Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

3 Undirected Graphical Models

As we have developed, directed graphical models are useful for capturing important forms of structure in a family of joint distributions. This structure took the form of a factorization of each constituent distribution into a collection of conditional distributions, which is expressed graphically. As we also observed, this structure could equivalently be expressed in terms of conditional independence constraints.

We now develop a different class of graphical models that very naturally express conditional independence structure in a family of distributions, and correspond to a different kind of factorization of the constituent distributions. These graphical models are referred to as *undirected* graphical models, since the edges will have no orientation. They are also referred to as *Markov random fields (MRF)* in light of the Markov (conditional independence) structure they are aimed at expressing. These models have a rich history, originating in the statistical physics community, and a wealth of applications in many different fields.

3.1 Distributions on Undirected Graphical Models

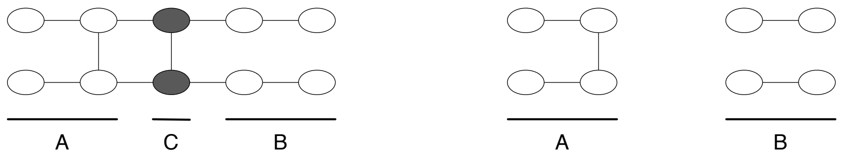
An undirected graphical model is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices $\mathcal{V}$ correspond to random variables and the undirected edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ tell us about the conditional independence structure among the variables.

For an undirected graph, there is a natural notion of *separation*, namely: Given three sets of nodes $A, B, C \subset \mathcal{V}$, A and B are *separated* by C if for any nodes $i \in A$ and $j \in B$, every path (in G) between i and j passes through a node in C. In particular, if $(i, j) \notin E$, then i and j are separated by $\mathcal{V} \setminus \{i, j\}$. The following Markov properties relate statements about graph separation to conditional independence relations.

Definition 1 (Undirected Global Markov Property)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. A distribution on $\mathcal{V}$ satisfies the global Markov property with respect to the undirected graph $\mathcal{G}$ if it satisfies all conditional independence statements $X_A \perp\!\!\!\perp X_B \mid X_C$ for all disjoint sets $A, B, C \subset \mathcal{V}$ such that $\mathcal{C}$ separates A and B in $\mathcal{G}$.*

Definition 2 (Pairwise Markov Property)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph. A distribution on $\mathcal{V}$ satisfies the the pairwise Markov property with respect to the undirected graph $\mathcal{G}$ if it satisfies all conditional independence statements $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$, where (i, j) is not an edge of $\mathcal{G}$;*

It is clear that the global Markov property implies the pairwise Markov property. However, the inverse implication does not hold in general. A simple counterexample can be constructed on 3 binary random variables $X = Y = Z$ with $P(X = 0) = P(X = 1) = \frac{1}{2}$. Note that (X, Y, Z) satisfies the pairwise Markov property with



(a) Conditioning nodes are shaded.

(b) Conditioning nodes and associated edges are removed.

Figure 1: This undirected graphical model expresses the conditional independence property $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}}$. When the shaded nodes corresponding to $\mathcal{C}$ are removed, the graph decomposes into multiple connected components, such that $\mathcal{A}$ and $\mathcal{B}$ belong to disjoint sets of components.

respect to the graph $\mathcal{G}$ containing a single edge $Y - Z$, but it does not satisfy the global Markov property with respect to $\mathcal{G}$.

The Markov properties ensure that an undirected graph defines a family of probability distributions via the *graph separation* property:

> The conditional independence relation $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}}$ holds whenever there is no path from a node in $\mathcal{A}$ to a node in $\mathcal{B}$ that does not pass through a node in $\mathcal{C}$.

In other words, the graph expresses conditional independencies in the family of distributions via *graph separation*. Conveniently, there are standard graph search algorithms for testing graph separation.

Testing for conditional independence using graph separation is illustrated in Fig. 1. The test is implemented as follows: delete all the nodes in $\mathcal{C}$ from the graph, as well as any edges touching them. If the resulting graph decomposes into multiple connected components, such that $\mathcal{A}$ and $\mathcal{B}$ belong to different components, then $x_{\mathcal{A}} \perp\!\!\!\perp x_{\mathcal{B}} \mid x_{\mathcal{C}}$.

