Chapter 2
BAYESIAN INFERENCE

The purpose I mean is, to show what reason
we have for believing that there are in the
constitution of things fixed laws according to
which events happen...
— Richard Price, 1763
(Introduction to Bayes' essay)

2.1 BASIC CONCEPTS

2.1.1 Probabilistic Formulation and Bayesian Inversion

Bayesian methods provide a formalism for reasoning about partial beliefs under conditions of uncertainty. In this formalism, propositions are given numerical parameters signifying the degree of belief accorded them under some body of knowledge, and the parameters are combined and manipulated according to the rules of probability theory. For example, if A stands for the statement "Ted Kennedy will seek the nomination for president in 1992," then $P(A \mid K)$ stands for a person's subjective belief in A given a body of knowledge K, which might include that person's assumptions about American politics, specific proclamations made by Kennedy, and an assessment of Kennedy's past and personality. In defining belief expressions, we often simply write $P(A)$ or $P(\neg A)$, leaving out the symbol K. This abbreviation is justified when K remains constant, since the main purpose of the quantifier P is to *summarize* K without explicating it. However, when the background information undergoes changes, we need to identify specifically the assumptions that account for our beliefs and articulate explicitly K or some of its elements.

In the Bayesian formalism, belief measures obey the three basic axioms of probability theory:

$$0 \le P(A) \le 1 \tag{2.1}$$

$$P(\mathit{Sure\ proposition}) = 1 \tag{2.2}$$

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive.} \tag{2.3}$$

The third axiom states that the belief assigned to any set of events is the sum of the beliefs assigned to its nonintersecting components. Hence, since any event A can be written as the union of the joint events (A and B) and (A and $\neg B$), their associated probabilities are given by

$$P(A) = P(A, B) + P(A, \neg B), \tag{2.4}$$

where $P(A, B)$ is short for $P(A$ and $B)$. More generally, if B_i, $i = 1, 2,...,n$, is a set of exhaustive and mutually exclusive propositions (called a *partition* or a *variable*), then $P(A)$ can be computed from $P(A, B_i)$, $i = 1, 2,...,n$, using the sum

$$P(A) = \sum_i P(A, B_i). \tag{2.5}$$

For example, the probability of A = "The outcomes of two dice are equal" can be computed by summing over the joint events (A and B_i) $i = 1, 2,...,6$, where B_i stands for the proposition "The outcome of the first die is i," yielding

$$P(A) = \sum_i P(A, B_i) = 6 \times \frac{1}{36} = \frac{1}{6}. \tag{2.6}$$

A direct consequence of Eqs. (2.2) and (2.4) is that a proposition and its negation must be assigned a total belief of unity,

$$P(A) + P(\neg A) = 1, \tag{2.7}$$

because one of the two statements is certain to be true.

The basic expressions in the Bayesian formalism are statements about *conditional probabilities*—e.g., $P(A \mid B)$—which specify the belief in A under the assumption that B is known with absolute certainty. If $P(A \mid B) = P(A)$, we say that A and B are *independent*. If $P(A \mid B,C) = P(A \mid C)$, we say that A and B are *conditionally independent* given C.

Contrary to the traditional practice of defining conditional probabilities in terms of joint events,

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \tag{2.8}$$

Bayesian philosophers see the conditional relationship as more basic than that of joint events, i.e., more compatible with the organization of human knowledge. In this view, B serves as a pointer to a context or frame of knowledge, and $A \mid B$ stands for an event A in the context specified by B (e.g., a symptom A in the context of a disease B). Consequently, empirical knowledge invariably will be encoded in conditional probability statements, while belief in joint events, if it is ever needed, will be computed from those statements via the product

$$P(A, B) = P(A \mid B)\, P(B), \tag{2.9}$$

which is equivalent to Eq. (2.8). For example, it was somewhat unnatural to assess

$$P(A, B_i) = \frac{1}{36}$$

directly in Eq. (2.6). The mental process underlying such assessment presumes that the two outcomes are independent, so to make this assumption explicit the probability of the joint event (*Equality*, B_i) should be assessed from the conditional event (*Equality* $\mid B_i$) via the product

$$P(Equality \mid B_i) P(B_i) = P(Outcome\ of\ second\ die\ is\ i \mid B_i) P(B_i) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}.$$

As in Eq. (2.5), the probability of any event A can be computed by conditioning it on any set of exhaustive and mutually exclusive events B_i, $i = 1, 2, ..., n$:

$$P(A) = \sum_i P(A \mid B_i)\, P(B_i). \tag{2.10}$$

This decomposition provides the basis for hypothetical or "assumption-based" reasoning in the Bayesian formalism. It states that the belief in any event A is a weighted sum over the beliefs in all the distinct ways that A might be realized. For example, if we wish to calculate the probability that the outcome X of the first die will be greater than the outcome Y of the second, we can condition the event $A: X > Y$ on all possible values of X and obtain

$$\begin{aligned} P(A) &= \sum_{i=1}^{6} P(Y < X \mid X = i)\, P(X = i) \\ &= \sum_{i=1}^{6} P(Y < i)\, \frac{1}{6} = \sum_{i=1}^{6} \sum_{j=1}^{i-1} P(Y = j)\, \frac{1}{6} \\ &= \frac{1}{6} \sum_{i=2}^{6} \frac{i-1}{6} = \frac{5}{12}. \end{aligned}$$

It is worth reemphasizing that formulas like Eq. (2.10) are always understood to apply in some larger context K, which defines the assumptions taken as common knowledge (e.g., the fairness of dice rolling). Eq. (2.10) is really a shorthand notation for the statement

$$P(A \mid K) = \sum_{i} P(A \mid B_i, K)\, P(B_i \mid K). \qquad \textbf{(2.11)}$$

Another useful generalization of the product rule (Eq. (2.9)) is the so-called *chain rule* formula. It states that if we have a set of n events, $E_1, E_2, \ldots, E_n$, then the probability of the joint event $(E_1, E_2, \ldots, E_n)$ can be written as a product of n conditional probabilities:

$$P(E_1, E_2, \ldots, E_n) = P(E_n \mid E_{n-1}, \ldots, E_2, E_1) \ldots P(E_2 \mid E_1)\, P(E_1). \qquad \textbf{(2.12)}$$

This product can be derived by repeated application of Eq. (2.9), in any convenient order.

The heart of Bayesian techniques lies in the celebrated inversion formula,

$$P(H \mid e) = \frac{P(e \mid H) P(H)}{P(e)}, \qquad \textbf{(2.13)}$$

which states that the belief we accord a hypothesis H upon obtaining evidence e can be computed by multiplying our previous belief $P(H)$ by the likelihood $P(e \mid H)$ that e will materialize if H is true. $P(H \mid e)$ is sometimes called the posterior probability (or simply *posterior*), and $P(H)$ is called the prior probability (or *prior*). The denominator $P(e)$ of Eq. (2.13) hardly enters into consideration because it is merely a normalizing constant $P(e) = P(e \mid H)P(H) + P(e \mid \neg H)P(\neg H)$, which can be computed by requiring that $P(H \mid e)$ and $P(\neg H \mid e)$ sum to unity.

Whereas a formal mathematician might dismiss Eq. (2.13) as a tautology stemming from the definition of conditional probabilities,

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \text{and} \quad P(B \mid A) = \frac{P(A, B)}{P(A)}, \tag{2.14}$$

the Bayesian subjectivist regards Eq. (2.13) as a normative rule for updating beliefs in response to evidence. In other words, while the mathematician views conditional probabilities as mathematical constructs, as in Eq. (2.14), the Bayes adherent views them as primitives of the language and as faithful translations of the English expression "..., given that I know A." Accordingly, Eq. (2.14) is not a definition but an empirically verifiable relationship between English expressions. It asserts, among other things, that the belief a person attributes to B after discovering A is never lower than that attributed to $A \wedge B$ before discovering A. Also, the ratio between these two beliefs will increase proportionally with the degree of surprise $[P(A)]^{-1}$ one associates with the discovery of A.

The importance of Eq. (2.13) is that it expresses a quantity $P(H \mid e)$—which people often find hard to assess—in terms of quantities that often can be drawn directly from our experiential knowledge. For example, if a person at the next gambling table declares the outcome ''Twelve,'' and we wish to know whether he was rolling a pair of dice or spinning a roulette wheel, our models of the gambling devices readily yield the quantities $P(Twelve \mid Dice)$ and $P(Twelve \mid Roulette)$—1/36 for the former and 1/38 for the latter. Similarly, we can judge the prior probabilities $P(Dice)$ and $P(Roulette)$ by estimating the number of roulette wheels and dice tables at the casino. Issuing a direct judgment of $P(Dice \mid Twelve)$ would have been much more difficult; only a specialist in such judgments, trained at the very same casino, could do it reliably.

To complete this brief introduction, we need to discuss the notion of *probabilistic models.* A probabilistic model is an encoding of probabilistic information that permits us to compute the probability of every well-formed sentence S in accordance with the axioms of Eqs. (2.1) through (2.3). Starting with a set of atomic propositions $A, B, C, \ldots$, the set of well-formed sentences consists of all Boolean formulas involving these propositions, e.g., $S = (A \vee B) \wedge \neg C$. The traditional method of specifying probabilistic models employs a joint distribution function, namely, a function that assigns nonnegative weights to every *elementary event* in the language (an elementary event being a conjunction in which every atomic proposition or its negation appears once), such that the sum of the weights adds up to 1. For example, if we have three atomic propositions, A, B, and C, a joint distribution function should assign nonnegative weights to all eight combinations: $(A \wedge B \wedge C)$, $(A \wedge B \wedge \neg C)$, ..., $(\neg A \wedge \neg B \wedge \neg C)$, such that the eight weights sum to 1.

It is sometimes convenient to view the conjunctive formulas corresponding to elementary events as points, and to regard other formulas as sets made up of these points. Since every Boolean formula can be expressed as a disjunction of elementary events, and since the elementary events are mutually exclusive, we can always compute $P(S)$ using the additive axiom (Eq. (2.3)). Conditional probabilities can be computed the same way, using Eq. (2.14). Thus, any joint probability function represents a complete probabilistic model.

Joint distribution functions are mathematical constructs of primarily theoretical use. They allow us to determine quickly whether we have sufficient information to specify a complete probabilistic model, whether the information we have is consistent, and at what point additional information is needed. The criterion is simply to check whether the information available is sufficient for uniquely determining the probability of every elementary event in the domain, and whether the probabilities add up to 1.

In practice, however, joint distribution functions are rarely specified explicitly. In the analysis of continuous random variables, the distribution functions are given by algebraic expressions such as those describing normal or exponential distributions, while for discrete variables, indirect representation methods have been developed, where the overall distribution is inferred from local relationships among small groups of variables. Network approaches, the most promising of these representations, provide the basis of discussion throughout this book. Their use will be illustrated in the following few sections, then given a more formal treatment in Chapter 3.

2.1.2 Combining Predictive and Diagnostic Supports

The essence of Bayes' Rule (Eq. (2.13)) is conveniently portrayed using the *odds* and *likelihood ratio* parameters. Dividing Eq. (2.13) by the complementary form for $P(\neg H \mid e)$, we obtain

$$\frac{P(H \mid e)}{P(\neg H \mid e)} = \frac{P(e \mid H)}{P(e \mid \neg H)} \frac{P(H)}{P(\neg H)} . \qquad (2.15)$$

Defining the *prior odds* on H as

$$O(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1-P(H)} \qquad (2.16)$$

and the *likelihood ratio* as

$$L(e \mid H) = \frac{P(e \mid H)}{P(e \mid \neg H)} , \qquad (2.17)$$

the *posterior odds*

$$O(H \mid e) = \frac{P(H \mid e)}{P(\neg H \mid e)} \tag{2.18}$$

are given by the product

$$O(H \mid e) = L(e \mid H)\, O(H). \tag{2.19}$$

Thus, Bayes' Rule dictates that the overall strength of belief in a hypothesis H, based on both our previous knowledge K and the observed evidence e, should be the product of two factors: the prior odds $O(H)$ and the likelihood ratio $L(e \mid H)$. The first factor measures the *predictive* or *prospective* support accorded to H by the background knowledge alone, while the second represents the *diagnostic* or *retrospective* support given to H by the evidence actually observed.

Strictly speaking, the likelihood ratio $L(e \mid H)$ might depend on the content of the tacit knowledge base K. However, the power of Bayesian techniques comes primarily from the fact that in causal reasoning the relationship $P(e \mid H)$ is fairly local, namely, given that H is true, the probability of e can be estimated naturally and is not dependent on many other propositions in the knowledge base. For example, once we establish that a patient suffers from a given disease H, it is natural to estimate the probability that he will develop a certain symptom e. The organization of medical knowledge rests on the paradigm that a symptom is a stable characteristic of the disease and should therefore be fairly independent of other factors, such as epidemic conditions, previous diseases, and faulty diagnostic equipment. For this reason the conditional probabilities $P(e \mid H)$, as opposed to $P(H \mid e)$, are the atomic relationships in Bayesian analysis. The former possess modularity features similar to logical production rules. They convey a degree of confidence in rules such as "If H then e," a confidence that persists regardless of what other rules or facts reside in the knowledge base.

EXAMPLE 1: Imagine being awakened one night by the shrill sound of your burglar alarm. What is your degree of belief that a burglary attempt has taken place? For illustrative purposes we make the following judgments: (a) There is a 95% chance that an attempted burglary will trigger the alarm system—$P(Alarm \mid Burglary) = 0.95$; (b) based on previous false alarms, there is a slight (1 percent) chance that the alarm will be triggered by a mechanism other than an attempted burglary—$P(Alarm \mid No\ burglary) = 0.01$; (c) previous crime patterns indicate that there is a one in ten thousand chance that a given house will be burglarized on a given night—$P(Burglary) = 10^{-4}$.

