An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 2

Conditional Independence and Factorization

A graphical model can be thought of as a probabilistic database, a machine that can answer "queries" regarding the values of sets of random variables. We build up the database in pieces, using probability theory to ensure that the pieces have a consistent overall interpretation. Probability theory also justifies the inferential machinery that allows the pieces to be put together "on the fly" to answer queries.

Consider a set of random variables $\{X_1, X_2, \ldots, X_n\}$ and let x_i represent the realization of random variable X_i. Each random variable may be scalar-valued or vector-valued. Thus x_i is in general a vector in a vector space. In this section, for concreteness, we assume that the random variables are discrete; in general, however, we make no such restriction. There are several kinds of query that we might be interested in making regarding such an ensemble. We might, for example, be interested in knowing whether one subset of variables is independent of another, or whether one subset of variables is conditionally independent of another subset of variables given a third subset. Or we might be interested in calculating conditional probabilities—the probabilities of one subset of variables given the values of another subset of variables. Still other kinds of queries will be described in later chapters. In principle all such queries can be answered if we have in hand the joint probability distribution, written $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Questions regarding independence can be answered by factoring the joint probability distribution, and questions regarding conditional probabilities can be answered by appropriate marginalization and normalization operations.

To simplify our notation, we will generally express discrete probability distributions in terms of the probability mass function $p(x_1, x_2, \ldots, x_n)$, defined as $p(x_1, x_2, \ldots, x_n) \triangleq P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. We also will often use X to stand for $\{X_1, \ldots, X_n\}$, and x to stand for $\{x_1, \ldots, x_n\}$, so that $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ can be written more succinctly as $P(X = x)$, or, more succinctly still, as $p(x)$. Note also that subsets of indices are allowed wherever single indices appear. Thus if $A = \{2, 4\}$ and $B = \{3\}$, then X_A is shorthand for $\{X_2, X_4\}$, X_B is shorthand for $\{X_3\}$, and $P(X_A = x_A \,|\, X_B = x_B)$ is shorthand for $P(X_2 = x_2, X_4 = x_4 \,|\, X_3 = x_3)$.

While it is in fact our goal to maintain and manipulate representations of joint probabilities, we must not be naive regarding the size of the representations. In the case of discrete random variables, one way to represent the joint probability distribution is as an n-dimensional table, in which each cell contains the probability $p(x_1, x_2, \ldots, x_n)$ for a specific setting of the variables $\{x_1, x_2, \ldots, x_n\}$. If each variable x_i ranges over r values, we must store and manipulate r^n numbers, a quantity exponential in n. Given that we wish to consider models in which n is in the hundreds or thousands, such a naive tabular representation is out.

Graphical models represent joint probability distributions more economically, using a set of "local" relationships among variables. To define what we mean by "local" we avail ourselves of graph theory.

2.1 Directed graphs and joint probabilities

Let us begin by considering directed graphical representations. A directed graph is a pair $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ a set of (oriented) edges. We will assume that $\mathcal{G}$ is acyclic.

Each node in the graph is associated with a random variable. Formally, we assume that there is a one-to-one mapping from nodes to random variables, and we say that the random variables are *indexed* by the nodes in the graph. Thus, for each $i \in \mathcal{V}$, there is an associated random variable X_i. Letting $\mathcal{V} = \{1, 2, \ldots, n\}$, as we often do for convenience, the set of random variables associated with the graph is given by $\{X_1, X_2, \ldots, X_n\}$.

Although nodes and random variables are rather different formal objects, we will find it convenient to ignore the distinction, letting the symbol "X_i" refer both to a node and to its associated random variable. Indeed, we will often gloss over the distinction between nodes and random variables altogether, using language such as "the marginal probability of node X_i."

Note that we will also sometimes use lower-case letters—that is, the realization variables x_i—to label nodes, further blurring distinctions. Given the strict one-to-one correspondence that we enforce between the notation for random variables (X_i) and their realizations (x_i), however, this is unlikely to lead to confusion.

It would be rather inconvenient to be restricted to the symbol "X" for random variables, and we often use other symbols as well. Thus, we may consider examples in which sets such as $\{W, X, Y, Z\}$ or $\{X_1, X_2, X_3, Y_1, Y_2, Y_3\}$ denote the set of random variables associated with a graph. As long as it is clear which random variable is associated with which node, then formally the random variables are "indexed" by the nodes in the graph as required, even though the indexing is not necessarily made explicit in the notation.

Each node has a set of *parent nodes*, which can be the empty set. For each node $i \in \mathcal{V}$, we let π_i denote the set of parents of node i. We also refer to the set of random variables X_{π_i} as the "parents" of the random variable X_i, exploiting the one-to-one relationship between nodes and random variables.

We use the locality defined by the parent-child relationship to construct economical representations of joint probability distributions. To each node $i \in \mathcal{V}$ we associate a function $f_i(x_i, x_{\pi_i})$. These functions are assumed to have the properties of conditional probability distributions: that is, $f_i(x_i, x_{\pi_i})$ is nonnegative and sums to one with respect to x_i for each value of x_{π_i}. We impose no additional constraint on these functions; in particular, there is no assumption of any relationship between the functions at different nodes.

Let $\mathcal{V} = \{1, 2, \ldots, n\}$. Given a set of functions $\{f_i(x_i, x_{\pi_i}) : i \in \mathcal{V}\}$, we define a joint probability distribution as follows:

$$p(x_1, x_2, \ldots, x_n) \triangleq \prod_{i=1}^{n} f_i(x_i, x_{\pi_i}). \tag{2.1}$$

That is, we define the joint probability as a product of the local functions at the nodes of the graph. To verify that the definition obeys the constraints on a joint probability, we check: (1) the right-hand side is clearly nonnegative; and (2) the assumption that each factor $f_i(x_i, x_{\pi_i})$ sums to one with respect to x_i, together with the assumption that the graph is acyclic, implies that the right-hand side sums to one with respect to $\{x_1, x_2, \ldots, x_n\}$. In particular, we can sum "backward" from the leaves of the graph, summing over the values of leaf nodes and removing the nodes from the graph, obtaining a value of one at each step.[1]

By choosing specific numerical values for the functions $f_i(x_i, x_{\pi_i})$, we generate a specific joint probability distribution. Ranging over all possible numerical choices for these functions, we define a *family of joint probability distributions associated with the graph* $\mathcal{G}$. It will turn out that this family is a natural mathematical object. In particular, as we will see later in this chapter, this family can be characterized not only in terms of products of local functions, but also more "graph-theoretically" in terms of the patterns of edges in the graph. It is this relationship between the different ways to characterize the family of probability distributions associated with a graph that is the key to the underlying theory of probabilistic graphical models.

With a definition of joint probability in hand, we can begin to address the problem of calculating conditional probabilities under this joint. Suppose in particular that we calculate $p(x_i \,|\, x_{\pi_i})$ under the joint probability in Eq. (2.1). What, if any, is the relationship between this conditional probability and $f_i(x_i, x_{\pi_i})$, a function which has the properties of a conditional probability but is otherwise arbitrary? As we ask the reader to verify in Exercise **??**, these functions are in fact one and the same. That is, under the definition of joint probability in Eq. (2.1), the function $f_i(x_i, x_{\pi_i})$ is necessarily the conditional probability of x_i given x_{π_i}. Put differently, we see that the functions $f_i(x_i, x_{\pi_i})$ must form a consistent set of conditional probabilities under a single joint probability. This is a pleasant and somewhat surprising fact given that we can define the functions $f_i(x_i, x_{\pi_i})$ arbitrarily.

Given that functions $f_i(x_i, x_{\pi_i})$ are in fact conditional probabilities, we henceforth drop the f_i notation and write the definition in terms of $p(x_i \,|\, x_{\pi_i})$:[2]

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, x_{\pi_i}). \tag{2.2}$$

[1]If this point is not clear now, it will be clear later when we discuss inference algorithms.

[2]Eq. (2.2) is often used as the definition of the joint probability for a directed graphical model. Such a definition risks circularity, however, because it is not clear in advance that an arbitrary collection of conditional probabilities, $\{p(x_i \,|\, x_{\pi_i})\}$, are necessarily conditionals under the same joint probability. Moreover, it is not clear in advance that an arbitrary collection of conditional probabilities is internally consistent. We thus prefer to treat Eq. (2.1) as the definition and view Eq. (2.2) as a consequence. Having made this cautionary note, however, for simplicity we refer to Eq. (2.2) as the "definition" of joint probability in the remainder of the chapter.

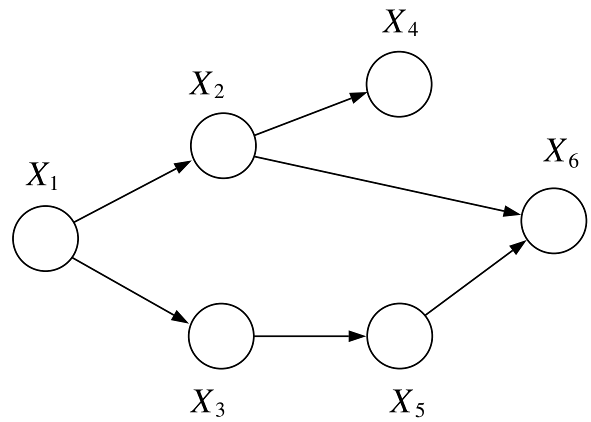


Figure 2.1: An example of a directed graphical model.

We refer to the conditional probabilities $p(x_i \,|\, x_{\pi_i})$ as the *local conditional probabilities* associated with the graph $\mathcal{G}$. These functions are the building blocks whereby we synthesize a joint distribution associated with the graph $\mathcal{G}$.

Figure 2.1 shows an example on six nodes. According to the definition, we obtain the joint probability as follows:

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2 \,|\, x_1)p(x_3 \,|\, x_1)p(x_4 \,|\, x_2)p(x_5 \,|\, x_3)p(x_6 \,|\, x_2, x_5), \tag{2.3}$$

by taking the product of the local conditional distributions.

Let us now return to the problem of representational economy. Are there computational advantages to representing a joint probability as a set of local conditional probabilities?

Each of the local conditional probabilities must be represented in some manner. In later chapters we will consider a number of possible representations for these probabilities; indeed, this representational issue is one of the principal topics of the book. For concreteness, however, let us make a simple choice here. For a discrete node X_i, we must represent the probability that node X_i takes on one of its possible values, for each combination of values for its parents. This can be done using a table. Thus, for example, the probability $p(x_1)$ can be represented using a one-dimensional table, and the probability $p(x_6 \,|\, x_2, x_5)$ can be represented using a three-dimensional table, one dimension for each of x_2, x_5 and x_6. The entire set of tables for our example is shown in Figure 2.2, where for simplicity we have assumed that the nodes are binary-valued. Filling these tables with specific numerical values picks out a specific distribution in the family of distributions defined by Eq. (2.3).

In general, if m_i is the number of parents of node X_i, we can represent the conditional probability associated with node X_i with an $(m_i + 1)$-dimensional table. If each node takes on r values, then we require a table of size r^{m_i+1}.

