5 STOCHASTIC METHODS

God does not play dice. . .

—ALBERT EINSTEIN (His answer to the credibility of Quantum Theory)

God not only plays dice, but he sometimes throws them where they can't be seen. . .

—STEPHEN HAWKING

Probable impossibilities are to be preferred to improbable possibilities. . .

—ARISTOTLE As quoted by Bobby Wolfe in "*Aces on Bridge*", 2003

5.0 Introduction

Chapter 4 introduced heuristic search as an approach to problem solving in domains where either a problem does not have an exact solution or where the full state space may be too costly to calculate. In this chapter we propose the stochastic methodology as appropriate for these situations. Probabilistic reasoning is also suitable for situations where state information is found from sampling an information base and causal models are learned from data.

One important application domain for the use of the stochastic methodology is diagnostic reasoning where cause/effect relationships are not always captured in a purely deterministic fashion, as is often possible in the knowledge-based approaches to problem solving that we saw in Chapters 2, 3, 4, and will see again in Chapter 8. A diagnostic situation usually presents evidence, such as fever or headache, without further causative justification. In fact, the evidence can often be indicative of several different causes, e.g., fever can be caused by either flu or an infection. In these situations probabilistic information can often indicate and prioritize possible explanations for the evidence.

Another interesting application for the stochastic methodology is gambling, where supposedly random events such as the roll of dice, the dealing of shuffled cards, or the spin of a roulette wheel produce possible player payoff. In fact, in the 18th century, the

attempt to provide a mathematical foundation for gambling was an important motivation for Pascal (and later Laplace) to develop a probabilistic calculus.

Finally, as observed in Section 1.1.4, a "situated" accounting of intelligence suggests that human decisions often emerge from complex, time-critical, and embodied environments where a fully mechanistic calculus may simply not be definable, or, if defined, may not compute answers in a usable time frame. In these situations intelligent actions may best be seen as stochastic responses to anticipated costs and benefits.

We next describe several problem areas, among many, where the stochastic methodology is often used in the computational implementation of intelligence; these areas will be major topics in later chapters.

1. **Diagnostic reasoning**. In medical diagnosis, for example, there is not always an obvious cause/effect relationship between the set of symptoms presented by the patient and the causes of these symptoms. In fact, the same sets of symptoms often suggest multiple possible causes. Probabilistic models are also important in complex mechanical situations, such as monitoring aircraft or helicopter flight systems. Rule-based (Chapter 8) and probabilistic (Chapters 9 and 13) systems have both been applied in these and other diagnostic domains.

2. **Natural language understanding**. If a computer is to understand and use a human language, that computer must be able to characterize how humans themselves use that language. Words, expressions, and metaphors are learned, but also change and evolve as they are used over time. The stochastic methodology supports the understanding of language; for example, when a computational system is trained on a database of specific language use (called a *corpus linguistics*). We consider these language issues further in this chapter as well as in Chapter 15.

3. **Planning and scheduling**. When an agent forms a plan, for example, a vacation trip by automobile, it is often the case that no deterministic sequence of operations is guaranteed to succeed. What happens if the car breaks down, if the car ferry is cancelled on a specific day, if a hotel is fully booked, even though a reservation was made? Specific plans, whether for humans or robots, are often expressed in probabilistic language. Planning is considered in Section 8.4.

4. **Learning**. The three previous areas mentioned can also be seen as domains for automated learning. An important component of many stochastic systems is that they have the ability to sample situations and learn over time. Some sophisticated systems are able to both sample data and predict outcomes as well as learn new probabilistic relationships based on data and outcomes. We consider learning in Part IV, and stochastic approaches to learning in Chapter 13.

The stochastic methodology has its foundation in the properties of counting. The probability of an event in a situation is described as the ratio of the number of ways the event can occur to the total number of possible outcomes of that event. Thus, the probability that an even number results from the roll of a fair die is the total number of even outcomes (here 2, 4, or 6) over the total number of outcomes (1, 2, 3, 4, 5, or 6), or 1/2. Again,

the probability of drawing a marble of a certain color from a bag of marbles is the ratio of the number of marbles of that color to the total number of marbles in the bag. In Section 5.1 we introduce basic counting techniques, including the sum and product rules. Because of their importance in counting, we also present the *permutations* and *combinations* of discrete events. This is an optional section that can be skipped by readers with a sufficient background in discrete mathematics.

In Section 5.2, we introduce a formal language for reasoning with the stochastic methodology. This includes the definitions of independence and different types of random variables. For example, for probabilistic propositions, random variables may be *boolean* (true or false), *discrete*, the integers 1 to 6 as in rolling a fair die, or *continuous*, a function defined on the real numbers.

In Section 5.3, we present Bayes' theorem which supports most approaches to stochastic modeling and learning (Section 9.3 and Chapter 13). Bayes' rule is important for interpreting new evidence in the context of the prior knowledge or experience. In Section 5.4 we present two applications of stochastic methods, including probabilistic finite state automata and a methodology for predicting English word patterns based on sampled data.

In Chapters 9 and 13 we continue presentation of probabilistic approaches to reasoning, including *Bayesian belief networks (BBNs)*, an introduction to *Markov models (HMMs)*, and first-order representation systems supporting stochastic modeling. These models use the directed acyclic graph (or DAG) formalism, and are referred to as *graphical models*. In Chaper 13 we present stochastic approaches to machine learning.

5.1 The Elements of Counting (optional)

The foundation for the stochastic methodology is the ability to count the elements of an application domain. The basis for collecting and counting elements is, of course, set theory, in which we must be able to unequivocally determine whether an element is or is not a member of a set of elements. Once this is determined, there are methodologies for counting elements of sets, of the complement of a set, and the union and intersection of multiple sets. We review these techniques in this section.

5.1.1 The Addition and Multiplication Rules

If we have a set A, the *number of the elements* in set A is denoted by |A|, called the *cardinality* of A. Of course, A may be empty (the number of elements is zero), finite, countably infinite, or uncountably infinite. Each set is defined in terms of a domain of interest or *universe*, U, of elements that might be in that set. For example, the set of male people in a classroom may be defined in the context, or or in the universe, of all the people in that room. Similarly, the roll of a 3 on a fair die may be seen as one of a set of six possible outcomes.

The domain or universe of a set A is thus also a set and is used to determine the *complement* of that set, $\overline{A}$. For example, the complement of the set of all males in the classroom just mentioned is the set of all females, and the complement of the {3} roll of the fair

die is {1, 2, 4, 5, 6}. One set A is a *subset* of another set B, $A \subseteq B$, if every element of the set A is also an element of the set B. Thus, trivially, every set is a subset of itself, any set A is a subset of its universe, and the empty set, denoted { } or ϕ, is a subset of every set.

The *union* of two sets A and B, $A \cup B$, may be described as the set of all elements in either set. The number of elements in the union of two sets is the total of all the elements in each of the sets minus the number of elements that are in both sets. The justification for this, of course, is the fact that each distinct element in a set may only be counted once. Trivially, if the two sets have no elements in common, the number of the elements in their union is the sum of the number of elements in each set.

The *intersection* of two sets A and B, $A \cap B$, is the set of all elements common to both sets. We now give examples of a number of the concepts just defined.

Suppose the universe, U, is the set {0, 1, 2, 3, 4, 5, 6, 7, 8, 9}
Let A be the set {1, 3, 5, 7, 9}
Let B be the set {0, 2, 4, 6, 8}
Let C be the set {4, 5, 6}
Then |A| is 5, |B| is 5, |C| is 3, and |U| is 10.
Also, $A \subseteq U$, $B \subseteq U$ and $A \cup B = U$, $|B| = |A|$, $A = \overline{B}$
Further, $|A \cup B| = |A| + |B| = 10$, since $A \cap B = \{\}$, but
$|A \cup C| = |A| + |C| - |A \cap C| = 7$, since $A \cap C = \{5\}$

We have just presented the major components for the *addition rule* for combining two sets. For any two sets A and C, the number of elements in the union of these sets is:

$$|A \cup C| = |A| + |C| - |A \cap C|$$

Note that this addition rule holds whether the two sets are disjoint or have elements in common. A similar addition rule holds for three sets A, B, and C, again whether or not they have elements in common:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

An argument similar to that made earlier can be used to justify this equation. Similar inclusion/exclusion equations are available, and easily demonstrated, for the addition of sets of elements of more than three sets.