Putting these assumptions together using Eq. (2.19), we obtain

$$O(Burglary \mid Alarm) = L(Alarm \mid Burglary)\, O(Burglary)$$

$$= \frac{0.95}{0.01}\, \frac{10^{-4}}{1-10^{-4}} = 0.0095.$$

So, from

$$P(A) = \frac{O(A)}{1 + O(A)}, \qquad (2.20)$$

we have

$$P(Burglary \mid Alarm) = \frac{0.0095}{1{+}0.0095} = 0.00941.$$

Thus, the retrospective support imparted to the burglary hypothesis by the alarm evidence has increased its degree of belief almost a hundredfold, from one in ten thousand to 94.1 in ten thousand. The fact that the belief in burglary is still below 1% should not be surprising, given that the system produces a false alarm almost once every three months. Notice that it was not necessary to estimate the absolute values of the probabilities $P(Alarm \mid Burglary)$ and $P(Alarm \mid No\ burglary)$. Only their ratio enters the calculation, so a direct estimate of this ratio could have been used instead.

2.1.3 Pooling of Evidence

Assume that the alarm system consists of a collection of N burglary detection devices, each one sensitive to a different physical mechanism (air turbulence, temperature variation, pressure, radar waves, etc.) and each one producing a distinct sound.

Let H stand for the event that a burglary took place and let e^k stand for the evidence obtained from the k-th detector, with e_1^k representing an activated detector and e_0^k representing a silent detector. The reliability (and sensitivity) of each detector is characterized by the probabilities $P(e_1^k \mid H)$ and $P(e_1^k \mid \neg H)$, or more succinctly by their ratio:

$$L(e_1^k \mid H) = \frac{P(e_1^k \mid H)}{P(e_1^k \mid \neg H)}. \qquad (2.21)$$

If some detectors are triggered while others remain silent, we have conflicting evidence on our hands, and the combined belief in the hypothesis H is computed by Eq. (2.19):

$$O(H \mid e^1, e^2, \ldots, e^N) = L(e^1, e^2, \ldots, e^N \mid H)\, O(H). \qquad (2.22)$$

Eq. (2.22) could require an enormous data base, because we need to specify the probabilities of activation for every subset of detectors, conditioned on H and on $\neg H$. Fortunately, reasonable assumptions of conditional independence can reduce this storage requirement drastically. Assuming that the state of each detector depends only on whether a burglary took place and is thereafter independent of the state of other detectors, we can write

$$P(e^1,e^2, ..., e^N | H) = \prod_{k=1}^{N} P(e^k | H) \tag{2.23}$$

and

$$P(e^1,e^2, ..., e^N | \neg H) = \prod_{k=1}^{N} P(e^k | \neg H), \tag{2.24}$$

which lead to

$$O(H | e^1,e^2, ..., e^N) = O(H) \prod_{k=1}^{N} L(e^k | H). \tag{2.25}$$

Thus, the individual characteristics of each detector are sufficient for determining the combined impact of any group of detectors.

2.1.4 Recursive Bayesian Updating

One of the attractive features of Bayes' updating rule is its amenability to recursive and incremental computation schemes. Let H denote a hypothesis, $\mathbf{e}_n = e^1, e^2,...,e^n$ denote a sequence of data observed in the past, and e denote a new fact. A brute-force way to calculate the belief in H, $P(H | \mathbf{e}_n, e)$ would be to append the new datum e to the past data $\mathbf{e}_n$ and perform a global computation of the impact on H of the entire data set $\mathbf{e}_{n+1} = \{\mathbf{e}_n, e\}$. Such a computation would be uneconomical for several reasons. First, the entire stream of past data must be available at all times. Also, as time goes on and the set $\mathbf{e}_n$ increases, the computation of $P(H | \mathbf{e}_n, e)$ becomes more and more complex. Under certain conditions, this computation can be significantly curtailed by incremental updating; once we have computed $P(H | \mathbf{e}_n)$, we can discard the past data and compute the impact of the new datum by the formula

$$P(H | \mathbf{e}_n, e) = P(H | \mathbf{e}_n) \frac{P(e | \mathbf{e}_n, H)}{P(e | \mathbf{e}_n)}. \tag{2.26}$$

Thus, comparing Eq. (2.26) and Eq. (2.13), we see that the old belief $P(H | \mathbf{e}_n)$ assumes the role of the prior probability in the computation of new impact; it completely summarizes the past experience and for updating need only be multiplied by the likelihood function $P(e | \mathbf{e}_n, H)$, which measures the probability of the new datum e, given the hypothesis and the past observations.

This recursive formulation still would be cumbersome but for the fact that the likelihood function is often independent of the past data and involves only e and H. For example, the likelihood that a patient will develop a certain symptom, given that he definitely suffers from a disease H, is normally independent of what symptoms the patient had in the past. This conditional independence condition, which gave rise to the product expression in Eqs. (2.23) through (2.25), allows us to write

$$P(e \mid \mathbf{e}_n, H) = P(e \mid H) \quad \text{and} \quad P(e \mid \mathbf{e}_n, \neg H) = P(e \mid \neg H), \tag{2.27}$$

and after dividing Eq. (2.26) by the complementary equation for $\neg H$, we obtain

$$O(H \mid \mathbf{e}_{n+1}) = O(H \mid \mathbf{e}_n)\, L(e \mid H), \tag{2.28}$$

which also is obtainable from the product form of Eq. (2.25).

Eq. (2.28) describes a simple recursive procedure for updating the posterior odds—upon the arrival of each new datum e, we multiply the current posterior odds $O(H \mid \mathbf{e}_n)$ by the likelihood ratio of e. This procedure sheds new light on the relationship between the prior odds $O(H)$ and the posterior odds $O(H \mid \mathbf{e}_n)$; the latter can be viewed as the prior odds relative to the next observation, while the former are nothing but posterior odds that have evolved from previous observations not included in $\mathbf{e}_n$.

If we take the logarithm of Eq. (2.28), the incremental nature of the updating process becomes more apparent. Writing

$$\log O(H \mid \mathbf{e}_n, e) = \log O(H \mid \mathbf{e}_n) + \log L(e \mid H), \tag{2.29}$$

we can view the log of the likelihood ratio as a weight, carried by the evidence e, which additively sways the belief in H one way or the other. Evidence supporting the hypothesis carries positive weight, and evidence that opposes it carries negative weight.

The simplicity and appeal of the log-likelihood calculation has led to a wide variety of applications, especially in intelligence-gathering tasks. For each new report, an intelligence analyst can estimate the likelihood ratio L. Using a log-log paper, the contribution of the report can easily be incorporated into the already accumulated overall belief in H. This method also facilitates retracting or revising beliefs in case a datum is found to be in error. If the erroneous datum is e, and the correct one is e', then to rectify the error one need only compute the difference

$$\Delta = \log L(e' \mid H) - \log L(e \mid H)$$

and add Δ to the accumulated log-odds of Eq. (2.29).

The ability to update beliefs recursively depends heavily on the conditional independence relation formulated in Eqs. (2.23) and (2.24) and will exist only when knowledge of H (or $\neg H$) renders past observations totally irrelevant with regard to future observations. It will not be applicable, for example, if the hypothesis H influences the observations only indirectly, via several causal links. For instance, suppose that in our burglar alarm example we cannot hear the alarm sound directly but must rely on the testimony of other people. Because the burglary hypothesis has an indirect influence on the witnesses, the testimony of one witness (regarding the alarm) affects our expectation of the next witness's testimony even when we are absolutely sure that a burglary has occurred. The two testimonies will, however, become independent once we know the actual state of the alarm system. For that reason, decision analysts (e.g., Kelly and Barclay [1973], Schum and Martin [1982]) have gone to great lengths to retain incremental updating in the context of "cascaded" inferencing. The issue will be discussed further in Section 2.2 and will be given full treatment, using network propagation techniques, in Chapter 4.

2.1.5 Multi-Valued Hypotheses

The assumption of conditional independence in Eqs. (2.23) and (2.24) is justified if both the failure of a detector to react to an attempted burglary and the factors that can cause it to be activated prematurely depend solely on mechanisms intrinsic to the individual detection systems, such as low sensitivity and internal noise. But if false alarms can be caused by external circumstances affecting a select group of sensors, such as a power failure or an earthquake, then the two hypotheses H = *Burglary* and $\neg H$ = *No burglary* may be too broad to allow sensor independence, and additional refinement of the hypothesis space may be necessary. This condition usually occurs when a proposition or its negation encompasses several possible states, each associated with a distinct set of evidence. For example, the hypothesis *Burglary* encompasses either *Break-in through the door* or *Break-in through a window*, and since each mode of entry has a distinct effect on the sensors, the modes ought to be spelled out separately. Similarly, the state *No burglary* allows the possibilities *Ordinary peaceful night, Night with earthquake*, and *Attempted entry by the neighbor's dog*, each influencing the sensors in a unique way. Eq. (2.24) might hold for each of these conditions, but not for their aggregate, *No burglary*. For this reason, it is often necessary to refine the hypothesis space beyond binary propositions and group the hypothesis into multi-valued *variables*, where each variable reflects a set of exhaustive and mutually exclusive hypotheses.

EXAMPLE 2: We assign the variable $H = \{H_1, H_2, H_3, H_4\}$ to the following set of conditions:

H_1 = *No burglary, animal entry.*
H_2 = *Attempted burglary, window break-in.*
H_3 = *Attempted burglary, door break-in.*
H_4 = *No burglary, no entry.*

Each evidence variable E^k can also be multi-valued (e.g., e_1^k = *No sound* , e_2^k = *Low sound*, e_3^k = *High sound*), in which case the causal link between H and E^k is quantified by an $m \times n$ matrix M^k, where m and n are the number of values that H and E^k, respectively, might take, and the (i, j)-th entry of M^k stands for

$$M_{ij}^k = P(e_j^k \mid H_i). \tag{2.30}$$

For example, the matrix below could represent the sensitivity of the k-th detector to the four conditions in H:

	e_1^k	e_2^k	e_3^k
	(no sound)	(low sound)	(high sound)
H_1	0.5	0.4	0.1
H_2	0.06	0.5	0.44
H_3	0.5	0.1	0.4
H_4	1	0	0.

Given a set of evidence readings $e^1, e^2, ..., e^k, ..., e^N$, the overall belief in the i-th hypothesis H_i is (by Eq. (2.13))

$$P(H_i \mid e^1, ..., e^N) = \alpha P(e^1, ..., e^N \mid H_i) P(H_i), \tag{2.31}$$

where $\alpha = [P(e^1,...,e^N)]^{-1}$ is a normalizing constant to be computed by requiring that Eq. (2.31) sum to unity (over i). Assuming conditional independence with respect to each H_i, we obtain

$$P(H_i \mid e^1, ..., e^N) = \alpha P(H_i) [\prod_{k=1}^{N} P(e^k \mid H_i)]. \tag{2.32}$$

Thus, the matrices $P(e^k \mid H_i)$ now play the role of the likelihood ratios in Eq. (2.25). If for each detector reading e^k we define the *likelihood vector*

$$\lambda^k = (\lambda_1^k, \lambda_2^k, ..., \lambda_m^k), \tag{2.33}$$

$$\lambda_i^k = P(e^k \mid H_i), \tag{2.34}$$

then Eq. (2.32) is computed by a simple vector-product process. First the individual likelihood vectors are multiplied together, term by term, to form an overall likelihood vector $\Lambda = \lambda^1, ..., \lambda^N$, namely,

$$\Lambda_i = \prod_{k=1}^{N} P(e^k | H_i). \tag{2.35}$$

Then we obtain the overall belief vector $P(H_i | e^1, ..., e^N)$ by the product

$$P(H_i | e^1, ..., e^N) = \alpha P(H_i) \Lambda_i, \tag{2.36}$$

which is reminiscent of Eq. (2.25).

Note that only the relative magnitudes of the conditional probabilities in Eq. (2.34) need be estimated; their absolute magnitudes do not affect the final result because α can be determined later, via the requirement $\sum_i P(H_i | e^1, ..., e^N) = 1$.

EXAMPLE 3: Let us assume that our alarm system contains two detectors having identical characteristics, given by the matrix of Example 2. Furthermore, let us represent the prior probabilities for the hypotheses in Example 2 with the vector $P(H_i) = (0.099, 0.009, 0.001, 0.891)$ and assume that detector 1 was heard to issue a high sound while detector 2 remained silent. From Eq. (2.34) we have

$$\lambda^1 = (0.1, 0.44, 0.4, 0), \quad \lambda^2 = (0.5, 0.06, 0.5, 1),$$

$$\Lambda = \lambda^1 \lambda^2 = (0.05, 0.0264, 0.2, 0),$$

$$P(H_i | e^1, e^2) = \alpha (4.95, 0.238, 0.20, 0) 10^{-3} = (0.919, 0.0439, 0.0375, 0),$$

from which we conclude that the chance of an attempted burglary (H_2 or H_3) is $0.0439 + 0.0375 = 8.14\%$.