It is worth emphasizing that, as in the case of directed models, a graph represents the list of every distribution that satisfies all the expressed conditional independencies. Some distributions on this list may, of course, satisfy additional conditional independencies.

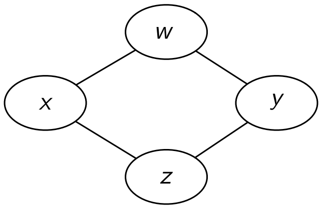
Like directed graphical models, undirected models are universal: for every distribution, there is an undirected graphical model that can represent it. In particular, a fully-connected undirected graph corresponds to the family of distributions with no conditional independence constraints—i.e., all possible distributions. Such universality is not of primary importance, however. What we want are graphical models that are effective at capturing structure in a distribution so as to allow for the development of efficient inference algorithms. From this perspective, directed and undirected models have distinguishing features, as we now develop.

3.2 Comparing Directed and Undirected Graphical Models

Choosing between directed and undirected models for a given application is not always straightforward. Ultimately, each is better than the other in capturing some forms of conditional independence structure, and, thus, factorization struture. In this section, we develop some preliminary insights into their differing capabilities. In particular, we demonstrate that there are families of distributions that can be represented using an undirected graph but not a directed one, and vice-versa, but also some that can be represented by both.

To start, consider an undirected graph with three nodes and two edges. Such a graph expresses a simple conditional independence (Markov) relation between the variables. As we saw, there are multiple ways to choose two directed edges to capture such conditional independence in a directed graphical model. Hence, this is an example of a family that is equally well represented by both types of models.

However, now consider the following graph.



Note that, among others, it expresses the independence constraint

$$x \perp\!\!\!\perp y \mid \{w, z\}. \tag{1}$$

Let's try to construct a directed graph with the same number of nodes to represent the same family of distributions (i.e., the same set of conditional independencies). First, note that it must contain at least the same set of edges as the undirected graph, because any pair of variables connected by an edge depend on each other regardless of whether or not any of the other variables are observed. In order for the graph to be acyclic, one of the nodes must come last in a topological ordering; without loss of generality, let's suppose it is node z. Then z has two incoming edges. Now, no matter what directions we assign to the remaining two edges, we cannot guarantee the property (1) that holds in the undirected graph, because the path $x \to z \leftarrow y$ is d-connected when z is observed. Therefore, there is no directed graph that expresses the same set of conditional independencies as the undirected one.

Next consider the following simple common-cause directed graphical model, which has V-structure.

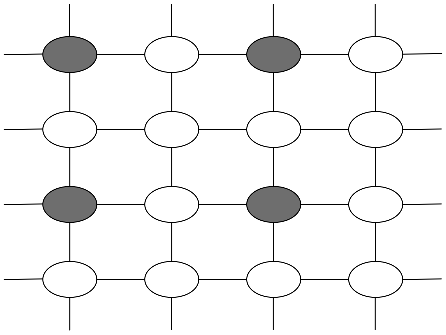
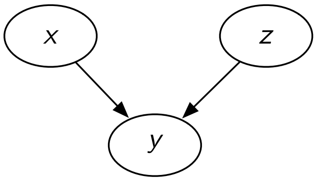


Figure 2: Part of an undirected graphical model for an image processing task, image superresolution. Nodes correspond to pixels, and every fourth pixel is observed.



As we developed, for this model we have $x \perp\!\!\!\perp z$, but not $x \perp\!\!\!\perp z \mid y$. By contrast, undirected graphical models have a certain monotonicity property: when additional nodes are observed, the new set of conditional independencies is a strict superset of the old one. Therefore, no undirected graph can represent the same family of distributions as a V-structure.

One example of a domain more naturally represented using undirected rather than directed graphs is computer vision. For instance, consider the problem of image superresolution, where we wish to double the number of pixels along each dimension. We formulate the graphical model shown in Fig. 2, where the nodes correspond to pixels, and undirected edges connect each pair of neighboring pixels. This graph represents a simple model for images whereby each pixel is independent of the rest of the image given its four neighboring pixels. In the superresolution task, a subset of the pixels in the image are observed, corresponding to the low-resolution image we have available, as shown by the shading in Fig. 2.