The *multiplication principle for counting* states that if we have two sets of elements A and B of size a and b respectively, then there are a x b unique ways of combining the elements of the sets together. The justification for this, of course, is that for each of the a elements of A there are b pairings for that element. The multiplication principle supports many of the techniques used in counting, including the Cartesian product of sets, as well as permutations and combinations of sets.

The *Cartesian product* of two sets A and B, denoted A X B, is the set of all ordered pairs (a, b) where a is an element of set A and b is an element of set B; or more formally:

$$A \times B = \{(a, b) \mid (a \in A) \wedge (b \in B)\}$$

and by the multiplication principle of counting:

$$|A \times B| = |A| \times |B|$$

The Cartesian product can, of course, be defined across any number of sets. The product for n sets will be the set of n-tuples where the first component of the n-tuple is any element of the first set, the second component of the n-tuple is any element of the second set, and so on. Again, the number of unique n-tuples that result is the product of the number of elements in each set.

5.1.2 Permutations and Combinations

A *permutation* of a set of elements is an arranged sequence of the elements of that set. In this arrangement of elements, each may be used only once. An example permutation is an arrangement or ordering of a set of ten books on a shelf that can hold all ten. Another example is the assignment of specific tasks to four of a group of six children.

We often wish to know how many (unique) permutations there are of a set of n elements. We use multiplication to determine this. If there are n elements in set A, then the set of permutations of these elements is a sequence of length n, where the first element of the sequence is any of the n elements of A, the second element of the sequence is any of the $(n - 1)$ remaining elements of A, the third element of the sequence is any of the $(n - 2)$ remaining elements, and so on.

The order in which the elements are placed in the permutation sequence is unimportant, i.e., any of the n elements of the set may be placed first in any location in the sequence, any of the $n - 1$ remaining elements may be placed second in any of the $n - 1$ remaining locations of the permutation sequence, and so on. Finally, by the multiplication principle, there are $n!$ permutation sequences for this set of n elements.

We can restrict the number of elements in a permutation of a set A to be any number greater than or equal to zero, and less than or equal to the number of elements n in the original set A. For example, we might want to know how many distinct orderings there are of ten possible books on a shelf that can only hold six of them at a time. If we wanted to determine the number of permutations of the n elements of A taken r at a time, where $0 \leq r \leq n$, we use multiplication as before, except that now we only have r places in each permutation sequence:

$$n \times (n - 1) \times (n - 2) \times (n - 3) \times \ldots \times (n - (r - 1))$$

Alternatively, we can represent this equation as:

$$\frac{n \times (n - 1) \times (n - 2) \times (n - 3) \times \ldots \times (n - (r - 1)) \times (n - r) \times (n - r - 1) \times \ldots \times 2 \times 1}{(n - r) \times (n - r - 1) \times \ldots \times 2 \times 1}$$

or equivalently, the number of permutations of n elements taken r at a time, which is symbolized as ${}_nP_r$, is:

$$ {}_nP_r = \frac{n!}{(n - r)!} $$

The *combination* of a set of n elements is any *subset* of these elements that can be formed. As with permutations, we often want to count the number of combinations of items that can be formed, given a set of items. Thus, there is only one combination of the n elements of a set of n items. The key idea here is that the number of combinations represents the number of *subsets of the full set of elements* that can be created. In the bookshelf example, combinations represent the different subsets of six books, the books on the shelf, that can be formed from the full set of ten books. Another example of combinations is the task of forming four-member committees from a group of fifteen people. Each person is either on the committee or not, and it doesn't make any difference whether they are the first or last member chosen. A further example is a five-card hand in a poker game. The order in which the cards are dealt makes no difference to the ultimate value of the hand. (It can make a huge difference to the value of the betting, if the last four cards are dealt face-up as in traditional stud poker, but the ultimate value of the hand is independent of the dealt order).

The number of combinations of n elements taken r at a time, where $0 \leq r \leq n$, is symbolized by ${}_nC_r$. A straightforward method for determining the number of these combinations is to take the number of permutations, ${}_nP_r$, as we did already, and then divide out the number of duplicate sets. Since any r element subset of n elements has r! permutations, to get the number of combinations of n elements taken r at a time we divide the number of permutations of n elements taken r at a time by r!. Thus we have:

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{(n - r)!\, r!} $$

There are many other variations of the counting principles just presented, some of which will be found in the Chapter 5 exercises. We recommend any discrete mathematics textbook for further development of these counting techniques.

5.2 Elements of Probability Theory

With the foundation in the counting rules presented in Section 5.1, we can now introduce probability theory. First, in Section 5.2.1, we consider some fundamental definitions, such as the notion of whether two or more events are independent of each other. In Section 5.2.2 we demonstrate how to infer explanations for particular data sets. This will set us up to consider several examples of probabilistic inference in Section 5.3 and Bayes' theorem in Section 5.4.

5.2.1 The Sample Space, Probabilities, and Independence

The following definitions, the foundation for a theory for probability, were first formalized by the French mathematician Laplace (1816) in the early nineteenth century. As

mentioned in the introduction to Chapter 5, Laplace was in the process of creating a calculus for gambling!

DEFINITION

ELEMENTARY EVENT

An *elementary* or *atomic event* is a happening or occurrence that cannot be made up of other events.

EVENT, E

An *event* is a set of elementary events.

SAMPLE SPACE, S

The set of all possible outcomes of an event E is the *sample space* S or *universe* for that event.

PROBABILITY, p

The *probability* of an event E in a sample space S is the ratio of the number of elements in E to the total number of possible outcomes of the sample space S of E. Thus, p(E) = |E| / |S|.

For example, what is the probability that a 7 or an 11 are the result of the roll of two fair dice? We first determine the sample space for this situation. Using the multiplication principle of counting, each die has 6 outcomes, so the total set of outcomes of the two dice is 36. The number of combinations of the two dice that can give a 7 is 1,6; 2,5; 3,4; 4,3; 5,2; and 6,1 - 6 altogether. The probability of rolling a 7 is thus 6/36 = 1/6. The number of combinations of the two dice that can give an 11 is 5,6; 6,5 - or 2, and the probability of rolling an 11 is 2/36 = 1/18. Using the additive property of distinct outcomes, there is 1/6 + 1/18 or 2/9 probability of rolling either a 7 or 11 with two fair dice.

In this 7/11 example, the two events are getting a 7 and getting an 11. The elementary events are the distinct results of rolling the two dice. Thus the event of a 7 is made up of the six atomic events (1,6), (2,5), (3,4), (4,3), (5,2), and (6,1). The full sample space is the union of all thirty-six possible atomic events, the set of all pairs that result from rolling the dice. As we see soon, because the events of getting a 7 and getting an 11 have no atomic events in common, they are *independent,* and the probability of their sum (union) is just the sum of their individual probabilities.

In a second example, how many four-of-a-kind hands can be dealt in all possible five-card poker hands? First, the set of atomic events that make up the full space of all five-card poker hands is the combination of 52 cards taken 5 at a time. To get the total number of four-of-a-kind hands we use the multiplication principle. We multiply the number of combinations of 13 cards taken 1 at a time (the number of different kinds of cards: ace, 2, 3..., king) times the number of ways to pick all four cards of the same kind (the combination of 4 cards taken 4 at a time) times the number of possible other cards that fill out the 5 card hand (48 cards remain). Thus, the probability of a four-of-a-kind poker hand is:

$$({}_{13}C_1 \times {}_4C_4 \times {}_{48}C_1) \; / \; {}_{52}C_5 = 13 \times 1 \times 48 \; / \; 2{,}598{,}960 \approx 0.00024$$

Several results follow immediately from the definitions just made. First, the probability of any event E from the sample space S is:

$$0 \leq p(E) \leq 1, \text{ where } E \subseteq S$$

A second result is that the sum of the probabilities of all possible outcomes in S is 1. To see this, note that the definition for sample space S indicates that it is made up of the union of all individual events E in the problem.

