Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

3 Undirected Graphical Models

In this lecture, we discuss undirected graphical models. Recall that directed graphical models were capable of representing any probability distribution (e.g. if the graph was a fully connected graph). The same is true for undirected graphs. However, the two formalisms can express different sets of conditional independencies and factorizations, and one or the other may be more intuitive for particular application domains.

Recall that we defined directed graphical models in terms of factorization into a product of conditional probabilites, and the Bayes Ball algorithm was required to test conditional independencies. In contrast, we define undirected graphical models in terms of conditional independencies, and then derive the factorization properties. In a sense, a directed graph more naturally represents conditional probabilities directly, whereas an undirected graph more naturally represents conditional independence properties.

3.1 Representation

An undirected graphical model is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertices (or nodes) $\mathcal{V}$ correpsond to variables and the undirected edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ tell us about the conditional independence structure. The undirected graph defines a family of probability distributions which satisfy the following graph separation property:

- $x_A \perp\!\!\!\perp x_B | x_C$ whenever there is no path from a node in A to a node in B which does not pass through a node in C.

As before, the graph represents the family of all distributions which satisfy this property; individual distributions in the family may satisfy additional conditional independence properties. An example of this definition is shown in Figure 1. We note that graph separation can be tested using a standard graph search algorithm. Because the graph separation property can be viewed as a spatial Markov property, undirected graphical models are sometimes called *Markov random fields*.

Another way to express this definition is as follows: delete all the nodes in C from the graph, as well as any edges touching them. If the resulting graph decomposes into multiple connected components, such that A and B belong to different components, then $x_A \perp\!\!\!\perp x_B | x_C$.

3.2 Directed vs. undirected graphs

We have seen that directed graphs naturally represent factorization properties and undirected graphs naturally represent conditional independence properties. Does this

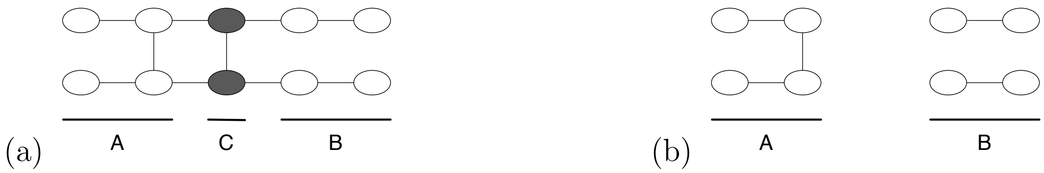
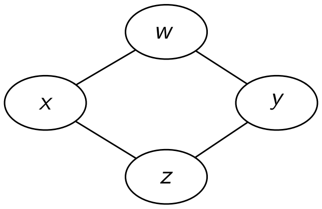


Figure 1: (a) This undirected graphical model expresses the conditional independence property $x_A \perp\!\!\!\perp x_B | x_C$. (b) When the shaded nodes are removed, the graph decomposes into multiple connected components, such that A and B belong to disjoint sets of components.

mean that we should always use directed graphs when we have conditional probability distributions and undirected graphs when we have conditional independencies? No —it turns out that each formalism can represent certain families of distributions that the other cannot.

For example, consider the following graph.



Let's try to construct a directed graph 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 $x \perp\!\!\!\perp y | w, z$ (which holds in the undirected graph), because the Bayes Ball can pass along the path $x \to z \leftarrow y$ when z is observed. Therefore, there is no directed graph that expresses the same set of conditional independencies as the undirected one.

What about the reverse case — can every directed graph be translated into an undirected one while preserving conditional independencies? No, as the example of the v-structure shows:

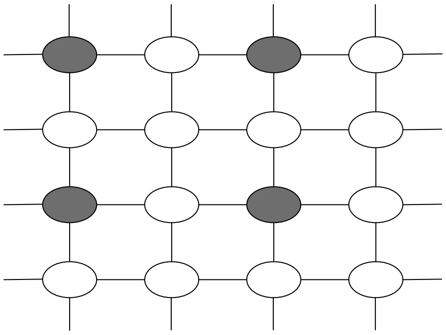
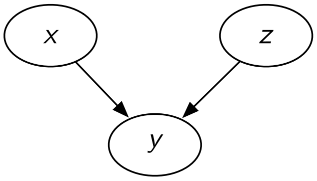


Figure 2: Part of an undirected graphical model for an image processing task, image superresolution. Nodes correspond to pixels, and every fourth pixel is observed.