We have exchanged exponential growth in n, the number of variables in the domain, for exponential growth in m_i, the number of parents of individual nodes X_i (the "fan-in"). This is very often a happy exchange. Indeed, in many situations the maximum fan-in in a graphical model is relatively small and the reduction in complexity can be enormous. For example, in hidden Markov

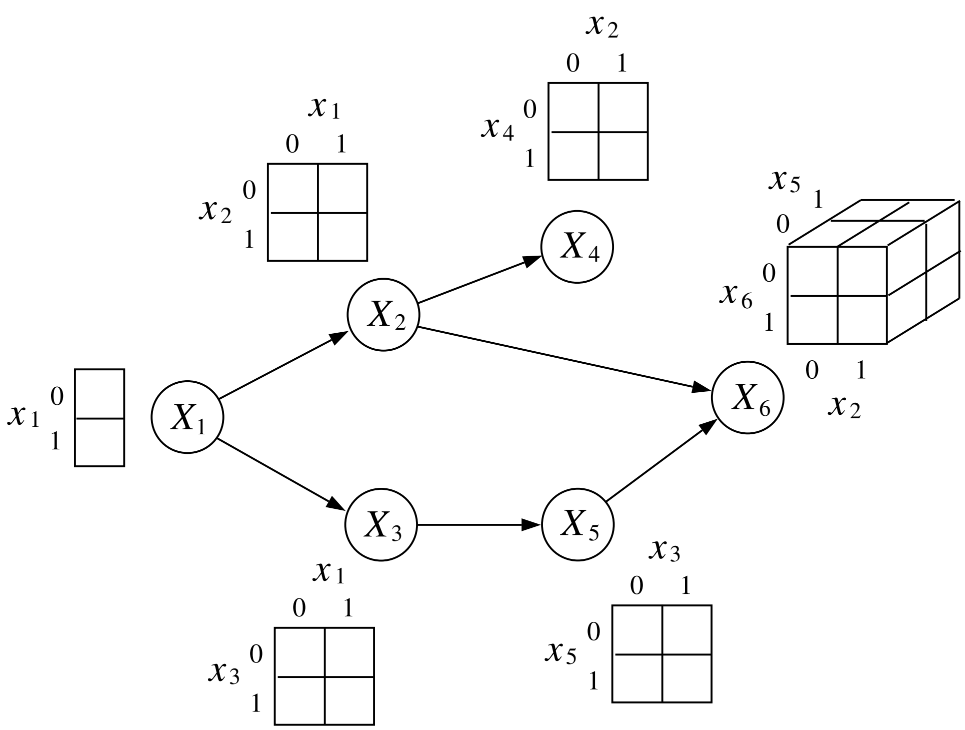


Figure 2.2: The local conditional probabilities represented as tables. Each of the nodes is assumed to be binary-valued. Each of these tables can be filled with arbitrary nonnegative numerical values, subject to the constraint that they sum to one for given fixed values of the parents of a node. Thus, each column in each table must sum to one.

models (see Chapter 12), each node has at most a single parent, while the number of nodes n can be in the thousands.

The fact that graphs provide economical representations of joint probability distributions is important, but it is only a first hint of the profound relationship between graphs and probabilities. As we show in the remainder of this chapter and in the following chapter, graphs provide much more than a data structure; in particular, they provide *inferential* machinery for answering questions about probability distributions.

2.1.1 Conditional independence

An important class of questions regarding probability distributions has to do with conditional independence relationships among random variables. We often want to know whether a set of variables is independent of another set, or perhaps conditionally independent of that set given a third set. Independence and conditional independence are important qualitative aspects of probability theory.

By definition, X_A and X_B are *independent*, written $X_A \perp\!\!\!\perp X_B$, if:

$$p(x_A, x_B) = p(x_A)p(x_B), \tag{2.4}$$

and X_A and X_C are *conditionally independent given* X_B, written $X_A \perp\!\!\!\perp X_C \mid X_B$, if:

$$p(x_A, x_C \mid x_B) = p(x_A \mid x_B)p(x_C \mid x_B), \tag{2.5}$$

or, alternatively,

$$p(x_A \mid x_B, x_C) = p(x_A \mid x_B), \tag{2.6}$$

for all x_B such that $p(x_B) > 0$. Thus, to establish independence or conditional independence we need to factor the joint probability distribution.

Graphical models provide an intuitively appealing, symbolic approach to factoring joint probability distributions. The basic idea is that representing a probability distribution within the graphical model formalism involves making certain independence assumptions, assumptions which are embedded in the structure of the graph. From the graphical structure other independence relations can be derived, reflecting the fact that certain factorizations of joint probability distributions imply other factorizations. The key advantage of the graphical approach is that such factorizations can be read off from the graph via simple graph search algorithms. We will describe such an algorithm in Section 2.1.2; for now let us try to see in general terms why graphical structure should encode conditional independence.

The *chain rule of probability theory* allows a probability mass function to be written in a general factored form, once a particular ordering for the variables is chosen. For example, a distribution on the variables $\{X_1, X_2, \ldots, X_6\}$ can be written as:

$$\begin{aligned} &p(x_1, x_2, x_3, x_4, x_5, x_6) \\ &\quad = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_1, x_2, x_3)p(x_5 \mid x_1, x_2, x_3, x_4)p(x_6 \mid x_1, x_2, x_3, x_4, x_5), \end{aligned}$$

where we have chosen the usual arithmetic ordering of the nodes. In general, we have:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}). \tag{2.7}$$

Comparing this expansion, which is true for an arbitrary probability distribution, with the definition in Eq. (2.2), we see that our definition of joint probability involves dropping some of the conditioning variables in the chain rule. Inspecting Eq. (2.6), it seems natural to try to interpret these missing variables in terms of conditional independence. For example, the fact that $p(x_4 \mid x_2)$ appears in Eq. (2.3) in the place of $p(x_4 \mid x_1, x_2, x_3)$ suggests that we should expect to find that X_4 is independent of X_1 and X_3 given X_2.

Taking this idea a step further, we might posit that *missing variables in the local conditional probability functions correspond to missing edges in the underlying graph.* Thus, $p(x_4 \mid x_2)$ appears as a factor in Eq. (2.3) because there are no edges from X_1 and X_3 to X_4. Transferring the interpretation from missing variables to missing edges we obtain a probabilistic interpretation for the missing edges in the graph in terms of conditional independence. Let us formalize this interpretation.

Define an ordering I of the nodes in a graph $\mathcal{G}$ to be *topological* if for every node $i \in \mathcal{V}$ the nodes in π_i appear before i in the ordering. For example, the ordering $I = (1, 2, 3, 4, 5, 6)$ is a topological ordering for the graph in Figure 2.1. Let ν_i denote the set of all nodes that appear earlier than i in the ordering I, excluding the parent nodes π_i. For example, $\nu_5 = \{1, 2, 4\}$ for the graph in Figure 2.1.

As we ask the reader to verify in Exercise **??**, the set ν_i necessarily contains all *ancestors* of node i (excluding the parents π_i), and may contain other *nondescendant* nodes as well.

Given a topological ordering I for a graph $\mathcal{G}$ we associate to the graph the following set of *basic conditional independence statements*:

$$\{X_i \perp\!\!\!\perp X_{\nu_i} \mid X_{\pi_i}\} \tag{2.8}$$

for $i \in \mathcal{V}$. Given the parents of a node, the node is independent of all earlier nodes in the ordering.

For example, for the graph in Figure 2.1 we have the following set of basic conditional independencies:

$$\begin{aligned}
X_1 \perp\!\!\!\perp \emptyset &\quad\mid\quad \emptyset && (2.9)\\
X_2 \perp\!\!\!\perp \emptyset &\quad\mid\quad X_1 && (2.10)\\
X_3 \perp\!\!\!\perp X_2 &\quad\mid\quad X_1 && (2.11)\\
X_4 \perp\!\!\!\perp \{X_1, X_3\} &\quad\mid\quad X_2 && (2.12)\\
X_5 \perp\!\!\!\perp \{X_1, X_2, X_4\} &\quad\mid\quad X_3 && (2.13)\\
X_6 \perp\!\!\!\perp \{X_1, X_3, X_4\} &\quad\mid\quad \{X_2, X_5\}, && (2.14)
\end{aligned}$$

where the first two statements are vacuous.

Is this interpretation of the missing edges in terms of conditional independence consistent with our definition of the joint probability in Eq. (2.2)? The answer to this important question is "yes," although proof will be again postponed until later. Let us refer to our example, however, to provide a first indication of the basic issues.

Let us verify that X_1 and X_3 are independent of X_4 given X_2 by direct calculation from the joint probability in Eq. (2.3). We first compute the marginal probability of $\{X_1, X_2, X_3, X_4\}$:

$$\begin{aligned}
p(x_1, x_2, x_3, x_4) &= \sum_{x_5} \sum_{x_6} p(x_1, x_2, x_3, x_4, x_5, x_6) && (2.15)\\
&= \sum_{x_5} \sum_{x_6} p(x_1) p(x_2 \,|\, x_1) p(x_3 \,|\, x_1) p(x_4 \,|\, x_2) p(x_5 \,|\, x_3) p(x_6 \,|\, x_2, x_5) && (2.16)\\
&= p(x_1) p(x_2 \,|\, x_1) p(x_3 \,|\, x_1) p(x_4 \,|\, x_2) \sum_{x_5} p(x_5 \,|\, x_3) \sum_{x_6} p(x_6 \,|\, x_2, x_5) && (2.17)\\
&= p(x_1) p(x_2 \,|\, x_1) p(x_3 \,|\, x_1) p(x_4 \,|\, x_2), && (2.18)
\end{aligned}$$

and also compute the marginal probability of $\{X_1, X_2, X_3\}$:

$$\begin{aligned}
p(x_1, x_2, x_3) &= \sum_{x_4} p(x_1) p(x_2 \,|\, x_1) p(x_3 \,|\, x_1) p(x_4 \,|\, x_2) && (2.19)\\
&= p(x_1) p(x_2 \,|\, x_1) p(x_3 \,|\, x_1). && (2.20)
\end{aligned}$$

Dividing these two marginals yields the desired conditional:

$$p(x_4 \,|\, x_1, x_2, x_3) = p(x_4 \,|\, x_2), \tag{2.21}$$

which demonstrates the conditional independence relationship $X_4 \perp\!\!\!\perp \{X_1, X_3\} \,|\, X_2$.

We can readily verify the other conditional independencies in Eq. (2.14), and indeed it is not hard to follow along the lines of the example to prove in general that the conditional independence statements in Eq. (2.8) follow from the definition of joint probability in Eq. (2.2). Thus we are licensed to interpret the missing edges in the graph in terms of a basic set of conditional independencies.

More interestingly, we might ask whether there are other conditional independence statements that are true of such joint probability distributions, and whether these statements also have a graphical interpretation.

For example, for the graph in Figure 2.1 it turns out that X_1 is independent of X_6 given $\{X_2, X_3\}$. This is not one of the basic conditional independencies in the list in Eq. (2.14), but it is *implied* by that list. We can verify this conditional independence by algebra. In general, however, such algebraic calculations can be tedious and it would be appealing to find a simpler method for checking conditional independencies. Moreover, we might wish to write down *all* of the conditional independencies that are implied by our basic set. Is there any way to do this other than by trying to factorize the joint distribution with respect to all possible triples of subsets of the variables?

A solution to the problem is suggested by examining the graph in Figure 2.3. We see that the nodes X_2 and X_3 *separate* X_1 from X_6, in the sense that all paths between X_1 and X_6 pass through X_2 or X_3. Moreover, returning to the list of basic conditional independencies in Eq. (2.14), we see that the parents X_{π_i} block all paths from the node X_i to the earlier nodes in a topological ordering. This suggests that the notion of *graph separation* can be used to derive a graphical algorithm for inferring conditional independence.

We will have to take some care, however, to make the notion of "blocking" precise. For example, X_2 is *not* necessarily independent of X_3 given X_1 and X_6, as would be suggested by a naive interpretation of "blocking" in terms of graph separation.

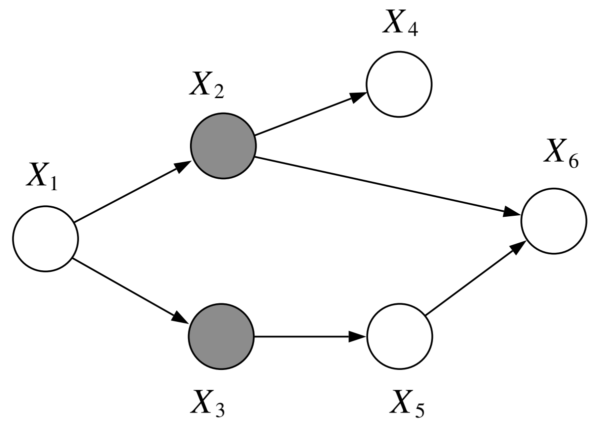


Figure 2.3: The nodes X_2 and X_3 separate X_1 from X_6.

We will pursue the analysis of blocking and conditional independence in the following section, where we provide a general graph search algorithm to solve the problem of finding implied independencies.

Let us make a final remark on the definition of the set of basic conditional independence statements in Eq. (2.8). Note that this set is dependent on both the graph $\mathcal{G}$ and on an ordering I. It is also possible to make an equivalent definition that is defined only in terms of the graph $\mathcal{G}$. In particular, recall that the set ν_i necessarily includes all ancestors of i (excluding the parents π_i). Note that the set of ancestors is independent of the ordering I. We thus might consider defining a basic set of independence statements that assert the conditional independence of a node from its ancestors, conditional on its parents. It turns out that the independence statements in this set hold if and only if the independence statements in Eq. (2.8) hold. As we ask the reader to verify in Exercise **??**, this equivalence follows easily from the "Bayes ball" algorithm that we present in the following section.

