4 Factor graphs and Comparing Graphical Model Types

We now introduce a third type of graphical model. Beforehand, let us summarize some key perspectives on our first two.

First, for some ordering of variables, we can write any probability distribution as

$$p_{\mathbf{x}}(x_1, \ldots, x_n) = p_{x_1}(x_1) p_{x_2|x_1}(x_2, x_1) \cdots p_{x_n|x_1,\ldots,x_{n-1}}(x_n|x_1, \ldots, x_{n-1}),$$

which can be expressed as a fully connected directed graphical model. When the conditional distributions involved do not depend on all the indicated conditioning variables, then some of the edges in the directed graphical model can be removed. This reduces the complexity of inference, since the associated conditional probability tables have a more compact description.

By contrast, undirected graphical models express distributions of the form

$$p_{\mathbf{x}}(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \Phi_c(x_c),$$

where the potential functions Φ_c are non-negative and $\mathcal{C}$ is the set of maximal cliques on an undirected graph $\mathcal{G}$. Remember that the Hammersley-Clifford theorem says that any distribution that factors in this way satisfies the Markov property on the graph and conversely that if p is strictly positive and satisfies the Markov property for $\mathcal{G}$ then it factors as above. Evidently, any distribution can be expressed using fully connected undirected graphical model, since this corresponds to a single potential involving all the variables—i.e., a joint probability distribution. Undirected graphical models efficiently represent conditional independencies in the distribution of interest, which are expressed by the removal of edges, and which similarly reduces the complexity of inference.

4.1 Factor graphs

Factor graphs are capable of capturing structure that the traditional directed and undirected graphical models above are not capable of capturing. A factor graph consists of a vector of random variables $\mathbf{x} = (x_1, \ldots, x_N)$ and a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$, which in addition to normal nodes also has factor nodes $\mathcal{F}$. Furthermore, the graph is constrained to be a bipartite graph between variable nodes and factor nodes.

The joint probability distribution associated to a factor graph is given by

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{j=1}^{m} f_j(x_{f_j}).$$

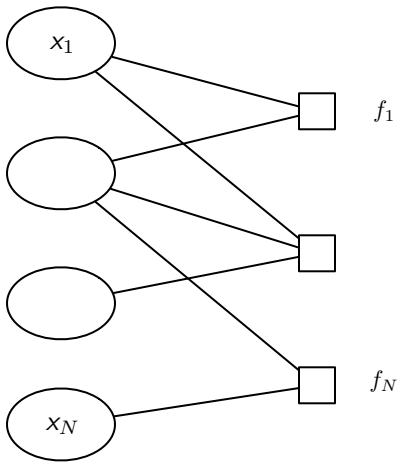

Figure 1: A general factor graph.

For example, in Figure 1, f_1 is a function of x_1 and x_2.

What constraints are imposed on the factors? The factors must be non-negative, but otherwise we're free to choose them. We could of course roll the partition function Z into one of the factors and that would constrain one of the factors.

The factor graph is not constrained to have factors only for maximal cliques, so we can more explicitly represent the factorization of the joint probability distribution.

It is very easy to encode certain kinds of (especially algebraic) constraints in the factor graph. One example is the Hamming code example from the first day of class, which you will see more of later in the subject. As a more basic example, consider the following.

Example 1. Suppose we have random variables representing taxes (x_1), social security (x_2), medicare (x_3), and foreign aid (x_4) with constraints

$$\begin{aligned} x_1 &\leq 3 \\ x_2 &\leq 0.5 \\ x_3 &\leq 0.25 \\ x_4 &\leq 0.01 \end{aligned}$$

and finally we need to decrease spending by 1, so $x_1 + x_2 + x_3 + x_4 \geq 1$. If we were interested in picking uniformly among the assignments that satisfy the constraints, we could encode this distribution conveniently with a factor graph in Figure 2.

The resulting distribution is given by

$$p_{\mathbf{x}}(\mathbf{x}) \propto \mathbb{1}_{x_1 \leq 3} \mathbb{1}_{x_2 \leq 0.5} \mathbb{1}_{x_3 \leq 0.25} \mathbb{1}_{x_4 \leq 0.01} \mathbb{1}_{x_1 + x_2 + x_3 + x_4 \geq 1}$$