3.3 Factorizations of Distributions over Undirected Graphs

In our development, directed graphical models were defined in terms of their factorizations, and from that we obtained their equivalent characterization in terms of conditional independence relations with some effort. By contrast, we defined undirected graphical models in terms of their conditional independence properties, and we will now obtain their equivalent characterization in terms of factorizations.

As will become apparent, there is not a natural factorization in terms of conditional probability distributions for such models. Instead, the factorization will be in terms of what are referred to as *potentials* (or sometimes *compatibility functions*).

To motivate this factorization, consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with no edge between nodes i and j. Then it follows from the pairwise Markov property that $\mathsf{x}_i \perp\!\!\!\perp \mathsf{x}_j \mid \mathsf{x}_{\mathcal{V}\setminus\{i,j\}}$, which in turn implies

$$\begin{aligned} p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) &= p_{\mathsf{x}_i,\mathsf{x}_j|\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}}(x_i, x_j | x_{\mathcal{V}\setminus\{i,j\}})\, p_{\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}}(x_{\mathcal{V}\setminus\{i,j\}}) \\ &= p_{\mathsf{x}_i|\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}}(x_i | x_{\mathcal{V}\setminus\{i,j\}})\, p_{\mathsf{x}_j|\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}}(x_j | x_{\mathcal{V}\setminus\{i,j\}})\, p_{\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}}(x_{\mathcal{V}\setminus\{i,j\}}), \end{aligned}$$

from which we conclude that a distribution missing such an edge can always be factorized in such a way that x_i and x_j are in different factors.

This example motivates a general factorization based on the concept of cliques, which we now develop.

Definition 3. *A* clique *in an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a fully connected subset of the nodes of $\mathcal{V}$, i.e., a subset of nodes such that between each pair there is an edge.*

We use $\mathrm{cl}(\mathcal{G})$ to denote the set of all cliques associated with an undirected graph $\mathcal{G}$.

Examples of cliques include individual nodes, pairs of nodes connected by an edge, and triples of nodes with three edges. In general, a clique with K nodes has $K(K-1)/2$ edges.

Given a set of variables $\mathsf{x}_1, \ldots, \mathsf{x}_N$, we define a family of distributions with respect to the graph $\mathcal{G}$ via

$$p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) = \frac{1}{Z} \prod_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \psi_\mathcal{C}(x_\mathcal{C}), \tag{2}$$

where Z normalizes the distribution and is referred to as the *partition function*. In this representation, the potentials $\psi_\mathcal{C}(\cdot)$ can be any nonnegative valued functions, i.e., they do not need to sum to 1.

The partition function Z can be written explicitly as

$$Z = \sum_{x_\mathcal{V}} \prod_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \psi_\mathcal{C}(x_\mathcal{C}).$$

Depending on the graph structure, this sum can be quite expensive to evaluate. Fortunately, for many calculations, such as computing conditional probabilities and finding most probable configurations, it is not needed. For other calculations, such as learning the parameters ψ of the representation from data, it is needed.

Note that member distributions in the family do not have a unique representation within the form (2). Indeed, any cliques that consist of more than one node must contain subsets of nodes that are cliques. So for example, a fully-connected three-node graph corresponding to variables x_1, x_2, and x_3 has a distribution that, according to (2), can be written in the form

$$\begin{aligned}
&p_{\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3}(x_1,x_2,x_3) \\
&\qquad \propto \psi_1(x_1)\,\psi_2(x_2)\,\psi_3(x_3)\,\psi_{12}(x_1,x_2)\,\psi_{13}(x_1,x_3)\,\psi_{23}(x_2,x_3)\,\psi_{123}(x_1,x_2,x_3). \quad (3)
\end{aligned}$$

However, we see that the potentials corresponding to lower-order cliques can always be incorporated into higher-order ones, so that (3) can be equivalently written as

$$p_{\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3}(x_1,x_2,x_3) \propto \psi'_{123}(x_1,x_2,x_3)$$

for a suitable choice of $\psi'_{123}(\cdot,\cdot,\cdot)$.

From this perspective, we conclude that without loss of generality we can always restrict our attention to the maximal cliques in a given graph, which are defined as follows.

Definition 4. *A* maximal clique *is a clique that is not a strict subset of another clique, i.e., a clique to which another node cannot be added without losing the clique property.*

We use $\mathrm{cl}^*(\mathcal{G})$ to denote the set of all maximal cliques of an undirected graph $\mathcal{G}$.

