6 Parameter Estimation in Directed and Undirected Graphical Models

In the first phase of this subject, we developed how to represent structure in distributions via graphical models. In the second phase, we will focus on how to learn graphical model representations from data.

6.1 Introductory Concepts and Perspectives

We begin with some basic principles, perspectives, and comments.

At the outset, it is important to emphasize that domain knowledge plays an important role in graphical model building. In particular, available application specific knowledge and the underlying governing principles of the discpline—such as from physics, chemistry, biology, economics, etc—can inform us of important aspects of the inherent structure in a model. For example, in the case of error-control decoding, the graphical model representing the code is prescribed by the system designer, and the channel model describing the error behavior is dictated by the physical properties of the measurement mechanism.

From the above perspective, the object of learning is to use data to reveal the remaining aspects. In this sense, learning should be thought of a method more of last rather than first resort.

It is equally important to emphasize that graphical model learning is a comparatively less mature topic, and as such an active area of research. As a result, our treatment is more introductory in nature, with some representative results, rather than a comprehensive treatment.

Our treatment will emphasize efficient learning methods. However, there are two kinds of efficiency that arise in the context of learning. In particular, in addition to the kind of *computational efficiency* that we have emphasized in earlier topics, we are also concerned with *sample efficiency*, which characterizes the number of samples of "training" data we need to learn a model to a desired level of accuracy.

Note, too, that both the relevant measure accuracy—as well as the required degree of such accuracy—in general depend strongly on the application and the associated inference task. For example, in an application involving self-driving cars, we might care about how closely posterior beliefs from the learned model match those from the true one. However, in a phylogenetics application where a tree represents evolutionary relationships among species, we might care about whether the learned edges in the associated model are correct with high probability.

When learning a distribution with factorization structure, there is a hierarchy of learning problems depending on the degree of prior knowledge of the distribution:

1. Given the factorization structure, learn (i.e., estimate) the parameters of the factors.

2. Given the relevant variables and their alphabets, learn the factorization structure (and, in turn, the parameters of the factors).

3. Given the variables, learn their alphabets, and, in turn, their (factorized) joint distribution.

4. Learn the number of variables in the system, their alphabets, and their joint distribution.

Our development will focus on the first two of these, which are building blocks and convey several of the key ideas. The latter two are scenarios for which only relatively weak inference is possible. More generally, it should be stressed that some degree of prior knowledge is invariably necessary: in the absence of any constraints, no model can be learned and no inference is possible.

It should also be emphasized that learning tasks can be loosely categorized into two classes: *supervised* learning and *unsupervised* learning. In supervised learning, training data is available for all the variables of interest. In unsupervised learning, training is available for only some of the variables of interest. Our initial development will focus on supervised learning, to develop the basic concepts.

6.2 Maximum Likelihood Parameter Estimation

Consider the problem of estimating the distribution of an arbitrary discrete random variable x over some alphabet $\mathcal{X}$. For example, this could represent a variable associated with individual node in a graphical model, or it could represent the collection of all variables in the graphical model viewed as a single supervariable, or it could represent any other subset of such variables in between.

We write the distribution for x as $p(\cdot;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ for some set Θ to reflect that it is parameterized by $\boldsymbol{\theta}$.

Example 1. Consider a Bernoulli random variable x over alphabet $\mathcal{X} = \{0, 1\}$ where $\mathbb{P}(\mathsf{x} = 1) = \theta$. Then

$$p(x;\theta) = \theta^{\mathbb{1}_{x=1}}(1-\theta)^{\mathbb{1}_{x=0}} = \begin{cases} \theta & x = 1 \\ 1-\theta & x = 0. \end{cases}$$

We treat $\boldsymbol{\theta}$ as nonrandom, but unknown. We can view $p(x;\boldsymbol{\theta}')$ either as a function of x or $\boldsymbol{\theta}'$. In particular, for a given $\boldsymbol{\theta}' \in \Theta$, the distribution $p(\cdot;\boldsymbol{\theta}')$ represents the *model* for the variable x. At the same time, for a given observed sample $x \in \mathcal{X}$ we refer to $p(x;\cdot)$ as the *likelihood function*, for which the notation $L(\cdot;x) \triangleq p(x;\cdot)$ is convenient.

In this setting, our (learning) goal is to estimate $\boldsymbol{\theta}$ based on a collection of K (independent) samples

$$\mathcal{D} = \{x^1, \ldots, x^K\}. \tag{1}$$

drawn from the distribution $p(\cdot; \boldsymbol{\theta})$. We refer to $\mathcal{D}$ as the *data*, and denote the resulting estimate using $\hat{\boldsymbol{\theta}}$.

A widely used approach to estimating unknown parameters from observations it to choose the parameter values so as to make the likelihood as large as possible. This is referred to as *maximum likelihood (ML)* estimation

Accordingly, if we have a single observation x^1, corresponding to $K = 1$, then the ML estimate is

$$\hat{\boldsymbol{\theta}}^{\mathrm{ML}}(x^1) = \arg\max_{\boldsymbol{\theta}} p(x^1; \boldsymbol{\theta}).$$

When $K > 1$, the likelihood function takes the form

$$L(\boldsymbol{\theta}; \mathcal{D}) = L(\boldsymbol{\theta}; x^1, \ldots, x^K) = \prod_{k=1}^{K} p(x^k; \boldsymbol{\theta}). \tag{2}$$

where the right-hand side is the joint distribution of the (independent) samples.