As a third result of the definitions, note that the *probability of the complement* of an event is:

$$p(\overline{E}) = (|S| - |E|) \; / \; |S| \; = \; (|S| \; / \; |S|) - \; (|E| \; / \; |S|) = 1 - \; p(E).$$

The complement of an event is an important relationship. Many times it is easier to determine the probability of an event happening as a function of it not happening, for example, determining the probability that at least one element of a randomly generated bit string of length n is a 1. The complement is that all the bits in the string are 0, with probability 2^{-n}, with n the length of the string. Thus, the probability of the original event is $1 - 2^{-n}$.

Finally, from the probability of the complement of an event set we work out the probability when no event occurs, sometimes referred to as a *contradictory* or *false* outcome:

$$p(\{\}) = 1 - \; p(\overline{\{\}}) = 1 - p(S) = 1 - 1 = 0, \text{ or alternatively,}$$
$$= |\{\}| \; / \; |S| = 0 \; / \; |S| = 0$$

A final important relationship, the probability of the union of two sets of events, may be determined from the principle of counting presented in Section 5.1, namely that for any two sets A and B: $|A \cup B| = |A| + |B| - |A \cap B|$. From this relationship we can determine the probability of the union of any two sets taken from the sample space S:

$$p(A \cup B) = |A \; \cup \; B| \; / \; |S| \; = \; (|A| \; + \; |B| \; - \; |A \; \cap \; B|) \; / \; |S|$$
$$= |A| \; / \; |S| \; + \; |B| \; / \; |S| \; - \; |A \cap B| \; / \; |S| \; = \; p(A) \; + \; p(B) \; - \; p(A \; \cap \; B)$$

Of course, this result may be extended to the union of any number of sets, along the line of the principle of inclusion/exclusion presented in Section 5.1.

We already presented an example of determining the probability of the union of two sets: The probability of rolling a 7 or an 11 with two fair dice. In this example, the formula just presented was used with the probability of the pairs of dice that gave a 7 disjoint from the pairs of dice that gave an 11. We may also use this formula in the more general case when the sets are not disjoint. Suppose we wanted to determine, rolling two fair dice, the probability of rolling an 8 or of rolling pairs of the same number. We would simply calculate the probability of this union where there is one elementary event - (4,4) - that is in the intersection of both desired outcome events.

We next consider the probability of two independent events. Suppose that you are a player in a four-person card game where all the cards are dealt out equally. If you do not have the queen of spades, you can conclude that each of the other players has it with probability 1/3. Similarly, you can conclude that each player has the ace of hearts with probability 1/3 and that any one player has both cards with probability $1/3 \times 1/3$, or 1/9. In this situation we assumed that the events of getting these two cards are independent, even though this is only approximately true. We formalize this intuition with a definition.

DEFINITION

INDEPENDENT EVENTS

Two events A and B are *independent* if and only if the probability of their both occurring is equal to the product of their occurring individually. This independence relation is expressed:

$$p(A \cap B) = p(A) * p(B)$$

We sometimes use the equivalent notation p(s,d) for $p(s \cap d)$. We clarify the notion of independence further in the context of conditional probabilities in Section 5.2.4.

Because the description of independence of events as just presented is an *if and only if* relationship, we can determine whether two events are independent by working out their probabilistic relationships. Consider the situation where bit strings of length four are randomly generated. We want to know whether the event of the bit string containing an even number of 1s is independent of the event where the bit string ends with a 0. Using the multiplication principle, each bit having 2 values, there are a total of $2^4 = 16$ bit strings of length 4.

There are 8 bit strings of length four that end with a 0: {1110, 1100, 1010, 1000, 0010, 0100, 0110, 0000}. There are also 8 bit strings that have an even number of 1s: {1111, 1100, 1010, 1001, 0110, 0101, 0011, 0000}. The number of bit strings that have both an even number of 1s and end with a 0 is 4: {1100, 1010, 0110, 0000}. Now these two events are independent since

$$p(\{\text{even number of 1s}\} \cap \{\text{end with 0}\}) = p(\{\text{even number of 1s}\}) \times p(\{\text{end with 0}\})$$
$$4/16 = 8/16 \times 8/16 = 1/4$$

Consider this same example of randomly generated bit strings of length four. Are the two following events independent: *the bit strings have an even number of 1s*, and *the bit strings end in a 1*? When two or more events are not independent, that is the probability of any one event affects the probability of the others, it requires the notion of *conditional probability* to work out their relationships. We see this in Section 5.2.4.

Before closing this section, we note that other axiom systems supporting the foundations of probability theory are possible, for instance, as an extension to the propositional calculus (Section 2.1). As an example of the set-based approach we have taken, the Rus-

sian mathematician Kolmogorov (1950) proposed a variant of the following axioms, equivalent to our definitions. From these three axioms Kolmogorov systematically constructed all of probability theory.

1. The probability of event E in sample space S is between 0 and 1, i.e., $0 \leq p(E) \leq 1$.
2. When the union of all E = S, $p(S) = 1$, and $p(\overline{S}) = 0$.
3. The probability of the union of two sets of events A and B is:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

In the next section we present a simple example of probabilistic reasoning.

5.2.2 Probabilistic Inference: An Example

We now demonstrate examples of reasoning with the ideas just presented. Suppose you are driving the interstate highway system and realize you are gradually slowing down because of increased traffic congestion. You begin to search for possible explanations of the slowdown. Could it be road construction? Has there been an accident? All you are aware of is that you are slowing down. But wait! You have access to the state highway statistics, and with your new automobile based GUI and inferencing system you can download to your car's computer the relevant statistical information. Okay, so you have the data; what can you do with them?

For this example we assume we have three true or false parameters (we will define this type parameter as a *boolean random variable* in Section 5.2.4). First, there is whether or not the traffic - and you - are slowing down. This situation will be labeled S, with assignment of t or f. Second, there is the probability of whether or not there is an accident, A, with assignments t or f. Finally, the probability of whether or not there is road construction at the time, C; again either t or f. We can express these relationships for the interstate highway traffic, thanks to our car-based data download system, in Table 5.1.

The entries of Table 5.1 are interpreted, of course, just like the truth tables of Section 2.1, except that the right hand column gives the probability of the situation on the left hand side happening. Thus, the third row of the table gives the probability of the traffic slowing down and there being an accident but with no construction as 0.16:

$$S \cap \overline{C} \cap A = 0.16$$

It should be noted that we have been developing our probabilistic calculus in the context of *sets* of events. Figure 5.1 demonstrates how the probabilities of Table 5.1 may be represented with the traditional Venn diagram. We could equally well have presented this situation as the probabilistic truth assignments of *propositions*, in which case the $\cap$ would be replaced by a $\wedge$ and Table 5.1 would be interpreted as the truth values of the conjunction of propositions.

Next, we note that the sum of all possible outcomes of the joint distribution of Table 5.1 is 1.0; this is as one would expect with the axioms of probability presented in Section

S	C	A	p
t	t	t	0.01
t	t	f	0.03
t	f	t	0.16
t	f	f	0.12
f	t	t	0.01
f	t	f	0.05
f	f	t	0.01
f	f	f	0.61

Table 5.1 The joint probability distribution for the traffic slowdown, S, accident, A, and construction, C, variables of the example of Section 5.3.2.

5.2.1. We can also work out the probability of any simple or complex set of events. For example, we can calculate the probability that there is traffic slowdown S. The value for slow traffic is 0.32, the sum of the first four lines of Table 5.1; that is, all the situations where S = t. This is sometimes called the *unconditional* or *marginal probability* of slow traffic, S. This process is called *marginalization* because all the probabilities other than slow traffic are summed out. That is, the distribution of a variable can be obtained by summing out all the other variables from the joint distribution containing that variable.

In a like manner, we can calculate the probability of construction C with no slowdown S - a phenomenon not uncommon in the State of New Mexico! This situation is captured by $p(C \cap S) = t$, as the sum of the 5th and the 6th lines of Table 5.1, or 0.06. If we consider the negation of the situation $C \cap \overline{S}$, we would get (using deMorgan's laws), $p(\overline{C} \cup S)$. Calculating the probability of the union of two sets, as presented in Section 5.2.1, we obtain:

$$0.16 + 0.12 + 0.01 + 0.61 + 0.01 + 0.03 + 0.16 + 0.12 - (0.16 + 0.12) = .94$$

And again the total probability of $C \cap \overline{S}$ and its complement (negation) is 1.0.