In turn, we can then rewrite (2) as

$$p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) = \frac{1}{Z} \prod_{\mathcal{C}\in\mathrm{cl}^*(\mathcal{G})} \psi_\mathcal{C}(x_\mathcal{C}).$$

It is worth noting, however, that such a representation is still not unique. For example, note that distributions corresponding to a graph with three nodes and two edges can be written in the form

$$\begin{aligned}
p_{\mathsf{x}_1,\mathsf{x}_2,\mathsf{x}_3}(x_1,x_2,x_3) &\propto \psi_{12}(x_1,x_2)\,\psi_{23}(x_2,x_3) \\
&= \underbrace{\Big[\psi_{12}(x_1,x_2)\,\phi(x_2)\Big]}_{\triangleq\psi'_{12}(x_1,x_2)} \cdot \underbrace{\left[\frac{1}{\phi(x_2)}\,\psi_{23}(x_2,x_3)\right]}_{\triangleq\psi'_{23}(x_2,x_3)}
\end{aligned}$$

for any choice of strictly positive function $\phi(\cdot)$.

The complexity of description (number of parameters) of members of this family follows, in turn, as

$$\sum_{\mathcal{C}\in\mathrm{cl}(\mathcal{G})} |\mathcal{X}|^{|\mathcal{C}|} \sim |\mathcal{X}|^{\max_{\mathcal{C}\in\mathrm{cl}(\mathcal{G})}|\mathcal{C}|}. \quad (4)$$

As with directed graphical models, this expresses that complexity of the model is largely determined by the number of variables involved in the largest table of the factorization.[1]

[1] Strictly speaking, this approximation does not always hold, as the number of maximal cliques may be exponential in the number of variables. An example of this phenomenon is developed as a homework exercise. Nevertheless, (4) is a good rule of thumb for graphs that arise in practice.

At this point, we have associated two families of distributions with a given undirected graph: one corresponding to the conditional independencies expressed by graph separation, and one corresponding to the factorization based on cliques in the graph. At this point, it is reasonable to ask what the relationships between these two families is. The answer, remarkably, is that they are essentially identical, which we soon establish formally with a fundamental theorem.

First, note that if the distribution of X factorizes according to an undirected graph $\mathcal{G}$, then it satisfies the global, and hence also the pairwise, Markov property with respect to $\mathcal{G}$. This can be seen as follows: Let $A, B, S \subset \mathcal{V}$ disjoint such that S separates A from B. Let $\tilde{A}$ denote the connected component in $\mathcal{V} \setminus S$ that contains A, and let $\tilde{B} = \mathcal{V} \setminus (\tilde{A} \cup S)$. Then any clique in $\mathcal{G}$ is either in $\tilde{A} \cup S$ or $\tilde{B} \cup S$. Hence

$$f(x) \quad = \quad \prod_{C \subseteq \mathcal{V} \text{ clique}} \psi_C(x_C) \quad = \quad \prod_{C' \subseteq \tilde{A} \cup S \text{ clique}} \psi_{C'}(x_{C'}) \quad \prod_{C'' \subseteq \tilde{B} \cup S \text{ clique}} \psi_{C''}(x_{C''}) \quad = \quad h(x_{\tilde{A} \cup S}) k(x_{\tilde{B} \cup S}),$$

which implies that $\tilde{A} \perp\!\!\!\perp \tilde{B} \mid S$ and hence also that $A \perp\!\!\!\perp B \mid S$.

Next, the following central result to the theory of undirected graphical models is known as *Hammersley-Clifford Theorem* and formally establishes the assumptions under which equivalence holds between the Markov properties and factorization.

Theorem 1 (Hammersley-Clifford)**.** *For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a strictly positive distribution p (i.e., $p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) > 0$ for all joint assignments $x_\mathcal{V}$) satisfies the conditional independencies implied graph separation if and only if it satisfies the factorization* (2) *in terms of the cliques of the graph. Moreover, the "if" statement holds even without the strictly positive constraint on $p_{\mathsf{x}_\mathcal{V}}$.*

Before proceeding to the proof, it is worth emphasizing what the theorem tells us. Specifically, it establishes that when the probability of every configuration of variables is strictly positive, the two families of distributions we can use an undirected graph to represent are identical. It further establishes that the second family (corresponding to factorizations) is more generally always a subset of the first (corresponding to conditional independencies).