In practice, it is often convenient to work in the logarithmic domain, operating on the (normalized) log-likelihood

$$\ell(\boldsymbol{\theta}; \mathcal{D}) \triangleq \frac{1}{K} \ln L(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} \ln p(x^k; \boldsymbol{\theta}). \tag{3}$$

Note that since the observations (20) are i.i.d., their joint distribution is permutation invariant. Hence, an equivalent representation for the data is the relative frequency of occurence of each symbol $a \in \mathcal{X}$, which is expressed by its *empirical distribution*[1]

$$\hat{p}(a) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{x^k = a}, \quad a \in \mathcal{X}$$

Note that in this representation $K\hat{p}(a)$ is the number of times the symbol a appears in the data. As a result, using the convenient identity

$$p(x^k; \boldsymbol{\theta}) = \prod_{a \in \mathcal{X}} p(a; \boldsymbol{\theta})^{\mathbb{1}_{x_k = a}}$$

[1]Note that the empirical distribution is a more compact representation for the data in this case, replacing the K-dimensional data with an $|\mathcal{X}| - 1$-dimensional empirical distribution. This is an instance of what is referred to as a *sufficient statistic* for the data. Also, our notation does not make explicit the dependency on the data. However, when necessary, we can express this dependency via the notation $\hat{p}(\cdot; \mathcal{D})$.

we can rewrite the log-likelihood function in terms of $\hat{p}(\cdot)$ as

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathcal{D}) &= \frac{1}{K}\sum_{k=1}^{K} \ln p(x^k; \boldsymbol{\theta}) \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{a\in\mathcal{X}} \mathbb{1}_{x^k=a} \ln p(a; \boldsymbol{\theta}) \\ &= \sum_{a\in\mathcal{X}} \ln p(a; \boldsymbol{\theta}) \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}_{x^k=a} \\ &= \sum_{a\in\mathcal{X}} \hat{p}(a) \ln p(a; \boldsymbol{\theta}). \end{aligned} \tag{4}$$

Example 2. Consider the case of a Bernoulli distribution with unknown parameter $\theta = p(1)$. Then given K samples of data, $\hat{p}(0)$ is the fraction of 0's in a data, and $\hat{p}(1) = 1 - \hat{p}(0)$ the fraction of 1's. Applying (4), the likelihood function is given by

$$\ell(\theta; \mathcal{D}) = \hat{p}(1) \ln \theta + \hat{p}(0) \ln(1 - \theta).$$

We find the maximum of this concave function by solving the stationary point equation

$$\frac{\partial \ell}{\partial \theta} = \frac{\hat{p}(1)}{\theta} - \frac{\hat{p}(0)}{(1-\theta)} = 0,$$

i.e.,

$$\frac{\hat{p}(1)}{1 - \hat{p}(1)} = \frac{\theta}{1-\theta},$$

whose solution is

$$\hat{\theta}^{\mathrm{ML}} = \hat{p}(1),$$

which is intuitively pleasing.

Example 3. Generalizing the preceding example, consider a fully-parameterized distribution p over an arbitrary finite alphabet

$$\mathcal{X} = \{a_1, \ldots, a_L\}. \tag{5}$$

Then $p(x; \boldsymbol{\theta})$ is a *categorical* distribution, i.e.,[2]

$$p(a_l; \boldsymbol{\theta}) = \theta_l, \quad l = 1, \ldots, L.$$

[2]The categorical distribution is sometimes referred to as a "multinoulli" distribution as it generalizes the Bernoulli distribution to larger alphabets. Independent samples from a categorical distribution produces a multinomial distribution, just as such samples from a Bernoulli distribution produce a binomial one. More specifically, for samples drawn from a categorical distribution, their representation in the form of an unnormalized empirical distribution vector $(K\hat{p}(a_1), \ldots, K\hat{p}(a_L))$ is multinomial distributed.

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$, which obviously satisfy the constraints

$$\sum_{l=1}^{L} \theta_l = 1 \qquad \text{and} \qquad \theta_l \in [0, 1]. \tag{6}$$

Specializing (4) to this case we obtain

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{l=1}^{L} \hat{p}(a_l) \ln \theta_l.$$

Expressing the normalizing constraints using Lagrange multipliers, only the first of which is active, we have

$$\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \left[\ell(\boldsymbol{\theta}; \mathcal{D}) + \lambda \Big(\sum_{l=1}^{L} \theta_l - 1 \Big) \right]$$

The stationary points then yield

$$\frac{\partial \ell}{\partial \theta_l} = \frac{\hat{p}(a_l)}{\theta_l} - \lambda = 0,$$

where λ is chosen to ensure (6) are satisfied, whence

$$\theta_l = \hat{p}(a_l), \qquad l = 1, \ldots, L.$$

Hence, the empirical distribution itself is the ML estimate of an unknown distribution. Furthermore, by the strong law of large numbers, the empirical distribution converges with probability one to the true distribution, i.e.,

$$\hat{p}(a) \xrightarrow{\text{a.s.}} p(a), \quad \text{as } K \to \infty, \quad \text{for all } a \in \mathcal{X} \ ,$$

so the ML estimate has a desirable asymptotic property.[3]

6.2.1 ML Estimation as Empirical Distribution Matching

ML estimation has another useful interpretation as well, which sheds additional light on the preceding example. In particular, note that we can rewrite the log-likelihood (4) in terms of entropy and information divergence via

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathcal{D}) &= \sum_{a \in \mathcal{X}} \hat{p}(a) \ln p(a; \boldsymbol{\theta}) \\ &= \mathbb{E}_{\hat{p}} \left[\ln p(\mathsf{x}; \boldsymbol{\theta}) \right] \\ &= \mathbb{E}_{\hat{p}} \left[\ln \hat{p}(\mathsf{x}) \right] - \mathbb{E}_{\hat{p}} \left[\ln \frac{\hat{p}(\mathsf{x})}{p(\mathsf{x}; \boldsymbol{\theta})} \right] \\ &= -H(\hat{p}) - D\big(\hat{p} \| p(\cdot; \boldsymbol{\theta})\big). \end{aligned} \tag{7}$$

[3]This property of an estimator is termed *consistency*.

