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 **??**.

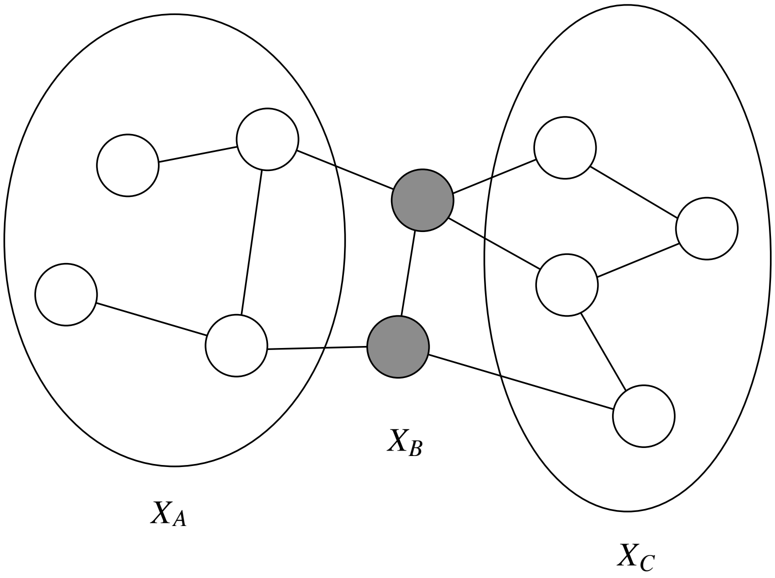


Figure 2.21: The set X_B separates X_A from X_C. All paths from X_A to X_C pass through X_B.

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

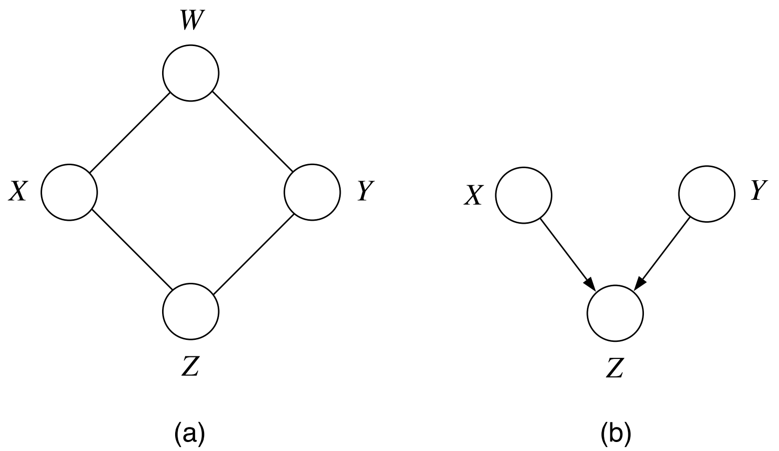


Figure 2.22: (a) An undirected graph whose conditional independence semantics cannot be captured by a directed graph on the same nodes. (b) A directed graph whose conditional independence semantics cannot be captured by an undirected graph on the same nodes.

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.

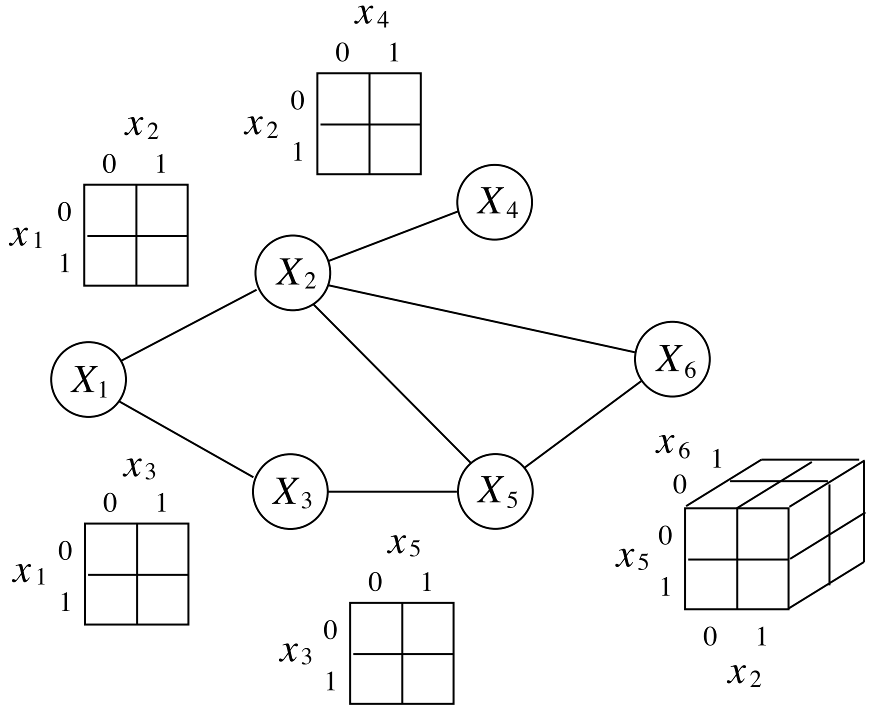


Figure 2.23: The maximal cliques in this graph in are $\{X_1, X_2\}$, $\{X_1, X_3\}$, $\{X_2, X_4\}$, $\{X_3, X_5\}$, and $\{X_2, X_5, X_6\}$. Letting all nodes be binary, we represent a joint distribution on the graph via the potential tables that are displayed.