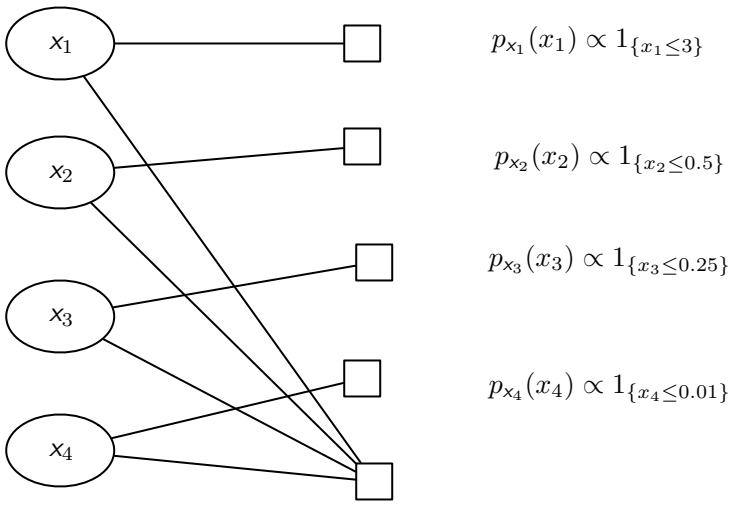

Figure 2: Factor graph encoding constraints on budget.

4.2 Converting Between Graphical Models Types

4.2.1 Converting Undirected Models to Factor Graphs

We can write down the probability distribution associated to the undirected graph

$$p_{\mathbf{x}}(\mathbf{x}) \propto f_{134}(x_1, x_3, x_4) f_{234}(x_2, x_3, x_4)$$

which naturally gives a factor graph representation using the potentials as factor nodes (Figure 3). In general, we can convert an undirected graphical model into a factor graph by defining a factor node for each maximal clique.

How many maximal cliques could an undirected graph have? We can construct an example where the number of maximal cliques scales like n^2 where n is the number of nodes in the undirected graph.

Consider a complete bipartite graph where the nodes are evenly split. Given any 3 nodes, 2 of the nodes must lie on the same side and hence be disconnected. Thus, there can't be any 3 node cliques, so all of the edges are maximal cliques, and there are $O(n^2)$ edges. In general there can be exponentially many maximal cliques in an undirected graph (See Problem Set 2).

4.2.2 Converting Factor Graphs to Directed Models

Take a topological ordering of the nodes say $\mathsf{x}_1, \ldots, \mathsf{x}_n$. For each node in turn, find a minimal set $U \subset \{x_1, \ldots, x_{i-1}\}$ such that $x_i \perp\!\!\!\perp \{x_1, \ldots x_{i-1}\} - U | U$ is satisfied and set x_i's parents to be the nodes in U. This amounts to reducing in turn each $p(x_i | x_1, \ldots, x_{i-1})$ as much as possible using the conditional independencies implied by the factor graph. An example is done in Figure 5.

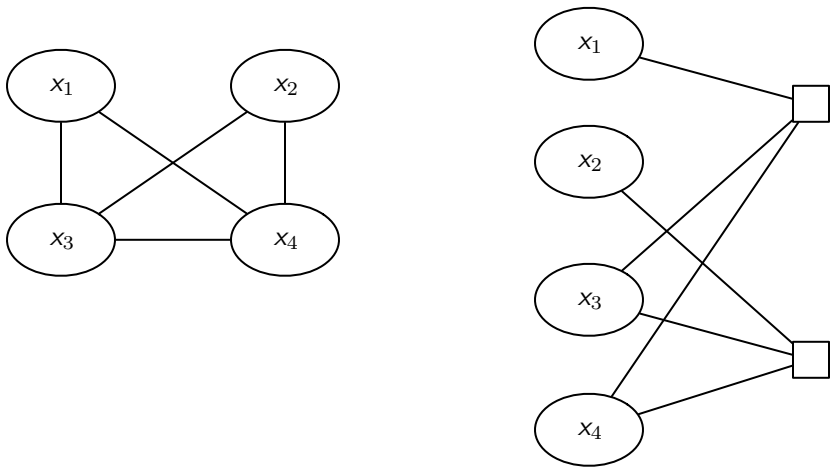

Figure 3: Representing an undirected graph as a factor graph.

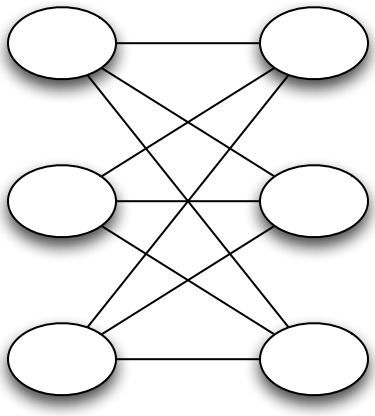

Figure 4: Complete bipartite graph has $O(n^2)$ maximal cliques.