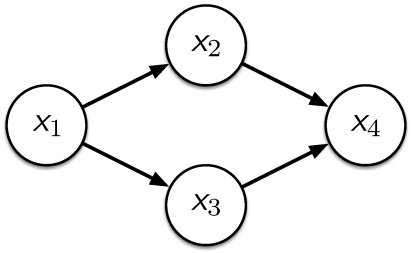

Figure 1: An example directed graphical model where the conditional distributions associated with each of the four nodes are parameterized by four sets of parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4$, respectively.

In this form, we see that the entropy term does not depend on the parameter $\boldsymbol{\theta}$, and so

$$\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{D}) = \arg\min_{\boldsymbol{\theta}} D\big(\hat{p} \| p(\cdot|\boldsymbol{\theta})\big). \tag{8}$$

Hence, the ML parameter estimate $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$ is the one that makes $p(\cdot|\hat{\boldsymbol{\theta}}^{\mathrm{ML}})$ as close as possible (in a divergence sense) to the empirical distribution. We refer to this as *empirical distribution matching*.

From this perspective, we see that Example 3 is an instance where the parameterization is such that the empirical distribution can be matched *exactly*, i.e., we can achieve $D(\hat{p} \| p(\cdot|\hat{\boldsymbol{\theta}}^{\mathrm{ML}})) = 0$. This happens whenever the distributions of interest are "fully parameterized."

More generally, however, such exact matching is not possible, and the partially-parameterized case corresponds to looking for the best match over the subset of possible distributions defined by $\boldsymbol{\theta}$. We explore an instance of such a scenario next.

6.3 Learning Directed Graphical Models Parameters

Let us now consider learning the parameters of a directed graphical model corresponding to a given graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Since the edges and nodes are given, the joint distribution for $\mathsf{x}_{\mathcal{V}}$ is constrained, or equivalently the distribution of the supervariable $\mathsf{z} = \mathsf{x}_{\mathcal{V}}$ is partially parameterized.

In some important cases of interest, it is straightforward to obtain ML estimates. To illustrate this, consider the example of a directed graphical model depicted in Fig. 1, which corresponds to the factorization

$$p(x_1, \ldots, x_4) = p_{\mathsf{x}_1}(x_1)\, p_{\mathsf{x}_2|\mathsf{x}_1}(x_2|x_1)\, p_{\mathsf{x}_3|\mathsf{x}_1}(x_3|x_1)\, p_{\mathsf{x}_4|\mathsf{x}_2,\mathsf{x}_3}(x_4|x_2, x_3). \tag{9}$$

Suppose that we no nothing about any of these four constituent conditional probability distributions (except the alphabets over which they are defined). Then

we need four sets of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4)$ to represent all possible joint distributions with the factorization (9). In particular, each $\boldsymbol{\theta}_i$ defines a categorical distribution corresponding to each joint assignment for the parents of x_i, so the log-likelihood takes the form

$$\ell(\mathbf{x}; \boldsymbol{\theta}) = \ln p_{\mathsf{x}_1}(x_1; \boldsymbol{\theta}_1) + \ln p_{\mathsf{x}_2|\mathsf{x}_1}(x_2|x_1; \boldsymbol{\theta}_2) + \ln p_{\mathsf{x}_3|\mathsf{x}_1}(x_3|x_1; \boldsymbol{\theta}_3) + \ln p_{\mathsf{x}_4|\mathsf{x}_2,\mathsf{x}_3}(x_4|x_2, x_3; \boldsymbol{\theta}_4).$$

Because each $\boldsymbol{\theta}_i$ appears in exactly one of these terms, the optimizations over $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4)$ are decoupled: each of the terms can be optimized separately. Moreover, since each individual term is simply the likelihood function for observations of a categorical distribution, the ML parameter estimate is the associated empirical conditional distribution. For example,

$$\boldsymbol{\theta}_2 = \hat{p}_{\mathsf{x}_2|\mathsf{x}_1}(\cdot|\cdot),$$

where

$$\hat{p}_{\mathsf{x}_2|\mathsf{x}_1}(\cdot|\cdot) = \frac{\hat{p}_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2)}{\hat{p}_{\mathsf{x}_1}(x_1)} = \frac{\sum_{x_{\mathcal{V}\setminus\{1,2\}}} \hat{p}(x_{\mathcal{V}})}{\sum_{x_{\mathcal{V}\setminus\{1\}}} \hat{p}(x_{\mathcal{V}})},$$

with $\hat{p}(\cdot)$ denoting the empirical distribution of the supervariable $\mathsf{z} = \mathsf{x}_{\mathcal{V}}$.

Remarks

Two aspects of our formulation enabled the convenient decoupling of parameter estimation in directed graphical models. First, we assumed the different conditional probability distributions do not have parameters in common. In practice, this is not always the case. As an example, consider learning the transition distribution of a homogeneous Markov chain. Here all edges in the associated graphical model share the same parameters. In such cases, the ML estimation problem is often still fairly straightforward, despite the coupling.

However, the second aspect of our formulation is our assumption that samples of the variables at all nodes in the graph were available, which corresponds to the case of supervised learning. When we develop unsupervised learning, we will see how the absence of complete observations leads to coupling in the optimization problem for the ML parameters, even when there are no shared parameters. An example of such a scenario would be learning the parameters of a hidden Markov model. In such cases, special techniques can be developed to obtain the desired estimates, as we will see in a subsequent installment of the notes.

6.4 Learning Undirected Graphical Model Parameters

As we developed, learning the parameters of a directed graphical model is relatively straightforward when there no coupling among parameters between different conditional probability distributions associated with the factorization of the distribution.