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 $p(x)$.

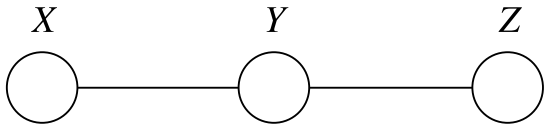


Figure 2.24: An undirected representation of a three-node Markov chain. The conditional independence associated with this graph is $X \perp\!\!\!\perp Z \mid Y$.

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 $p(x_C)$?

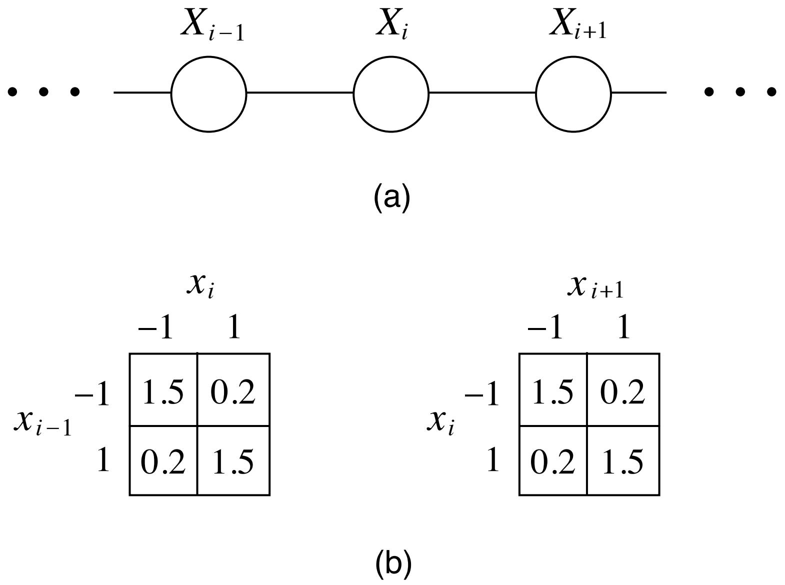
An example will readily show that this approach is infeasible. Consider the model shown in Figure 2.24. The conditional independence that is associated with this graph is $X \perp\!\!\!\perp Z \mid Y$. This independence statement implies (by definition) that the joint must factorize as:

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

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.



(a)

		x_i	
		−1	1
x_{i-1}	−1	1.5	0.2
	1	0.2	1.5

		x_{i+1}	
		−1	1
x_i	−1	1.5	0.2
	1	0.2	1.5

(b)

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:

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

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

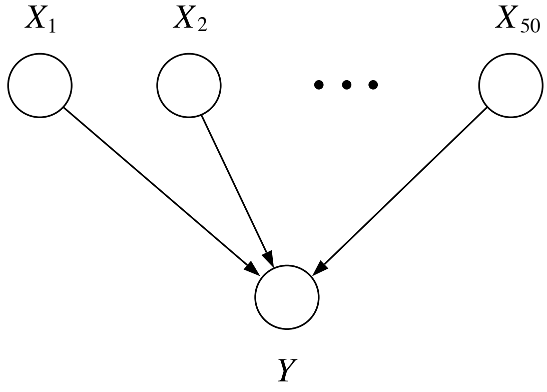


Figure 2.26: An example in which a node has many parents. In such a graph, a general specification of the local conditional probability distribution requires an impractically large number of parameters.

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:

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

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

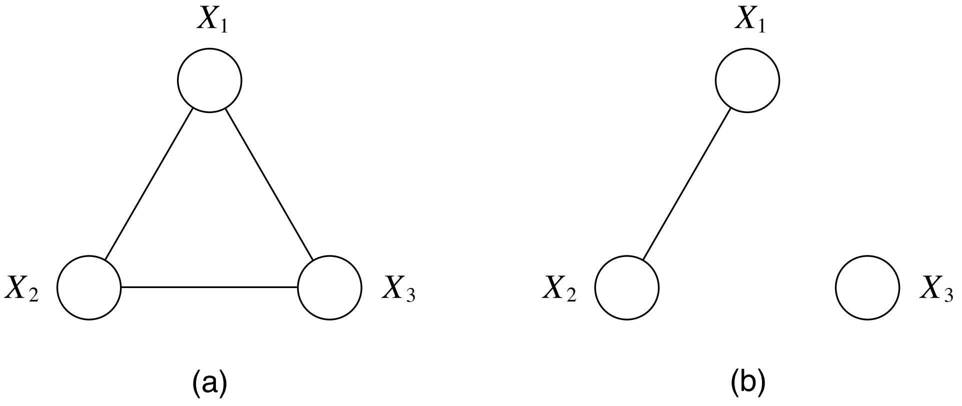


Figure 2.27: (a) An undirected graph which makes no independence assertions. (b) An undirected graph which asserts $X_3 \perp\!\!\!\perp \{X_1, X_2\}$.

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