The definition in Eq. (2.8) was chosen so as to be able to contrast the definition of the joint probability in Eq. (2.2) with the general chain rule in Eq. (2.7). An order-independent definition of the basic set of conditional independencies is, however, an arguably more elegant characterization of conditional independence in a graph, and it will take center stage in our more formal treatment of conditional independence and Markov properties in Chapter 16.

2.1.2 Conditional independence and the Bayes ball algorithm

The algorithm that we describe is called the *Bayes ball algorithm*, and it has the colorful interpretation of a ball bouncing around a graph. In essence it is a "reachability" algorithm, under a particular definition of "separation."

Our approach will be to first discuss the conditional independence properties of three canonical, three-node graphs. We then embed these properties in a protocol for the bouncing ball; these are the local rules for a graph-search algorithm.

Two final remarks before we describe the algorithm. In our earlier discussion in Section 2.1.1,

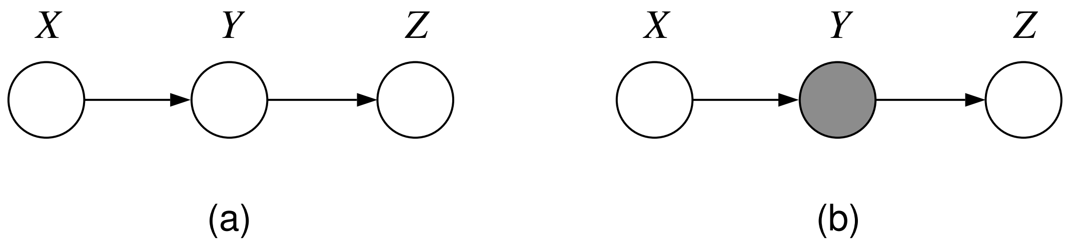


Figure 2.4: (a) The missing edge in this graph corresponds to the conditional independence statement $X \perp\!\!\!\perp Z \mid Y$. As suggested in (b), conditioning on Y has the graphical interpretation of blocking the path between X and Z.

and also in the current section, we presented conditional independence as being subservient to the basic definition in Eq. (2.2) of the joint probability. That is, we justified an assertion of conditional independence by factorizing Eq. (2.2) or one of its marginals. This is not the only point of view that we can take, however. Indeed it turns out that this relationship can be reversed, with Eq. (2.2) being derived from a characterization of conditional independence, and we will also introduce this point of view in this section. By the end of the current section we hope to have clarified what is meant by a "characterization of conditional independence."

On a related note, let us recall a remark that was made earlier, which is that to each graph we associate a *family of joint probability distributions*. In terms of the definition of joint probability in Eq. (2.2), this family arises as we range over different choices of the numerical values of the local conditional probabilities $p(x_i \mid x_{\pi_i})$. Our work in the current section can be viewed as providing an alternative, more qualitative, characterization of a family of probability distributions associated to a given graph. In particular we can view the conditional independence statements generated by the Bayes ball algorithm as generating a list of constraints on probability distributions. Those joint probabilities that meet all of the constraints in this list are in the family, and those that fail to meet one or more constraints are out. It is then an interesting question as to the relationship between this characterization of a family of probability distributions in terms of conditional independence and the more numerical characterization of a family in terms of local conditional probabilities. This is the topic of Section 2.1.3.

Three canonical graphs

As we discussed in Section 2.1.1, the missing edges in a directed graphical model can be interpreted in terms of conditional independence. In this section, we flesh out this interpretation for three simple graphs.

Consider first the graph shown in Figure 2.4, in which X, Y, and Z are connected in a chain. There is a missing edge between X and Z, and we interpret this missing edge to mean that X and Z are conditionally independent given Y; thus:

$$X \perp\!\!\!\perp Z \mid Y. \tag{2.22}$$

Moreover, we assert that there are no other conditional independencies associated with this graph.

Let us justify the first assertion, showing that $X \perp\!\!\!\perp Z \mid Y$ can be derived from the assumed form of the joint distribution for directed models Eq. (2.2). We have:

$$p(x, y, z) = p(x)p(y \mid x)p(z \mid y), \tag{2.23}$$

which implies:

$$\begin{aligned} p(z \mid x, y) &= \frac{p(x, y, z)}{p(x, y)} && (2.24) \\ &= \frac{p(x)p(y \mid x)p(z \mid y)}{p(x)p(y \mid x)} && (2.25) \\ &= p(z \mid y), && (2.26) \end{aligned}$$

which establishes the independence.

The second assertion needs some explanation. What do we mean when we say that "there are no other conditional independencies associated with this graph"? It is important to understand that this does *not* mean that no further conditional independencies can arise in any of the distributions in the family associated with this graph (that is, distributions that have the factorized form in Eq. (2.23)). There are certainly some distributions which exhibit additional independencies. For example, we are free to choose any local conditional probability $p(y \mid x)$; suppose that we choose a distribution in which the probability of y happens to be the same no matter the value of x. We can readily verify that with this particular choice of $p(y \mid x)$, we obtain $X \perp\!\!\!\perp Y$.

The key point, then, is that Figure 2.4 does not assert that X and Y are necessarily dependent (i.e., not independent). That is, edges that are present in a graph do not necessarily imply dependence (whereas edges that are missing do necessarily imply independence). But the "lack of independence" does have a specific interpretation: the general theory that we present in Chapter 16 will imply that if a statement of independence is not made, then there exists at least one distribution for which that independence relation does not hold. For example, it is easy to find distributions that factorize as in Eq. (2.23) and in which X is not independent of Y.

In essence, the issue comes down to a difference between universally quantified statements and existentially quantified statements, with respect to the family of distributions associated with a given graph. Asserted conditional independencies *always* hold for these distributions. Non-asserted conditional independencies *sometimes* fail to hold for the distributions associated with a given graph, but sometimes they do hold. This of course has important consequences for algorithm design. In particular, if we build an algorithm that is based on conditional independencies, the algorithm will be correct for *all* of the distributions associated with the graph. An algorithm based on the absence of conditional independencies will *sometimes* be correct, sometimes not.

For an intuitive interpretation of the graph in Figure 2.4, let X be the "past," Y be the "present," and Z be the "future." Thus our conditional independence statement $X \perp\!\!\!\perp Z \mid Y$ translates into the statement that the past is independent of the future given the present, and we can interpret the graph as a simple classical Markov chain.

Our second canonical graph is shown in Figure 2.5. We associate to this graph the conditional independence statement:

$$X \perp\!\!\!\perp Z \mid Y, \tag{2.27}$$

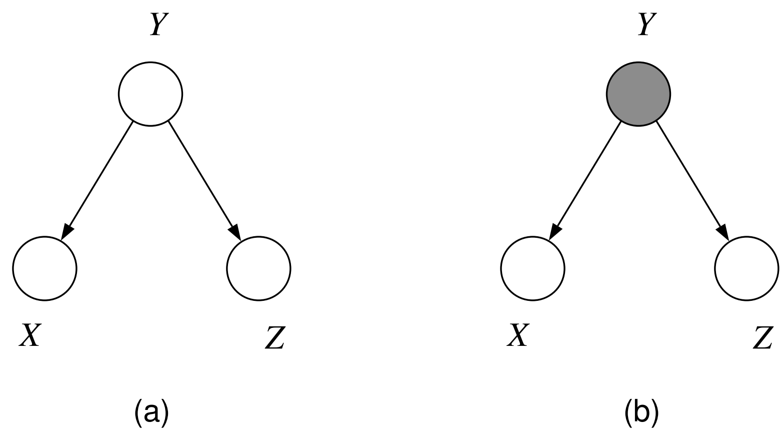


Figure 2.5: (a) The missing edge in this graph corresponds to the conditional independence statement $X \perp\!\!\!\perp Z \mid Y$. As suggested in (b), conditioning on Y has the graphical interpretation of blocking the path between X and Z.

and once again we assert that no other conditional independencies associated with this graph.

A justification of the conditional independence statement follows from the factorization rule. Thus:

$$p(x, y, z) = p(y)p(x \mid y)p(z \mid y) \tag{2.28}$$

implies:

$$\begin{aligned} p(x, z \mid y) &= \frac{p(y)p(x \mid y)p(z \mid y)}{p(y)} && (2.29) \\ &= p(x \mid y)p(z \mid y), && (2.30) \end{aligned}$$

which means that X and Z are independent given Y.

An intuitive interpretation for this graph can be given in terms of a "hidden variable" scenario. Let X be the variable "shoe size," and let Z be the variable "amount of gray hair." In the general population, these variables are strongly dependent, because children tend to have small feet and no gray hair. But if we let Y be "chronological age," then we might be willing to assert that $X \perp\!\!\!\perp Z \mid Y$; that is, given the age of a person, there is no further relationship between the size of their feet and the amount of gray hair that they have. The hidden variable Y "explains" all of the observed dependence between X and Z.

Note once again we are making no assertions of dependence based on Figure 2.5. In particular, we do not necessarily assume that X and Z are dependent because they both "depend" on the variable Y. (But we can assert that there are at least some distributions in which such dependencies are to be found).

Finally, the most interesting canonical graph is that shown in Figure 2.6. Here the conditional independence statement that we associate with the graph is actually a statement of marginal independence:

$$X \perp\!\!\!\perp Z, \tag{2.31}$$

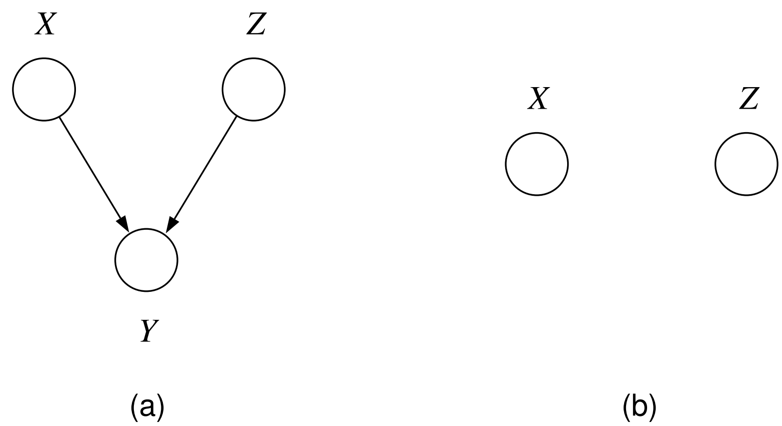


Figure 2.6: (a) The missing edge in this graph corresponds to the marginal independence statement $X \perp\!\!\!\perp Z$. As shown in (b), this is a statement about the subgraph defined on X and Z. Note moreover that conditioning on Y does not render X and Z independent, as would be expected from a naive characterization of conditional independence in terms of graph separation.

which we leave to the reader to verify in terms of the form of the joint probability. Once again, we assert that no other conditional independencies hold. In particular, note that we do not assert any conditional independence involving all three of the variables.

The fact that we do not assert that X is independent of Z given Y in Figure 2.6 is an important fact that is worthy of some discussion. Based on our earlier discussion, we should expect to be able to find scenarios in which a variable X is independent of another variable Z, given no other information, but once a third variable Y is observed these variables become dependent. Indeed, such a scenario is provided by a "multiple, competing explanation" interpretation of Figure 2.6.

Suppose that Bob is waiting for Alice for their noontime lunch date, and let $\{\textbf{late} = \text{"yes"}\}$ be the event that Alice does not arrive on time. One explanation of this event is that Alice has been abducted by aliens, which we encode as $\{\textbf{aliens} = \text{"yes"}\}$ (see Figure 2.7). Bob uses Bayes' theorem to calculate the probability $P(\textbf{aliens} = \text{"yes"} \mid \textbf{late} = \text{"yes"})$ and is dismayed to find that it is larger than the base rate $P(\textbf{aliens} = \text{"yes"})$. Alice has perhaps been abducted by aliens. Now let $\{\textbf{watch} = \text{"no"}\}$ denote the event that Bob forgot to set his watch to reflect daylight savings time. Bob now calculates $P(\textbf{aliens} = \text{"yes"} \mid \textbf{late} = \text{"yes"}, \textbf{watch} = \text{"no"})$ and is relieved to find that the probability of $\{\textbf{aliens} = \text{"yes"}\}$ has gone down again. The key point is that $P(\textbf{aliens} = \text{"yes"} \mid \textbf{late} = \text{"yes"}) \neq P(\textbf{aliens} = \text{"yes"} \mid \textbf{late} = \text{"yes"}, \textbf{watch} = \text{"no"})$, and thus **aliens** is not independent of **watch** given **late**.

On the other hand, it is reasonable to assume that **aliens** is *marginally* independent of **watch**; that is, Bob's watch-setting behavior and Alice's experiences with aliens are presumably unrelated and we would evaluate their probabilities independently, outside of the context of the missed lunch date.