5.2.3 Random Variables

In the theory of probability, individual probabilities are either computed analytically, through combinatorial methods, or empirically, by sampling a population of events. To this point most of our probabilities have been determined analytically. For example, there are six sides to a die, and two sides to a coin. When the die or coin is "fair" we say that each outcome, from rolling or flipping, is equally likely. As a result, it is straightforward to determine the event space for these problem situations we later call *parametric*.

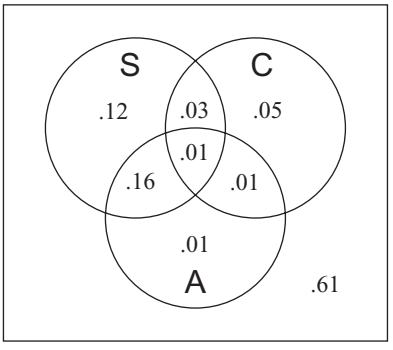

Figure 5.1 A Venn diagram representation of the probability distributions of Table 5.1; S is traffic slowdown, A is accident, C is construction.

More interesting probabilistic reasoning, however, results from the sampling based analysis of situations in the actual world of events. These situations often lack a well defined specification that supports the analytical calculation of probabilities. It is also the case that some situations, even when an analytic foundation exists, are so complex that time and computation costs are not sufficient for the deterministic calculation of probabilistic outcomes. In these situations we usually adopt an empirical sampling methodology.

Most importantly, we assume that all outcomes of an experiment are not equally likely. We retain, however, the basic axioms or assumptions we have made in the previous sections; namely, that the probability of an event is a number between (and including) 0 and 1, and that the summed probabilities of all outcomes is 1. We also retain our rule for the probability of unioned sets of events. We define the idea of a *random variable* as a method for making this calculus precise.

D E F I N I T I O N

RANDOM VARIABLE
A *random variable* is a function whose domain is a sample space and range a set of outcomes, most often real numbers. Rather than using a problem-specific event space, a random variable allows us to talk about probabilities as numerical values that are related to an event space.

BOOLEAN, DISCRETE, and CONTINUOUS RANDOM VARIABLES
A *boolean random variable* is a function from an event space to {true, false} or to the subset of real numbers {0.0, 1.0}. A boolean random variable is sometimes called a *Bernoulli trial*.

A *discrete random variable*, which includes boolean random variables as a subset, is a function from the sample space to (a countable subset of) real numbers in [0.0, 1.0].

A *continuous random variable* has as its range the set of real numbers.

An example using a discrete random variable on the domain of Season, where the atomic events of Season are {spring, summer, fall, winter}, assigns .75, say, to the domain element Season = spring. In this situation we say p(Season = spring) = .75. An example of a boolean random variable in the same domain would be the mapping p(Season = spring) = true. Most of the probabilistic examples that we consider will be of discrete random variables.

Another example of the use of a boolean random variable would be to calculate the probability of obtaining 5 heads in 7 flips of a fair coin. This would be the combination of 5 of 7 flips being heads times the 1/2 probability of heads to the 5th power times the 1/2 probability of not getting heads to the 2nd power, or:

$$_7C_5 \times (1/2)^5 \times (1/2)^2$$

This coin flip situation is an example of what is called the *binomial distribution*. In fact, the outcome of any situation where we want to measure r successes in n trials, where p is the known probability of success may be represented as:

$$_nC_r \times p^r \times (1 - p)^{(n - r)}$$

An important natural extension to associating probabilistic measures to events is the notion of the expected cost or payoff for that outcome. For example, we can calculate the prospective payback from betting specific money values of the draw of a card or the spin of a roulette wheel. We define the *expectation of a random variable or event*, *ex(E)*:

DEFINITION

EXPECTATION OF AN EVENT

If the reward for the occurrence of an event E, with probability p(E), is r, and the cost of the event not occurring, 1 - p(E), is c, then the *expectation* derived from an event occurring, ex(E), is:

$$ex(E) = r \times p(E) + c \times (1 - p(E))$$

For example, suppose that a fair roulette wheel has integers 0 through 36 equally spaced on the slots of the wheel. In the game each player places $1 on any number she chooses: if the wheel stops on the number chosen, she wins $35; otherwise she loses the dollar. The reward of a win is $35; the cost of a loss, $1. Since the probability of winning is 1/37, of losing, 36/37, the expected value for this event, ex(E), is:

$$ex(E) = 35\ (1/37)\ +\ (-1)\ (36/37)\ \approx\ -0.027$$

Thus the player loses, on average, about $0.03 per play!

We conclude this subsection with a brief discussion and summary of the origins of the values of probabilities used in stochastic reasoning. As noted above, in most of our examples so far, we considered probabilistic values that can be determined by reasoning about known situations such as the flip of a fair coin or the spin of a fair roulette wheel. When we have these situations, we can draw many conclusions about aspects of the probability space, such as its *mean*, the statistically-based probability measure, and how far from this mean the sampled values usually vary, the *standard deviation* of the outcomes in this domain.

We refer to this well understood situation as the *parametric* approach to the generation of a sample outcome space. The parametric approach is justified in stochastic situations where there exist a priori expectations for the structure of the results of experimental trials. Our task is to "fill in" the parameters of this well understood situation. An example is flipping a fair coin with the outcome as the binomial distribution. We then can build our expectations of the situation with the binomial model for possible outcomes.

There are a number of advantages of parametric approaches. The first is that fewer data points are needed to calibrate the expected outcomes, since the shape of the outcome curve is known in advance. A further advantage is that it is often possible to determine a priori the number of outcomes or the amount of training data sufficient to make quality probability estimates. In fact, in the parametric situation, besides calculating the mean and standard deviation of the expectations, we can make an accurate determination of when certain data points are outside normal expectations.

Of course, many, if not most, interesting situations do not have clear expected outcomes. One example is diagnostic reasoning, in medicine, say. A second example is the use and interpretation of a natural language expression. With language it is quite common to take a *non parametric* approach to expectations by sampling a large number of situations, as one might have for example, in a language *corpus*. By analyzing collected examples of language use in newspapers, say, or in conversations from a help-desk for computer support, it is possible to infer meaning for ambiguous expressions in these domains. We demonstrate this methodology analyzing possible phoneme relationships in Section 5.3.

With sufficient data points, the resulting discrete distribution in non parametric environments can often be smoothed by interpolation to be continuous. Then new situations can be inferred in the context of this created distribution. A major disadvantage of non parametric methods is that, with the absence of the constraints from prior expectations, a large amount of training data is often required to compensate. We present examples of this type reasoning in Section 5.3.

5.2.4 Conditional Probability

The probability measures discussed to this point in Chapter 5 are often called *prior probabilities*, because they are worked out prior to having any new information about the

expected outcomes of events in a particular situation. In this present section we consider the *conditional probability* of an occurrence of an event, that is, the probability of the event, given some new information or constraint on that event.

As seen earlier in this chapter, the *prior probability* of getting a 2 or a 3 on the roll of a fair die is the sum of these two individual results divided by the total number of possible outcomes of the roll of a fair die, or 2/6. The prior probability of a person having a disease is the number of people with the disease divided by the number of people in the domain of concern.

An example of a *conditional* or *posterior probability* is the situation of a patient entering the doctor's office with a set of symptoms, headaches and nausea, say. The experienced doctor will know a set of prior expectations for different diseases based on symptoms, but will want to determine a specific diagnosis for this patient currently suffering from the headaches and nausea. To make these ideas more precise we make two important definitions.

DEFINITION

PRIOR PROBABILITY
The *prior probability*, generally an *unconditioned probability,* of an event is the probability assigned based on all knowledge supporting its occurrence or absence, that is, the probability of the event prior to any new evidence. The prior probability of an event is symbolized: p(event).

POSTERIOR PROBABILITY
The *posterior* (after the fact) *probability*, generally a *conditional probability,* of an event is the probability of an event given some new evidence. The posterior probability of an event given some evidence is symbolized: p(event | evidence).