We now turn to the problem of learning the parameters of undirected graphical models. At first glance, this is a comparatively more difficult problem. To see this, recall that for given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the corresponding factorization of the joint distribution over variables $x_{\mathcal{V}}$ can be expressed in the form

$$p_{x_{\mathcal{V}}}(x_{\mathcal{V}}) = \frac{1}{Z} \exp \left\{ \sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}) \right\}, \tag{10}$$

where $\mathrm{cl}^*(\mathcal{G})$ denotes, as usual, the maximum cliques of $\mathcal{G}$, and where $\tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}$ denote the log-potentials.

Now analogous to our treatment for directed graphical models, suppose we assign a separate set of parameters $\boldsymbol{\theta}_{\mathcal{C}}$ to each log-potential $\tilde{\psi}_{\mathcal{C}}$. Then we can rewrite (10) in the form

$$p_{x_{\mathcal{V}}}(x_{\mathcal{V}}) = \frac{1}{Z(\boldsymbol{\Theta})} \exp \left\{ \sum_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \tilde{\psi}_{\mathcal{C}}(x_{\mathcal{C}}; \boldsymbol{\theta}_{\mathcal{C}}) \right\}, \tag{11}$$

where

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_C), \qquad C = |\,\mathrm{cl}^*(\mathcal{G})|, \tag{12}$$

and where our notation for the partition function makes explicit its dependence on the parameters (12) of all the log-potentials.

As (11) reveals, the difficulty in parameter estimation in this case is that the dependence of the partition function on all the parameters in the model means that we cannot solve for each $\boldsymbol{\theta}_{\mathcal{C}}$ separately (such as by using the associated marginals $p_{x_{\mathcal{C}}}$). Without some degree of decoupling, the parameter estimation problem is computationally difficult.

Our approach to addressing this problem is to recognize that there are many different equivalent parameterization of a distribution, and that some are more natural than others for the development of efficient learning algorithms.

6.5 Parameter Transformations

To begin, suppose we have an arbitrary distribution $p(\cdot; \boldsymbol{\theta})$ over some alphabet $\mathcal{X}$. Then as we developed earlier, the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$ based on training data $\mathcal{D} = (x^1, \ldots, x^K)$ is given by

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{D}),$$

where

$$\ell(\boldsymbol{\theta}; \mathcal{D}) = \sum_{a \in \mathcal{X}} \hat{p}(a) \ln p(a; \boldsymbol{\theta}).$$

Now suppose that we consider a different but equivalent parameterization $\boldsymbol{\xi}$, i.e., there exists an invertible function $\mathbf{g}(\cdot)$ such that $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\xi})$. Then

$$\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{D}) = \max_{\boldsymbol{\xi}} \ell(\mathbf{g}(\boldsymbol{\xi}); \mathcal{D}), \tag{13}$$

and

$$\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{g}(\hat{\boldsymbol{\xi}}^{\mathrm{ML}}), \tag{14}$$

i.e., ML estimation is invariant to (invertible) reparameterizations.

6.6 Canonical Parameterization

We now develop an invertible transformation of the parameterization (11) that is well matched to the parameter estimation task.

We illustrate the methodology for graphs whose node variables are defined over a binary alphabet $\mathcal{X} = \{0, 1\}$, but note in advance that larger alphabets are readily accommodated. In particular, if our alphabet $\mathcal{X} = \{0, \ldots, 2^M - 1\}$ for some integer, then we can treat any variable $\mathsf{x} \in \mathcal{X}$ as a supervariable $\mathsf{x} = (\mathsf{u}_1, \ldots, \mathsf{u}_M)$ where $\mathsf{u}_i \in \mathcal{U} = \{0, 1\}$ is the ith bit in the binary expanion of x.

We focus on complete parameterizations, so the parameters are the collection of tables representing the collection of the (log) maximal clique potentials in (10). As our canonical parameterization, we rewrite (10) in the form

$$p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) = \exp\left\{ \bar{Z} + \sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \bar{\psi}_\mathcal{C}(x_\mathcal{C}) \right\}, \tag{15}$$

where the factorization over all cliques $\mathrm{cl}(\mathcal{G})$, and where we constrain the form of the constituent log-potentials such that

$$\bar{\psi}_\mathcal{C}(x_\mathcal{C}) = 0 \quad \text{when} \quad x_\mathcal{C} \neq \mathbf{1}, \qquad \text{for all } \mathcal{C} \in \mathrm{cl}(\mathcal{G}), \tag{16}$$

where $\mathbf{1}$ denotes the sequence of all 1's.

We now verify that this is possible. It suffices to show that for any maximal clique $\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})$ we can find a one-to-one mapping of the form

$$\tilde{\psi}_\mathcal{C}(x_\mathcal{C}) = \bar{Z}_\mathcal{C} + \sum_{\mathcal{S}:\, \mathcal{S} \subset \mathcal{C}} \bar{\psi}_\mathcal{S}(x_\mathcal{S}), \tag{17}$$

since if (17) holds for each maximal clique, then when we combine the lower-order log clique potentials via (10) their sum will continue to satisfy the constraint (16).