Proof. The "if" part of the theorem is a comparatively simpler exercise. Suppose we have sets of nodes $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ such that $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$. Then let

$$\mathcal{A}_+ \triangleq \{i \in \mathcal{V} \colon i \in \mathcal{A} \text{ or } i \text{ is connected to } \mathcal{A} \text{ by paths not including nodes in } \mathcal{C}\}, \quad (5)$$

from which we see $\mathcal{A} \subset \mathcal{A}_+$ and, due to graph separation, $\mathcal{B} \cap \mathcal{A}_+ = \varnothing$. Analogously, we let

$$\mathcal{B}_+ \triangleq \{i \in \mathcal{V} \colon i \in \mathcal{B} \text{ or } i \text{ is connected to } \mathcal{B} \text{ by paths not including nodes in } \mathcal{C}\}, \quad (6)$$

so $\mathcal{B} \subset \mathcal{B}_+$, and $\mathcal{V} = \mathcal{A}_+ \cup \mathcal{C} \cup \mathcal{B}_+$ is a partition of $\mathcal{V}$.

Now the key observation is that any clique in the graph must be a subset of either $\mathcal{A}_+ \cup \mathcal{C}$ or $\mathcal{B}_+ \cup \mathcal{C}$, since $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$. Indeed, if there were a clique that included a node in $\mathcal{A}_+$ and one in $\mathcal{B}_+$, then there would have to be an edge between these nodes, which would mean there would be a path between a node in $\mathcal{A}$ and one in $\mathcal{B}$ that doesn't go through a node in $\mathcal{C}$. Hence, in the representation (2) we can partition the potentials into two groups: one involving the cliques in $\mathcal{A}_+ \cup \mathcal{C}$, and the other involving the cliques in $\mathcal{B}_+ \cup \mathcal{C}$. Combining the constituent potentials within each, (2) can then be expressed in the form

$$p_{\mathsf{x}_{\mathcal{V}}}(x_{\mathcal{V}}) \propto \xi_{\mathcal{A}_+\cup\mathcal{C}}(x_{\mathcal{A}_+\cup\mathcal{C}})\, \xi_{\mathcal{B}_+\cup\mathcal{C}}(x_{\mathcal{B}_+\cup\mathcal{C}}). \tag{7}$$

But as developed in the homework, any random vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ satisfy $\mathbf{u} \perp\!\!\!\perp \mathbf{v} \mid \mathbf{w}$ if and only if $p_{\mathbf{u},\mathbf{v},\mathbf{w}}(\mathbf{u},\mathbf{v},\mathbf{w}) \propto \xi_1(\mathbf{u},\mathbf{w})\,\xi_2(\mathbf{v},\mathbf{w})$. Moreover, if $\mathbf{u} \perp\!\!\!\perp \mathbf{v} \mid \mathbf{w}$, then $f(\mathbf{u}) \perp\!\!\!\perp g(\mathbf{v}) \mid \mathbf{w}$ for any functions $f(\cdot)$ and $g(\cdot)$. Using these two facts together with (7), we conclude that $\mathsf{x}_{\mathcal{A}} \perp\!\!\!\perp \mathsf{x}_{\mathcal{B}} \mid \mathsf{x}_{\mathcal{C}}$ as required.

We now develop the "only if" part of the theorem, which is more subtle. We restrict our attention to the case of variables x_i over a binary alphabet $\mathcal{X} = \{0, 1\}$, which simplifies the development, while revealing the main ideas.[2] There are various ways to establish the result for more general alphabets, as the literature reflects, but the so-called Möbius Inversion Formula is typically involved, as it is implicitly in the proof below.