This kind of scenario is known as "explaining-away" and it is commonplace in real-life situations. Moreover, there are other such scenarios (e.g., those involving multiple, synergistic explanations)

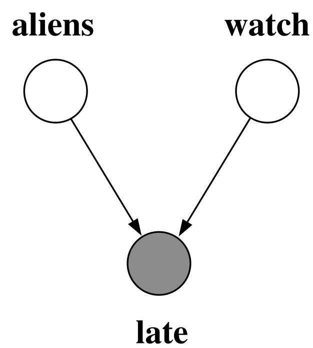


Figure 2.7: A graph representing the fact that Alice is late for lunch with Bob, with two possible explanations—that she has been abducted by aliens and that Bob has forgotten to set his watch to reflect daylight savings time.

in which variables that are marginally independent become dependent when a third variable is observed. We clearly do not want to assume in general that X is independent of Z given Y in Figure 2.6.

Graph separation

We would like to forge a general link between graph separation and assertions of conditional independence. Doing so would allow us to use a graph-search algorithm to answer queries regarding conditional independence.

Happily, the graphs in Figure 2.4 and Figure 2.5 exhibit situations in which naive graph separation corresponds directly to conditional independence. Thus, as shown in Figure 2.4(b), shading the Y node blocks the path from X to Z, and this can be interpreted in terms of the conditional independence of X and Z given Y. Similarly, in Figure 2.5(b), the shaded Y node blocks the path from X to Z, and this can be interpreted in terms of the conditional independence of X and Z given Y.

On the other hand, the graph in Figure 2.6 involves a case in which naive graph separation and conditional independence are opposed. It is when the node Y is unshaded that X and Z are independent; when Y is shaded they become dependent. If we are going to use graph-theoretic ideas to answer queries about conditional independence, we need to pay particular attention to this case.

The solution is straightforward. Rather than relying on "naive" separation, we define a new notion of separation that is more appropriate to our purposes. This notion is known as *d-separation*, for "directed separation." We provide a formal discussion of d-separation in Chapter 16; in the current chapter we provide a simple operational definition of d-separation in terms of the Bayes ball algorithm.

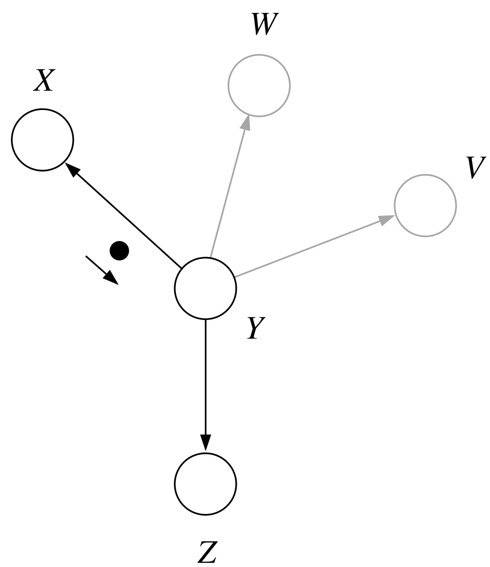


Figure 2.8: We develop a set of rules to specify what happens when a ball arrives from a node X at a node Y, en route to a node Z.

The Bayes ball algorithm

The problem that we wish to solve is to decide whether a given conditional independence statement, $X_A \perp\!\!\!\perp X_B \mid X_C$, is true for a directed graph $\mathcal{G}$. Formally this means that the statement holds for every distribution that factors according to $\mathcal{G}$, but let us not worry about formal issues for now, and let our intuition—aided by the three canonical graphs that we have already studied—help us to define an algorithm to decide the question.

The algorithm is a "reachability" algorithm: we shade the nodes X_C, place a ball at each of the nodes X_A, let the balls bounce around the graph according to a set of rules, and ask whether any of the balls reach one of the nodes in X_B. If none of the balls reach X_B, then we assert that $X_A \perp\!\!\!\perp X_B \mid X_C$ is true. If a ball reaches X_B then we assert that $X_A \perp\!\!\!\perp X_B \mid X_C$ is not true.

The basic problem is to specify what happens when a ball arrives at a node Y from a node X, en route to a node Z (see Figure 2.8). Note that we focus on a particular candidate destination node Z, ignoring the other neighbors that Y may have. (We will be trying all possible neighbors, but we focus on one at a time). Note also that the balls are allowed to travel in either direction along the edges of the graph.

We specify these rules by making reference to our three canonical graphs. In particular, referring to Figure 2.4, suppose that ball arrives at Y from X along an arrow oriented from X to Y, and we are considering whether to allow the ball to proceed to Z along an arrow oriented from Y to Z. Clearly, if the node Y is shaded, we do not want the ball to be able to reach Z, because $X \perp\!\!\!\perp Z \mid Y$ for this graph. Thus we require the ball to be "blocked" in this case. Similarly, if a ball arrives at Y from Z, we do not allow the ball to proceed to X; again the ball is blocked. We summarize these rules with the diagram in Figure 2.9(a).

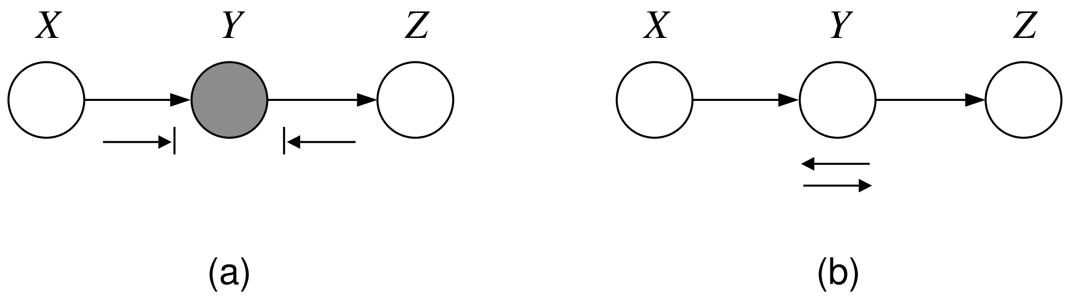


Figure 2.9: The rules for the case of one incoming arrow and one outgoing arrow. (a) When the middle node is shaded, the ball is blocked. (b) When the middle node is unshaded, the ball passes through.

On the other hand, if Y is not shaded, then we want to allow the ball to reach Z from X (and similarly X from Z), because we do not want to assert conditional independence in this case. Thus we have the diagram in Figure 2.9(b), which shows the ball "passing through" when Y is not shaded.

Similar considerations apply to the graph in Figure 2.5, where the arrows are oriented outward from the node Y. Once again, if Y is shaded we do not want the ball to pass between X and Z, thus we require it to be blocked at Y. On the other hand, if Y is unshaded we allow the ball to pass through. These rules are summarized in Figure 2.10.

Finally, we consider the graph in Figure 2.6 in which both of the arrows are oriented towards node Y (this is often referred to as a "v-structure"). Here we simply reverse the rules. Thus, if Y is not shaded we require the ball to be blocked, reflecting the fact that X and Z are marginally independent. On the other hand, if Y is shaded we allow the ball to pass through, reflecting the fact that we do not assert that X and Z are conditionally independent given Y. The rules for this graph are given in Figure 2.11.

We also intend for these rules to apply to the case in which the source node and the destination node (X and Z, respectively) are the same. That is, when a ball arrives at a node, we consider each possible outgoing edge in turn, including the edge the ball arrives on.

Consider first the case in which the ball arrives along an edge that is oriented from X to Y. In this case, the situation is effectively one in which a ball arrives on the head of an arrow and departs on the head of an arrow. This situation is covered by Figure 2.11. We see that the ball should be blocked if the node is unshaded and should "pass through" if the node is shaded, a result that is summarized in Figure 2.12. Note that the action of "passing through" is better described in this case as "bouncing back."

The remaining situation is the one in which the ball arrives along an edge that is oriented from Y to X. The ball arrives on the tail of an arrow and departs on the tail of an arrow, a situation which is covered by Figure 2.10. We see that the ball should be blocked if the node is shaded and should bounce back if the node is unshaded, a result that is summarized in Figure 2.13.

Let us consider some examples. Figure 2.14 shows a chain-structured graphical model (a Markov chain) on a set of nodes $\{X_1, X_2, \ldots, X_n\}$. The basic conditional independencies for this graph (cf.

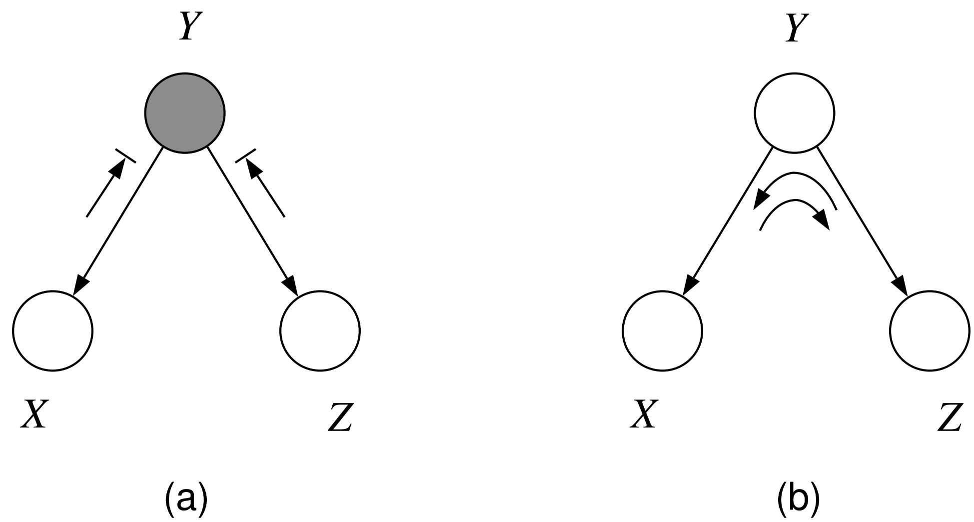


Figure 2.10: The rules for the case of two outgoing arrows. (a) When the middle node is shaded, the ball is blocked. (b) When the middle node is unshaded, the ball passes through.

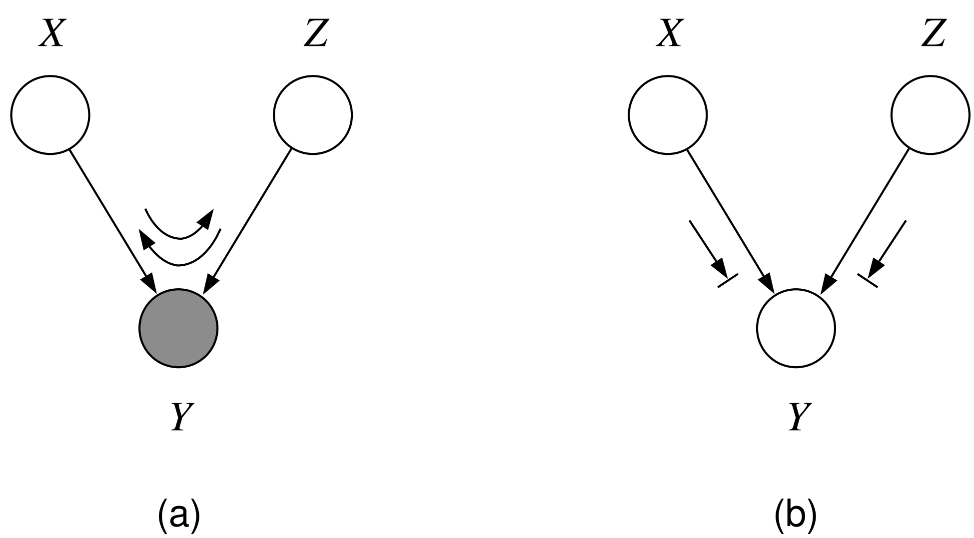


Figure 2.11: The rules for the case of two outgoing arrows. (a) When the middle node is shaded, the ball passes through. (b) When the middle node is unshaded, the ball is blocked.