For pairwise maximum clique potentials, this mapping is easily expressed. To see this, without loss of generality, suppose our clique potential is $\tilde{\psi}_{12}(x_1, x_2)$. Then, the

mapping is

$$\begin{aligned}
\bar{Z} &= \tilde{\psi}_{12}(0,0) \\
\bar{\psi}_1(1) &= \tilde{\psi}_{12}(1,0) - \tilde{\psi}_{12}(0,0) \\
\bar{\psi}_2(1) &= \tilde{\psi}_{12}(0,1) - \tilde{\psi}_{12}(0,0) \\
\bar{\psi}_{12}(1,1) &= \tilde{\psi}_{12}(1,1) - \tilde{\psi}_{12}(1,0) - \tilde{\psi}_{12}(0,1) + \tilde{\psi}_{12}(0,0).
\end{aligned}$$

As this makes clear there is an invertible linear mapping between the 4 parameters in each of the two representations, viz.,

$$\begin{bmatrix} \bar{Z} \\ \bar{\psi}_1(1) \\ \bar{\psi}_2(1) \\ \bar{\psi}_{12}(1,1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\psi}_{12}(0,0) \\ \tilde{\psi}_{12}(1,0) \\ \tilde{\psi}_{12}(0,1) \\ \tilde{\psi}_{12}(1,1) \end{bmatrix}.$$

For higher-order cliques, we construct the transformation iteratively. Again without loss of generality, suppose our maximal clique potential is $\tilde{\psi}_{\mathcal{C}}(x_1, \ldots, x_C)$. Then first we must have

$$\tilde{Z} = \tilde{\psi}_{\mathcal{C}}(\mathbf{0}),$$

where $\mathbf{0}$ denotes the sequence of all 0's.

Next, consider configurations of the form $\mathbf{e}_i$, which is a sequence that is all 0's except for a 1 in the ith position. In this case, all log potentials $\bar{\psi}_{\mathcal{S}}$ must be zero except for singleton ones due to our constraint. Hence,

$$\bar{\psi}_i(1) = \tilde{\psi}_{\mathcal{C}}(\mathbf{e}_i) - \tilde{Z}, \qquad i = 1, \ldots, C.$$

There are $\binom{C}{1}$ such singleton log potentials.

Proceeding, we consider configurations of the form $\mathbf{e}_{ij}$, which is a sequence that is all 0's except for 1's in the ith and jth positions. Now all log potentials $\bar{\psi}_{\mathcal{S}}$ must be zero except for singleton and pairwise ones, but the singleton log potentials are already fixed, whence

$$\bar{\psi}_{ij}(1,1) = \tilde{\psi}_{\mathcal{C}}(\mathbf{e}_{ij}) - \bar{\psi}_i(1) - \bar{\psi}_j(1) - \tilde{Z}, \qquad i, j = 1, \ldots, C.$$

There are $\binom{C}{2}$ such pairwise log potentials.

We continue this process, considering configurations of the form $\mathbf{e}_{ijk}$ with 1's at only ith, jth, and kth positions, resulting in

$$\begin{aligned}
\bar{\psi}_{ijk}(1,1,1) = \tilde{\psi}_{\mathcal{C}}(\mathbf{e}_{ijk}) &- \bar{\psi}_{ij}(1,1) - \bar{\psi}_{ik}(1,1) - \bar{\psi}_{jk}(1,1) \\
&- \bar{\psi}_i(1) - \bar{\psi}_j(1) - \bar{\psi}_k(1) - \tilde{Z}, \qquad i, j, k = 1, \ldots, C,
\end{aligned}$$

and continue to configurations with progressively more 1's, using previously generated log potentials, until we have solved for all the potentials.

Hence, we obtain all the $\bar{\psi}_{\mathcal{S}},\ \mathcal{S} \subset \mathcal{C}$ from $\tilde{\psi}_{\mathcal{C}}$. The total number of parameters we obtain is

$$\sum_{c=0}^{C} \binom{C}{c} = 2^C,$$

which is precisely the number of entries in the table describing $\tilde{\psi}_{\mathcal{C}}$. Moreover, by reversing the process we can reconstruct the $\tilde{\psi}_{\mathcal{C}}$ from the $\bar{\psi}_{\mathcal{S}},\ \mathcal{S} \subset \mathcal{C}$.

Finally, we note that the value of expressing the parameterization in canonical form is that it yields a convenient form for the log potentials. In particular, any log potential $\bar{\psi}_{\mathcal{C}}(x_{\mathcal{C}})$ satisfying the constraint (16) can be expressed in the form

$$\bar{\psi}_{\mathcal{C}}(x_{\mathcal{C}}) = \theta_{\mathcal{C}} \prod_{c \in \mathcal{C}} x_c, \tag{18}$$

where $\theta_{\mathcal{C}}$ denotes the canonical parameter. As a result, we can rewrite (15) in the canonical form

$$p_{x_{\mathcal{V}}}(x_{\mathcal{V}}; \boldsymbol{\theta}) = \frac{1}{Z} \exp\left\{ \sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \theta_{\mathcal{C}} \prod_{c \in \mathcal{C}} x_c \right\}, \tag{19a}$$

with

$$\boldsymbol{\theta} = \{\theta_{\mathcal{C}} \colon \mathcal{C} \in \mathrm{cl}(\mathcal{G})\}, \tag{19b}$$

which we now exploit.

6.7 Estimation of Canonical Parameters

We estimate the parameters $\boldsymbol{\theta}$ in the canonical form (19) based on training data consisting of K i.i.d. samples from p, which we denote using

$$\mathcal{D} = \{x_{\mathcal{V}}^1, \ldots, x_{\mathcal{V}}^K\}. \tag{20}$$

Directly applying the method of maximum likelihood is difficult. However, we will describe a method that uses the empirical distribution $\hat{p}_{x_{\mathcal{V}}}$, and exploit the fact that

$$\hat{p}_{x_{\mathcal{V}}}(a_{\mathcal{V}}) \xrightarrow{\text{a.s.}} p_{x_{\mathcal{V}}}(a_{\mathcal{V}}), \quad \text{as } K \to \infty, \qquad \text{for all } a_{\mathcal{V}} \in \mathcal{X}^{|\mathcal{V}|}. \tag{21}$$

6.7.1 Case: Pairwise Maximal Cliques

To develop the basic idea, we first restrict our attention to the case in which the cliques of $\mathcal{G}$ are of at most size 2. In this case, (19) specializes to

$$p(x_{\mathcal{V}}) = \frac{1}{Z} \exp\left\{ \sum_{i \in \mathcal{V}} \theta_i x_i + \sum_{(i,j) \in \mathcal{E}} \theta_{ij} x_i x_j \right\}, \tag{22}$$

with $\theta_i, \theta_{ij} \in \mathbb{R}$ for all $i \in \mathcal{V}$ and $(i,j) \in \mathcal{E}$, with Z the partition function.