Note, in particular, that in the binary case, any vector $\mathbf{x} = x_{\mathcal{V}} = (x_1, \dots, x_N) \in \mathcal{X}^N$ can be equivalently represented by its index set

$$\mathcal{I}(\mathbf{x}) \triangleq \{i \in \mathcal{V} \colon x_i = 1\} \subset \mathcal{V},$$

so

$$p_{\mathsf{x}}(\mathbf{x}) = \mathbb{P}\left(\mathsf{x} = \mathbf{x}\right) = \mathbb{P}\left(\mathsf{x}_{\mathcal{I}(\mathbf{x})} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus\mathcal{I}(\mathbf{x})} = \mathbf{0}\right), \tag{8}$$

where $\mathbf{1}$ and $\mathbf{0}$ denote vectors of all 1's and all 0's, respectively, of appropriate length.

As our proof, we establish that

$$p_{\mathsf{x}}(\mathbf{x}) \propto \exp\left\{ \sum_{\mathcal{C}\in\mathrm{cl}(\mathcal{G}) \colon \mathcal{C}\subset\mathcal{I}(\mathbf{x})} H(\mathcal{C}) \right\}, \tag{9a}$$

where for $\mathcal{S} \subset \mathcal{V}$

$$H(\mathcal{S}) \triangleq \sum_{\mathcal{A} \colon \mathcal{A}\subset\mathcal{S}} (-1)^{|\mathcal{S}\setminus\mathcal{A}|} \ln \mathbb{P}\left(\mathsf{x}_{\mathcal{A}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus\mathcal{A}} = \mathbf{0}\right). \tag{9b}$$

[2] This proof is adapted from A. Grimmet, "Theorem about Random Fields," *Bull. London Math. Soc.*, vol. 5, pp. 81–84, 1973.

To obtain (9a), note first that using (9b) we have

$$\begin{aligned}
\sum_{\mathcal{B}\colon\, \mathcal{B}\subset\mathcal{S}} H(\mathcal{B}) &= \sum_{\mathcal{B}\colon\, \mathcal{B}\subset\mathcal{S}} \; \sum_{\mathcal{A}\colon\, \mathcal{A}\subset\mathcal{B}} (-1)^{|\mathcal{B}\backslash\mathcal{A}|} \ln \mathbb{P}\left(x_{\mathcal{A}} = \mathbf{1}, x_{\mathcal{V}\backslash\mathcal{A}} = \mathbf{0}\right) \\
&= \sum_{\mathcal{A}\colon\, \mathcal{A}\subset\mathcal{S}} \; \sum_{\mathcal{B}\colon\, \mathcal{A}\subset\mathcal{B}\subset\mathcal{S}} (-1)^{|\mathcal{B}\backslash\mathcal{A}|} \ln \mathbb{P}\left(x_{\mathcal{A}} = \mathbf{1}, x_{\mathcal{V}\backslash\mathcal{A}} = \mathbf{0}\right) \\
&= \sum_{\mathcal{A}\colon\, \mathcal{A}\subset\mathcal{S}} \ln \mathbb{P}\left(x_{\mathcal{A}} = \mathbf{1}, x_{\mathcal{V}\backslash\mathcal{A}} = \mathbf{0}\right) \left(\sum_{\mathcal{B}\colon\, \mathcal{A}\subset\mathcal{B}\subset\mathcal{S}} (-1)^{|\mathcal{B}\backslash\mathcal{A}|} \right) \\
&= \ln \mathbb{P}\left(x_{\mathcal{S}} = \mathbf{1}, x_{\mathcal{V}\backslash\mathcal{S}} = \mathbf{1}\right), \qquad (10)
\end{aligned}$$

where to obtain (10) we have used that

$$\sum_{\mathcal{B}\colon\, \mathcal{A}\subset\mathcal{B}\subset\mathcal{S}} (-1)^{|\mathcal{B}\backslash\mathcal{A}|} = \mathbb{1}_{\mathcal{A}=\mathcal{S}}. \qquad (11)$$

To verify (11), note first that when $\mathcal{A} = \mathcal{S}$ we have $\mathcal{B} = \mathcal{A}$ so $|\mathcal{B}\backslash\mathcal{A}| = 0$ and the left-hand side of (11) is trivially 1. Next, when $\mathcal{A} \neq \mathcal{S}$, the collection of all $\mathcal{B}$ such that $\mathcal{A} \subset \mathcal{B} \subset \mathcal{S}$ can be partitioned into $k+1$ subcollections with $k = |\mathcal{S}\backslash\mathcal{A}|$, where the jth subcollection contains all the sets $\mathcal{B}$ such that $|\mathcal{B}\backslash\mathcal{A}| = j$, for $j = 0, 1, \ldots, k$. Since the number of sets in the jth subcollection is $\binom{k}{j}$, it follows that