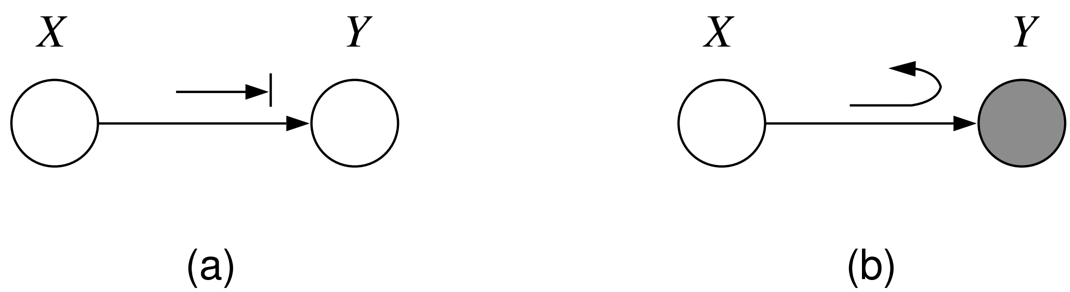


Figure 2.12: The rules for this case follow from the rules in Figure 2.11. (a) When the ball arrives at an unshaded node, the ball is blocked. (b) When the ball arrives at a shaded node, the ball "passes through," which effectively means that it bounces back.

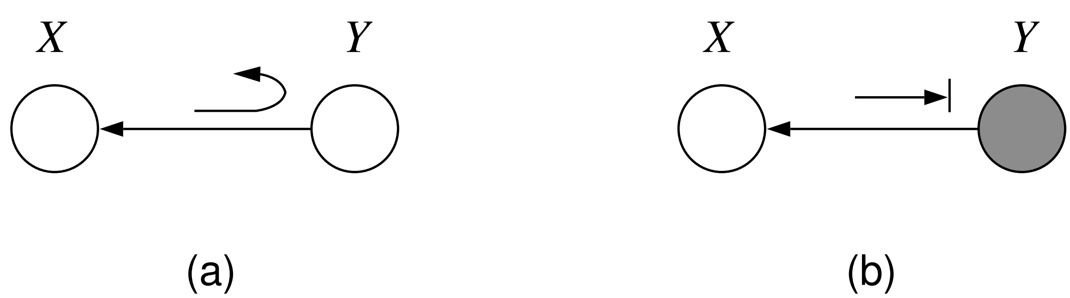


Figure 2.13: The rules for this case follow from the rules in Figure 2.10. (a) When the ball arrives at an unshaded node, the ball "passes through," which effectively means that it bounces back. (b) When the ball arrives at a shaded node, the ball is blocked.

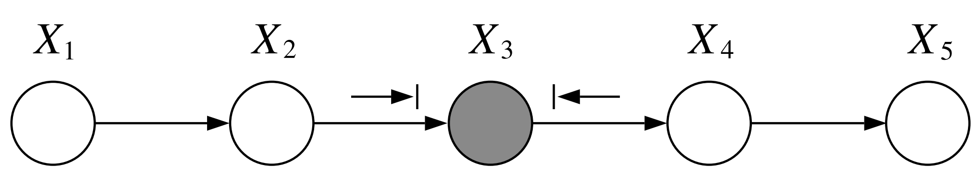


Figure 2.14: The separation of X_3 from X_1, given its parent, X_2, is a basic independence statement for this graph. But conditioning on X_3 also separates any subset of X_1, X_2 from any subset of X_4, X_5, and all of these separations also correspond to conditional independencies.

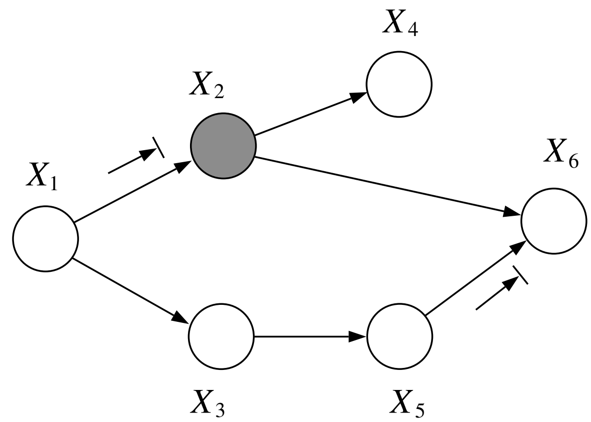


Figure 2.15: A ball arriving at X_2 from X_1 is blocked from continuing on to X_4. Also, a ball arriving at X_6 from X_5 is blocked from continuing on to X_2.

Eq. (2.8)) are the conditional independencies:

$$X_{i+1} \perp\!\!\!\perp \{X_1, X_2, \ldots, X_{i-1}\} \mid X_i. \tag{2.32}$$

There are, however, many other conditional independencies that are implied by this basic set, such as:

$$X_1 \perp\!\!\!\perp X_5 \mid X_4, \qquad X_1 \perp\!\!\!\perp X_5 \mid X_2, \qquad X_1 \perp\!\!\!\perp X_5 \mid \{X_2, X_4\}, \tag{2.33}$$

each of which can be established from algebraic manipulations starting from the definition of the joint probability. Indeed, in general we can obtain the conditional independence of any subset of "future" nodes from any subset of "past" nodes given any subset of nodes that separates these subsets. This is clearly the set of conditional independence statements picked out by the Bayes ball algorithm; the ball is blocked when it arrives at X_3 from either the left or the right.

Consider the graph in Figure 2.1 and consider the conditional independence $X_4 \perp\!\!\!\perp \{X_1, X_3\} \mid X_2$ which we demonstrated to hold for this graph (this is one of the basic set of conditional independencies for this graph; recall Eqs. 2.9 through eq:example-set-of-basic-CI). Using the Bayes ball approach, let us consider whether it is possible for a ball to arrive at node X_4 from either node X_1 or node X_3, given that X_2 is shaded (see Figure 2.15). To arrive at X_4, the ball must pass through X_2. One possibility is to arrive at X_2 from X_1, but the path through to X_4 is blocked because of Figure 2.9(a). The other possibility is to arrive at X_2 via X_6. However, any ball arriving at X_6 must do so via X_5, and such a ball is blocked at X_6 because of Figure 2.11(b).

Note that balls can also bounce back at X_2 and X_6, but this provides no help with respect to arriving at X_4.

We claimed in Section 2.1.1 that $X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$, a conditional independence that is not in the basic set. Consider a ball starting at X_1 and traveling to X_3 (see Figure 2.16). Such a ball cannot pass through to X_5 because of Figure 2.9(a). Similarly, a ball cannot pass from X_1 through X_2 (to either X_4 or X_6) because of Figure 2.9(a).

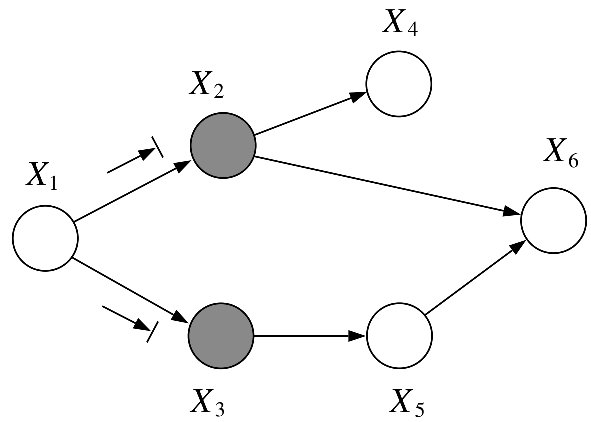


Figure 2.16: A ball cannot pass through X_2 to X_6 nor through X_3.

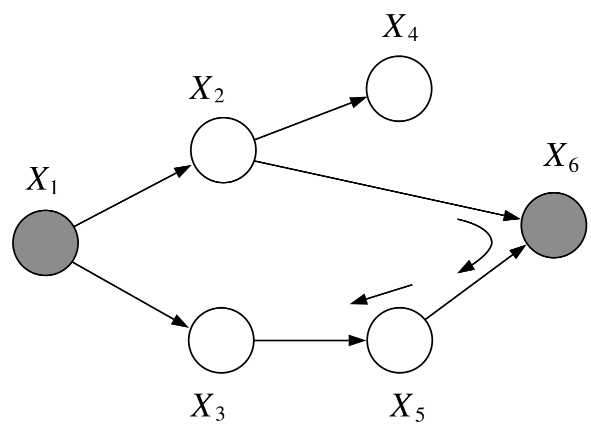


Figure 2.17: A ball can pass from X_2 through X_6 to X_5, and thence to X_3.

We also claimed in Section 2.1.1 that it is not the case that $X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$. To establish this claim we note that a ball can pass through X_2 to X_6 because of Figure 2.9(b), and (see Figure 2.17) can then pass from through X_6 to X_5, on the basis of Figure 2.11(a). The ball then passes through X_5 and arrives at X_3. Intuitively (and loosely), the observation of X_6 implies the possibility of an "explaining-away" dependency between X_2 and X_5. Clearly X_5 and X_3 are dependent, and thus X_2 and X_3 are dependent.

Finally, consider again the scenario with Alice and Bob, and suppose that Bob does not actually observe that Alice fails to show at the hour that he expects her. Suppose instead that Bob is an important executive and there is a security guard for Bob's building who reports to Bob whether a guest has arrived or not. We augment the model to include a node **report** for the security guard's report and, as shown in Figure 2.18, we hang this node off of the node **late**. Now observation of **report** is essentially as good as observation of **late**, particularly if we believe that the security guard is reliable. That is, we should still have **aliens** $\perp\!\!\!\perp$ **watch**, and moreover we should not assert

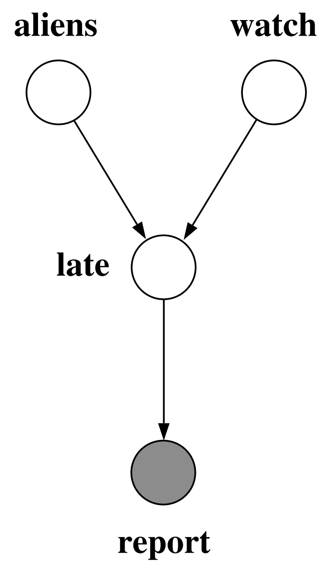


Figure 2.18: An extended graphical model for the Bob-Alice scenario, including a node **report** for the security guard's report.

aliens$\perp\!\!\!\perp$**watch** | **report**. That is, if the security guard reports that Alice has not arrived, then Bob worries about aliens and subsequently has his worries alleviated when he realizes that he has forgotten about daylight savings time.

This pattern is what the Bayes ball algorithm delivers. Consider first the marginal independence **aliens**$\perp\!\!\!\perp$**watch**. As can be verified from Figure 2.19(a), a ball that starts at **aliens** is blocked from passing though **late** directly to **watch**. Moreover, although a ball can pass through **late** to **report**, such a ball dies at **report**. Thus the ball cannot arrive at **watch**.

Consider now the situation when **report** is observed (Figure 2.19(b)). As before a ball that starts at **aliens** is blocked from passing though **late** directly to **watch**; however, a ball can pass through **late** to **report**. At this point Figure 2.12(b) implies implies that the ball bounces back at **report**. The ball can then pass through **late** on the path from **report** to **watch**. Thus we cannot conclude independence of **aliens** and **watch** in the case that **report** is observed.

Some further thought will show that it suffices for any descendant of **late** to be observed in order to enable the explaining-away mechanism and render **aliens** and **watch** dependent.

Remarks

We hope that the reader agrees that the Bayes ball algorithm is a simple, intuitively-appealing algorithm for answering conditional independence queries. Of course, we have not yet provided a fully-specified algorithm, because there are many implementational details to work out, including how to represent multiple balls when X_A and X_B are not singleton sets, how to make sure that the algorithm considers all possible paths in an efficient way, how to make sure that the algorithm doesn't loop, etc. But these details are just that—details—and with a modicum of effort the reader

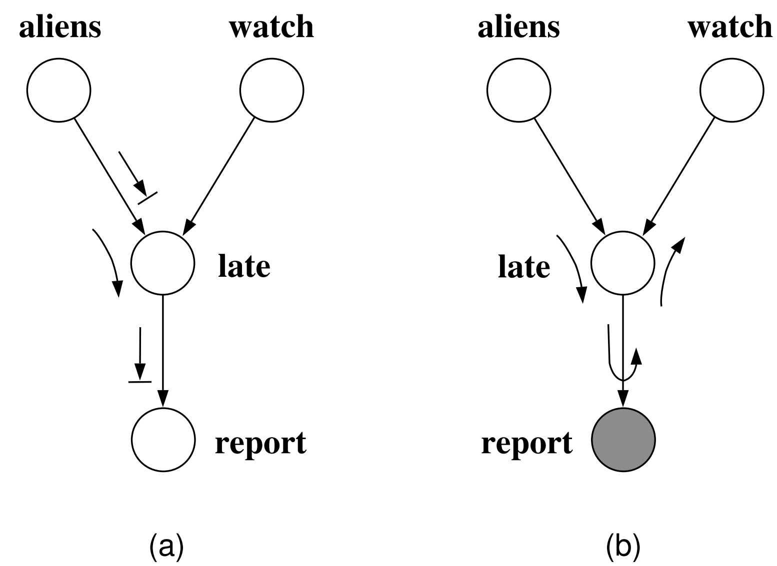


Figure 2.19: (a) A ball cannot pass from **aliens** to **watch** when no observations are made on **late** or **report**. (b) A ball can pass from **aliens** to **watch** when **report** is observed.

can work out such an implementation. Our main interest in the Bayes ball algorithm is to provide a handy tool for quick evaluation of conditional independence queries, and to provide concrete support for the more formal discussion of conditional independence that we undertake in the next section.