Our goal is to produce estimates of the parameters

$$\boldsymbol{\theta} = \{\theta_i, i \in \mathcal{V} \quad \text{and} \quad \theta_{ij}, (i,j) \in \mathcal{E}\}$$

based on the training data (20).

From our development in Section 6.6 we see that

$$\theta_i = \ln p_{x_\mathcal{V}}(\mathbf{e}_i) - \ln p(\mathbf{0}) \tag{23}$$

where $\mathbf{e}_i$ and $\mathbf{0}$ are as defined earlier. Hence, we can estimate θ_i using

$$\hat{\theta}_i = \ln \hat{p}_{x_\mathcal{V}}(\mathbf{e}_i) - \ln \hat{p}_{x_\mathcal{V}}(\mathbf{0}). \tag{24}$$

Comparing (24) to (23) and using (21), we see that

$$\hat{\theta}_i \xrightarrow{\text{a.s.}} \theta_i \quad \text{as } K \to \infty, \qquad \text{for all } i \in \mathcal{V}, \tag{25}$$

so the estimator $\hat{\theta}_i$ is good at least for sufficiently large K.

In a similar manner, our development in Section 6.6 also establishes that

$$\theta_{ij} = \ln p_{x_\mathcal{V}}(\mathbf{e}_{ij}) - \ln p_{x_\mathcal{V}}(\mathbf{e}_i) - \ln p_{x_\mathcal{V}}(\mathbf{e}_j) + \ln p_{x_\mathcal{V}}(\mathbf{0}), \tag{26}$$

where $\mathbf{e}_{ij}$ is also as defined earlier. Accordingly, we estimate θ_{ij} using

$$\hat{\theta}_{ij} = \ln \hat{p}_{x_\mathcal{V}}(\mathbf{e}_{ij}) - \ln \hat{p}_{x_\mathcal{V}}(\mathbf{e}_i) - \ln \hat{p}_{x_\mathcal{V}}(\mathbf{e}_j) + \ln \hat{p}_{x_\mathcal{V}}(\mathbf{0}).\text{,} \tag{27}$$

Analogously comparing (26) to (27) and using (21) we see that

$$\hat{\theta}_{ij} \xrightarrow{\text{a.s.}} \theta_{ij} \quad \text{as } K \to \infty, \qquad \text{for all } i,j \in \mathcal{V}, \tag{28}$$

so the estimator $\hat{\theta}_{ij}$ is also good at least for sufficiently large K.

While the parameter estimates (24) and (27) are asymptotically good, note that the number of samples K required is extremely large. Indeed, for $\hat{p}_{x_\mathcal{V}}(a_\mathcal{V})$ to be close to $p_{x_\mathcal{V}}(a_\mathcal{V})$ for any particular configuration $a_\mathcal{V}$, the data $\mathcal{D}$ must include many realizations of the configuration. Letting $N = |\mathcal{V}|$, since there 2^N configurations, this means that in practice we would need $K = O(2^N)$ samples (i.e., an exponentially large data set) before reliable parameter estimates are obtained.

While the parameter estimation strategy just described has high sample complexity, this can often be dramatically improved by exploiting the conditional independence structure in the model. In particular, we will exploit that

$$x_i \perp\!\!\!\perp x_{\mathcal{V}\setminus\mathcal{N}(i)\cup\{i\}} \mid x_{\mathcal{N}(i)}, \qquad \text{for any } i \in \mathcal{V}, \tag{29}$$

where, as usual, $\mathcal{N}(i)$ denotes the neighbors of node i, i.e.,

$$\mathcal{N}(i) = \{j \in \mathcal{V} \colon (i,j) \in \mathcal{E}\}.$$

This implies, for example, that

$$p_{x_i|x_{\mathcal{N}(i)}}(\cdot|\mathbf{0}) = p_{x_i|x_{\mathcal{V}\setminus\{i\}}}(\cdot|\mathbf{0}).$$

Therefore, for any $i \in \mathcal{V}$,

$$\begin{aligned}
\ln p_{x_i|x_{\mathcal{N}(i)}}(1|\mathbf{0}) - \ln p_{x_i|x_{\mathcal{N}(i)}}(0|\mathbf{0}) &= \ln p_{x_i|x_{\mathcal{V}\setminus\{i\}}}(1|\mathbf{0}) - \ln p_{x_i|x_{\mathcal{V}\setminus\{i\}}}(0|\mathbf{0}) \\
&= \ln p_{x_i,x_{V\setminus\{i\}}}(1,\mathbf{0}) - \ln p_{x_i,x_{V\setminus\{i\}}}(0,\mathbf{0}) \\
&= \ln p_{x_\mathcal{V}}(\mathbf{e}_i) - \ln p_{x_\mathcal{V}}(\mathbf{0}) \\
&= \theta_i.
\end{aligned}$$

Thus, we can replace (24) with the estimator

$$\hat{\theta}_i = \ln \hat{p}_{x_i|x_{\mathcal{N}(i)}}(1|\mathbf{0}) - \ln \hat{p}_{x_i|x_{\mathcal{N}(i)}}(0|\mathbf{0}), \qquad i \in \mathcal{V}. \tag{30}$$

The estimator (30) will typically get close to θ_i with many fewer samples. Indeed, if $d_i - |\mathcal{N}(i)|$ is the degree of node i, then 2^{d_i+1} of the 2^N possible configurations will have $x_i = 0$ and $x_{\mathcal{N}(i)} = \mathbf{0}$, and likewise for $x_i = 1$ and $x_{\mathcal{N}(i)} = \mathbf{0}$. Thus, the sample complexity of estimating each θ_i is $O(2^{d_{\max}})$, where $d_{\max}$ is the maximum degree of the graph.