We next begin the presentation of Bayes' theorem, who's general form is seen in Section 5.4. The idea supporting Bayes is that the probability of a new (posterior) situation of an hypothesis given evidence can be seen as a function of known probabilities for the evidence given that hypothesis. We can say that we wish to determine the function *f*, such that $p(h|e) = f(p(e|h))$. We usually want to determine the value on the left side of this equation given that it is often much easier to compute the values on the right hand side.

We now prove Bayes' theorem for one symptom and one disease. Based on the previous definitions, the posterior probability of a person having disease d, from a set of diseases D, with symptom or evidence, s, from a set of symptoms S, is:

$$p(d|s) = |d \cap s| / |s|$$

As in Section 5.1, the "|"s surrounding a set is the cardinality or number of elements in that set. The right side of this equation is the number of people having both (intersection) the disease d and the symptom s divided by the total number of people having the symptom s. Figure 5.2 presents a Venn diagram of this situation. We expand the right hand side

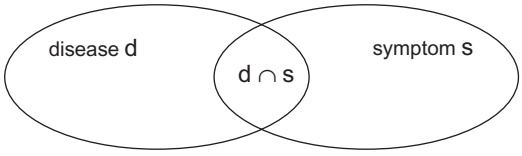

Figure 5.2 A Venn diagram illustrating the calculations of p(d|s) as a function of p(s|d).

of this equation. Since the sample space for determining the probabilities of the numerator and denominator are the same, we get:

$$p(d|s) = p(d \cap s) / p(s).$$

There is an equivalent relationship for $p(s|d)$; again, see Figure 5.2:

$$p(s|d) = p(s \cap d) / p(d).$$

We next solve the $p(s|d)$ equation to determine the value for $p(s \cap d)$:

$$p(s \cap d) = p(s|d)\, p(d).$$

Substituting this result in the previous equation for $p(d|s)$ produces Bayes' rule for one disease and one symptom:

$$p(d|s) = \frac{p(s|d)p(d)}{p(s)}$$

Thus, the posterior probability of the disease given the symptom is the product of the likelihood of the symptom given the disease and the likelihood of the disease, normalized by the probality of that symptom. We generalize this rule in Section 5.4.

We next present the *chain rule*, an important technique used across many domains of stochastic reasoning, especially in natural language processing. We have just developed the equations for any two sets, A_1 and A_2:

$$p(A_1 \cap A_2) = p(A_1 \mid A_2)\, p(A_2) = p(A_2 \mid A_1)\, p(A_1).$$

and now, the generalization to multiple sets A_i, called the chain rule:

$$p(A_1 \cap A_2 \cap \ldots \cap A_n) = p(A_1)\, p(A_2 \mid A_1)\, p(A_3 \mid A_1 \cap A_2) \ldots p(A_n \mid \bigcap_{i=1}^{n-1} A_i)$$

We make an inductive argument to prove the chain rule, consider the nth case:

$$p(A_1 \cap A_2 \cap ... \cap A_{n-1} \cap A_n) = p((A_1 \cap A_2 \cap ... \cap A_{n-1}) \cap A_n),$$

We apply the intersection of two sets rule to get:

$$p((A_1 \cap A_2 \cap ... \cap A_{n-1}) \cap A_n) = p(A_1 \cap A_2 \cap ... \cap A_{n-1})\, p(A_n \mid A_1 \cap A_2 \cap ... \cap A_{n-1})$$

and then reduce again, considering that:

$$p(A_1 \cap A_2 \cap ... A_{n-1}) = p((A_1 \cap A_2 \cap ... \cap A_{n-2}) \cap A_{n-1})$$

until $p(A_1 \cap A_2)$ is reached, the base case, which we have already demonstrated.

We close this section with several definitions based on the use of the chain rule relationship. First, we redefine independent events (see Section 5.2.1) in the context of conditional probabilities, and then we define conditionally independent events, or the notion of how events can be independent of each other, given some third event.

D E F I N I T I O N

INDEPENDENT EVENTS
Two events A and B are *independent* of each other if and only if $p(A \cap B) = p(A)\, p(B)$. When $p(B) \neq 0$ this is the same as saying that $p(A) = p(A|B)$. That is, knowing that B is true does not affect the probability of A being true.

CONDITIONALLY INDEPENDENT EVENTS
Two events A and B are said to be *conditionally independent* of each other, given event C if and only if $p((A \cap B) \mid C) = p(A \mid C)\, p(B \mid C)$.

As a result of the simplification of general chain rule offered by conditionally independent events, larger stochastic systems can be built with smaller computational cost; that is, conditionally independent events simplify joint distributions. An example from our slow traffic situation: suppose that as we slow down we notice orange traffic control barrels along the side of the road. Besides suggesting that the cause of our slowdown is now more likely to be from road construction than a traffic accident, the presence of orange barrels will have its own probabilistic measure. In fact, the variables representing traffic slowdown and the presence of orange barrels are conditionally independent since they are both caused by road construction. Thus, we say that the variable road construction *separates* traffic slowdown from orange barrels.

Because of the statistical efficiencies gained by conditional independence, a major task in the construction of large stochastic computational systems is to break out a complex problem into more weakly connected subproblems. The subproblem interaction relationships are then controlled by the various conditional separation relationships. We see this idea further formalized with the definition of *d-separation* in Section 9.3.

Next, in Section 5.3, we present the general form of Bayes' theorem and demonstrate how, in complex situations, the computation necessary to support full Bayesian inference

can become intractable. Finally, in Section 5.4, we present examples of reasoning based on Bayesian-based probability measures.

5.3 Bayes' Theorem

The Reverend Thomas Bayes was a mathematician and a minister. His famous theorem was published in 1763, four years after his death. His paper, entitled *Essay towards Solving a Problem in the Doctrine of Chances* was published in the *Philosophical Transactions of the Royal Society of London*. Bayes' theorem relates cause and effect in such a way that by understanding the effect we can learn the probability of its causes. As a result Bayes' theorem is important both for determining the causes of diseases, such as cancer, as well as useful for determining the effects of some particular medication on that disease.

5.3.1 Introduction

One of the most important results of probability theory is the general form of Bayes' theorem. First, we revisit one of the results of Section 5.2.4, Bayes' equation for one disease and one symptom. To help keep our diagnostic relationships in context, we rename the variables used previously to indicate individual hypotheses, h_i, from a set of hypotheses, H, and a set of evidence, E. Furthermore, we will now consider the set of individual hypotheses h_i as disjoint, and having the union of all h_i to be equal to H.

$$p(h_i|E) = (p(E|h_i) \times p(h_i)) / p(E)$$

This equation may be read, "The probability of an hypothesis h_i given a set of evidence E is . . ." First, note that the denominator $p(E)$ on the right hand side of the equation is very much a normalizing factor for any of the hypotheses h_i from the set H. It is often the case that Bayes' theorem is used to determine which hypothesis out of a set of possible hypotheses is strongest, given a particular evidence set E. In this case we often drop the $p(E)$ denominator which is identical for all the h_i, with the saving of a possibly large computational cost. Without the denominator, we have created the *maximum a posteriori* value for an hypothesis:

$$\arg\max(h_i)\ p(E|h_i)\ p(h_i)$$

We read this expression as "The maximum value over all h_i of $p(E|h_i)\ p(h_i)$". The simplification just described is highly important for diagnostic reasoning as well as in natural language processing. Of course, the *arg max*, or maximum likelihood hypothesis, is no longer a random variable as defined previously.

Next consider the calculation of the denominator $p(E)$ in the situation where the entire sample space *is partitioned by* the set of hypotheses h_i. The partition of a set is defined as the split of that set into disjoint non overlapping subsets, the union of which

make up the entire set. Assuming that the set of hypotheses h_i partition the entire sample space, we get:

$$p(E) = \sum_i p(E|h_i)\, p(h_i)$$

This relationship is demonstrated by considering the fact that the set of hypotheses h_i forms a partition of the full set of evidence E along with the rule for the probability of the intersection of two sets. So:

$$E = (E \cap h_1) \cup (E \cap h_2) \cup \ldots \cup (E \cap h_n)$$

But by the generalized union of sets rule developed in Section 5.2.1:

$$\begin{aligned} p(E) &= p((E \cap h_1) \cup (E \cap h_2) \cup \ldots \cup (E \cap h_n)) \\ &= p(E \cap h_1) + p(E \cap h_2) + \ldots + p(E \cap h_n) - p(E \cap h_1 \cap E \cap h_2 \cap \ldots \cap h_n) \\ &= p(E \cap h_1) + p(E \cap h_2) + \ldots + p(E \cap h_n) \end{aligned}$$

since the set of hypotheses h_i partition E and their intersection is empty.