2.1.3 Characterization of directed graphical models

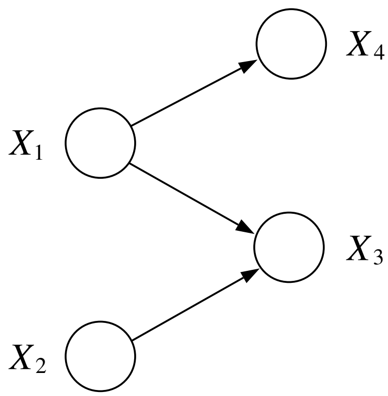
A key idea that has emerged in this chapter is that a graphical model is associated with a *family* of probability distributions. Moreover, as we now discuss, this family can be characterized in two equivalent ways.

Let us define two families and then show that they are equivalent. Actually we defer the proof of equivalence until Chapter 16, but we state the theorem here and discuss its consequences.

The first family is defined via the definition of joint probability for directed graphs, which we repeat here for convenience. Thus for a directed graph $\mathcal{G}$, we have:

$$p(x_1, x_2, \ldots, x_n) \triangleq \prod_{i=1}^{n} p(x_i \,|\, x_{\pi_i}). \tag{2.34}$$

Let us now consider ranging over all possible numerical values for the local conditional probabilities $\{p(x_i \,|\, x_{\pi_i})\}$, imposing only the restriction that these functions are nonnegative and normalized. For discrete variables this would involve ranging over all possible real-valued tables on nodes x_i and their parents. While in practice, we often want to choose simplified parameterizations instead of these tables, for the general theory we must range over all possible tables.



(a)

$$X_1 \perp\!\!\!\perp X_2$$

$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$

$$X_3 \perp\!\!\!\perp X_4 \mid X_1$$

$$X_2 \perp\!\!\!\perp X_4 \mid \{ X_1, X_3 \}$$

$$\{ X_2, X_3 \} \perp\!\!\!\perp X_4 \mid X_1$$

(b)

Figure 2.20: The list in (b) shows all of the conditional independencies that hold for the graph in (a).

For each choice of numerical values for the local conditional probabilities we obtain a particular probability distribution $p(x_1, \ldots, x_n)$. Ranging over all such choices we obtain a family of distributions that we refer to as $\mathcal{D}_1$.

Let us now consider an alternative way to generate a family of probability distributions associated with a graph $\mathcal{G}$. In this approach we will make no use of the numerical parameterization of the joint probability in Eq. (2.34)—this approach will be more "qualitative."

Given a graph $\mathcal{G}$ we can imagine making a list of all of the conditional independence statements that characterize the graph. To do this, imagine running the Bayes ball algorithm for all triples of subsets of nodes in the graph. For any given triple X_A, X_B and X_C, the Bayes ball algorithm tells us whether or not $X_A \perp\!\!\!\perp X_B \mid X_C$ should be included in the list associated with the graph.

For example, Figure 2.20 shows a graph, and all of its associated conditional independence statements. In general such lists can be significantly longer than the list in this example, but they are always finite.

Now consider all possible joint probability distributions $p(x_1, \ldots, x_n)$, where we make no restrictions at all. Thus, for discrete variables, we consider all possible n-dimensional tables. For each such distribution, imagine testing the distribution against the list of conditional independencies associated with the graph $\mathcal{G}$. Thus, for each conditional independence statement in the list, we test whether the distribution factorizes as required. If it does, move to the next statement. If it does not, throw out this distribution and try a new distribution. If a distribution passes all of the tests in the list, we include that distribution in a family that we denote as $\mathcal{D}_2$.

In Chapter 16, we state and prove a theorem that shows that the two families $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same family. This theorem, and an analogous theorem for undirected graphs, provide a strong and important link between graph theory and probability theory and are at the core of the graphical model formalism. They show that the characterizations of probability distributions via numerical parameterization and conditional independence statements are one and the same, and allow us to use these characterizations interchangeably in analyzing models and defining algorithms.

2.2 Undirected graphical models

The world of graphical models divides into two major classes—those based on directed graphs and those based on undirected graphs.[3] In this section we discuss undirected graphical models, also known as *Markov random fields*, and carry out a development that parallels our discussion of the directed case. Thus we will present a factorized parameterization for undirected graphs, a conditional independence semantics, and an algorithm for answering conditional independence queries. There are many similarities to the directed case—and much of our earlier work on directed graphs carries over—but there are interesting and important differences as well.

An undirected graphical model is a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is a set of nodes that are in one-to-one correspondence with a set of random variables, and where $\mathcal{E}$ is a set of undirected edges. The random variables can be scalar-valued or vector-valued, discrete or continuous. Thus we will be concerned with graphical representations of a joint probability distribution, $p(x_1, x_2, \ldots, x_n)$—a mass function in the discrete case and a density function in the continuous case.

2.2.1 Conditional independence

As we saw in Section 2.1.3, there are two equivalent characterizations of the class of joint probability distributions associated with a directed graph. Our presentation of directed graphical models began (in Section 2.1) with the factorized parameterization and subsequently motivated the conditional independence characterization. We could, however, have turned this discussion around and started with a set of conditional independence axioms, subsequently deriving the parameterization. In the case of undirected graphs, indeed, this latter approach is the one that we will take. For undirected graphs, the conditional independence semantics is the more intuitive and straightforward of the two (equivalent) characterizations.

To specify the conditional independence properties of a graph, we must be able to say whether $X_A \perp\!\!\!\perp X_C \mid X_B$ is true for the graph, for arbitrary index subsets A, B, and C. For directed graphs we defined the conditional independence properties operationally, via the Bayes ball algorithm (we provide a corresponding declarative definition in Chapter 16). For undirected graphs we go straight to the declarative definition.

We say that X_A is independent of X_C given X_B if the set of nodes X_B separates the nodes X_A from the nodes X_C, where by "separation" we mean naive graph-theoretic separation (see Figure 2.21). Thus, if every path from a node in X_A to a node in X_C includes at least one node in X_B, then we assert that $X_A \perp\!\!\!\perp X_C \mid X_B$ holds; otherwise we assert that $X_A \perp\!\!\!\perp X_C \mid X_B$ does not hold.

[3]There is also a generalization known as *chain graphs* that subsumes both classes. We will discuss chain graphs in Chapter **??**.

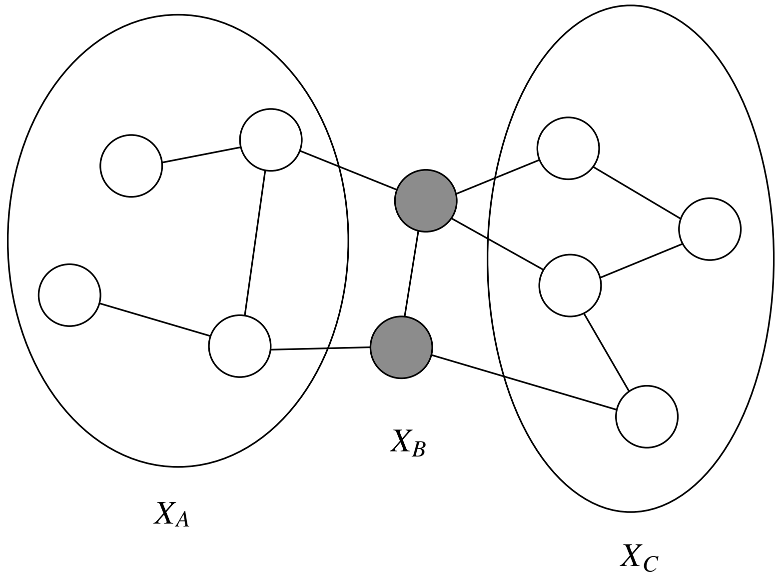


Figure 2.21: The set X_B separates X_A from X_C. All paths from X_A to X_C pass through X_B.

As before, the meaning of the statement "$X_A \perp\!\!\!\perp X_C \mid X_B$ holds for a graph $\mathcal{G}$" is that every member of the family of probability distributions associated with $\mathcal{G}$ exhibits that conditional independence. On the other hand, the statement "$X_A \perp\!\!\!\perp X_C \mid X_B$ does not hold for a graph $\mathcal{G}$" means—in its strong form—that some distributions in the family associated with $\mathcal{G}$ do not exhibit that conditional independence.

Given this definition, it is straightforward to develop an algorithm for answering conditional independence queries for undirected graphs. We simply remove the nodes X_B from the graph and ask whether there are any paths from X_A to X_C. This is a "reachability" problem in graph theory, for which standard search algorithms provide a solution.

Comparative semantics

Is it possible to reduce undirected models to directed models, or vice versa? To see that this is not possible in general, consider Figure 2.22.

In Figure 2.22(a) we have an undirected model that is characterized by the conditional independence statements $X \perp\!\!\!\perp Y \mid \{W, Z\}$ and $W \perp\!\!\!\perp Z \mid \{X, Y\}$. If we try to represent this model in a directed graph on the same four nodes, we find that we must have at least one node in which the arrows are inward-pointing (a "v-structure"). (Recall that our graphs are acyclic). Suppose without loss of generality that this node is Z, and that this is the only v-structure. By the conditional independence semantics of directed graphs, we have $X \perp\!\!\!\perp Y \mid W$, and we do not have $X \perp\!\!\!\perp Y \mid \{W, Z\}$. We are unable to represent both conditional independence statements, $X \perp\!\!\!\perp Y \mid \{W, Z\}$ and $W \perp\!\!\!\perp Z \mid \{X, Y\}$, in the directed formalism.

On the other hand, in Figure 2.22(b) we have a directed graph characterized by the singleton

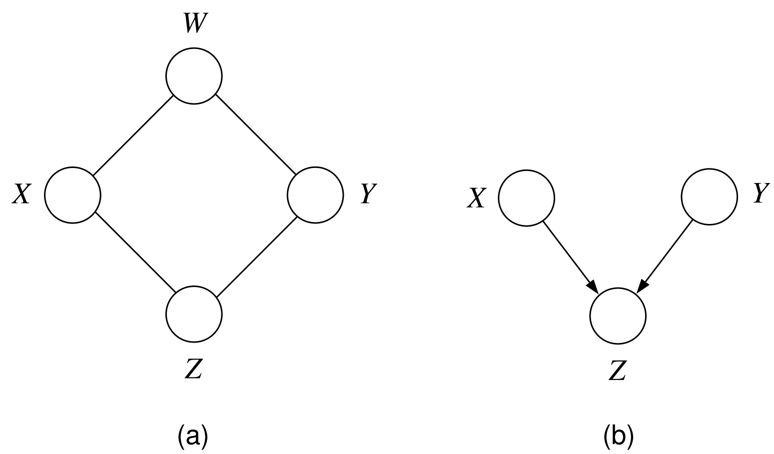


Figure 2.22: (a) An undirected graph whose conditional independence semantics cannot be captured by a directed graph on the same nodes. (b) A directed graph whose conditional independence semantics cannot be captured by an undirected graph on the same nodes.

independence statement $X \perp\!\!\!\perp Y$. No undirected graph on three nodes is characterized by this singleton set. A missing edge in an undirected graph only between X and Y captures $X \perp\!\!\!\perp Y \mid Z$, not $X \perp\!\!\!\perp Y$. An additional missing edge between X and Z captures $X \perp\!\!\!\perp Y$, but implies $X \perp\!\!\!\perp Z$.

We will show in Chapter 16 that there are some families of probability distributions that can be represented with either directed or undirected graphs. There is no good reason to restrict ourselves to these families, however. In general, directed models and undirected models are different modeling tools, and have different strengths and weaknesses. The two together provide modeling power beyond that which could be provided by either alone.