Of course, the updating of belief need not be delayed until all the evidence is collected but can be carried out incrementally. For example, if we first observe $e^1 = \textit{High sound}$, our belief in H calculates to

$$P(H_i | e^1) = \alpha (0.0099, 0.00396, 0.0004, 0) = (0.694, 0.277, 0.028, 0).$$

This probability now serves as a prior belief with respect to the next datum, and after we observe $e^2 = \textit{No sound}$, it updates to

$$P(H_i | e^1, e^2) = \alpha' \lambda_i^2 \cdot P(H_i | e^1) = \alpha' (0.347, 0.0166, 0.014, 0)$$

$$= (0.919, 0.0439, 0.0375, 0),$$

as before. Thus, the quiescent state of detector 2 lowers the probability of an attempted burglary from 30.5% to 8.14%.

2.2 HIERARCHICAL MODELING

2.2.1 Uncertain Evidence (Cascaded Inference)

One often hears the claim that Bayesian techniques cannot handle uncertain evidence because the basic building block in these techniques is the relationship $P(A \mid B)$, which requires that the conditioning event B be known with certainty. To see the difficulties that led to this myth, let us modify slightly the alarm scenario.

EXAMPLE 4: Mr. Holmes receives a telephone call from his neighbor Dr. Watson, who states that he hears the sound of a burglar alarm from the direction of Mr. Holmes's house. While preparing to rush home, Mr. Holmes recalls that Dr. Watson is known to be a tasteless practical joker, and he decides to first call another neighbor, Mrs. Gibbon, who, despite occasional drinking problems, is far more reliable.

Since the evidence variable $S = Sound$ is now uncertain, we cannot use it as evidence in Eq. (2.19) but instead must apply Eq. (2.19) to the actual evidence at hand, W = *Dr. Watson's testimony*, and write

$$O(H \mid W) = L(W \mid H) O(H). \tag{2.37}$$

Unfortunately, the task of estimating $L(W \mid H)$ will be more difficult than estimating $L(S \mid H)$, because it requires mentally tracing a two-step process, as shown in Figure 2.1. Even if we obtain $L(W \mid H)$, we will not be able to combine it with other possible testimonies, say Mrs. Gibbon's (G), through a simple process of multiplication as in Eq. (2.35), because those testimonies will no longer be conditionally independent with respect to H. What Mrs. Gibbon is about to say depends only on whether an alarm sound can be heard in the neighborhood, not on whether a burglary actually took place. Thus, we cannot assume $P(G \mid Burglary, W) = P(G \mid Burglary)$; the joint event of a burglary and Dr. Watson's testimony constitutes stronger evidence for the occurrence of the alarm sound than does the burglary alone.

Given the level of detail used in our story, it is more reasonable to assume that the testimony (W and G) and the hypothesis (H) are mutually independent once we know whether the alarm sound was actually triggered. In other words, each neighbor's testimony depends directly on the alarm sound (S) and is influenced only indirectly by the possible occurrence of a burglary (H) or by the other testimony (see Figure 2.1).

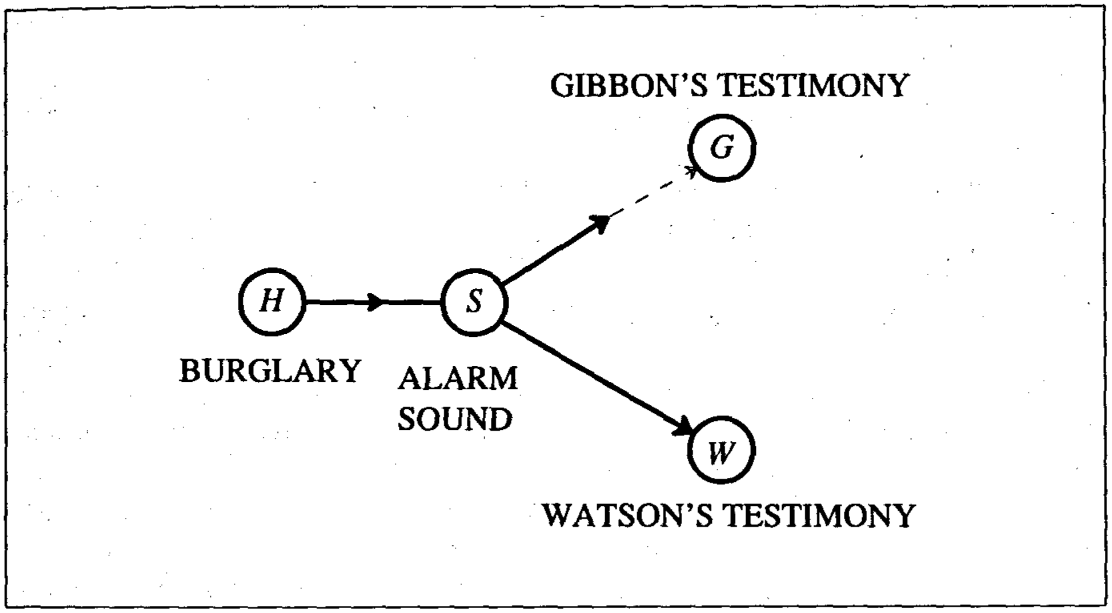

Figure 2.1. *The alarm sound (S), supported by unreliable testimonies (W and G), represents an uncertain evidence for a burglary (H).*

These considerations can easily be incorporated into the Bayesian formalism. Using Eq. (2.11), we simply condition and sum Eq. (2.31) over all possible states of the intermediate variable S and obtain

$$\begin{aligned} P(H_i \mid G, W) &= \alpha P(G, W \mid H_i)P(H_i) \\ &= \alpha P(H_i)\sum_j P(G, W \mid H_i, S_j)P(S_j \mid H_i), \end{aligned} \qquad \textbf{(2.38)}$$

where S_j, $j=1, 2$ stands for the two possible states of the alarm system, namely, $S_1 = \mathit{Sound\ ON}$ and $S_2 = \mathit{Sound\ OFF}$. Moreover, the conditional independence of G, W, and H_i with respect to the mediating variable S allows us to state

$$P(G, W \mid H_i, S_j) = P(G \mid S_j)P(W \mid S_j), \qquad \textbf{(2.39)}$$

and Eq. (2.38) becomes

$$P(H_i \mid G, W) = \alpha P(H_i)\sum_j P(G \mid S_j)P(W \mid S_j)P(S_j \mid H_i). \qquad \textbf{(2.40)}$$

The final computation can be interpreted as a three-stage process. First, the local likelihood vectors $P(G \mid S_j)$ and $P(W \mid S_j)$ are multiplied to obtain a combined likelihood vector

$$\Lambda_j(S) = P(e \mid S_j) = P(G \mid S_j)\, P(W \mid S_j), \qquad \textbf{(2.41)}$$

where e stands for the total evidence collected (G and W). Second, the vector $\Lambda_j(S)$ is multiplied by the link matrix $M_{ij} = P(S_j \mid H_i)$ to form the likelihood vector of the top hypothesis $\Lambda_i(H) = P(e \mid H_i)$. Finally, using the product rule of Eq. (2.24), we multiply $\Lambda_i(H)$ by the prior probability $P(H_i)$ to compute the overall belief in H_i.

This process demonstrates the psychological and computational roles of the mediating variable S. The conditional independence associated with S makes it a convenient anchoring point from which reasoning "by assumptions" can proceed effectively, because it decomposes the reasoning task into a set of independent subtasks. It permits us to use local chunks of information taken from diverse domains (e.g., $P(H_i)$, $P(G \mid S_j)$, $P(W \mid S_j)$, $P(S_j \mid H_i)$) and fit them together to form a global inference $P(H \mid e)$ in stages, using simple, local vector operations. It is this role which prompts us to posit that conditional independence is not a grace of nature for which we must wait passively, but rather a psychological necessity which we satisfy actively by organizing our knowledge in a specific way. An important tool in such organization is the identification of intermediate variables that induce conditional independence among observables; if such variables are not in our vocabulary, we create them. In medical diagnosis, for instance, when some symptoms directly influence each other, the medical profession invents a name for that interaction (e.g., "syndrome," "complication," "pathological state") and treats it as a new auxiliary variable that induces conditional independence; dependency between any two interacting symptoms is fully attributed to the dependencies of each on the auxiliary variable. It may be to reap the computational advantages associated with such independence that we organize most of our knowledge in causal hierarchies (see Chapter 8).

2.2.2 Virtual (Intangible) Evidence

Let us imagine a new development in the story of Mr. Holmes.

EXAMPLE 5: When Mr. Holmes calls Mrs. Gibbon, he soon realizes that she is somewhat tipsy. Instead of answering his question directly, she goes on and on about her latest back operation and about how terribly noisy and crime-ridden the neighborhood has become. When he finally hangs up, all Mr. Holmes can glean from the conversation is that there is probably an 80% chance that Mrs. Gibbon did hear an alarm sound from her window.

The Holmes-Gibbon conversation is the kind of evidence that is hard to fit into any formalism. If we try to estimate the probability $P(e \mid Alarm\ sound)$ we will get ridiculous numbers because it entails anticipating, describing, and assigning probabilities to all the possible paths Mrs. Gibbon's conversation might have taken under the circumstances. Alternatively, if we try to directly estimate $P(Alarm\ sound \mid e)$, we must be careful to clearly specify what other information was consulted in producing the estimate.

These difficulties arise whenever the task of gathering evidence is delegated to autonomous interpreters who, for various reasons, cannot explicate their interpretive process in full detail but nevertheless often produce informative conclusions that summarize the evidence observed. In our case, Mr. Holmes provides us with a direct mental judgment, based on Mrs. Gibbon's testimony, that the hypothesis *Alarm sound* should be accorded a confidence measure of 80%. The interpretation process remains hidden, however, and we cannot tell how much of the previously obtained evidence was considered in the process. Thus, it is impossible to integrate this probabilistic judgment with previously established beliefs unless we make additional assumptions.

The prevailing convention in the Bayesian formalism is to assume that probabilistic summaries of virtual evidence are produced independently of previous information; they are interpreted as local binary relations between the evidence and the hypothesis upon which it bears, independent of other information in the system. For this reason, we cannot interpret Mr. Holmes's summary as literally stating $P(S \mid G) = 0.80$. $P(S \mid G)$ should be sensitive to variations in crime rate information—$P(H)$—or equipment characteristics—$P(S \mid H)$. The impact of Gibbon's testimony should be impervious to such variations. Therefore, the measure $P(S \mid G)$ cannot represent the impact the phone conversation has on the truth of *Alarm sound*.

The likelihood ratio, on the other hand, meets this locality criterion, and for that reason probabilistic summaries of virtual evidence are interpreted as conveying likelihood information.† For example, Mr. Holmes's summary of attributing 80% credibility to the *Alarm sound* event can be interpreted as

$$P(G \mid Alarm\ sound) : P(G \mid No\ alarm\ sound) = 4{:}1. \qquad \textbf{(2.42)}$$

More generally, if the variable upon which the tacit evidence e impinges most directly has several possible states, $S_1, S_2, ..., S_i, ...$, we instruct the interpreter to estimate the relative magnitudes of the terms $P(e \mid S_i)$, perhaps by eliciting estimates of the ratios $P(e \mid S_i) : P(e \mid S_1)$. Since the absolute magnitudes do not

† It is interesting to note that an identical assumption has been tacitly incorporated into the calculus of certainty factors [Shortliffe 1976] if one interprets *CF* to stand for $(\lambda - 1) / (\lambda + 1)$ [Heckerman 1986b].

affect the calculations, we can update the beliefs as though this likelihood vector originated from an ordinary, logically definable event e.

For example, assuming that Mr. Watson's phone call already contributed a likelihood ratio of 9:1 in favor of the hypothesis *Alarm sound*, the combined weight of Watson's and Gibbon's testimonies would yield a likelihood vector $\Lambda_i(S) = P(W, G|S_i) = (36, 1)$. Now we can integrate this vector into the computation of Eq. (2.38). Using the numbers given in Example 1, we get

$$\Lambda_i(H) = \sum_j \Lambda_j(S) P(S_j|H_i) = \begin{pmatrix} 0.95 & 0.05 \\ 0.01 & 0.99 \end{pmatrix} \begin{pmatrix} 36 \\ 1 \end{pmatrix} = \begin{pmatrix} 34.25 \\ 1.35 \end{pmatrix},$$

$$\begin{aligned} P(H_i|G, W) &= \alpha\, \Lambda_i(H)\; P(H_i) = \alpha\,(34.25, 1.35)\;\; (10^{-4}, 1 - 10^{-4}) \\ &= (0.00253, 0.99747). \end{aligned} \tag{2.43}$$

It is important to verify that Mr. Holmes's 80% summarization is indeed based only on Mrs. Gibbon's testimony and not on prejudicial beliefs borrowed from the previous evidence (e.g., Watson's testimony or crime rate information); otherwise we are in danger of counting the same information twice. The likelihood ratio is in fact the only reasonable interpretation of Mr. Holmes's summarization that reflects a local binary relationship between the hypothesis and the evidence, unaffected by previous information [Heckerman 1986b].

An effective way of eliciting pure likelihood ratio estimates is to present the interpreter with a direct query: "How much more likely are we to obtain such an evidence under H, compared with the denial of H?" Alternatively, we can ask the interpreter to imagine that the evidence arrives in some standard state of belief, then request an estimate of how much the degree of belief in the hypothesis would be modified because of the evidence. In our example, if Mr. Holmes had a "neutral" belief in S before conversing with Mrs. Gibbon—$P(Alarm) = P(No\ alarm) = 1/2$—then the after-conversation estimate $P(Alarm|G) = 80\%$ would indeed correspond to a likelihood ratio of 4:1 in favor of *Alarm*. Bayesian practitioners claim that people are capable of retracing the origins of their beliefs and of entertaining hypothetical questions such as "What if you didn't receive Watson's call?" or "What is the increase in belief due to Gibbon's testimony alone?" This explains why interpretations of virtual evidence often are cast in terms of absolute probabilities, rather than probability changes or probability ratios. Evidently, the interpreter begins with some standard level of belief in the hypothesis (not necessary 50%), mentally assimilates the impact of the observed evidence, and then reports the updated posterior probability that emerges. However, it is not the final value but the ratio between the initial value and the final value that characterizes the impact of the evidence on the hypothesis, as this ratio is the only quantity that remains impervious to changes in the initial standard chosen. This issue will be discussed further in Section 2.3.3.