$$\sum_{\{\mathcal{B}\colon\, \mathcal{A}\subset\mathcal{B}\subset\mathcal{S}\}} (-1)^{|\mathcal{B}\backslash\mathcal{A}|} = \sum_{j=0}^{k} (-1)^j \binom{k}{j} = \left(1 + (-1)\right)^k = 0, \qquad (12)$$

where to obtain the middle equality we have used the binomial theorem

$$(u+v)^k = \sum_{j=0}^{k} \binom{k}{j} u^{k-j} v^j, \quad u, v \in \mathbb{R}$$

with $u = 1$ and $v = -1$.

Using (10) with (8) establishes that for any $\mathbf{x} \in \mathcal{X}^N$,

$$\ln p_{\mathbf{x}}(\mathbf{x}) = \sum_{\mathcal{S}\colon\, \mathcal{S}\subset\mathcal{I}(\mathbf{x})} H(\mathcal{S}), \qquad (13)$$

Comparing (13) with (9), we see it remains only to show that for $\mathcal{S} \subset \mathcal{I}(\mathbf{x})$

$$H(\mathcal{S}) = 0 \text{ when } \mathcal{S} \text{ is not a clique}, \qquad (14)$$

which we now establish using the conditional independence (Markov) structure in the graph. Proceeding, when $\mathcal{S}$ is not a clique, there exists $i, j \in \mathcal{S}$ so that $(i, j) \notin \mathcal{E}$. We can always rewrite $H(\mathcal{S})$ as

$$\begin{aligned} H(\mathcal{S}) &= \sum_{\mathcal{A}\colon \mathcal{A}\subset\mathcal{S}} (-1)^{|\mathcal{S}\setminus\mathcal{A}|} \ln \mathbb{P}\left(\mathsf{x}_{\mathcal{A}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus\mathcal{A}} = \mathbf{0}\right) && (15) \\ &= \sum_{\mathcal{B}\colon \mathcal{B}\subset\mathcal{S},\, i,j\notin\mathcal{B}} \Big[(-1)^{|\mathcal{S}\setminus\mathcal{B}|}\, \tau_{\varnothing} + (-1)^{|\mathcal{S}\setminus(\mathcal{B}\cup\{i\})|}\, \tau_{\{i\}} \\ &\qquad\qquad + (-1)^{|\mathcal{S}\setminus(\mathcal{B}\cup\{j\})|}\, \tau_{\{j\}} + (-1)^{|\mathcal{S}\setminus(\mathcal{B}\cup\{i,j\})|}\, \tau_{\{i,j\}}\Big] && (16) \\ &= \sum_{\mathcal{B}\colon \mathcal{B}\subset\mathcal{S},\, i,j\notin\mathcal{B}} (-1)^{|\mathcal{S}\setminus\mathcal{B}|} \ln \frac{\tau_{\varnothing}\, \tau_{\{i,j\}}}{\tau_{\{i\}}\, \tau_{\{j\}}}, && (17) \end{aligned}$$

with

$$\tau_{\mathcal{A}} = \mathbb{P}\left(\mathsf{x}_{\mathcal{B}\cup\mathcal{A}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\mathcal{A})} = \mathbf{0}\right), \tag{18}$$

where to obtain (16) we have used that the collection of sets in $\mathcal{S}$ in the summation in (15) can be partitioned into four subcollections: one that contains the sets that exclude both i and j, one that contains the sets that include i but not j, one that contains the sets that include j but not i, and one that contains the sets that include both i and j. As our expansion reflects, the last three collections can be generated from the first by taking each set in the first, and augmenting it with either i, j, or both.