In a similar manner, we can exploit conditional independence structure in the distribution to improve on the estimator (27) of θ_{ij}, further replacing $p_{x_i,x_j|x_{\mathcal{V}\setminus\{i,j\}}}(1,1|\mathbf{0})$ with $p_{x_i,x_j|x_{\mathcal{N}(i)\cup\mathcal{N}(j)\setminus\{i,j\}}}(1,1|\mathbf{0})$. Here we find that estimating each θ_{ij} can be accomplished with a sample complexity of $O(2^{2d_{\max}})$. Thus, the overall sample complexity is $O(2^{2d_{\max}}) + O(2^{d_{\max}}) = O(2^{2d_{\max}})$, which is much lower than $O(2^N)$ when $d_{\max} \ll N$.

6.7.2 Beyond Pairwise Potentials

The extension of the foregoing estimation procedure to general models of the form (19) is conceptually straightforward, but is notationally slightly more cumbersome. In particular, for any clique $\mathcal{C} \in \mathrm{cl}(\mathcal{G})$,

$$\ln p_{x_\mathcal{C}|x_{\mathcal{N}(\mathcal{C})}}(\mathbf{1}|\mathbf{0}) - \ln p_{x_i|x_{\mathcal{N}(\mathcal{C})}}(0|\mathbf{0}) = \sum_{\mathcal{C}'\subset\mathcal{C}} \theta_{\mathcal{C}'}, \tag{31}$$

where $\mathcal{N}(\mathcal{C})$ denotes the neighbors of all nodes in $\mathcal{C}$. Hence, if we construct parameter estimates according to

$$\ln \hat{p}_{x_\mathcal{C}|x_{\mathcal{N}(\mathcal{C})}}(\mathbf{1}|\mathbf{0}) - \ln \hat{p}_{x_i|x_{\mathcal{N}(\mathcal{C})}}(0|\mathbf{0}) = \sum_{\mathcal{C}'\subset\mathcal{C}} \hat{\theta}_{\mathcal{C}'}, \tag{32}$$

they will converge to the correct values as $K \to \infty$. To solve for the parameters, we apply (32) iteratively starting with singleton cliques and progressing up to maximal cliques, using the earlier estimates in the construction of later ones. In this way, we obtain an estimation method with sample complexity $O(2^{d_\mathrm{c}})$, where d_c is the maximum over all cliques of the total number of neighbors of all vertices in a clique (including the vertices in the clique) for the graph $\mathcal{G}$.

6.8 Bayesian Parameter Estimation (optional)

We can interpret the problem formulation of Section 6.2 as one of *model selection*. In particular, the set of candidate models for the distribution of an variable x is

$$\{p(\cdot;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$$

for some set Θ. From this perspective, the method of maximum likelihood selects the model $p(\cdot;\hat{\boldsymbol{\theta}}^{\mathrm{ML}})$ based on the data (20), which can be used subsequently to make inferences about x.

A different strategy is to avoid explicitly selecting one of the models, but use a weighted combination of the candidate models when making inferences about x, where the weights are chosen in a suitable way according to the data.

One way to realize this kind of inference is based on a *Bayesian* formulation of the problem. In this formulation, we treat $\boldsymbol{\theta}$ as a random variable to which we assign some prior $p_{\Theta}(\boldsymbol{\theta})$, which is generally strictly positive but otherwise somewhat arbitrary. Then in the absence of data, the induced model for x is a mixture of the candidate models; specifically,

$$p_{\mathsf{x}}(x) = \int_{-\infty}^{+\infty} p_{\Theta}(\boldsymbol{\theta})\, p_{\mathsf{x}|\Theta}(x|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}., \tag{33}$$

where we recognize the weights as the prior $p_{\Theta}(\boldsymbol{\theta})$.

However, when we observe data (20) the induced mixture model for x becomes

$$\begin{aligned} p_{\mathsf{x}|\mathcal{D}}(x|\mathcal{D}) &= \int_{-\infty}^{+\infty} p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})\, p_{\mathsf{x}|\Theta,\mathcal{D}}(x|\boldsymbol{\theta},\mathcal{D})\, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{-\infty}^{+\infty} p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})\, p_{\mathsf{x}|\Theta}(x|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}, \end{aligned} \tag{34}$$

where we have used that $\mathsf{x} \leftrightarrow \Theta \leftrightarrow \mathcal{D}$ form a Markov chain, and where we see the weights are now the posterior $p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D})$. These relationships between x, Θ, and $\mathcal{D}$ are naturally expressed via the directed graph of Fig. 2.

The weight revision required for going from (33) to (34) is given by Bayes' Rule:

$$p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}) = \frac{p_{\Theta}(\boldsymbol{\theta})\, p_{\mathcal{D}|\Theta}(\mathcal{D}|\boldsymbol{\theta})}{\int_{\Theta} p_{\Theta}(\boldsymbol{\theta}')\, p_{\mathcal{D}|\Theta}(\mathcal{D}|\boldsymbol{\theta}')\, \mathrm{d}\boldsymbol{\theta}'}. \tag{35}$$

Clearly, for an arbitrary choice of prior, this weight revision is quite cumbersome. However, when the prior is chosen from a family that appropriately "matches" the form of the model family, the need to explicitly carry out the integration in (35) whenever new data is acquired can be avoided. Such priors are referred to as *conjugate priors*.