2.2.3 *Predicting Future Events*

One of the attractive features of causal models in the Bayesian formulation is the ease they lend to the prediction of future events such as the denouement of a social episode, the outcome of a given test, and the prognosis of a given disease. The need to facilitate such predictive tasks may in fact be the very reason that human beings have adopted causal schema for encoding experiential knowledge.

EXAMPLE 6: Immediately after his conversation with Mrs. Gibbon, as Mr. Holmes is preparing to leave his office, he recalls that his daughter is scheduled to arrive home at any minute. If greeted by an alarm sound, she probably ($P = 0.70$) would phone him for instructions. Now he wonders whether he should wait a few more minutes in case she calls.

To estimate the likelihood of our new target event, D = *Daughter will call*, we have to add a new causal link to the graph of Figure 2.1. Assuming that hearing an alarm sound is the only event that would induce Mr. Holmes's daughter to call, the new link, shown in Figure 2.2, should emanate from the variable S and be quantified by the following $P(D \mid S)$ matrix:

		D	
		will call	will not call
S	on	0.7	0.3
	off	0.0	1.0

Accordingly, to compute $P(D \mid \textit{All evidence})$ we write

$$P(D \mid e) = \sum_j P(D \mid S_j, e)\, P(S_j \mid e) = \sum_j P(D \mid S_j)\, P(S_j \mid e), \tag{2.44}$$

which means that the lengthy episodes with Mr. Watson and Mrs. Gibbon impart their influence on D only via the belief $P(S_j \mid e)$ that they induce on S.

It is instructive to see how $P(S_j \mid e)$ can be obtained from the previous calculation of $P(H_i \mid e)$. A natural temptation would be to use the updated belief

$P(H_i \mid e)$ as a new prior probability and, through rote, to write the conditioning equation

$$P(S_j \mid e) = \sum_i P(S_j \mid H_i)\, P(H_i \mid e). \tag{2.45}$$

This equation, however, is valid only in a very special set of circumstances. It would be wrong in our example because the changes in the belief of H actually originated from corresponding changes in S; reflecting these back to S would amount to counting the same evidence twice. The correct conditioning equation should be

$$P(S_j \mid e) = \sum_i P(S_j \mid H_i, e)\, P(H_i \mid e) \tag{2.46}$$

instead of Eq. (2.45). Since $P(S_j \mid H_i)$ may be different than $P(S_j \mid H_i, e)$, it follows that the evidence obtained affects not only the belief in H and S but also the strength of the causal link between H and S. At first glance, this realization makes Bayesian methods appear to be useless in handling a large number of facts; having to recalculate all the link matrices each time a new piece of evidence arrives would be an insurmountable computational burden.

Fortunately, there is a simple way of updating beliefs that circumvents this difficulty and uses only the original link matrices (see Chapter 4 for elaboration). The calculation of $P(S_j \mid e)$, for instance, can be performed as follows: Treating S as an intermediate hypothesis, Eq. (2.13) dictates .

$$P(S_j \mid e) = \alpha P(e \mid S_j)\, P(S_j) \tag{2.47}$$

The term $P(e \mid S_j)$ is the likelihood vector $\Lambda_j(S)$, which earlier was calculated as (36, 1), while the prior $P(S_j)$ is given by the matrix multiplication

$$P(S_j) = \sum_i P(S_j \mid H_i)\, P(H_i) = (10^{-4}, 1-10^{-4}) \begin{bmatrix} 0.95 & 0.05 \\ 0.01 & 0.99 \end{bmatrix} = (0.0101, 0.9899).$$

Together, we have

$$P(S_j \mid e) = \alpha\,(36, 1)\ \ (0.0101, 0.9899) = (0.2686, 0.7314),$$

which gives the event S_1 = *Alarm sound on* a credibility of 26.86% and gives the predicted event D = *Daughter will call* the probability

$$P(D \mid e) = \sum_i P(D \mid S_i)\, P(S_i \mid e) = (0.2686, 0.7314) \begin{bmatrix} 0.7 \\ 0 \end{bmatrix} = 0.188. \tag{2.48}$$

2.2.4 *Multiple Causes and "Explaining Away"*

Consider the following situation:

EXAMPLE 7: As he is debating whether or not to rush home, Mr. Holmes remembers reading in the instruction manual of his alarm system that the device is sensitive to earthquakes and can be accidentally ($P = 0.20$) triggered by one. He realizes that if an earthquake had occurred, it surely ($P = 0.40$) would be on the news. So he turns on his radio and waits for either an announcement over the air or a call from his daughter.

Mr. Holmes perceives two episodes as potential causes for the alarm sound—an attempted burglary and an earthquake. Though burglaries can be safely assumed to be independent of earthquakes, a positive radio announcement reduces the likelihood of a burglary, since it "explains away" the alarm sound. It does this even though the two causal events are perceived as individual variables (see Figure 2.2); general knowledge about earthquakes rarely intersects knowledge about burglaries.

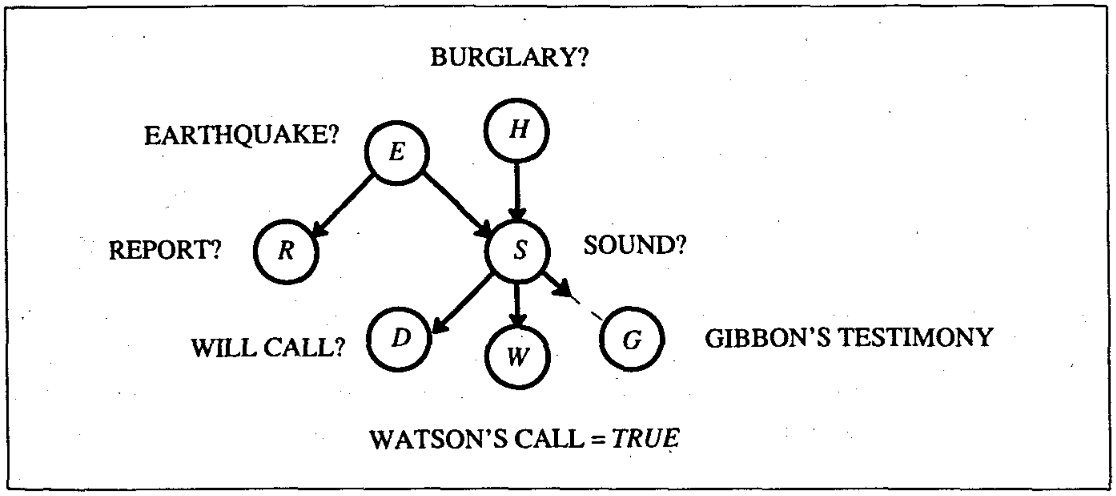

Figure 2.2. *A network depicting predicted events (D), explanatory variables (E and H) and evidence variables (W, G and R).*

This interaction among multiple causes is a prevailing pattern of human reasoning. (See Section 1.2.2.) When a physician discovers evidence in favor of one disease, it reduces the perceived likelihood of other diseases, although the patient may well be suffering from two or more disorders simultaneously. A suspect who provides an alternative explanation for being present at the scene of the crime appears less likely to be guilty, even though the explanation furnished does not preclude his having committed the crime.

To model this "sideways" interaction a matrix M should be assessed, giving the distribution of the consequence variable as a function of every possible combination of the causal variables. In our example, we should specify $M = P(S \mid E, H)$ where E stands for the variable $E = \{Earthquake,\ No\ earthquake\}$ and H stands for the hypothesis variable $H = \{Burglary,\ No\ Burglary\}$. Although this matrix is identical in form to the one described in Eq. (2.30), where several causal variables from example 2 were combined into one compound variable $\{H_1, H_2, H_3, H_4\}$, treating E and H as two separate entities has an advantage: it allows us to relate each of the variables to a separate set of evidence without consulting the other. For example, we can quantify the relation between E and R = *Radio announcement* by the probabilities $P(R \mid E)$ without having to consider the irrelevant event of burglary, as would be required by compounding the pair (E, H) into one variable. Moreover, upon confirmation of R, we can update the beliefs of E and H in two separate steps, mediated by the updating of S. This more closely resembles the local process used by people in tracing lines of evidence. (An updating scheme for networks with multiple-parent nodes is described in Section 4.3.)

If the number of causal factors k is large, estimating M may be troublesome because in principle it requires a table of size 2^{k+1}. In practice, however, people conceptualize causal relationships by creating hierarchies of small clusters of variables, and the interactions among the factors in each cluster are normally categorized into prestored, prototypical structures, each requiring about k parameters. Common examples of such prototypical structures are noisy OR-gates (i.e., any one of the factors is likely to trigger the effect), noisy AND-gates, and various enabling mechanisms (i.e., factors identified as having no influence of their own except that they enable other influences to become effective). In Example 7, it is reasonable to assume that the influences of burglaries and earthquakes on alarm systems is of the noisy OR-type; accordingly, only two parameters are needed, one describing the sensitivity of the alarm to earthquakes (in the absence of burglaries), the other describing its sensitivity to burglaries (in the absence of earthquakes). These prototypical structures will be treated formally in Section 4.3.2.

2.2.5 Belief Networks and the Role of Causality

In the preceding discussion we twice resorted to the use of diagrams. Figures 2.1 and 2.2 were not, however, presented merely for mnemonic or illustrative purposes. We will see that they convey important conceptual information, far more meaningful than the numerical estimates of the probabilities involved. The formal properties of such diagrams, called *Bayesian belief networks*, will be discussed in Section 3.3; here, we briefly outline their salient features.

Formally, Bayesian networks are directed acyclic graphs in which each node represents a random variable, or uncertain quantity, which can take on two or more possible values. The arcs signify the existence of direct causal influences between the linked variables, and the strengths of these influences are quantified by conditional probabilities. Informally, the structure of a Bayesian network can be determined by a simple procedure: We assign a vertex to each variable in the domain and draw arrows toward each vertex X_i from a select set Π_{X_i} of vertices perceived to be direct causes of X_i. The strengths of these direct influences are then quantified by assigning to each variable X_i a link matrix $P(x_i \mid \pi_{X_i})$, which represents judgmental estimates of the conditional probabilities of the event $X_i = x_i$, given any value combination π_{X_i} of the parent set Π_{X_i}. The conjunction of these local estimates specifies a complete and consistent global model (i.e., a joint distribution function) on the basis of which all probabilistic queries can be answered. The overall joint distribution function over the variables $X_1, ..., X_n$, is given by the product

$$P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i \mid \pi_{X_i}). \tag{2.49}$$

So, for example, the joint distribution corresponding to the network of Figure 2.2 is given by

$$P(h, e, r, s, d, w, g) = P(h)\, P(e)\, P(r \mid e)\, P(s \mid e, h)\, P(d \mid s) \tag{2.50}$$
$$P(w \mid s)\, P(g \mid s),$$

where lowercase symbols stand for the particular values (*TRUE* or *FALSE*) of the corresponding variables.

The advantage of network representation is that it allows people to express directly the fundamental qualitative relationship of "direct dependency." The network then displays a consistent set of additional direct and indirect dependencies and preserves it as a stable part of the model, independent of the numerical estimates. For example, Figure 2.2 demonstrates that the radio report (R) does not change the prospects of Holmes's daughter phoning (D), once we verify the actual state of the alarm system (S). This fact is conveyed by the network topology—showing S blocking the path between R and D—even though it was not considered explicitly during the construction of the network. It can be inferred visually from the linkages used to put the network together, and it will remain part of the model regardless of the numerical estimates of the link matrices.

The directionality of the arrows is essential for displaying *nontransitive* dependencies, i.e., S depends on both E and H, yet E and H are marginally independent (they become dependent only if S or any of its descendants are known). If the arcs were stripped of their arrows, some of these relationships would be misrepresented. It is this computational role of identifying what information is or is not relevant in any given situation that we attribute to the mental construct of causation. Causality modularizes our knowledge as it is cast from experience. By displaying the irrelevancies in the domain, causal schemata minimize the number of relationships that need to be considered while a model is constructed, and in effect legitimizes many future local inferences. The prevailing practice in rule-based expert systems of encoding knowledge by evidential rules (i.e., if effect then cause) is deficient in this respect. It usually fails to account for induced dependencies between causes (e.g., an earthquake explaining away the alarm sound), and if one ventures to encode these by direct rules, the number of rules becomes unmanageable [Shachter and Heckerman 1987].

In Chapter 3, we will present a formal characterization of dependencies expressible in both causal and non-causal networks. In Chapters 4 and 5 we will show that belief networks can also be used as inference engines, where the network topology provides both the storage locations and the timing information to sequence the computational steps involved in answering probabilistic queries. Examples of such queries are "What are the chances of a burglary, given that the radio announced an earthquake and my daughter did not call?" and "What is the most likely explanation of Watson's phone call?" Answers to such queries will be assembled by local, parallel message-passing processes, with minimal external supervision. The essential role of causality will be explored further in Chapters 8 and 10. Before advancing to these topics, we will use the next few sections to further elaborate on the philosophy of Bayesian inference and the role of networks in shaping human judgment.