This calculation of $p(E)$ produces the general form of Bayes' theorem, where we assume that the set of hypotheses, h_i partition the evidence set E:

$$p(h_i|E) = \frac{p(E|h_i) \times p(h_i)}{\sum_{k=1}^{n} p(E|h_k) \times p(h_k)}$$

$p(h_i \mid E)$ is the probability that h_i is true given evidence E.
$p(h_i)$ is the probability that h_i is true overall.
$p(E \mid h_i)$ is the probability of observing evidence E when h_i is true.
n is the number of possible hypotheses.

Bayes' theorem provides a way of computing the probability of a hypothesis h_i, given a particular piece of evidence, given only the probabilities with which the evidence follows from actual causes (the hypotheses).

As an example, suppose we want to examine the geological evidence at some location to see whether or not it is suited to finding copper. We must know in advance the probability of finding each of a set of minerals and the probability of certain evidence being present when any particular mineral is found. Then we can use Bayes' theorem, with evidence found at the particular location, to determine the likelihood of copper. This approach is used by PROSPECTOR, built at Stanford University and SRI International and employed in mineral exploration (copper, molybdenum, and other minerals). PROSPECTOR has found commercially significant mineral deposits at several sites (Duda et al. 1979*a*).

We next present a simple numerical example demonstrating Bayes' theorem. Suppose that you go out to purchase an automobile. The probability that you will go to dealer 1, d_1, is 0.2. The probability of going to dealer 2, d_2, is 0.4. There are only three dealers you are considering and the probability that you go to the third, d_3, is also 0.4. At d_1 the probability of purchasing a particular automobile, a_1, is 0.2; at dealer d_2 the probability of purchasing a_1 is 0.4. Finally, at dealer d_3, the probability of purchasing a_1 is 0.3. Suppose you purchase automobile a_1. What is the probability that you purchased it at dealer d_2?

First, we want to know, given that you purchased automobile a_1, that you bought it from dealer d_2, i.e., to determine $p(d_2|a_1)$. We present Bayes' theorem in variable form for determining $p(d_2|a_1)$ and then with variables bound to the situation in the example.

$$p(d_2|a_1) = (p(a_1|d_2)\ p(d_2)) / (p(a_1|d_1)\ p(d_1) + p(a_1|d_2)\ p(d_2) + p(a_1|d_3)\ p(d_3))$$
$$= (0.4)\ (0.4) / ((0.2)\ (0.2) + (0.4)\ (0.4) + (0.4)\ (0.3))$$
$$= 0.16 / 0.32$$
$$= 0.5$$

There are two major commitments in using Bayes' theorem: first, all the probabilities on the relationships of the evidence with the various hypotheses must be known, as well as the probabilistic relationships among the pieces of evidence. Second, and sometimes more difficult to determine, all the relationships between evidence and hypotheses, or $p(E|h_k)$, must be estimated or empirically sampled. Recall that calculation of $p(E)$ for the general form of Bayes' theorem also required that the hypothesis set h_i partitioned the set of evidence E. In general, and especially in areas such as medicine and natural language processing, an assumption of this partition cannot be justified a priori.

It is an interesting to note, however, that many situations that violate this assumption (that the individual pieces of evidence partition the evidence set) behave quite well! Using this partition assumption, even in situations where it is not justified, is called using *naive Bayes* or a *Bayes classifier.* With naive Bayes, the assumption is, for an hypothesis h_j:

$$p(E|h_j) \approx \prod_{i=1}^{n} p(e_i|\ h_j)$$

i.e., we assume the pieces of evidence are independent, given a particular hypothesis.

Using Bayes' theorem to determine the probability of some hypothesis h_i given a set of evidence E, $p(h_i|E)$, the numbers on the right-hand side of the equation are often easily obtainable. This is especially true when compared to obtaining the values for the left-hand side of the equation, i.e., determining $p(h_i|E)$ directly. For example, because the population is smaller, it is much easier to determine the number of meningitis patients who have headaches than it is to determine the percentage of headache sufferers with meningitis. Even more importantly, for the simple case of a single disease and a single symptom, not very many numbers are needed. Troubles begin, however, when we consider multiple diseases h_i from the domain of diseases H and multiple symptoms e_n from a set E of possible symptoms. When we consider each disease from H and each symptom from E singly, we have $m \times n$ measures to collect and integrate. (Actually $m \times n$ posterior probabilities plus $m + n$ prior probabilities.)

Unfortunately, our analysis is about to get much more complex. To this point, we considered each symptom e_i individually. In actual situations, single isolated symptoms are rarely the case. When a doctor is considering a patient, for instance, there are often many combinations of symptoms she must consider. We require a form of Bayes' theorem to consider any single hypothesis h_i in the context of the union of multiple possible symptoms e_i.

$$p(h_i|e_1 \cup e_2 \cup ... \cup e_n) = (p(h_i)\ p(e_1 \cup e_2 \cup ... \cup e_n|h_i)) / p(e_1 \cup e_2 \cup ... \cup e_n)$$

With one disease and a single symptom we needed only $m \times n$ measurements. Now, for every pair of symptoms e_i and e_j and a particular disease hypothesis h_i, we need to know both $p(e_i \cup e_j \mid h_i)$ and $p(e_i \cup e_j)$. The number of such pairs is $n \times (n - 1)$, or approximately n^2, when there are n symptoms in E. Now, if we want to use Bayes, there will be about ($m \times n^2$ conditional probabilities) + (n^2 symptom probabilities) + (m disease probabilities) or about $m \times n^2 + n^2 + m$ pieces of information to collect. In a realistic medical system with 200 diseases and 2000 symptoms, this value is over 800,000,000!

There is some hope, however. As was discussed when we presented conditional independence, many of these symptom pairs will be independent, that is $p(e_i|e_j) = p(e_i)$. Independence means, of course, that the probability of e_i is not affected by the presence of e_j. In medicine, for example, most symptoms are not related, e.g., hair loss and sore elbow. But even if only ten percent of our example symptoms are not independent, there are still about 80,000,000 relationships remaining to consider.

In many diagnostic situations, we must also deal with negative information, e.g., when a patient does not have a symptom such as bad blood pressure. We require both:

$$p(\overline{e_i}) = 1 - p(e_i) \text{ and } p(\overline{h_i}|e_i) = 1 - p(h_i|e_i).$$

We also note that $p(e_i|h_i)$ and $p(h_i|e_i)$ are not the same and will almost always have different values. These relationships, and the avoidance of circular reasoning, is important for the design of *Bayesian belief networks* considered in Section 9.3.1.

A final problem, which again makes keeping the statistics of complex Bayesian systems virtually intractable, is the need to rebuild probability tables when new relationships between hypotheses and evidence sets are discovered. In many active research areas such as medicine, new discoveries happen continuously. Bayesian reasoning requires complete and up-to-date probabilities, including joint probabilities, if its conclusions are to be correct. In many domains, such extensive data collection and verification are not possible, or if possible, quite expensive.

Where these assumptions are met, however, Bayesian approaches offer the benefit of a mathematically well-founded handling of uncertainty. Most expert system domains do not meet these requirements and must rely on heuristic approaches, as presented in Chapter 8. Furthermore, due to complexity issues, we know that even fairly powerful computers cannot use full Bayesian techniques for successful real-time problem solving. We end this chapter with two examples that show how a Bayesian approach might work to organize hypothesis/evidence relationships. But first we must define probabilistic finite state machines/acceptors

5.4 Applications of the Stochastic Methodology

You say [t ow m ey t ow] and I say [t ow m aa t ow]. . .