We saw in Lecture 2 that $x \perp\!\!\!\perp z$, but not $x \perp\!\!\!\perp z|y$. 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.

An example of a domain more naturally represented using undirected rather than directed graphs is image processing. 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 Figure 2, where the nodes correspond to pixels, and undirected edges connect each pair of neighboring pixels. This graph represents the assumption that each pixel is independent of the rest of the image given its four neighboring pixels. In the superresolution task, we may treat every fourth pixel as observed, as shown in the Figure 2.

3.3 Parameterization

Like directed graphical models, undirected graphical models can be characterized either in terms of conditional independence properties or in terms of factorization. Unlike directed graphical models, undirected graphical models do not have a natural factorization into a product of conditional probabilities. Instead, we represent the distribution as a product of funcations called *potentials*, times a normalization constant.

To motivate this factorization, consider a graph with no edge between nodes x_i

and x_j. By definition, $\mathsf{x}_i \perp\!\!\!\perp \mathsf{x}_j | \mathsf{x}_{rest}$, where x_{rest} is shorhand for all the other nodes in the graph. We find that

$$\begin{aligned} p_{\mathsf{x}_{all}} &= p_{\mathsf{x}_i, \mathsf{x}_j | \mathsf{x}_{rest}} p_{\mathsf{x}_{rest}} \\ &= p_{\mathsf{x}_1 | \mathsf{x}_{rest}} p_{\mathsf{x}_2 | \mathsf{x}_{rest}} p_{x_{rest}}. \end{aligned}$$

Therefore, we conclude that the joint distribution can be factorized in such a way that x_i and x_j are in different factors.

This motivates the following factorization criterion. A *clique* is a fully connected set of nodes. A *maximal clique* is a clique that is not a strict subset of another clique. Given a set of variables $\mathsf{x}_1, \ldots, \mathsf{x}_N$ and a set $\mathcal{C}$ of maximal cliques, define the following representation of the joint distribution:

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &\propto \prod_{C \in \mathcal{C}} \psi_{\mathsf{x}_C}(x_C), \\ &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{\mathsf{x}_C}(x_C). \end{aligned} \tag{1}$$

In this representation, Z is called the *partition function*, and is chosen such that the probabilities corresponding to all joint assignments sum to 1. The functions ψ can be any nonnegative valued functions (i.e. do not need to sum to 1), and are sometimes referred to as *compatibilities*.

The partition function Z can be written explicitly as

$$Z = \sum_{\mathbf{x}} \prod_{C \in \mathcal{C}} \psi_{\mathsf{x}_C}(x_C).$$

This sum can be quite expensive to evaluate. Fortunately, for many calculations, such as conditional probabilities and finding the most likely joint assignment, we do not need it. For other calculations, such as learning the parameters ψ, we do need it.

The complexity of description (number of parameters) is given by:

$$\sum |\mathcal{X}|^{|C|} \approx |\mathcal{X}|^{\max |C|}.$$

As with directed graphical models, the main determinant of the complexity is the number of variables involved in each term of the factorization.[1]

So far, we have defined an arbitrary factorization property based on the graph structure. Is this related to our earlier definition of undirected graphical models in terms of conditional independencies? The relationship was formally established by the following theorem.

[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 is given in one of the homeworks. However, it is a good rule of thumb for graphs that arise in practice.

Theorem 1 (Hammersley-Clifford) *A strictly positive distribution p (i.e. $p_{\mathbf{x}}(\mathbf{x}) > 0$ for all joint assignments $\mathbf{x}$) satisfies the graph separation property from our definition of undirected graphical models if and only if it can be represented in the factorized form* (1)*.*

Proof:

One direction, the factorization (1) implying satisfaction of the graph separation criterion, is straightforward. The other direction requires non-trivial arguments. For it, we shall provide proof for binary Markov random field, i.e. $\mathbf{x} \in \{0,1\}^N$. This proof is adapted from [*Grimmet, A Theorem about Random Fields, BULL. LONDON MATH. SOC, 5 (1973), 81-84*].