2.2.2 Parameterization

As in the case of directed graphs, we would like to obtain a "local" parameterization for undirected graphical models. For directed graphs the parameterization was based on local conditional probabilities, where "local" had the interpretation of a set $\{i, \pi_i\}$ consisting of a node and its parents. The definition of the joint probability as a product of such local probabilities was motivated via the chain rule of probability theory.

In the undirected case it is rather more difficult to utilize conditional probabilities to represent the joint. One possibility would be to associate to each node the conditional probability of the node given its neighbors. This approach falls prey to a major consistency problem, however—it is hard to ensure that the conditional probabilities at different nodes are consistent with each other and thus with a single joint distribution. We are not able to choose these functions independently and arbitrarily, and this poses problems both in theory and in practice.

A better approach turns out to be to abandon conditional probabilities altogether. By so doing we will lose the ability to give a local probabilistic interpretation to the functions used to represent the joint probability, but we will retain the ability to choose these functions independently and arbitrarily, and we will retain the all-important representation of the joint as a *product* of local functions.

A key problem is to decide the domain of the local functions; in essence, to decide the meaning of "local" for undirected graphs. It is here that the discussion of conditional independence in the previous section is helpful. In particular, consider a pair of nodes X_i and X_j that are not linked in the graph. The conditional independence semantics imply that these two nodes are conditionally independent given all of the other nodes in the graph (because upon removing this latter set there can be no paths from X_i to X_j). Thus it must be possible to obtain a factorization of the joint probability that places x_i and x_j in different factors. This implies that we can have no local function that depends on both x_i and x_j in our representation of the joint. Such a local function, say $\psi(x_i, x_j, x_k)$, would not factorize with respect to x_i and x_j in general—recall that we are assuming that the local functions can be chosen arbitrarily.

Recall that a *clique* of a graph is a fully-connected subset of nodes. Our argument thus far has suggested that the local functions should not be defined on domains of nodes that extend beyond the boundaries of cliques. That is, if X_i and X_j are not directly connected, they do not appear together in any clique, and correspondingly there should be no local function that refers to both nodes. We now consider the flip side of the coin. Should we allow arbitrary functions that are defined on all of the cliques? Indeed, an interpretation of the edges that are present in the graph in terms of "dependence" suggests that we should. We have not defined dependence, but heuristically, dependence is the "absence of independence" in one or more of the distributions associated with a graph. If X_i and X_j are linked, and thus appear together in a clique, we can achieve dependence between them by defining a function on that clique.

The *maximal cliques* of a graph are the cliques that cannot be extended to include additional nodes without losing the property of being fully connected. Given that all cliques are subsets of one or more maximal cliques, we can restrict ourselves to maximal cliques without loss of generality. Thus, if X_1, X_2, and X_3 form a maximal clique, then an arbitrary function $\psi(x_1, x_2, x_3)$ already captures all possible dependencies on these three nodes; we gain no generality by also defining functions on sub-cliques such as $\{X_1, X_2\}$ or $\{X_2, X_3\}$.[4]

In summary, our arguments suggest that the meaning of "local" for undirected graphs should be "maximal clique." More precisely, the conditional independence properties of undirected graphs imply a representation of the joint probability as a product of local functions defined on the maximal cliques of the graph. This argument is in fact correct, and we will establish it rigorously in Chapter 16. Let us proceed to make the definition and explore some of its consequences.

Let C be a set of indices of a maximal clique in an undirected graph G, and let $\mathcal{C}$ be the set of all such C. A *potential function*, $\psi_{X_C}(x_C)$, is a function on the possible realizations x_C of the maximal clique X_C.

Potential functions are assumed to be nonnegative, real-valued functions, but are otherwise arbitrary. This arbitrariness is convenient, indeed necessary, for our general theory to go through,

[4]While there is no need to consider non-maximal cliques in developing the general theory relating conditional independence and factorization—our topic in this section—in practice it is often convenient to work with potentials on non-maximal cliques. This issue will return in Section 2.3 and in later chapters. Let us define joint probabilities in terms of maximal cliques for now, but let us be prepared to relax this definition later.

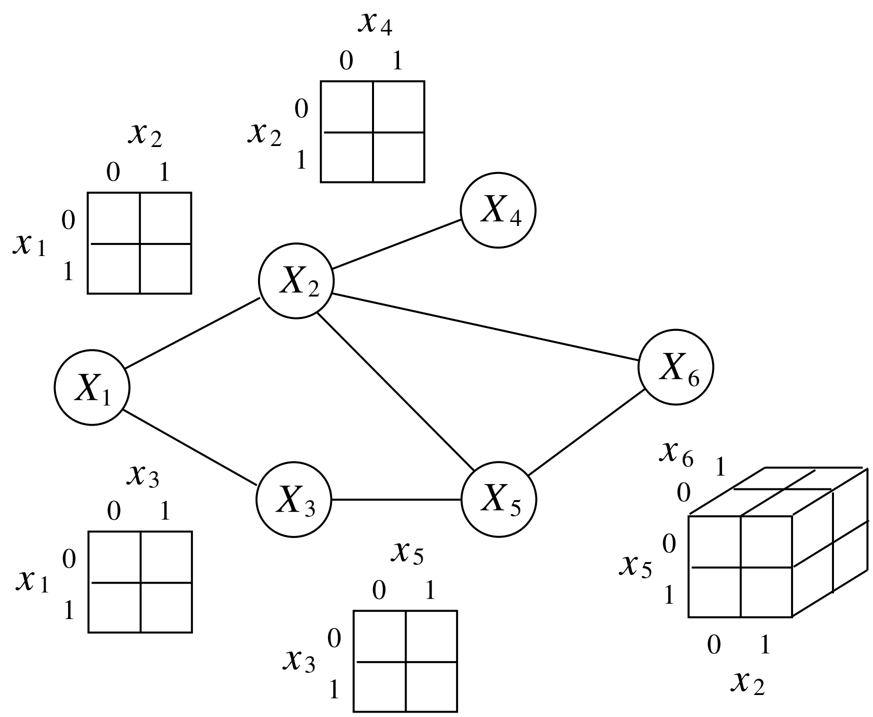


Figure 2.23: The maximal cliques in this graph in are $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_2, X_4\}$, $\{X_3, X_5\}$, and $\{X_2, X_5, X_6\}$. Letting all nodes be binary, we represent a joint distribution on the graph via the potential tables that are displayed.

but it also presents a problem. There is no reason for a product of arbitrary functions to be normalized and thus define a joint probability distribution. This is a bullet which we simply have to bite if we are to achieve the desired properties of arbitrary, independent potentials and a product representation for the joint.

Thus we define:

$$p(x) \triangleq \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C), \tag{2.35}$$

where Z is the normalization factor:

$$Z \triangleq \sum_x \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C), \tag{2.36}$$

obtained by summing the product in Eq. (2.35) over all assignments of values to the nodes X.

An example is shown in Figure 2.23. The nodes in this example are assumed discrete, and thus tables can be used to represent the potential functions. An overall configuration x picks out subvectors x_C, which determine particular cells in each of the potential tables. Taking the product of the numbers in these cells yields an unnormalized representation of the joint probability $p(x)$.

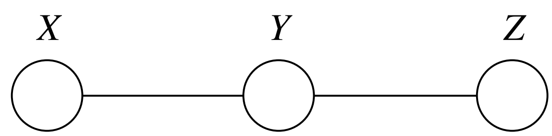


Figure 2.24: An undirected representation of a three-node Markov chain. The conditional independence associated with this graph is $X \perp\!\!\!\perp Z \mid Y$.

The normalization factor Z is obtained by summing over all configurations x. There are an exponential number of such configurations and it is unrealistic to try to perform such a sum by naively enumerating all of the summands. Note, however, that the expression being summed over is a factored expression, in which each factor refers to a local set of variables, and thus we can exploit the distributive law. This is an issue that is best discussed in the context of the more general discussion of probabilistic inference, and we return to it in Chapter 3.

Note, however, that we do not necessarily have to calculate Z. In particular, recall that a conditional probability is a ratio of two marginal probabilities. The factor Z appears in both of the marginal probabilities, and cancels when we take the ratio. Thus we calculate conditionals by summing across unnormalized probabilities—the numerator in Eq. (2.35)—and taking the ratio of these sums.

The interpretation of potential functions

Although local conditional probabilities do not provide a satisfactory approach to the parameterization of undirected models, it might be thought that marginal probabilities could be used instead. Thus, why not replace the potential functions $\psi_{X_C}(x_C)$ in Eq. (2.35) with marginal probabilities $p(x_C)$?

An example will readily show that this approach is infeasible. Consider the model shown in Figure 2.24. The conditional independence that is associated with this graph is $X \perp\!\!\!\perp Z \mid Y$. This independence statement implies (by definition) that the joint must factorize as:

$$p(x, y, z) = p(y)p(x \mid y)p(z \mid y). \tag{2.37}$$

The cliques in Figure 2.24 are $\{X, Y\}$ and $\{Y, Z\}$. We can multiply the first two factors in Eq. (2.37) together to obtain a potential function $p(x, y)$ on the first clique, leaving $p(z \mid y)$ as the potential function on the second clique. Alternatively, we can multiply $p(z \mid y)$ by $p(y)$ to yield a potential $p(y, z)$ on the second clique, leaving $p(x \mid y)$ as the potential on the first clique. Thus we can obtain a factorization in which one of the potentials is a marginal probability, and the other is a conditional probability. But we are unable to obtain a representation in which both potentials are marginal probabilities. That is:

$$p(x, y, z) \neq p(x, y)p(y, z). \tag{2.38}$$

In fact, it is not hard to see that $p(x, y, z) = p(x, y)p(y, z)$ implies $p(y) = 0$ or $p(y) = 1$, and that this representation is thus a rather limited and unnatural one.

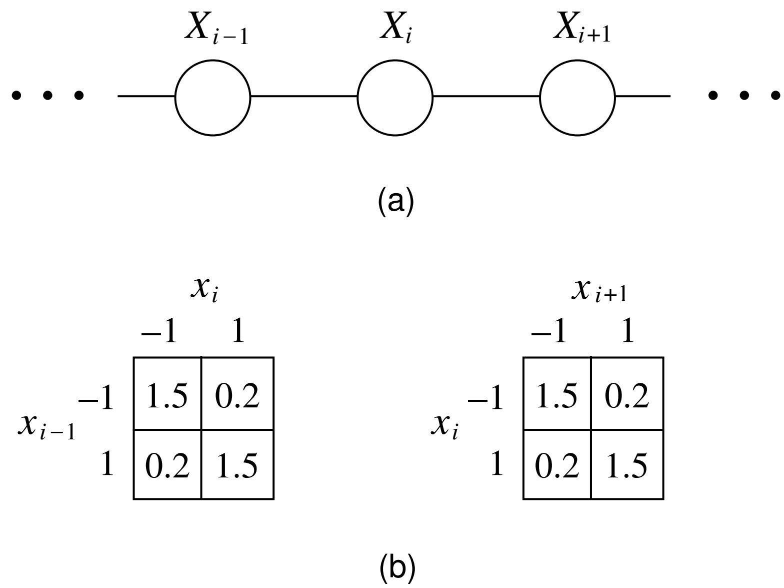


Figure 2.25: (a) A chain of binary random variables X_i, where $X_i \in \{-1, 1\}$. (b) A set of potential tables that encode a preference for neighboring variables to have the same values.

In general, potential functions are neither conditional probabilities nor marginal probabilities, and in this sense they do not have a local probabilistic interpretation. On the other hand, potential functions do often have a natural interpretation in terms of pre-probabilistic notions such as "agreement," "constraint," or "energy," and such interpretations are often useful in choosing an undirected model to represent a real-life domain. The basic idea is that a potential function favors certain local configurations of variables by assigning them a larger value. The global configurations that have high probability are, roughly, those that satisfy as many of the favored local configurations as possible.

Consider a set of binary random variables, $X_i \in \{-1, 1\}, i = 0, \ldots, n$, arrayed on a one-dimensional lattice as shown in Figure 2.25(a). In physics, such lattices are used to model magnetic behavior of crystals, where the binary variables have an interpretation as magnetic "spins." All else being equal, if a given spin X_i is "up"; that is, if $X_i = 1$, then its neighbors X_{i-1} and X_{i+1} are likely to be "up" as well. We can easily encode this in a potential function, as shown in Figure 2.25(b). Thus, if two neighboring spins agree, that is, if $X_i = 1$ and $X_{i-1} = 1$, or if $X_i = -1$ and $X_{i-1} = -1$, we obtain a large value for the potential on the clique $\{X_{i-1}, X_i\}$. If the spins disagree we obtain a small value.