2.3 EPISTEMOLOGICAL ISSUES OF BELIEF UPDATING

2.3.1 Patterns of Plausible Inference: Polya vs. Bayes?

In our previous discussion we suggested that once we encode knowledge in probabilistic terms and adhere to the rules of probability calculus, we are guaranteed never to produce paradoxical or counterintuitive conclusions. This raises an interesting question about how people produce intuitively acceptable conclusions using mechanisms that seem to involve only qualitative, nonnumerical relationships. If such mechanisms work for people, can we simulate them on digital machines and thus facilitate commonsense reasoning? This is indeed the ultimate objective of many works in AI, most notably nonmonotonic logics. The goal is to capture the patterns of plausible reasoning in nonnumerical terms, as principles governing English sentences that contain linguistic hedges such as "typically," "likely," and "surely." In this subsection we discuss some of the difficulties associated with using the logical approach instead of the probabilistic approach. A more detailed discussion will be given in Chapter 10.

POLYA'S PATTERNS OF PLAUSIBLE INFERENCE

George Polya (1887–1985) was one of the first mathematicians to attempt a formal characterization of qualitative human reasoning. In his 1954 book *Mathematics and Plausible Reasoning,* Polya argued that the process of discovery, even in as formal a field as mathematics, is guided by nondeductive inference mechanisms, entailing a lot of guesswork. "Patterns of plausible inference" was his term for the principles governing this guesswork.

Among the conspicuous patterns listed by Polya, we find the following four:

1. ***Inductive patterns:*** "The verification of a consequence renders a conjecture more credible."

   For example, the conjecture "It rained last night" becomes more credible when we verify the consequence "The ground is wet."

2. ***Successive verification of several consequences:*** "The verification of a new consequence counts more or less if the new consequence differs more or less from the former, verified consequences."

   For example, if in trying to substantiate the conjecture "All ravens are black," we observe *n* Australian ravens, all of them black, our subsequent confidence in the conjecture will be increased substantially if the $(n + 1)$-th raven is a black Brazilian raven rather than another black Australian raven.

3. ***Verification of improbable consequences:*** "The verification of a consequence counts more or less according as the consequence is more or less improbable in itself."

   For example, the conjecture "It rained last night" obtains more support from "The roof is leaking" than from the more common observation "The grass is wet."

4. ***Inference from analogy:*** "A conjecture becomes more credible when an analogous conjecture turns out to be true."

   For example, the conjecture "Of all objects displacing the same volume, the sphere has the smallest surface" becomes more credible when we prove the related theorem "Of all curves enclosing the same area, the circle has the shortest perimeter."

Polya also identified three main sub-patterns of inductive reasoning:

1. ***Examining a consequence:*** same as (1) above.

2. ***Examining a possible ground:*** "Our confidence in a conjecture can only diminish when a possible ground for the conjecture is exploded."

3. ***Examining a conflicting conjecture:*** "Our confidence in a conjecture can only increase when an incompatible rival conjecture is exploded."

These patterns can be further refined depending on whether propositions are verified categorically or just become more credible (Polya called this *shaded* verification).

Polya summarized the patterns and subpatterns by the following table:

	(1) Demonstrative	(2) Shaded Demonstrative	(3) Shaded Inductive	(4) Inductive
1. Examining a	$A \rightarrow B$	$A \rightarrow B$	$A \rightarrow B$	$A \rightarrow B$
consequnce	*B* false	*B* less cr.	*B* more cr.	*B* true
	A false	*A* less cr.	*A*s. more cr.	*A* more cr.
2. Examining a possible	$A \leftarrow B$	$A \leftarrow B$	$A \leftarrow B$	$A \leftarrow B$
ground	*B* true	*B* more cr.	*B* less cr.	*B* false
	A true	*A* more cr.	*A*s. less cr.	*A* less cr.
3. Examining a conflicting	$A \vert B$	$A \vert B$	$A \vert B$	$A \vert B$
conjecture	*B* true	*B* more cr.	*B* less cr.	*B* false
	A false	*A* less cr.	*A*s. more cr.	*A* more cr.

In this table, $A \rightarrow B$ means that *A* implies *B*, *cr.* is short for "credible," *s.* is short for "somewhat," and $A \mid B$ means that *A* is incompatible with *B*, i.e., *A* and *B* cannot both be true at the same time.

The patterns for "Examining a possible ground" are logically equivalent to those for "Examining a consequence." For example, entry (2,2) follows from (1,2) because $A \rightarrow B$ is logically equivalent to $(\neg B) \rightarrow (\neg A)$ and "*B* more cr." is equivalent to "$\neg B$ less cr." It still makes sense to restate row 2 separately since people do not readily perceive logical identities as psychological necessities; redundant inference rules are useful for dealing with logically equivalent but syntactically different situations.

WHY POLYA PREFERRED PROBABILITIES OVER LOGIC

When stated individually, each pattern in Polya's table appears plausible and is supported by many examples. However, after extracting many such conspicuous primitive patterns, Polya stopped short of proposing them as syllogistic axioms (or inference rules) for a new logic, capable of manipulating concepts such as "credible," "more credible," and "somewhat credible." Instead, Polya shelved this promising prospect and retreated to the safety of probability calculus—from which, supposedly, all the qualitative patterns of plausible inference should follow naturally and automatically, leaving no need to express them in symbolic terms.

The reason for Polya's sharp retreat is explained in Chapter 15 of his book and is based on the realization that primitive patterns of plausible reasoning, as reasonable as they appear and as syntactically similar as they are to logical syllogisms, are of basically different character than those syllogisms. Polya identified four basic differences between the two modes of reasoning, the most important being a feature he called *self-sufficiency* (today we use the term *monotonicity*)—new information, as long as it does not conflict with the premises, will never change the conclusions reached by demonstrative inferences.

> Nothing is needed beyond the premises to validate the conclusion and nothing can invalidate it if the premises remain solid.

By contrast, credibility levels established by plausible inferences are not "durable," as they may change with new information and are sensitive to the entire content of one's knowledge base. In Polya's words:

> In opposition to demonstrative inference, plausible inference leaves indeterminate a highly relevant point: the "strength" or the "weight" of the conclusion. This weight may depend not only on clarified grounds such as those expressed in the premises, but also on unclarified unexpressed grounds somewhere in the background of the person who draws the conclusion.

This is indeed the violation of modularity discussed in Chapter 1. Polya claimed, however, that in each inferential step the direction of change depends only on the premises considered at that step. For example, in the inductive pattern above, the credibility of the hypothesis can only increase with the discovery of its consequence, regardless of what background information we possess. This, we shall soon demonstrate, is not entirely correct (see also Figure 1.2). The gap between demonstrative and plausible inferences is, in fact, wider than that identified by Polya, i.e., not only the strength of the conclusions but also their "direction" depends on "unclarified unexpressed grounds somewhere in the background...."

Notwithstanding this oversight, Polya apparently chose the calculus of probability as a surrogate for logic because he believed that if things are set up properly, probability calculus will preserve all the qualitative patterns of plausible reasoning and, as a bonus, will provide the correct strengths of the conclusions. Polya, in fact, showed that all the patterns of his table follow from probability theory. For example, here is Polya's probabilistic proof of the inductive pattern

$$(A \rightarrow B)\ \&\ B \Longrightarrow A \textit{ more credible}: \tag{2.51}$$

Assume that in knowledge state S_1, A and B accrue the credibility measures $P(A)$ and $P(B)$, respectively, and that in state S_2, B is known to be true, i.e., $P_2(B) = 1$. One can defend the validity of Eq. (2.51) by showing that the inequality $P(A \mid B) > P(A)$ holds in all cases. Indeed, using Bayes' Rule (Eq. (2.13)) and the fact that $A \rightarrow B$ implies $P(B \mid A) = 1$, we obtain

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)} = \frac{P(A)}{P(B)}, \tag{2.52}$$

and, since $P(B) \leq 1$, we have

$$P(A \mid B) \geq P(A), \tag{2.53}$$

with equality holding iff either $P(A) = 0$ or $P(B) = 1$. Thus, it appears as though probability calculus lends unqualified confirmation to the inductive pattern (Eq. (2.51)).

Unfortunately, the above proof has a major flaw. The inequality in Eq. (2.53) is valid only in the rare and uninteresting case when B is the only new piece of information by which S_2 differs from S_1. To be used as a syllogistic rule of inference, the inductive pattern of Eq. (2.51) must be universally applicable to any two knowledge states S_1 and S_2. Yet, if S_2 differs from S_1 by two facts, say B and C, Eq. (2.51) no longer holds. An extreme case is when C directly opposes A. For example, consider the following three events:

A = "It rained last night."
B = "My grass is wet."
C = "My neighbor's grass is dry."

Any reasonable probabilistic model would yield

$$P(A \mid B) > P(A) \qquad \text{but} \qquad P(A \mid B, C) < P(A).$$

Although the left-hand side of Eq. (2.51) is satisfied in this example, the right-hand side of Eq. (2.51) contradicts our expectations whenever S_2 entails both B and C.

This might be construed as an artificial antl harmless example, because the knowledge base should also contain the rule $C \rightarrow \neg A$, which eventually will establish the falsity of A after Eq. (2.51) temporarily raises its credibility. A more convincing criticism would be to demonstrate the failure of Eq. (2.51) when C has no relation whatsoever to A. For example:

A = "It rained last night."
B = "My grass is wet."
C = "The sprinkler was on last night."

Here, the falsity of Eq. (2.51) could produce paradoxical and irreversible consequences. Perhaps it was this realization that prevented Polya from proposing his patterns as inference rules for a logic of plausible reasoning.

IF BAYES NEVER ERRS, WHY DID POLYA?

It is instructive, at this point, to reiterate the fundamental difference between the role of premises in logic and that of conditioning events in probability calculus (see Chapter 1). In logic, the truth of a premise B is all that is required for deducing the conclusion A. In probability calculus, the expression $P(A \mid B)$ specifically identifies B as the *only* information available—aside from the tacit knowledge base K, which we assume to be constant. This distinction is also reflected in significant computational differences between the two formalisms. The statement

$$P(A \mid B) = p$$

denotes totally different operational semantics than the production rule

$$\text{If } B \text{ then } A \quad (\text{with certainty } p). \tag{2.54}$$

The latter constitutes a carte blanche to execute a certain transformation on the database *whenever* it entails the truth of B, regardless of what other information it contains. The former permits us to draw certain conclusions (about the probability of A) *only when* the database entails B and no other information that can affect A once we know B.

This difference may explain why the designers of first-generation expert systems preferred the rule-based approach over straightforward Bayes' conditioning. The latter seems to require that we inspect the entire database at each step of the computation to see if it contains any new information that is relevant to A and not fully accounted for in B. In subsequent chapters, we shall see that networks provide an effective scheme for indexing this information so that local inspections are sufficient. On the other hand, systems based on rules such as Eq. (2.54) invariably run into the same paradoxical difficulties that plagued Polya's patterns. For example, such systems would draw the same conclusion from Eq. (2.54) whether B was established by $C' =$ "My shoes are muddy" or by $C =$ "The sprinkler was on last night." This is a clear violation of common sense. Section 10.3 provides a remedy to this problem, within the framework of rule-based systems.

It is also interesting to inquire why Polya's patterns are considered plausible if they are not supported by probability theory and they lead to paradoxical conclusions. The answer lies in the type of assumptions we all make when asked to judge the plausibility of an argument. Apparently, the inductive pattern (Eq. (2.51)) appears plausible to most people, because we tacitly assume that the truth of B is the *only* relevant change known to have taken place in the world. In other words, unless otherwise stated, all belief values, especially of events that precede B, are presumed to persist unaltered. Since changes in the belief of other propositions (e.g., "The sprinkler was on") are not mentioned in Eq. (2.51), we presume that in the transition from S_1 to S_2 the truth of B ("The grass is wet") was established by direct observation or reliable testimony and not as a consequence of other, unmentioned changes.

So far, we have discussed the difficulties associated with the nonmodularity of plausible inferences, i.e., the impropriety of drawing conclusions from certain truths in the database without checking other truths that may reside there. The following discussion will focus on an even tougher problem, *query sensitivity*, which stems not from neglecting facts that were learned but from neglecting to specify which facts could have been learned. In other words, plausible reasoning, unlike logical deduction, is sensitive not only to the information at hand but also to the query process by which the information was obtained.

2.3.2 The Three Prisoners Paradox: When the Bare Facts Won't Do

Three prisoners, *A*, *B*, and *C*, have been tried for murder, and their verdicts will be read and their sentences executed tomorrow morning. They know only that one of them will be declared guilty and will be hanged to die while the other two will be set free; the identity of the condemned prisoner is revealed to the very reliable prison guard, but not to the prisoners themselves.

In the middle of the night, Prisoner *A* calls the guard over and makes the following request: "Please give this letter to one of my friends—to one who is to be released. You and I know that at least one of them will be freed." The guard takes the letter and promises to do as told. An hour later Prisoner *A* calls the guard again and asks, "Can you tell me which of my friends you gave the letter to? It should give me no clue regarding my own status because, regardless of my fate, each of my friends had an equal chance of receiving my letter." The guard answers, "I gave the letter to Prisoner *B*; he will be released tomorrow." Prisoner *A* returns to his bed and thinks, "Before I talked to the guard, my chances of being executed were one in three. Now that he has told me that *B* will be released, only *C* and I remain, and my chances of dying have gone from 33.3% to 50%. What did I do wrong? I made certain not to ask for any information relevant to my own fate...."