But

$$\begin{aligned} \frac{\tau_{\{i,j\}}}{\tau_{\{j\}} + \tau_{\{i,j\}}} &= \frac{\mathbb{P}\left(\mathsf{x}_i = 1, \mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right)}{\mathbb{P}\left(\mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right)} && (19) \\ &= \mathbb{P}\left(\mathsf{x}_i = 1 \mid \mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right) \\ &= \mathbb{P}\left(\mathsf{x}_i = 1 \mid \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right), && (20) \end{aligned}$$

where to obtain the denominator on the right-hand side of (19) we have used that

$$\begin{aligned} \tau_{\{j\}} &= \mathbb{P}\left(\mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{j\})} = \mathbf{0}\right) \\ &= \mathbb{P}\left(\mathsf{x}_i = 0, \mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right) \end{aligned}$$

and

$$\tau_{\{i,j\}} = \mathbb{P}\left(\mathsf{x}_i = 1, \mathsf{x}_j = 1, \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right),$$

and where to obtain (20) we have used that $(i,j) \notin \mathcal{E}$ and hence x_i is independent of x_j conditioned on all the remaining variables $\mathsf{x}_{\mathcal{V}\setminus\{i,j\}}$. In an analogous manner, we obtain

$$\frac{\tau_{\{i\}}}{\tau_{\varnothing} + \tau_{\{i\}}} = \mathbb{P}\left(\mathsf{x}_i = 1 \mid \mathsf{x}_{\mathcal{B}} = \mathbf{1}, \mathsf{x}_{\mathcal{V}\setminus(\mathcal{B}\cup\{i,j\})} = \mathbf{0}\right), \tag{21}$$

which upon comparison with (20) reveals that

$$\frac{\tau_{\varnothing}}{\tau_{\{i\}}} = \frac{\tau_{\{j\}}}{\tau_{\{i,j\}}},$$

so the argument of the logarithm in (17) is 1 and $H(\mathcal{S}) = 0$, yielding (9) as required. □

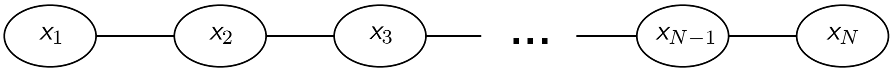


Figure 3: An undirected graph corresponding to a one-dimensional Ising model. In the Ising model, the states x_i from a binary alphabet $\mathcal{X}$.

3.4 Connections to Statistical Mechanics

There are rich connections between undirected graphical models and statistical mechanics. As a glimpse of this relationship, note that we can rewrite (2) in the form of a Gibbs (or Boltzmann) distribution

$$p_{x_\mathcal{V}}(x_\mathcal{V}) \triangleq \frac{1}{Z} \exp \left\{ - \underbrace{\sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} H_\mathcal{C}(x_\mathcal{C})}_{\triangleq H(x_\mathcal{V})} \right\}. \tag{22}$$

In this representation, $x_\mathcal{V}$ corresponds to the state of a collection of $N = |\mathcal{V}|$ particles, and $H(x_\mathcal{V})$ is referred to as the *Hamiltonian*, which relates to the *energy* of the configuration $x_\mathcal{V}$ of states. Via (22) we see that global configurations with low energy are favored over those with high energy.

One well-studied example is the *Ising* model, a one-dimensional version of which is depicted in Fig. 3. In this Ising model, the variables $x_1, \ldots, x_N$ represent *spins* and are arranged in a chain, taking on values in the alphabet $\mathcal{X} = \{+, -\}$. Moreover, there is a common pairwise potential function that either favors or discourages configurations of states in which neighboring spins are identical. Specifically,

$$p_{x_\mathcal{V}}(x_\mathcal{V}) \propto \exp \left\{ G(x_N) - \sum_{i=1}^{N-1} H(x_i, x_{i+1}) \right\}$$

for some $G(\cdot)$ and $H(\cdot, \cdot)$. For instance, one (symmetric) system that favors neighboring spins to be the same has

$$H(+,+) = H(-,-) = -\frac{3}{2}\beta \qquad \text{and} \qquad H(+,-) = H(-,+) = -\frac{1}{5}\beta,$$

where $\beta = 1/kT$, with k denoting Boltzmann's constant and T denoting the temperature of the system.

3.5 Beyond Directed Acyclic Graphs and Markov Random Fields

In spite of their flexibility, there are important families of distributions whose factorization structure cannot be expressed and ultimately exploited by either the directed or undirected graphical models we have discussed thus far. So in order to model some such distributions, we need to introduce yet a third class of graphical models: factor graphs. As we will see, factor graphs are very flexible and powerful, but at a price. In some sense, the existence of three distinct formalisms for representing families of distributions as graphs is a sign of the still rather immature state of this important and growing field.