joint distribution.
- **Conditional independence** brought about by direct causal relationships in the domain allows the full joint distribution to be factored into smaller, conditional distributions. The **naive Bayes** model assumes the conditional independence of all effect variables, given a single cause variable; its size grows linearly with the number of effects.
- A wumpus-world agent can calculate probabilities for unobserved aspects of the world, thereby improving on the decisions of a purely logical agent. Conditional independence makes these calculations tractable.

Bibliographical and Historical Notes

Probability theory was invented as a way of analyzing games of chance. In about 850 CE the Indian mathematician Mahaviracarya described how to arrange a set of bets that can't lose (what we now call a Dutch book). In Europe, the first significant systematic analyses were produced by Girolamo Cardano around 1565, although publication was posthumous (1663). By that time, probability had been established as a mathematical discipline due to a series of results from a famous correspondence between Blaise Pascal and Pierre de Fermat in 1654. The first published textbook on probability was *De Ratiociniis in Ludo Aleae* (On Reasoning in a Game of Chance) by Huygens (1657). The "laziness and ignorance" view of uncertainty was described by John Arbuthnot in the preface of his translation of Huygens (Arbuthnot, 1692): "It is impossible for a Die, with such determin'd force and direction, not to fall on such determin'd side, only I don't know the force and direction which makes it fall on such determin'd side, and therefore I call it Chance, which is nothing but the want of art."

The connection between probability and reasoning dates back at least to the nineteenth century: in 1819, Pierre Laplace said, "Probability theory is nothing but common sense reduced to calculation." In 1850, James Maxwell said, "The true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind."

There has been endless debate over the source and status of probability numbers. The **frequentist** position is that the numbers can come only from *experiments*: if we test 100 people and find that 10 of them have a cavity, then we can say that the probability of a cavity is approximately 0.1. In this view, the assertion "the probability of a cavity is 0.1" means that 0.1 is the fraction that would be observed in the limit of infinitely many samples. From any finite sample, we can estimate the true fraction and also calculate how accurate our estimate is likely to be.

The **objectivist** view is that probabilities are real aspects of the universe—propensities of objects to behave in certain ways—rather than being just descriptions of an observer's degree of belief. For example, the fact that a fair coin comes up heads with probability 0.5 is a propensity of the coin itself. In this view, frequentist measurements are attempts to observe these propensities. Most physicists agree that quantum phenomena are objectively probabilistic, but uncertainty at the macroscopic scale—e.g., in coin tossing—usually arises from ignorance of initial conditions and does not seem consistent with the propensity view.

Frequentist

Objectivist

The **subjectivist** view describes probabilities as a way of characterizing an agent's beliefs, rather than as having any external physical significance. The subjective **Bayesian** view allows any self-consistent ascription of prior probabilities to propositions, but then insists on proper Bayesian updating as evidence arrives.

Subjectivist

Even a strict frequentist position involves subjectivity because of the **reference class** problem: in trying to determine the outcome probability of a *particular* experiment, the frequentist has to place it in a reference class of "similar" experiments with known outcome frequencies. But what's the right class? I. J. Good wrote, "every event in life is unique, and every real-life probability that we estimate in practice is that of an event that has never occurred before" (Good, 1983, p. 27).

Reference class

For example, given a particular patient, a frequentist who wants to estimate the probability of a cavity will consider a reference class of other patients who are similar in important ways—age, symptoms, diet—and see what proportion of them had a cavity. If the dentist considers everything that is known about the patient—hair color, weight to the nearest gram, mother's maiden name—then the reference class becomes empty. This has been a vexing problem in the philosophy of science.

Pascal used probability in ways that required both the objective interpretation, as a property of the world based on symmetry or relative frequency, and the subjective interpretation, based on degree of belief—the former in his analyses of probabilities in games of chance, the latter in the famous "Pascal's wager" argument about the possible existence of God. However, Pascal did not clearly realize the distinction between these two interpretations. The distinction was first drawn clearly by James Bernoulli (1654–1705).

Leibniz introduced the "classical" notion of probability as a proportion of enumerated, equally probable cases, which was also used by Bernoulli, although it was brought to prominence by Laplace (1816). This notion is ambiguous between the frequency interpretation and the subjective interpretation. The cases can be thought to be equally probable either because of a natural, physical symmetry between them, or simply because we do not have any knowledge that would lead us to consider one more probable than another. The use of this latter, subjective consideration to justify assigning equal probabilities is known as the **principle of indifference**. The principle is often attributed to Laplace (1816), but he never used the name explicitly; Keynes (1921) did. George Boole and John Venn both referred to it as the **principle of insufficient reason**.