SEARCHING FOR THE BARE FACTS

So far, we have the classical Three Prisoners story as described in many books of mathematical puzzles (e.g., Gardner [1961]). Students are asked to test which of the two values, 1/3 or 1/2, reflects prisoner *A*'s updated chances of perishing at dawn.† Let us attempt to resolve the issue using formal probability theory.

† A survey conducted in the author's class in 1984 showed 23 students in favor of 1/2 and 3 students in favor of 1/3. (The proportion was reversed in 1987, when class notes became available.)

Let I_B stand for the proposition "Prisoner B will be declared innocent," and let G_A stand for the proposition "Prisoner A will be declared guilty." Our task is to compute the probability of G_A given all the information obtained from the guard, i.e., to compute $P(G_A \mid I_B)$. Since $G_A \supset I_B$, we have $P(I_B \mid G_A) = 1$, and we can write

$$P(G_A \mid I_B) = \frac{P(I_B \mid G_A)\, P(G_A)}{P(I_B)} = \frac{P(G_A)}{P(I_B)} = \frac{1/3}{2/3} = 1/2. \qquad (2.55)$$

Thus, when facts are wrongly formulated, even the tools of probability calculus are insufficient safeguards against drawing counterintuitive or false conclusions. (Readers who are not convinced that the answer 50% is false are invited to eavesdrop on Prisoner A's further reflections: "... Worse yet, by sheer symmetry, my chances of dying would also have risen to 50% if the guard had named C instead of B—so my chances must have been 50% to begin with. I must be hallucinating....")

The fallacy in the preceding formulation arose from omitting the full context in which the answer was obtained by Prisoner A. By *context* we mean the entire range of answers one could possibly obtain (as in Eq. (2.30)), not just the answer actually obtained. In our example, it is important to know not only that the guard said, "B will be released," but also that the only other possible reply was "C will be released." Had the guard's answer, "B will be released," been a reply to the query "Will B die tomorrow?" the preceding analysis would have been correct.

A useful way of ensuring that we have considered the full context is to condition our analysis on events actually observed, not on their implications. In our example, the information in

$$I_B = \text{"}B\text{ will be declared innocent."}$$

was inferred from a more direct observation,

$$I'_B = \text{"Guard said that }B\text{ will be declared innocent."}$$

If we compute $P(G_A \mid I'_B)$ instead of $P(G_A \mid I_B)$, we get the correct answer:

$$P(G_A \mid I'_B) = \frac{P(I'_B \mid G_A)P(G_A)}{P(I'_B)} = \frac{1/2 \cdot 1/3}{1/2} = 1/3. \qquad (2.56)$$

The calculations in Eq. (2.56) differ from those in Eq. (2.55) in two ways. First, G_A subsumed I_B but does not subsume I'_B, because it is possible for A to be the condemned man and hear the guard report, "C will be released." Second, $P(I'_B)$ is 1/2, whereas $P(I_B)$ was 2/3. These differences exist because I'_B implies I_B but not vice versa; even if B is to be released, the guard can truthfully report, "C will be released"—if A is slated to die.

The lesson of the Three Prisoners paradox is that we cannot assess the impact of new information by considering only propositions implied by the information; we must also consider what information *could have* been reported.

THE THOUSAND PRISONER PROBLEM

Here is an extreme example, in which knowledge of the query context is even more important. Imagine you are one of one thousand prisoners awaiting sentencing with the knowledge that only one of you has been condemned. By sheer luck, you find a computer printout (with a court seal on it) listing 998 prisoners; each name is marked "innocent," and yours is not among them. Should your chances of dying increase from 1/1000 to 1/2? Most people would say yes, and rightly so.

Imagine, however, that while poring anxiously over the list you discover the query that produced it: "Print the names of any 998 innocent right-handed prisoners." If you are the only left-handed person around, would you not breathe a sigh of relief? Again, most people would.

Though the discovery of the query adds no logical conclusions to our knowledge base, it alters drastically the relative likelihood of events that remain unsettled. In other words, the range of possibilities is the same before and after you discover the query: Either you or the other unlisted prisoner will die. Yet the query renders the death of the other prisoner much more likely, because while you can blame your exclusion from the list on being left-handed, the other prisoner has no explanation except being found guilty. If the list contained 999 names marked "innocent," knowledge of the query would have no impact on your beliefs, because the only possible conclusion would be that you had been found guilty.

Again we see the computational virtues and epistemological weaknesses of crisp logic: It allows us to dispose of the query once we learn its ramifications but prevents the ramifications learned from altering the likelihood of uncertain events. Indeed, if we wish to determine merely which events are possible we need not retain the queries; the bare information will suffice. But if we are concerned also with the relative likelihood of these possible events, then the query process is necessary. If the process is unknown, then several likely processes can be conjectured and their average computed (see next subsection).

But first, let us return to the jail cell. Mathematically, the discovery of the query should restore your confidence of innocence to its original value of 99.9%, but psychologically you are more frightened than you were before you found the list. In your intuition, the realization that you are one of the only two potentially guilty individuals evidently carries more weight than Bayesian arithmetic does. Still, intuition is a multifaceted resource, and pondering further, you should muster intuitive support for the Bayesian conclusion as well: Finding the query after seeing the list should have the same effect as seeing the list after the query. In the second case, once you know the query, the list is useless to you, because it can contain neither your name nor the name of the guilty prisoner. Consequently, your chances of being found guilty should revert to 1/1000.

WHAT IF WE DON'T KNOW THE QUERY?

In the Three Prisoners story, we assumed that if both B and C were pardoned, the guard would give the letter to one or the other with equal (½) probability. What if we do not know the process by which the letter recipient is chosen, when A is condemned? The conditional probability $P(I'_B|G_A)$ can vary from 0 (the guard avoids B), to 1 (the guard avoids C). Likewise, the marginal probability $P(I'_B)$ can vary from $^1/_3$ to $^2/_3$. Treating $q = P(I'_B|G_A)$ as a variable, Eq. (2.56) can be written as follows:

$$P(G_A|I'_B) = \frac{P(I'_B|G_A)\, P(G_A)}{P(I'_B|G_A)\, P(G_A) + P(I'_B|G_B)\, P(G_B) + P(I'_B|G_C)\, P(G_C)}$$
$$= \frac{q\, ^1/_3}{q\, ^1/_3 + 0 + 1 \cdot {}^1/_3} = \frac{q}{1+q}\,. \qquad (2.57)$$

Thus, as q varies from 0 to 1, $P(G_A|I'_B)$ varies from 0 to ½.

Philosophers disagree on how to treat ignorance of this sort. Some favor the use of probability intervals, where the upper and lower probabilities represent the boundaries of our convictions, while others prefer an interpolation rule that selects a single probability model having some desirable properties. The Dempster-Shafer (D-S) formalism (see Chapter 9) is an example of the interval-based approach, while maximum-entropy techniques [Tribus 1969, Jaynes 1979] represent the single model approach.

Bayesian technique lies somewhere in between. For example, in the absence of information about the selection process used by the guard, several plausible models of the process are articulated, and their likelihoods are assessed. In our example, we may treat the critical parameter q as a random variable ranging from 0 to 1 and assess a probability distribution $f(q)$ on q, reflecting the likelihood that the guard will exhibit a bias q in favor of selecting B. This method yields a unique distribution on the variables previously considered, via

$$P(G_A|I'_B) = \int_0^1 \frac{q}{1+q}\, f(q|I'_B)\, dq = \frac{\int_0^1 q\, f(q) dq}{1 + \int_0^1 q\, f(q) dq}\,, \qquad (2.58)$$

but the method simultaneously maintains a distinction between conclusions based on definite models and conclusions based on uncertain models. For example, the knowledge that the choice between B and C is made at random is modeled by $q = ½$, while total lack of knowledge about the process is represented by $f(q) = 1\,,\ 0 \le q \le 1$. Though both models yield the same point values of $^1/_3$ for

$P(G_A \mid I'_B)$, they differ substantially in the way they allow new facts to be assimilated. Suppose Prisoner A recalls that the guard had a fistfight with C yesterday. This fact can easily be incorporated if q is a random variable (by updating $f(q)$), but not if q is a fixed value. The problem of representing uncertainty about probabilities will be discussed further in Section 7.3.

2.3.3 Jeffrey's Rule and the Problem of Autonomous Inference Agents

The Three Prisoners puzzle shows that before we can determine the implications of a new fact in our knowledge base, we must know the process by which the fact was learned—in particular, what other facts could have been gathered in that process. Such detailed knowledge is not always available; we often must respond to new information without having the slightest idea how it was collected. These situations occur when the gathering of information is delegated to autonomous agents, each using private procedures which for various reasons cannot be explicated in full detail.

OBSERVATION BY CANDLE LIGHT

Richard Jeffrey was the first to recognize the importance of this problem, and he devised a rule for handling it [Jeffrey 1965]. The autonomous agents used in Jeffrey's original example are our sensory organs, as described in the following passage:

> The agent inspects a piece of cloth by candlelight and gets the impression that it is green, although he concedes that it might be blue or, even (but very improbably), violet. If G, B and V are the propositions that the cloth is green, blue and violet, respectively, then the outcome of the observation might be that, whereas originally his degrees of belief in G, B and V were 0.30, 0.30 and 0.40, his degrees of belief in those same propositions after the observation are 0.70, 0.25 and 0.05. If there were a proposition E in his preference ranking [i.e., knowledge base] which described the precise quality of his visual experience in looking at the cloth, one would say that what the agent learned from the observation was that E is true. If his original subjective probability assignment was *prob*, his new assignment should then be $prob_E$, and we would have
>
> $$prob\ G = .30 \quad prob\ B = .30 \quad prob\ V = .40$$
>
> representing his opinions about the color of the cloth before the observation, but would have
>
> $$prob(G \mid E) = .70 \quad prob(B \mid E) = .25 \quad prob(V \mid E) = .05$$
>
> representing his opinions about the color of the cloth after the observation.... When the agent looks at the piece of cloth by candlelight there is a particular complex

> pattern of physical stimulation of his retina, on the basis of which his beliefs about the possible colors of the cloth change in the indicated ways. However, the pattern of stimulation need not be describable in the language he speaks; and even if it is, there is every reason to suppose that the agent is quite unaware of what that pattern is, and is quite incapable of uttering or identifying a correct description of it. Thus, a complete description of the pattern of stimulation includes a record of the firing times of all the rods and cones in the outer layer of retinal neurons during the period of the observation. Even if the agent is an expert physiologist, he will be unable to produce or recognize a correct record of this sort on the basis of his experience during the observation.

With this story in mind, Jeffrey wonders how the new information should be used to influence other propositions that depend on the color of the cloth:

> Then the problem is this: Given that a passage of experience has led the agent to change his degrees of belief in certain propositions $B_1, B_2, ..., B_n$ from their original values,
>
> $$prob\ B_1, prob\ B_2, ..., prob\ B_n$$
>
> to new values,
>
> $$PROB\ B_1, PROB\ B_2, ..., PROB\ B_n,$$
>
> how should these changes be propagated over the rest of the structure of his beliefs? If the original probability measure was *prob*, and the new one is *PROB*, and if A is a proposition in the agent's preference ranking [i.e., knowledge base] but is not one of the n propositions whose probabilities were directly affected by the passage of experience, how shall $PROB\ A$ be determined?

Jeffrey's solution is based on the critical assumption that the propositions B selected to summarize the experience possess a special property: "...while the observation changed the agent's degree of belief in B and in certain other propositions, it did not change the *conditional degree of belief* in any propositions on the evidence B or on the evidence $\neg B$" (italics added). Thus, if $B_1, B_2, ..., B_n$ are exhaustive and mutually exclusive propositions (like *Green*, *Blue*, and *Violet* in the candlelight example), Jeffrey maintains that, for every proposition A not "directly affected by the passage of experience," we should write

$$PROB\ (A \mid B_i) = prob\ (A \mid B_i) \quad i = 1, 2, ..., n. \qquad \textbf{(2.59)}$$

This, together with the additivity of *PROB*, leads directly to

$$PROB(A) = \sum_i prob\ (A \mid B_i)\ PROB(B_i), \qquad \textbf{(2.60)}$$

a formula now known as *Jeffrey's Rule* of updating, or the rule of *probability kinematics*.

The convenience of the rule is enticing in a way that is reminiscent of the logical rules of deduction; we need not know anything about how $prob(B_i)$ was updated to $PROB(B_i)$—only the net result matters. We simply take $PROB(B_i)$ as a new set of priors and apply the textbook formula of Eq. (2.10). Unfortunately, the rule is applicable only in situations where the criterion of Eq. (2.59) holds, and this condition, as we shall soon see, is not easy to test.

Traditional probabilistic analysis gives us a way to decide when Eqs. (2.59) and (2.60) are applicable, based on Bayes' conditioning. If we denote by e the evidence actually observed and equate $PROB(A)$ with $prob(A \mid e)$, we get the *Bayes conditionalization formula,*

$$prob(A \mid e) = \sum_i prob(A \mid B_i, e)\, prob(B_i \mid e), \qquad \textbf{(2.61)}$$

which coincides with Eq. (2.60) only when A and e are conditionally independent given B_i, i.e., only when

$$prob(A \mid B_i, e) = prob(A \mid B_i). \qquad \textbf{(2.62)}$$

However, philosophers might argue that it sometimes makes no sense to equate $PROB(A)$ with $prob(A \mid e)$ or even to talk about $prob(A \mid e)$, e being an elusive, non-propositional experience. Indeed, the textbook definition of conditional probability, $P(A \mid e) = P(A, e) / P(e)$, suggests that before $P(A \mid e)$ can be computed one must have the joint probability $P(A, e)$, so e must already be integrated in one's knowledge base as a proposition that might later be an object of attention. This condition clearly is not met in the candlelight story; the sensory experience responsible for the color judgment cannot have been anticipated in anyone's knowledge base. In such cases, so the argument goes, Bayes conditionalization is not applicable and should give way to the more general Jeffrey's Rule. Likewise, the conditional independence criterion of Eq. (2.62) is a quality ascertainable only by Bayes conditionalization and therefore is clearly inadequate for delineating the class of propositions A to which Jeffrey's Rule applies.