—IRA GERSHWIN, "*Let's Call the Whole Thing Off"*

In this section we present two examples that use probability measures to reason about the interpretation of ambiguous information. First, we define an important modeling tool based on the finite state machine of Section 3.1, *the probabilistic finite state machine*.

DEFINITION

PROBABILISTIC FINITE STATE MACHINE

A *probabilistic finite state machine* is a finite state machine where the next state function is a probability distribution over the full set of states of the machine.

PROBABILISTIC FINITE STATE ACCEPTOR

A *probabilistic finite state machine* is an *acceptor*, when one or more states are indicated as the *start* states and one or more as the *accept* states.

It can be seen that these two definitions are simple extensions to the finite state and the Moore machines presented in Section 3.1. The addition for non-determinism is that the next state function is no longer a function in the strict sense. That is, there is no unique range state for each input value and each state of the domain. Rather, for any state, the next state function is a probability distribution over all possible next states.

5.4.1 So How Is "Tomato" Pronounced?

Figure 5.3 presents a probabilistic finite state acceptor that represents different pronunciations of the word "tomato". A particular acceptable pronunciation for the word tomato is

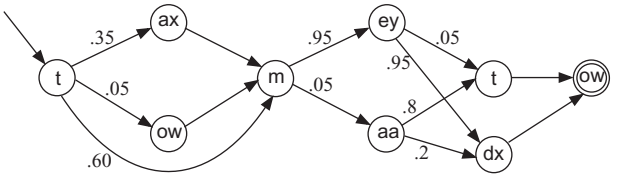

Figure 5.3 A probabilistic finite state acceptor for the pronunciation of tomato , adapted from Jurafsky and Martin (2009).

characterized by a path from the start state to the accept state. The decimal values that label the arcs of the graph represent the probability that the speaker will make that particular transition in the state machine. For example, 60% of all speakers in this data set go directly from the t phone to the m without producing any vowel phone between.

Besides characterizing the various ways people in the pronunciation database speak the word "tomato", this model can be used to help interpret ambiguous collections of phonemes. This is done by seeing how well the phonemes match possible paths through the state machines of related words. Further, given a partially formed word, the machine can try to determine the probabilistic path that best completes that word.

In a second example, also adapted from Jurafsky and Martin (2009), we consider the *phoneme recognition* problem, often called *decoding*. Suppose a phoneme recognition algorithm has identified the phone ni (as in "knee") that occurs just after the recognized word (phone) I, and we want to associate ni with either a word or the first part of a word. In this case we have linguistic *corpora*, the *Brown* and *Switchboard* corpora, to assist us.

The *Brown corpus* is a one million word collection of sentences from 500 written texts, including newspapers, novels, and academic writings collected at Brown University in the 1960s (Kucera and Francis 1967, Francis 1979). The *Switchboard corpus* is a 1.4 million word collection of telephone conversations. These corpora together contain about 2,500,000 words that let us sample both written and spoken information bases.

There are a number of ways to proceed in identifying the most likely word to associate with the ni phone. First we can determine which word, with this phone first, is the most likely to be used. Table 5.2 presents the raw frequencies of these words along with the probability of their occurrence, that is, the frequency of the word divided by the total number of words in these combined corpora. (This table is adapted from Jurafsky and Martin (2009); see their book for a justification that "the" belongs to this collection.) From this data, the word "the" would seem to be the first choice for matching ni.

We next apply a form of Bayes' theorem we presented in the previous section. Our second attempt at analyzing the phone ni following I, uses a simplification of that formula that ignores its denominator:

$$p(\text{word} \mid [ni]) \propto p([ni] \mid \text{word}) \times p(\text{word})$$

word	frequency	probability
knee	61	.000024
the	114834	.046
neat	338	.00013
need	1417	.00056
new	2625	.001

Table 5.2. The *ni* words with their frequencies and probabilities from the Brown and Switchboard corpora of 2.5M words, adapted from Jurafsky and Martin (2009).

word	p([ni] | word)	p(word)	p([ni] | word) x p(word)
new	0.36	0.001	0.00036
neat	0.52	0.00013	0.000068
need	0.11	0.00056	0.000062
knee	1.0	0.000024	0.000024
the	0.0	0.046	0.0

Table 5.3. The *ni* phone/word probabilities from the Brown and Switchboard corpora (Jurafsky and Martin 2009).

The results of this calculation, ordered from most recommended to least, are found in Table 5.3 and explain why p(ni | the) is impossible. The results of Table 5.3 also suggest that new is the most likely word for decoding ni. But the two-word combination I new doesn't seem to make much sense, whereas other combinations, such as I need does. Part of the problem in this situation is that we are still reasoning on the phone level, that is, determining the probability p(ni | new). There is, in fact, a straightforward way of addressing this issue, and that is to look for explicit two word combinations in the corpora. Following this line of reasoning, it turns out that I need is a much more likely pair of consecutive words than is I new, or any of the other I-word combinations, for that matter.

The methodology for deriving probabilities from pairs, or triples, of word combinations in corpora is called *n-gram analysis*. With two words, we were using *bigrams*, with three, *trigrams*. The probabilities derived from word combinations using n-grams is important, as we see again in Chapter 15.

5.4.2 Extending the Road/Traffic Example

We present again and extend the example of Section 5.2.2. Suppose you are driving the interstate highway system and realize you are gradually slowing down because of increased traffic congestion. You begin to search for possible explanations of the slowdown. Could it be road construction? Has there been an accident? Perhaps there are other possible explanations. After a few minutes you come across orange barrels at the side of the road that begin to cut off the outside lane of traffic. At this point you determine that the best explanation of your slow down is road construction. At the same time the alternative hypothesis of accident is *explained away.* Similarly if you would have seen flashing lights in the distance ahead, such as from a police vehicle or an ambulance, the best explanation given this evidence would be a traffic accident and road construction would have been *explained away.* When an hypothesis is explained away that does not mean that it is no longer possible. Rather, in the context of new evidence, it is simply less likely.

Figure 5.4 presents a Bayesian account of what we have just seen. road construction is correlated with orange barrels and bad traffic. Similarly, accident correlates with flashing lights and bad traffic. We examine Figure 5.4 and build a joint probability distribution for the road construction and bad traffic relationship. We simplify both of these variables to be either true (t) or false (f) and represent the probability distribution in Table

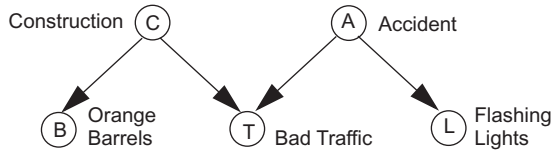

Figure 5.4 The Bayesian representation of the traffic problem with potential explanations.

C is true = .5

C	T	p
t	t	.3
t	f	.2
f	t	.1
f	f	.4

T is true = .4

Table 5.4 The joint probability distribution for the traffic and construction variables of Figure 5.4.

5.4. Note that if construction is f there is not likely to be bad traffic and if it is t then bad traffic is likely. Note also that the probability of road construction on the interstate, C = true, is .5 and the probability of having bad traffic, T = true, is .4 (this is New Mexico!).

We next consider the change in the probability of road construction given the fact that we have bad traffic, or p(C|T) or p(C = t | T = t).

$$p(C|T) = p(C = t , T = t) \ / (p(C = t , T = t) + p(C = f , T = t)) = .3 / (.3 + .1) = .75$$

So now, with the probability of road construction being .5, given that there actually is bad traffic, the probability for road construction goes up to .75. This probability will increase even further with the presence of orange barrels, explaining away the hypothesis of accident.

Besides the requirement that we may have knowledge or measurements for any of our parameters being in a particular state, we also must address, as noted in Section 5.4.1, complexity issues. Consider the calculation of the joint probability of all the parameters of Figure 5.4 (using the chain rule and a topologically sorted order of variables):

$$p(C,A,B,T,L) = p(C) \times p(A|C) \times p(B|C,A) \times p(T|C,A,B) \times p(L|C,A,B,T)$$

This result is a general decomposition of the probability measures that is always true. The cost of producing this joint probability table is exponential in the number of parameters involved, and in this case requires a table of size 2^5 or 32. We are considering a toy problem, of course, with only five parameters. A situation of interesting size, with thirty or

more parameters say, requires a joint distribution table of roughly a billion elements. As we will see in Section 9.3, Bayesian belief networks and d-separation give us further tools for addressing this representational and computational complexity.