Now any $\mathbf{x} \in \{0,1\}^N$ is equivalent to a set $S(\mathbf{x}) \equiv S \subseteq V = \{1,\ldots,N\}$, where $S(\mathbf{x}) = \{i \in V : x_i = 1\}$. With this in mind, the probability distribution of $\mathbf{X}$ over $\{0,1\}^N$ is equivalent to probability distribution over the set of all subsets of V, 2^V. To that end, let us start by defining

$$Q(S) = \sum_{A \subseteq S} (-1)^{|S-A|} \log p\big(\mathbf{X}_A = \mathbf{1}, \mathbf{X}_{V\backslash A} = \mathbf{0}\big) \tag{2}$$

where $\mathbf{1}$ and $\mathbf{0}$ are vectors of all ones and zeros respectively (of appropriate length). Accordingly,

$$Q(\emptyset) = \log p(\mathbf{X} = \mathbf{0}).$$

We claim that if $S \subset V$ is not a clique of the graphical model G, then

$$Q(S) = 0. \tag{3}$$

To prove this, we shall use the fact that $\mathbf{X}$ is a Markov Random Field with respect to G. Now, since S is not a clique, there exists $i, j \in S$ so that $(i,j) \notin E$. Now consider

$$\begin{aligned} Q(S) &= \sum_{A \subseteq S} (-1)^{|S-A|} \log p\big(\mathbf{X}_A = \mathbf{1}, \mathbf{X}_{V\backslash A} = \mathbf{0}\big) \\ &= \sum_{B \subset S : i,j \notin B} (-1)^{|S-B|} \log \frac{p\big(\mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B} = \mathbf{0}\big) \times p\big(\mathbf{X}_{B\cup\{i,j\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big)}{p\big(\mathbf{X}_{B\cup\{i\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i\}} = \mathbf{0}\big) \times p\big(\mathbf{X}_{B\cup\{j\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{j\}} = \mathbf{0}\big)}. \end{aligned} \tag{4}$$

With notation $a_{i,j} = p\big(\mathbf{X}_{B\cup\{i,j\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big)$, $a_i = p\big(\mathbf{X}_{B\cup\{i\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i\}} = \mathbf{0}\big)$, $a_j = p\big(\mathbf{X}_{B\cup\{j\}} = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{j\}} = \mathbf{0}\big)$, and $a_0 = p\big(\mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B} = \mathbf{0}\big)$, we have

$$\begin{aligned} \frac{a_{i,j}}{a_j + a_{i,j}} &= \frac{p\big(X_i = 1, X_j = 1, \mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big)}{p\big(X_j = 1, \mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big)} \\ &= p\big(X_i = 1 | X_j = 1, \mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big) \\ &= p\big(X_i = 1 | \mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V\backslash B\cup\{i,j\}} = \mathbf{0}\big), \end{aligned} \tag{5}$$

where in the last equality, we have used the fact that $(i,j) \notin E$ and hence X_i is independent of X_j condition on the assignment of all other variables fixed. In a very similar manner,

$$\frac{a_i}{a_0 + a_i} = p\big(X_i = 1 | \mathbf{X}_B = \mathbf{1}, \mathbf{X}_{V \backslash B \cup \{i,j\}} = \mathbf{0}\big). \tag{6}$$

From (5)-(6), we conclude that

$$\frac{a_0}{a_i} = \frac{a_j}{a_{i,j}}$$

therefore in (4) we have that $Q(S) = 0$. This establishes the claim that $Q(S) = 0$ if $S \subset V$ is not a clique. From (2) and with notation $G(A) = \log p\big(\mathbf{X}_A = \mathbf{1}, \mathbf{X}_{V\backslash A} = \mathbf{0}\big)$ for all $A \subset V$ and $\mu(S,A) = (-1)^{|}S - A|$ any $S, A \subset V$ such that $A \subset S$, we can re-write (2) as