While no alternate criterion for testing Eq. (2.59) is formulated in Jeffrey's book, some hint is provided by the requirement that A "is not one of the n propositions whose probabilities were directly affected by the passage of experience." Jeffrey apparently believed that the question of whether a proposition A is affected directly or indirectly can be decided on qualitative grounds, prior to defining joint distributions. In this sense, he pioneered the idea that dependence relationships are the fundamental building blocks of probabilistic knowledge, more basic than numerical distributions (a position that will be developed further in Chapter 3).

In a subsequent publication [Jeffrey 1968], Jeffrey replaced the notion of directness with that of a *basis,* where a basis B for an observation is defined as the set of propositions $B_1, B_2, \ldots, B_n$ that satisfy Eq. (2.59) for every A not in B. This way, the validity of Eq. (2.60) is automatically guaranteed to hold for every A not in B, but from a practical viewpoint the problem of determining the basis associated with a given observation remains unresolved.

To demonstrate the type of information required for determining the applicability of Jeffrey's Rule, let us return to the candlelight example and assign two alternative meanings to proposition A.

Case 1 e — B — A: Assume that the proposition A stands for the statement "The cloth will be sold the next day," and we know the chances of selling the cloth depend solely on its color:

$$P(A \mid Green) = 0.40, \quad P(A \mid Blue) = 0.40, \quad \text{and}$$
$$P(A \mid Violet) = 0.80. \qquad \textbf{(2.63)}$$

Eq. (2.60), then, allows us to calculate the updated belief in the salability of the cloth, based only on the color inspection (see Figure 2.3). Prior to the test, our belief in selling the cloth measured

$$prob(A) = (0.4)(0.3) + (0.4)(0.3) + (0.8)(0.4) = 0.56\,,$$

and once the test results become known, our belief should change to

$$PROB(A) = (0.4)(0.7) + (0.4)(0.25) + (0.8)(0.05) = 0.42\,.$$

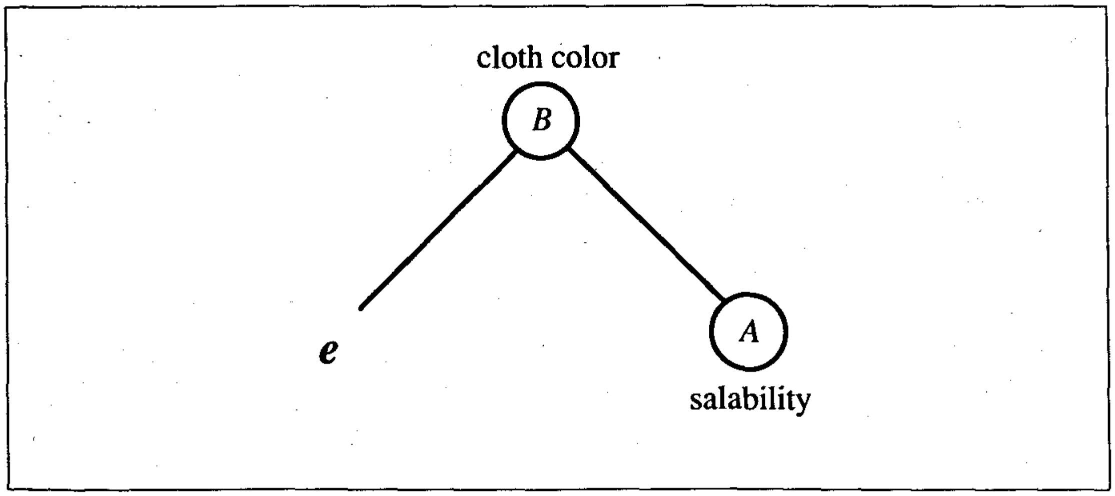

Figure 2.3. *A network representing the conditional independence of A and e, given B.*

Bayes conditionalization would yield the same result, because the salability of the cloth depending only on its color is interpreted as A and e being conditionally independent, and therefore

$$P(A \mid Color,\, e) = P(A \mid Color), \qquad \textbf{(2.64)}$$

which legitimizes Jeffrey's assumption that

$$PROB(A \mid B_i) = prob(A \mid B_i),$$

as long as we identify $PROB(A \mid B_i)$ with $P(A \mid B_i, e)$. In other words, modern Bayesians take the liberty of writing equations such as Eq. (2.64) even though $P(A \mid Color, e)$ is available nowhere and cannot be computed numerically. The equation does convey the qualitative information expressed in the story—that color is the only factor relevant to salability—and it thus draws legitimacy not from numerical probability values but from a more reliable knowledge source: people's qualitative reasoning about dependencies.

Note that Jeffrey's Rule is equivalent to the Bayesian treatment of virtual evidence (Section 2.2.2), using the likelihood vector

$$\Lambda_i(B) \triangleq P(e \mid B_i) = \alpha \frac{PROB(B_i)}{prob(B_i)} = \alpha \left(\frac{0.70}{0.30}, \frac{0.25}{0.30}, \frac{0.05}{0.40}\right)$$

$$= \alpha\,(2.330, 0.833, 0.125). \qquad \textbf{(2.65)}$$

Indeed, in Section 2.2.2 we saw that the likelihood vector requires no absolute probability assessments and therefore avoids the difficulties associated with non-propositional evidence (e.g., the visual stimulus in the candlelight story). We also argued that the assumption of conditional independence means that the likelihood vector is the only stable component in the relation between the evidence and the impacted variable B, making it more reliable to assess than the final product $PROB(B_i)$. Thus, an alternate way of viewing the impact of sensory experience on one's knowledge is to replace the former by a likelihood vector impinging on the basis B. (A similar idea was advanced by Field [1978].)

To demonstrate the volatility of the assumption in Eq. (2.59), let us choose an example where it is obviously violated.

Case 2 A — e — B: Imagine that the main interest of our candlelight observer lies not in the color of the cloth but rather in the chemical composition of the candle wax. The agent inspects the color of the cloth, adjusts his belief from $prob(B_i)$ to $PROB(B_i)$, and then wonders how to update $prob(A)$, where A is the proposition that the wax is a notoriously cheap brand known to produce flames deficient in violet content.

Are we justified in using Jeffrey's Rule? Since the color of the cloth (B_i) is of no relevance to A prior to the observation, we have $prob(A \mid B_i) = prob\ A$. If we blindly apply Eq. (2.60), we obtain a paradoxical result,

$$PROB\ (A) = \sum_i prob(A)\, PROB\ (B_i) = prob\ (A)\,, \qquad \textbf{(2.66)}$$

which states that no matter how violet or greenish the cloth looks under the candlelight, the observer's belief regarding the makeup of the wax ought to remain unaltered.

Is there any information in the story that should warn us against applying Jeffrey's Rule here? Modern Bayesians claim that even though we lack the knowledge required for precise description of the measurement process, our qualitative understanding of the process is sufficient to alert us to the falsity of $P(A \mid B_i, e) = P(A \mid B_i)$ and thus protect us from drawing a false conclusion like Eq. (2.66). Colloquially, we say that in Case 1, the color of the cloth "stood between" the evidence and A (the salability of the cloth), while in Case 2 it was the evidence that mediated between the colors and A (the brand of wax), as shown in Figure 2.4.

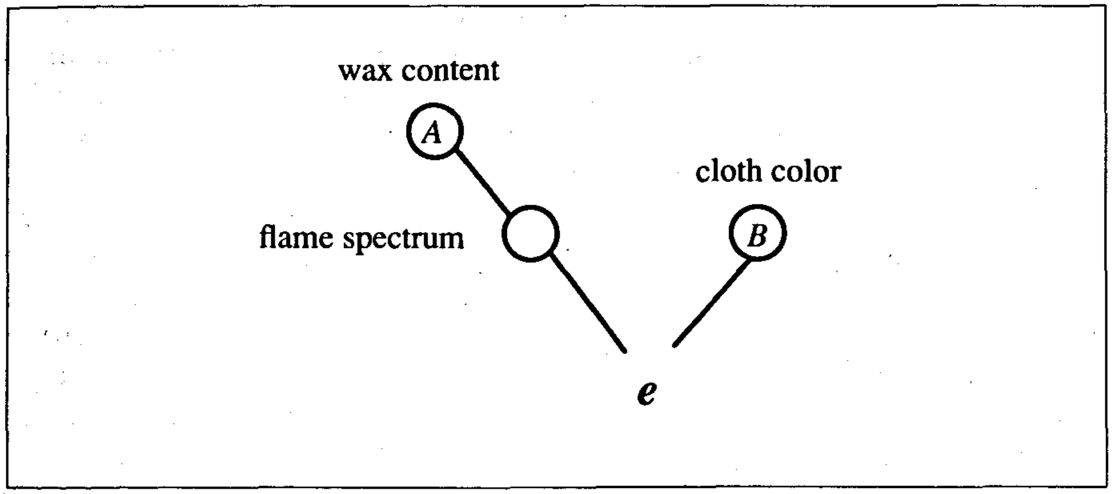

Figure 2.4. *A network representing an evidence (e) mediating between A and B.*

One might argue that Jeffrey's original account also prevents us from applying his rule to Case 2 because A presumably should qualify as "one of the n propositions whose probabilities were directly affected by the passage of experience." But the criterion by which this passage of experience can be termed "direct" is rather hard to define. In other words, it is hard to see how the visual experience bears directly on the nature of the wax (A) when it is the flame that mediates between the two (see Figure 2.4). If anything, B seems more directly affected by e than A is; the agent's judgment about the color was reported first, and color bears a closer semantic relation to visual experience than wax chemistry does.

If the road map outlining one's passage of experience is so crucial for understanding the structure of stories (i.e., which propositions should be affected by the evidence and how), it is unfortunate that the philosophical literature on probability kinematics does not provide a more complete analysis of this crucial source of information. Evidently, some believed that this road map is so deeply entrenched in human intuition that no further explication is required.

Neo-Bayesian philosophers go one step beyond Jeffrey. They say any assertions one wishes to make about "passage of experience" ought to be explicated formally, using the familiar syntax of probability calculus. For example, one's intuition that A is not directly affected by the passage of experience ought to be written in the format of Eq. (2.62), treating e as a genuine propositional entity. On the surface, this requirement seems vacuous. If one interprets Eq. (2.62) merely as a notation for expressing intuitions about the "passage of experience," then Bayes conditionalization—$P(A \mid e)$—ceases to be a statement about the numeric magnitudes of $P(A)$ and $P(A \mid e)$ and becomes no more informative than the verbal, intuitive sentences it purports to replace. However, there is a profound significance to the use of the $P(* \mid *)$ syntax instead of some other notation.

First, it embodies the claim that passages of experience have traffic laws of their own and that these laws are similar, if not identical, to those governing Bayes conditionalization. For example, one traffic law states that it is inconsistent for an agent to assert, "B stands between e and a pair of propositions $\{A_1, A_2\}$" without also asserting, "B and A_1 together stand between e and A_2." This consistency requirement holds both in Bayes conditionalization and in the road map metaphor. Thus, even if one insists that statements such as Eq. (2.62) represent qualitative facts about the passage of experience, not conditional probabilities, by agreeing to manipulate these sentences by the rules of Bayes conditionalization one is guaranteed never to violate any of the traffic laws•that govern the roadmaps of experience. The question of whether graphical representation of dependencies can yield similar guarantees is treated in Chapter 3.

Second, the use of the $P(* \mid *)$ syntax to define criteria such as Eq. (2.62) suggests procedures a person should use to test mentally the validity of the criterion in any given situation. Eq. (2.62) instructs a person to imagine first that the cloth has a definite color, say $B_i = Green$, then test whether any visual experience e could significantly sway the belief in A one way or the other. In Case 1 the answer is clearly no, because the salability was proclaimed to be a function only of the cloth color. In Case 2, however, this mental exercise would evoke some vivid scenarios that could sway our belief. For example, a green cloth that appears totally violet under the candlelight would induce a different opinion about the candle's wax than a green cloth that appears totally yellow under candlelight. Thus, Bayes conditionalization has syntactic and psychological merits beyond the numerical definition

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

that appears in most textbooks on probability theory.

Case 2 carries two messages. First, we demonstrated again that even when we cannot describe precisely the observed evidence e, the qualitative elements of the story are sufficient for judging whether the situation meets Jeffrey's criterion, or the conditional independence requirement $P(A \mid B_i, e) = P(A \mid B_i)$. Second, we demonstrated that Jeffrey's Rule is invalid not only when A is directly affected by the passage of experience; it is enough that A branches off someplace on the path from e to B, as in Figure 2.4. A more striking example is provided by the diamond structure of Figure 2.5. Here, B is clearly more directly affected by e than A is, as B stands between e and A, yet Eq. (2.62) will be violated.