5.5 Epilogue and References

Games of chance date back, at least, to the Greek and Roman civilizations. It wasn't until the European Renaissance, however, that the mathematical analysis of probability theory began. As noted in the chapter, probabilistic reasoning begins with determining principles of counting and combinatorics. Actually, one of the first combinatorial "machines" is attributed to Ramon Llull (Ford et al. 1995), a Spanish philosopher and Franciscan monk, who created his device, said to automatically enumerate the attributes of God, with the intention of converting heathens. The first publication on probabilities, *De Ratiociniis Ludo Aleae*, was authored by Christian Huygens (1657). Huygens describes earlier results by Blaise Pascal, including a methodology for calculating probabilities, as well as conditional probabilities. Pascal targeted both the "objective" analysis of the world of games as well as the more "subjective" analysis of belief systems, including the existence of God.

The definitions of probabilities presented in Section 5.2.1 are based on the formalism proposed by the French mathematician Pierre Simon Laplace. Laplace's book *Theorie Analytique des Probabilitees* (1816) documents this approach. Laplace's work was based on earlier results published by Gotlob Leibnitz and James Bernoulli.

Thomas Bayes was a mathematician and a minister. His famous theorem was published in 1764, after his death. His paper, entitled *Essay towards Solving a Problem in the Doctrine of Chances* was published in the *Philosophical Transactions of the Royal Society of London*. Ironically, Bayes' theorem is never explicitly stated in this paper, although it is there! Bayes also has extensive discussion of the "reality" of statistical measures.

Bayes' research was partly motivated to answer the philosophical skepticism of the Scots philosopher David Hume. Hume's dismissal of causality destroyed any foundation for arguments through miracles for the existence of God. In fact, in a 1763 paper presented to the British Royal Society, the minister Richard Price used Bayes' theorem to show there was good evidence in favor of the miracles described in the New Testament.

The mathematicians of the early twentieth century, including Fisher (1922), Popper (1959), and Carnap (1948) completed the foundation of modern probability theory continuing the "subjective/objective" debates on the nature of probabilities. Kolmogorov (1950, 1965) axiomatized the foundations of probabilistic reasoning (see Section 5.2.1).

It is possible to present the topic of probabilistic reasoning from several vantage points. The two most popular approaches are based on the propositional calculus and set theory. For the propositional calculus, Section 2.1, propositions are assigned a confidence or probabilistic truth value in the range [0.0, 1.0]. This approach offers a natural extension to the semantics of propositional calculus. We have chosen the second approach to probabilistic reasoning, feeling this orientation is a bit more intuitive, bringing to bear all the counting and other techniques of set theory, as seen in Section 5.1. In Chapters 9 and 13, we extend our presentation of stochastic systems to first-order (variable based) representation schemes.

Bayes' theorem has offered a foundation for several expert systems of the 1970s and 1980s, including an extensive analysis of acute abdominal pain at the University of Glasgow Hospital (de Dombal et al. 1974), and PROSPECTOR, the system from Stanford University supporting mineral exploration (Duda et al. 1979a). The naive Bayes approach has been used in a number of classification problems, including pattern recognition (Duda and Hart 1973), natural language processing (Mooney 1996), and elsewhere. Domingos and Pazzani (1997) offer justification for the successes of naive Bayes classifiers in situations which do not meet Bayesian independence assumptions.

There is a large amount of current research in probabilistic reasoning in artificial intelligence, some presented in Chapters 9, 13, and 16. Uncertain reasoning makes up a component of AI conferences, including AAAI, IJCAI, NIPS, and UAI. There are several excellent introductory texts in probabilistic reasoning (Ross 1988, DeGroot 1989), as well as several introductions to the use of stochastic methods in artificial intelligence applications (Russell and Norvig 2003, Jurafsky and Martin 2009, Manning and Schutze 1999).

I am indebted to Dan Pless of Sandia National Laboratories for the general approach taken in this chapter, for several of the examples, and also for editorial suggestions.

5.6 Exercises

1. A fair six-sided die is tossed five times and the numbers up are recorded in a sequence. How many different sequences are there?
2. Find the number of distinguishable permutations of the letters in the word MISSISSIPPI, of the letters of the word ASSOCIATIVE.
3. Suppose that an urn contains 15 balls, of which eight are red and seven are black. In how many ways can five balls be chosen so that:
   a. all five are red? all five are black?
   b. two are red and three are black
   c. at least two are black
4. How many ways can a committee of three faculty members and two students be selected from a group of five faculty and seven students?
5. In a survey of 250 television viewers, 88 like to watch news, 98 like to watch sports, and 94 like to watch comedy. 33 people like to watch news and sports, 31 like to watch sports and comedy, and 35 like to watch news and comedy. 10 people like to watch all three. Suppose a person from this group is picked at random:
   a. What is the probability that they watch news but not sports?
   b. What is the probability that they watch news or sports but not comedy?
   c. What is the probability that they watch neither sports nor news?
6. What is the probability that a four digit integer with no beginning zeros:
   a. Has 3, 5, or 7 as a digit?
   b. Begins with 3, ends with 5, or has 7 as a digit?
7. Two dice are tossed. Find the probability that their sum is:
   a. 4

b. 7 or an even number

c. 10 or greater

8. A card is drawn from the usual fifty-two card deck. What is the probability of:

   a. drawing a face card (jack, queen, king or ace)

   b. drawing a queen or a spade

   c drawing a face card or a club

9. What is the probability of being dealt the following hands in a five card poker game (from the normal deck of fifty-two cards)?

   a. A "flush" or all cards from the same suit.

   b. A "full house" or two cards of the same value and three cards of another value.

   c. A "royal flush" or the ten, jack, queen, king, and ace all of the same suit.

10. The *expectation* is the *mean* or average of the value of a random variable. In throwing a die, for example, it can be calculated by totaling up the resulting values from a large number of throws and then dividing by the number of throws.

    a. What is the expectation from throwing a fair die?

    b. The value of a roulette wheel with 37 equally likely results?

    c. The value of a draw from a set of cards (ace is valued at 1, all other face cards as 10)?

11. Suppose that we are playing a game where we toss a die and then receive the amount of dollars equal to the value of the die. For example, if a 3 comes up we receive $3. If it costs us $4 to play this game, is this reasonable?

12. Consider the situation where bit strings of length four are randomly generated. Demonstrate whether or not the event of production of bit strings containing an even number of 1s is independent of the event of producing bit strings that end in a 1.

13. Show that the statement $p(A,B|C) = p(A|C)\ p(B|C)$ is equivalent to both $p(A|B,C) = p(A|C)$ and $p(B|A,C) = p(B|C)$.

14. In manufacturing a product, 85% of the products that are produced are not defective. Of the products inspected, 10% of the good ones are seen as defective and not shipped whereas only 5% of the defective products are approved and shipped. If a product is shipped, what is the probability that it is defective?

15. A blood test is 90% effective in detecting a disease. It also falsely diagnoses that a healthy person has the disease 3% of the time. If 10% of those tested have the disease, what is the probability that a person who tests positive will actually have the disease?

16. Suppose an automobile insurance company classifies a driver as good, average, or bad. Of all their insured drivers, 25% are classified good, 50% are average, and 25% are bad. Suppose for the coming year, a good driver has a 5% chance of having an accident, and average driver has 15% chance of having an accident, and a bad driver has a 25% chance. If you had an accident in the past year what is the probability that you are a good driver?

17. Three prisoners, A, B, C are in their cells. They are told that one of them will be executed the next day and the others will be pardoned. Only the governor knows who will be executed. Prisoner A asks the guard a favor. "Please ask the governor who will be executed, and then tell either prisoner B or C that they will be pardoned." The guard does as was asked and then comes back and tells prisoner A that he has told prisoner B that he (B) will be pardoned. What are prisoner A's chances of being executed, given this message? Is there more information than before his request to the guard? (Adapted from Pearl 1988).