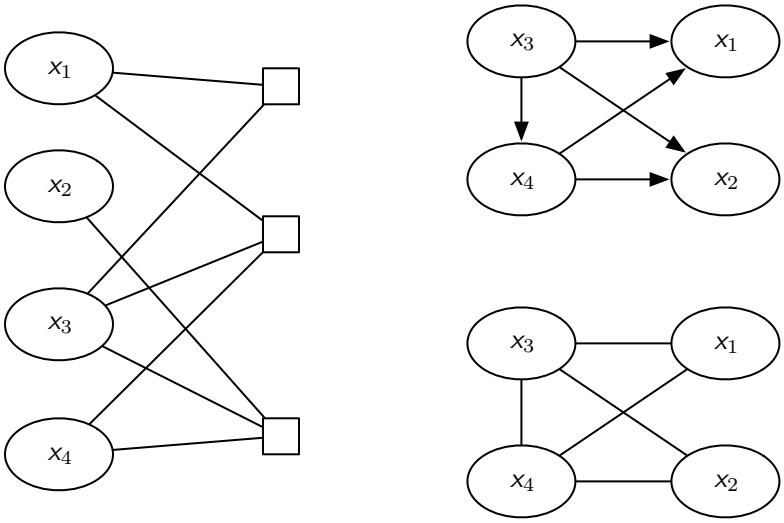

Figure 5: Converting a factor graph into a directed graph and then into an undirected graph.

4.2.3 Converting from Directed to Undirected Models

This is done through *moralization*, which says to completely connect the parents of each node and then remove the direction of arrows.

Moralization "marries" the parents by connecting them together. See Figure 5 for an example.

The important thing to recognize is that the conversion process is not lossless. In these constructions, any conditional independence implied by the converted graph is satisfied by the original graph. However, in general, some of the conditional independences implied in the original graph are not implied by the converted graph. How do we know that we've come up with "good" conversions. We want the converted graph to be close to the original graph for some definition of closeness. We'll explore these notions through I-maps, D-maps, and P-maps.

4.3 Measuring Goodness of Graphical Representations

4.3.1 I-map

Consider a probability distribution D and a graphical model $\mathcal{G}$. Let $CI(D)$ denote the set of conditional independencies satisfied by D and let $CI(\mathcal{G})$ denote the set of all conditional independencies implied by $\mathcal{G}$.

Definition 1 (I-map)**.** *We say $\mathcal{G}$ is an* independence map *or* I-map *for D if $CI(\mathcal{G}) \subset CI(D)$. In other words, every conditional independence implied by $\mathcal{G}$ is satisfied by*

Figure 6: Bot (i.e. $p_{x,y} = p_x p_y$).

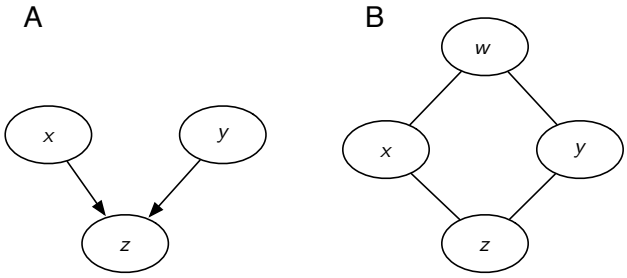

Figure 7: Two example graphs

D if $\mathcal{G}$ is an I-map for D.

The complete graph is always an example of an I-map for any distribution because it implies no conditional independencies.

4.3.2 D-map

Definition 2 (D-map)**.** *We say $\mathcal{G}$ is a dependence map or* D-map *for D if $CI(\mathcal{G}) \supset CI(D)$ that is every conditional independence that D satisfies is implied by $\mathcal{G}$.*

The graph with no edges is an example of a D-map for any distribution because it implies every conditional independence.

4.3.3 P-map

Definition 3 (P-map)**.** *We say $\mathcal{G}$ is a perfect map or* P-map *for D if $CI(\mathcal{G}) = CI(D)$, i.e., if every conditional independence implied by $\mathcal{G}$ is satisfied by D and vice versa.*

Example 2. Consider three distributions that factor as follows

$$p_1 = p_x p_y p_z$$
$$p_2 = p_{z|x,y} p_x p_y$$
$$p_3 = p_{z|x,y} p_{x|y} p_y$$

Then the graph in Figure 7(a) is an I-map for p_1, a D-map for p_3 and a P-map for p_2. The graph in Figure 7(b) is an I-map for

$$p(x, y, w, z) = \frac{1}{Z} f_1(x, w) f_2(w, y) f_3(z, y) f_4(x, z)$$

by the Hammersley-Clifford theorem.

6.438 Algorithms for Inference

Fall 2014