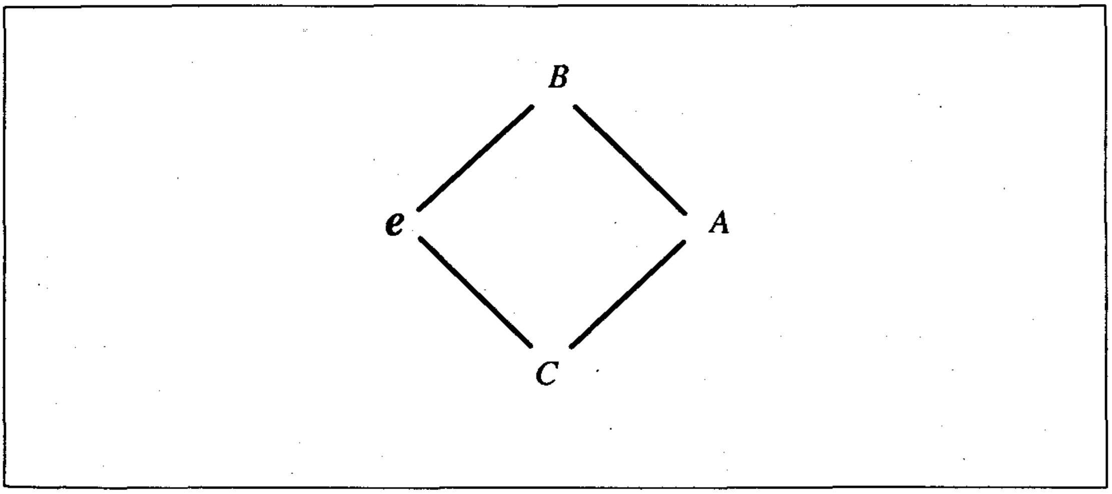

Figure 2.5. *A is not affected directly by the passage of experience, yet the observation e changes the conditional degree of belief in A given B.*

So far, we have used the diagrams in Figures 2.3 through 2.5 primarily as mnemonic devices to distinguish among the cases discussed and to make an occasional association with Jeffrey's "passage of experience" notion. However, the preceding discussion also demonstrates a rather useful pattern produced by graphical representations (Figures 2.3 through 2.5): Jeffrey's Rule is applicable if and only if *B separates A* from e. This may be what Jeffrey meant by requiring that A not be "one of the n propositions whose probabilities were directly affected by the passage of experience." The notion of *separation* and its relation to information independence will be given formal treatment in Chapter 3.

SUMMARY

Jeffrey's Rule of belief updating was devised to replace Bayes conditioning in cases where the evidence cannot be articulated propositionally. Our analysis shows that to determine whether the rule is valid in any specific case, one must have topological knowledge about one's belief structure, namely, which beliefs are directly related and which are only indirectly related. If such knowledge is available, it can be faithfully represented by the syntax of conditional independence sentences, and traditional Bayes' methods can be used to update beliefs. Thus, the question arises whether it is *ever* necessary to avoid conditionalization in formal belief updating.

Since simple criteria based on graphical considerations lead to conclusions that match our intuition, perhaps human intuition itself can be represented by networks of relations, and perhaps intuitive judgments are really mental tracings of those networks. These suggestions motivate the discussion of dependency graphs in Chapter 3.

2.4 BIBLIOGRAPHICAL AND HISTORICAL REMARKS

The Italian mathematician Gerolamo Cardano (1501-1576) is believed to be the first to have formulated the notion of probability in gambling in terms of the number of distinguishable ways that events may occur. This development marks a radical (if somewhat tardy) change in cultural attitudes toward uncertainty. Although fascination with the unpredictability of gambling devices goes back to the time of the Pharaohs [David 1962], these devices were not perceived as possessing inherent elements of uncertainty; instead, they were seen as means of communicating with a source of knowledge (e.g., deity) that was basically deterministic [Hacking 1975].

Cardano's "objective" view of probability developed into a rather sophisticated mathematical theory of combinatorics, in the hands of Fermat (1601-1665), Pascal (1623-1662), Huygens (1629-1695), James Bernoulli (1654-1705), DeMoivre (1667- 1754), and LaPlace (1749-1827), until in 1837 Denis Poisson gave it a new twist by defining probability as a limit of a long-run relative frequency. Emile Borel (1871-1956) and A. N. Kolmogorov are credited with developing the modern axiomatic foundations of mathematical probability, of which Eqs. (2.1) through (2.3) are a simplified version [Kolmogorov 1950]. Kolmogorov's axiomatization of probability is responsible for the unfortunate tradition of treating Eq. (2.8) as a definition of conditional probability, rather than a theorem that follows from more primitive axioms about conditioning.

In parallel to these mathematical developments, an alternative view of probability came into being with Bernoulli's suggestion that probability is a "degree of confidence" that an individual attaches to an uncertain event. This concept, aided by Bayes' Rule [Bayes 1763], blossomed in the writings of LaPlace and De Morgan and later in the works of Keynes [1921] and Jeffreys [1939]. However, the established communities of statisticians and mathematical probabilists viewed this "subjectivist" intrusion with suspicion. It was not until the 1950s, with the development of statistical decision theory (see Section 6.5), that

Bayesian methods gained their current momentum. The two defining attributes of the Bayesian school are (1) willingness to accept subjective opinions as an expedient substitute for raw data and (2) adherence to Bayes conditionalization as the primary mechanism for updating beliefs in light of new information. The articles in Kyburg and Smokler [1980] deal with the philosophical underpinning of the Bayesian revival.

A critical analysis of Bayes conditionalization can be found in Shafer [1982, 1985, 1986b]: According to Shafer, it was DeMoivre who first formulated the idea that the occurrence of one event can change the probability of another and who proved the multiplication rule of Eq. (2.9) using the method of expectation. Bayes gave a version of DeMoivre's proof for his rule (Eq. (2.13)), while interpreting it as providing the subjective probabilities of past events. Exercise 2.2 gives a modern version of the example used in Bayes' original essay [Bayes 1763]. Alternatives to Bayes conditioning—including Jeffrey's rule and Dempster's rule (see Chapter 9)—have been discussed by Diaconis and Zabell [1986]. Jeffrey's rule constitutes the minimum entropy extension of *prob* (·), and Lemmer and Barth [1982] first proposed it for belief updating in expert systems. The formal identity between Jeffrey's rule and virtual conditionalization (as in Eq. (2.65)) renders the two semantically equivalent, i.e., beliefs updated by Jeffrey's rule cannot be distinguished from those updated by Bayes' conditionalization on some virtual evidence. Another alternative to Bayes' conditionalization, called *imaging,* was introduced by Lewis [1976] and was used to represent counterfactual conditionals.

The Three Prisoners story is one of many well-known puzzles that illustrate the need for specifying the query process in tasks involving inference from observations (see Exercise 2.6). Shafer [1985] calls this query process a *protocol* and views it as a disadvantage of Bayes conditioning, since we must assign probabilities for all possible ways information may be obtained. Our discussion in Section 2.3.2 attempts to convince the reader that formalisms that ignore the query process altogether (see Chapter 9 for examples) are bound to be insensitive to an important component of human reasoning. In the Thousand Prisoners story, for example, such systems will not attach any significance to discovering the query after seeing the list; beliefs will remain the same, based solely on the one-in-two model (see Exercise 9.5b).

Our treatment of virtual evidence (Section 2.2.2), using the vector of likelihood-ratios, sidesteps the requirement of specifying a full protocol in advance (see Exercise 2.7). This option expands the repertoire of Bayes analysis by permitting us to assimilate evidence by means other than straight conditioning, and it simultaneously facilitates the manipulation of belief updates within the traditional syntax of probability calculus.

There are, of course, items of information that cannot and should not be handled as evidential data, but must be treated as constraints on—or specificational adjustments to—the probabilistic model we currently possess. Conditional sentences are typical examples of such information. For example, the sentence "If Joe goes to the party Mary will not go" must be treated as a meta-level constraint in the form of conditional probability and not as evidence to be conditioned upon (see Exercises 10.1 and 10.2). On the other hand, the sentence "Joe and Mary will not both go to the party," though logically equivalent to the previous sentence, is a form of information that can be treated as evidence for conditionalization. The difference is that conditionalization changes the probability of Joe's going to the party while constraint-based updating leaves this probability intact. The purpose of the English word *if* is to convey a distinction between these two modes of assimilating information and to instruct the listener to refrain from straight conditioning.

The papers in Harper et al. [1981] provide a cross section of the philosophical literature dealing with conditionals. Section 10.2 illustrates how conditional information can be absorbed in the form of specification constraints, following the work of Adams [1975].

The treatment of Jeffrey's rule (Section 2.3) is further expanded in Pearl [1990](1).

Recent works on foundational issues of probability have focused on higher-order probabilities(2)(3) (see also Section 7.3) and on the development of logics for reasoning about probabilities.(4)(5)(6)

(1) Pearl, J. Jeffrey's Rule, Passage of Experience, and *Neo*-Bayesianism. H.E. Kyburg, Jr. et al., (eds.), *Knowledge Representation and Defeasible Reasoning*, 1990, Kluwer Academic Publishers, 245-264.

(2) Haddawy, P., and Frisch, A.M. Modal logics of higher-order probability. In Shachter et al., (eds.), *Uncertainty in AI* 4, North Holland, 1990, 133-148.

(3) Fagin, R., and Halpern, J. Y., Reasoning about Knowledge and Probability: Preliminary Report, in *Proceedings, 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, Morgan-Kaufmann, 1988, 277-293.

(4) Fagin, R., Halpern, J.Y., and Megiddo, N., A logic for reasoning about probabilities, *Information and Computation*, 87 (1/2), 1990, 78-128.

(5) Bacchus, F., *Representing and reasoning with probabilistic knowledge*, Cambridge, MA: The MIT Press, 1990.

(6) Kyburg, H., Evidential Probability, *Proceedings IJCAI-91*, Sydney, Australia, 1991, 1196-1203.

Exercises

2.1. There are three urns labeled one, two, and three. These urns contain, respectively, three white and three black balls, four white and two black balls and one white and two black balls. An experiment consists of selecting an urn at random, then drawing a ball from it.

a. Find the probability of selecting urn 2 and drawing a black ball.

b. Find the probability of drawing a black ball.

c. Find the conditional probability that urn 2 was selected, given that a black ball was drawn.

It may be helpful to label the possible outcomes $(1, B)$, $(1, W)$, $(2, B)$, $(2, W)$, $(3, B)$, $(3, W)$.

2.2. A billiard table has unit length, measured from left to right. A ball is rolled on this table, and when it stops, a partition is placed at its stopping position, a distance x from the left end of the table. A second ball is now rolled between the left end of the table and the partition, and its stopping position, y, is measured.

a. Answer qualitatively: How does knowledge of y affect our belief about x? Is x more likely to be near y, far from y, or near the midpoint between y and 1?

b. Justify your answer for (**a**) by quantitative analysis. Assume each stopping position is uniformly distributed over its feasible range.

2.3. Let the hypothesis variable $H = \{H_1, H_2, H_3, H_4\}$ stand for the following set of exhaustive and mutually exclusive conditions

H_1 = *No burglary, animal entry.*
H_2 = *Attempted burglary, window break-in.*
H_3 = *Attempted burglary, door break-in.*
H_4 = *No burglary, no entry.*

with prior probabilities $P(H_i) = (0.099, 0.009, 0.001, 0.891)$. Let the alarm system contain two detectors, E^1 and E^2, with the following sensitivity matrices:

	e_1^1	e_2^1	e_3^1
H_1	0.5	0.4	0.1
H_2	0.06	0.5	0.44
H_3	0.5	0.1	0.4
H_4	1.0	0	0

	e_1^2	e_2^2	e_3^2
H_1	0.8	0.1	0.1
H_2	0.8	0.1	0.1
H_3	0.1	0.1	0.8
H_4	0.9	0.05	0.05 .

a. What is the probability of burglary if detector E^1 is *OFF* ($E^1 = e_1^1$) and E^2 is *HIGH* ($E^2 = e_3^2$)?

b. Repeat problem (**a**) under the following conditions:

- A reliable witness claims to have heard detector E^1, but she cannot tell whether it was *High sound* (e_3^1) or *Low sound* (e_2^1).
- A second reliable witness claims detector E^2 was definitely not in *High sound* state but there is a slight (5%) chance that it issued a *Low sound* (e_2^2).

c. You are considering adding to your alarm system a new detector E^3, with the following sensitivity matrix:

	OFF	*ON*
H_1	0.1	0.9
H_2	0.9	0.1
H_3	0.9	0.1
H_4	1	0

What is the probability that E^3 will be activated under the conditions described in problem (**b**)?

d. You are considering installing a monitor E^4 at your office, connected directly to detector E^1. The relation between E^1 and E^4 is characterized by the matrix

	$E^4 = OFF$	$E^4 = ON$
$E^1 = OFF$	0.9	0.1
$E^1 = LOW$	0.2	0.8
$E^1 = HIGH$	0.1	0.9 .

What is the probability that E^4 will turn on under the conditions of problem (**b**).

2.4. **a.** Verify which entries in Table 1 (page 54) are unconditionally supported by probability theory and which must be qualified with additional assumptions about context.

b. Which of the entries are violated in the Three Prisoners story.

2.5. How would Jeffrey's rule handle the Three Prisoners problem?

2.6. I have three cups and one ball. I put the ball under one of the cups and mix up the cups. You must pick the cup with the ball under it. You choose one without inspecting its content. Then I remove one of the other cups and show you that it does not have a ball under it. Now I give you the chance to change your choice of cups. Should you do it? How is this puzzle related to the Three Prisoners story?

2.7. **a.** Formulate Case 2 of the candlelight story using a Bayesian approach, and determine what additional information is required for computing $P(A \mid e)$. (Recall: e is non-propositional, so the absolute value of $P(e \mid \cdot)$ is meaningless).

b. Assume reasonable values for the missing information and compute $P(A \mid e)$.