Principle of indifference

Principle of insufficient reason

The debate between objectivists and subjectivists became sharper in the 20th century. Kolmogorov (1963), R. A. Fisher (1922), and Richard von Mises (1928) were advocates of the relative frequency interpretation. Karl Popper's "propensity" interpretation (1959, first published in German in 1934) traces relative frequencies to an underlying physical symmetry. Frank Ramsey (1931), Bruno de Finetti (1937), R. T. Cox (1946), Leonard Savage (1954), Richard Jeffrey (1983), and E. T. Jaynes (2003) interpreted probabilities as the degrees of belief of specific individuals. Their analyses of degree of belief were closely tied to utilities and to behavior—specifically, to the willingness to place bets.

Rudolf Carnap offered a different interpretation of probability—not as the degree of belief that an individual actually has, but as the degree of belief that an idealized reasoner *should* have in a particular proposition *a*, given a particular body of evidence *e*. Carnap attempted to make this notion of degree of **confirmation** mathematically precise, as a logical relation between *a* and *e*. Currently it is believed that there is no unique logic of this kind; rather, any

such logic rests on a subjective prior probability distribution whose effect is diminished as more observations are collected.

The study of this relation was intended to constitute a mathematical discipline called **inductive logic**, analogous to ordinary deductive logic (Carnap, 1948, 1950). Carnap was not able to extend his inductive logic much beyond the propositional case, and Putnam (1963) showed by adversarial arguments that some difficulties were inherent. More recent work by Bacchus, Grove, Halpern, and Koller (1992) extends Carnap's methods to first-order theories.

The first rigorously axiomatic framework for probability theory was proposed by Kolmogorov (1950, first published in German in 1933). Rényi (1970) later gave an axiomatic presentation that took conditional probability, rather than absolute probability, as primitive.

In addition to de Finetti's arguments for the validity of the axioms, Cox (1946) showed that any system for uncertain reasoning that meets his set of assumptions is equivalent to probability theory. This gave renewed confidence to probability fans, but others were not convinced, objecting to the assumption that belief must be represented by a single number. Halpern (1999) describes the assumptions and shows some gaps in Cox's original formulation. Horn (2003) shows how to patch up the difficulties. Jaynes (2003) has a similar argument that is easier to read.

The Rev. Thomas Bayes (1702–1761) introduced the rule for reasoning about conditional probabilities that was posthumously named after him (Bayes, 1763). Bayes only considered the case of uniform priors; it was Laplace who independently developed the general case. Bayesian probabilistic reasoning has been used in AI since the 1960s, especially in medical diagnosis. It was used not only to make a diagnosis from available evidence, but also to select further questions and tests by using the theory of information value (Section 15.6) when available evidence was inconclusive (Gorry, 1968; Gorry *et al.*, 1973). One system outperformed human experts in the diagnosis of acute abdominal illnesses (de Dombal *et al.*, 1974). Lucas *et al.* (2004) provide an overview.

These early Bayesian systems suffered from a number of problems. Because they lacked any theoretical model of the conditions they were diagnosing, they were vulnerable to unrepresentative data occurring in situations for which only a small sample was available (de Dombal *et al.*, 1981). Even more fundamentally, because they lacked a concise formalism (such as the one to be described in Chapter 13) for representing and using conditional independence information, they depended on the acquisition, storage, and processing of enormous tables of probabilistic data. Because of these difficulties, probabilistic methods for coping with uncertainty fell out of favor in AI from the 1970s to the mid-1980s. Developments since the late 1980s are described in the next chapter.

The naive Bayes model for joint distributions has been studied extensively in the pattern recognition literature since the 1950s (Duda and Hart, 1973). It has also been used, often unwittingly, in information retrieval, beginning with the work of Maron (1961). The probabilistic foundations of this technique, described further in Exercise 12.BAYS, were elucidated by Robertson and Sparck Jones (1976). Domingos and Pazzani (1997) provide an explanation for the surprising success of naive Bayesian reasoning even in domains where the independence assumptions are clearly violated.

There are many good introductory textbooks on probability theory, including those by Bertsekas and Tsitsiklis (2008), Ross (2015), and Grinstead and Snell (1997). DeGroot and Schervish (2001) offer a combined introduction to probability and statistics from a Bayesian

standpoint, and Walpole *et al.* (2016) offer an introduction for scientists and engineers. Jaynes (2003) gives a very persuasive exposition of the Bayesian approach. Billingsley (2012) and Venkatesh (2012) provide more mathematical treatments, including all the complications with continuous variables that we have left out. Hacking (1975) and Hald (1990) cover the early history of the concept of probability, and Bernstein (1996) gives a popular account.