$$Q(S) = \sum_{A \subset S} \mu(S,A) G(A). \tag{7}$$

Therefore,

$$\begin{aligned} \sum_{A \subset S} Q(A) &= \sum_{A \subset S} \sum_{B \subset A} \mu(A,B) G(B) \\ &= \sum_{B \subset S} \sum_{B \subset A \subset S} \mu(A,B) G(B) \\ &= \sum_{B \subset S} G(B) \big(\sum_{B \subset A \subset S} \mu(A,B) \big) \\ &= \sum_{B \subset S} G(B) \delta(B,S) \\ &= G(S), \end{aligned} \tag{8}$$

where $\delta(B,S)$ is 1 if $B = S$ and 0 otherwise. To see the second last equality, note that given $B \subset S$ and $B \neq S$, all A such that $B \subset A \subset S$ can be decomposed amongst sets so that $|A - B| = \ell$, for $0 \leq \ell \leq k \equiv |S - B|$. The number A with $|A - B| = \ell$, is $\binom{k}{\ell}$. Therefore,

$$\begin{aligned} \sum_{B \subset A \subset S} \mu(A,B) &= \sum_{\ell=0}^{k} (-1)^{\ell} \binom{k}{\ell} \\ &= \big(1 - 1\big)^k = 0. \end{aligned} \tag{9}$$

Of course, when $B = S$, the above is equal to 1 trivially. Thus, we have that for any $\mathbf{x} \in \{0,1\}^N$,

$$\begin{aligned} \log p\big(\mathbf{X} = \mathbf{x}\big) &= G(N(\mathbf{x})) \\ &= \sum_{S \subset N(\mathbf{x}) : S \text{ clique}} Q(S), \end{aligned} \tag{10}$$

where $N(\mathbf{x}) = \{i \in V : x_i = 1\}$. In summary,

$$p\big(\mathbf{X} = \mathbf{x}\big) \propto \exp\big(\sum_{S \subset V : S \text{ clique}} P_S(\mathbf{x}) \big), \tag{11}$$

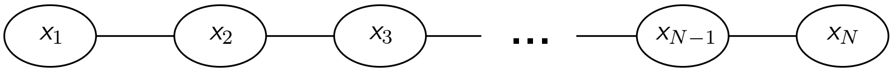


Figure 3: A one dimensional Ising model.

where the potential function $P_S : \{0,1\}^{|S|} \to \mathbb{R}$, for each clique $S \subset V$ is defined as

$$P_S(\mathbf{x}) = \begin{cases} Q(S) & \text{if } S \subset N(\mathbf{x}) \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

This completes the proof. ■

3.4 Energy interpretation

We now note some connections to statistical physics. The factorization (1) can be rewritten as

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{Z}\exp\Big(-H(\mathbf{x})\Big). \\ &\triangleq \frac{1}{Z}\exp\Big(-\sum_{C\in\mathcal{C}} H_C(x_C)\Big), \end{aligned}$$

a form known as the *Boltzmann distribution.* $H(\mathbf{x})$ is sometimes called the *Hamiltonian*, and relates to the *energy* of the state $\mathbf{x}$. Effectively, global configurations with low energy are favored over those with high energy. One well-studied example is the *Ising* model, for which the one-dimensional case is shown in Figure 3. In the one-dimensional case, the variables $x_1, \ldots, x_N$, called *spins*, are arranged in a chain, and take on values in $\{+,-\}$. The pairwise compatibilities either favor or punish states where neighboring spins are identical. For instance, we may define

$$H_{x_i,x_{i+1}} = \begin{bmatrix} 3/2 & 1/5 \\ 1/5 & 3/2 \end{bmatrix}.$$

There exist factorizations of distributions that cannot be represented by either directed or undirected graphical models. In order to model some such distributions, we will introduce the factor graph in the next lecture. (Note that this does *not* imply that there exist *distributions* which cannot be represented in either formalism. In fact, in both formalisms, fully connected graphs can represent any distribution. This observation is uninteresting, because the very point of graphical models is to compactly represent distributions in ways that support efficient learning and inference.) The existence of three separate formalisms for representing families of distributions as graphs is a sign of the immaturity of the field.

6.438 Algorithms for Inference
Fall 2014