The fact that potentials must be nonnegative can be inconvenient, and it is common to exploit the fact that the exponential function, $f(x) = \exp(x)$, is a nonnegative function, to represent potentials in an unconstrained form. We let:

$$\psi_{X_C}(x_C) = \exp\{-H_C(x_C)\}, \tag{2.39}$$

for a real-valued function $H_C(x_C)$, where the negative sign is a standard convention. Thus if we range over arbitrary $H_C(x_C)$, we can range over legal potentials.

The exponential representation has another useful feature. In particular, products of exponentials behave nicely, and from Eq. (2.35) we obtain:

$$p(x) = \frac{1}{Z}\prod_{C\in\mathcal{C}} \exp\{-H_C(x_C)\} \tag{2.40}$$

$$= \frac{1}{Z}\exp\{-\sum_{C\in\mathcal{C}} H_C(x_C)\} \tag{2.41}$$

as an equivalent representation of the joint probability for undirected models. The sum in the latter expression is generally referred to as the "energy":

$$H(x) \triangleq \sum_{C\in\mathcal{C}} H_C(x_C) \tag{2.42}$$

and we have represented the joint probability of an undirected graphical model as a *Boltzmann distribution*:

$$p(x) = \frac{1}{Z}\exp\{-H(x)\}. \tag{2.43}$$

Without going too far astray into the origins of the Boltzmann representation in statistical physics, let us nonetheless note that the representation of a model in terms of energy, and in particular the representation of the total energy as a sum over local contributions to the energy, is exceedingly useful. Many physical theories are specified in terms of energy, and the Boltzmann distribution provides a translation from energies into probabilities.

Quite apart from any connection to physics, the undirected graphical model formalism is often quite useful in domains in which global constraints on probabilities are naturally decomposable into sets of local constraints, and the undirected representation is apt at capturing such situations.

2.2.3 Characterization of undirected graphical models

In Section 2.1.3 we discussed a theorem that shows that the two different characterizations of the family of probability distributions associated with a directed graphical model—one based on local conditional probabilities and the other based on conditional independence assertions—were the same. A formally identical theorem holds for undirected graphs.

For a given undirected graph $\mathcal{G}$, we define a family of probability distributions, $\mathcal{U}_1$, by ranging over all possible choices of positive potential functions on the maximal cliques of the graph.

We define a second family of probability distributions, $\mathcal{U}_2$, via the conditional independence assertions associated with $\mathcal{G}$. Concretely, we make a list of all of the conditional independence statements, $X_A \perp\!\!\!\perp X_B \mid X_C$, asserted by the graph, by assessing whether the subset of nodes X_A is separated from X_B when the nodes X_C are removed from the graph. A probability distribution is in $\mathcal{U}_2$ if it satisfies all such conditional independence statements, otherwise it is not.

In Chapter 16 we state and prove a theorem, the Hammersley-Clifford theorem, that shows that $\mathcal{U}_1$ and $\mathcal{U}_2$ are identical. Thus the characterization of probability distributions in terms of potentials on cliques and conditional independence are equivalent. As in the directed case, this is an important and profound link between probability theory and graph theory.

2.3 Parameterizations

We have introduced two kinds of graphical model representations in this chapter—directed graphical models and undirected graphical models. In each of these cases we have defined conditional independence semantics and corresponding parameterizations. Thus, in the directed case, we have:

$$p(x) \triangleq \prod_{i=1}^{n} p(x_i \mid x_{\pi_i}), \tag{2.44}$$

and in the undirected case, we have:

$$p(x) \triangleq \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C). \tag{2.45}$$

By ranging over all possible conditional probabilities in Eq. (2.44) or all possible potential functions in Eq. (2.45) we obtain certain families of probability distributions, in particular exactly those distributions which respect the conditional independence statements associated with a given graph.

Conditional independence is an exceedingly useful constraint to impose on a joint probability distribution. In practical settings conditional independence can sometimes be assessed by domain experts, and in such cases it provides a powerful way to embed qualitative knowledge about the relationships among random variables into a model. Moreover, as we will discuss in the following chapter, the relationship between conditional independence and factorization allows the development of powerful general inference algorithms that use graph-theoretic ideas to compute marginal probabilities of interest. We often impose conditional independence as a rough, tentative assumption in a domain so as to be able to exploit the efficient inference algorithms and begin to learn something about the domain.

On the other hand, conditional independence is by no means the only kind of constraint that one can impose on a probabilistic model. Another large class of constraints arise from assumptions about the algebraic structure of the conditional probabilities or potential functions that define a model. In particular, rather than ranging over all possible conditional probabilities or potential functions, we may wish to range over a proper subset of these functions, thus defining a proper subset of the family of probability distributions associated with a graph. Thus, in practice we often work with *reduced parameterizations* that impose constraints on probability distributions beyond the structural constraints imposed by conditional independence.

We will present many examples of reduced parameterizations in later chapters. Let us briefly consider two such examples in the remainder of this section to obtain a basic appreciation of some of the issues that arise.

Directed graphical models require conditional probabilities, and if the number of parents of a given node is large, then the specification of the conditional probability can be problematic. Consider in particular the graph shown in Figure 2.26(a), where all of the variables are assumed binary (for simplicity), and where the number of parents of Y is assumed large. In particular, if Y has 50 parents, then ranging over "all possible conditional probabilities" means specifying 2^{50} numbers, one probability for each configuration of the parents. Clearly such a specification cannot

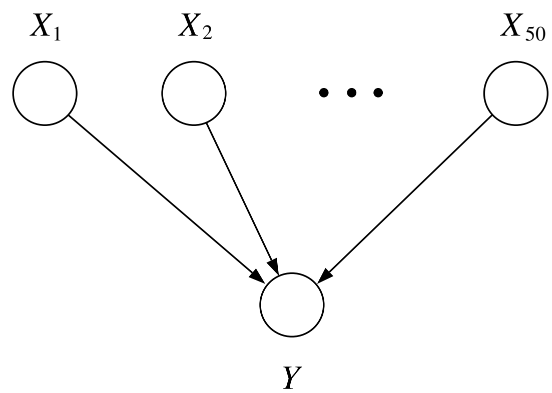


Figure 2.26: An example in which a node has many parents. In such a graph, a general specification of the local conditional probability distribution requires an impractically large number of parameters.

be stored on a computer, and, equally problematically, it would be impossible to collect enough data to be able to estimate these numbers with any degree of precision. We are forced to consider "reduced parameterizations." One such parameterization, discussed in detail in Chapter 8, is the following:

$$p(Y = 1 \mid x) = f(\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m), \tag{2.46}$$

for a given function $f(\cdot)$ whose range is the interval $(0, 1)$ (we will provide examples of such functions in Chapter 8). Here, we need only specify the 50 numbers θ_i to specify a distribution.

In general, we can consider directed graphical models in which each node is parameterized as shown in Eq. (2.46). The family of probability distributions associated with the model as a whole is that obtained by ranging over all possible values of θ_i in the defining conditional probabilities. This is a proper sub-family of the family of distributions associated with the graph.

If practical considerations often force us to work with reduced parameterizations, of what value is the general definition of "the family of distributions associated with a graph"? As we will see in Chapter 4 and Chapter 17, given a graph, efficient probabilistic inference algorithms can be defined that operate on the graph. These algorithms are based solely on the graph structure and are correct for any distribution that respects the conditional independencies encoded by the graph. Thus such algorithms are correct for any distribution in the family of distributions associated with a graph, including those in any proper sub-family associated with a reduced parameterization.

Similar issues arise in undirected models. Consider in particular the graph shown in Figure 2.27(a). From the point of view of independence, there is little to say—there are no independence assertions associated with this graph. Equivalently, the family of probability distributions associated with the graph is the set of all possible probability distributions on the three variables, obtained by ranging over all possible potential functions $\psi(x_1, x_2, x_3)$. Suppose, however, that we are interested in models in which the potential function is defined algebraically as a product of

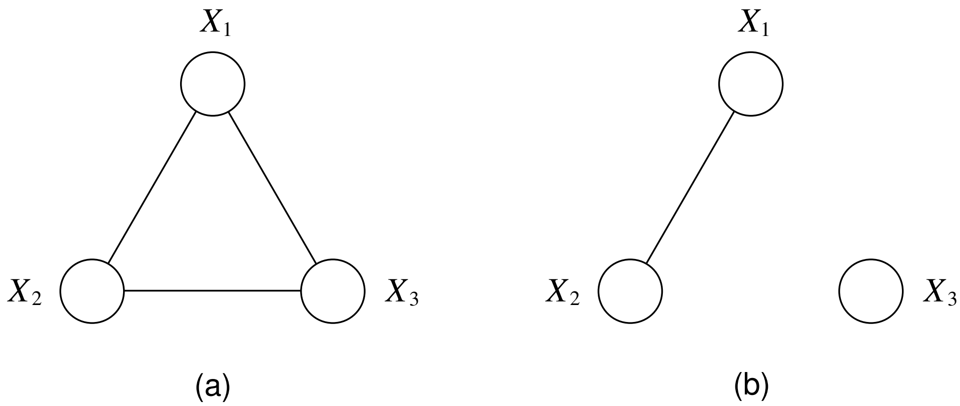


Figure 2.27: (a) An undirected graph which makes no independence assertions. (b) An undirected graph which asserts $X_3 \perp\!\!\!\perp \{X_1, X_2\}$.

factors on smaller subsets of variables. Thus, we might let:

$$\psi(x_1, x_2, x_3) = f(x_1, x_2)g(x_3), \tag{2.47}$$

or let:

$$\psi(x_1, x_2, x_3) = r(x_1, x_2)s(x_2, x_3)t(x_1, x_3), \tag{2.48}$$

for given functions f, g, r, s and t. Ranging over all possible choices of these functions, we obtain potentials that are necessarily members of the family associated with the graph in Figure 2.27(a)—because all such potentials respect the (vacuous) conditional independence requirement. The potential in Eq. (2.47), however, also respects the (non-vacuous) conditional independence requirement of the graph in Figure 2.27(b). We would normally use this latter graph if we decide (a priori) to restrict our parameterization to the form given in Eq. (2.47). On the other hand, the potential given in Eq. (2.48) is problematic in this regard—there is no smaller graph that represents this class of potentials. Any graph with a missing edge makes an independence assertion regarding one or more pairs of variables, and $\psi(x_1, x_2, x_3) = r(x_1, x_2)s(x_2, x_3)t(x_1, x_3)$ does not respect such an assertion, when we range over all functions r, s and t.

Thus we see that "factorization" is a richer concept than "conditional independence." There are families of probability distributions that can be defined in terms of certain factorizations of the joint probability that cannot be captured solely within the undirected or directed graphical model formalism. From the point of view of designing inference algorithms, this might not be viewed as a problem, because an algorithm that is correct for the graph is correct for a distribution in any sub-family defined on the graph. However, by ignoring the algebraic structure of the potential, we may be missing opportunities for simplifying the algebraic operations of inference.

In Chapter 4 we introduce *factor graphs*, a graphical representation of probability distributions in which such reduced parameterizations are made explicit. Factor graphs allow a more fine-grained representation of probability distributions than is provided by either the directed or the undirected graphical formalism, and in particular allow the factorization of the potential in Eq. (2.48) to be represented explicitly in the graph. While factor graphs provide nothing new in terms of representing and exploiting conditional independence relationships—the main theme of the current chapter—they do provide a way to represent and exploit algebraic relationships, an issue that will return in Chapter 4.

2.4 Summary

In this chapter we have presented some of the basic definitions and basic issues that arise when one associates probability distributions with graphs. A key idea that we have emphasized is that a graphical model is a representation of a *family* of probability distributions. This family is characterized in one of two equivalent ways—either in terms of a numerical parameterization or in terms of a set of conditional independencies. Both of these characterizations are important and useful, and it is the interplay between these characterizations that gives the graphical models formalism much of its distinctive flavor.

Directed graphs and undirected graphs have different parameterizations and different conditional independence semantics, but the key concept of using graph theory to capture the notion of a joint probability distribution being constructed from a set of "local" pieces is the same in the two cases.

We have also introduced simple algorithms that help make the problem of understanding conditional independence in graphical models more concrete. The reader should be able to utilize the Bayes ball algorithm to read off conditional independence statements from directed graphs. Similarly, for undirected graphs the reader should understand that naive graph separation encodes conditional independence. Conditional independence assertions in undirected graphs can be assessed via a graph reachability algorithm.

2.5 Historical remarks and bibliography