Given our emphasis on fully parameterized distributions, it suffices to focus on the class of conjugate priors for the categorical distribution, which are the *Dirichlet*

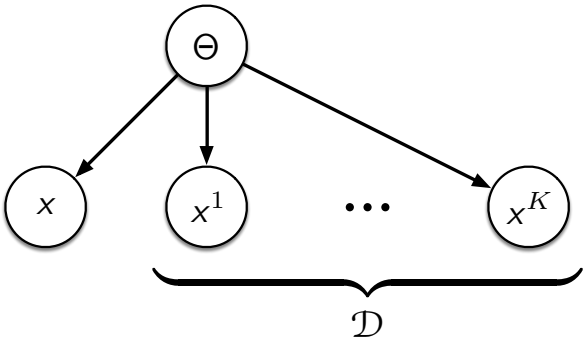

Figure 2: The directed graphical model expressing the factorization in the joint distribution for x, Θ, and $\mathcal{D}$.

priors.[4] In particular, we say p_Θ is a Dirichlet distribution with hyperparameters $\alpha_1, \ldots, \alpha_L$, for which we use the notation $\mathtt{D}(\alpha_1, \ldots, \alpha_L)$, when

$$p_\Theta(\boldsymbol{\theta}) \propto \prod_{l=1}^{L} \theta_l^{\alpha_l - 1}$$

for all $\boldsymbol{\theta}$ such that

$$0 \leq \theta_l \leq 1, \quad l \in \{1, \ldots, L\} \qquad \text{and} \qquad \sum_{l=1}^{L} \theta_l = 1,$$

and $p_\Theta(\boldsymbol{\theta}) = 0$ otherwise.

Indeed, with this prior, a straightforward evaluation of (35) reveals that

$$\Theta|\mathcal{D} \sim \mathtt{D}(\alpha_1 + K\hat{p}(a_1), \ldots, \alpha_L + K\hat{p}(a_L)).$$

where, as usual, our alphabet $\mathcal{X}$ is (5) and $\hat{p}(\cdot)$ denotes the empirical distribution of the data.

As a result, for such priors, we can express the revised mixture model (34) in the form

$$p_{\mathsf{x}|\mathcal{D}}(a_l|\mathcal{D}) = \frac{K\hat{p}(a_l) + \alpha_l}{K + \sum_{l=1}^{L} \alpha_l}. \tag{36}$$

[4]For the special case of the Bernoulli distribution, the specialization of the Dirichlet prior is referred to as the *beta* prior

$$p_\Theta(\theta) \propto \theta^{\alpha_2 - 1}(1-\theta)^{\alpha_1 - 1},$$

which we denote using $\mathtt{Beta}(\alpha_1, \alpha_2)$.

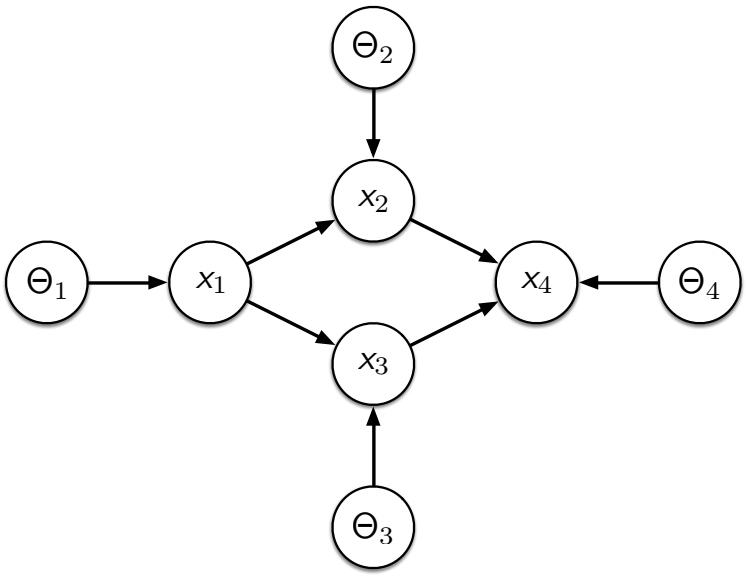

Figure 3: The graph of Fig. 1 augmented with parameter nodes to express the Bayes' parameter estimation setting.

Remarks

We conclude with a couple of comments. First, it should be emphasized that such Bayesian formulations are easy to use to learn distributions represented by parameterized graphical models. Indeed, it suffices to augment the graph with auxiliary nodes. For example, in the case of the graphical model of Fig. 1, we add four nodes representing the latent variables $\Theta_1, \ldots, \Theta_4$ since

$$\begin{aligned} &p(x_1, \ldots, x_4, \boldsymbol{\theta}) \\ &\qquad = p_{x_1|\Theta_1}(x_1|\boldsymbol{\theta}_1)\, p_{x_2|x_1,\Theta_2}(x_2|x_1, \boldsymbol{\theta}_2)\, p_{x_3|x_1,\Theta_3}(x_3|x_1, \boldsymbol{\theta}_3)\, p_{x_4|x_2,x_3,\Theta_4}(x_4|x_2, x_3, \boldsymbol{\theta}_4) \\ &\qquad\quad \cdot p_{\Theta_1}(\boldsymbol{\theta}_1)\, p_{\Theta_2}(\boldsymbol{\theta}_2)\, p_{\Theta_3}(\boldsymbol{\theta}_3)\, p_{\Theta_4}(\boldsymbol{\theta}_4), \end{aligned}$$

where we note that we have chosen the four parameter sets to be independent.

And as a final comment, note that when needed we can also use the Bayesian formulation to select a particular model, rather than using a mixture of them. Indeed, in such a setting the natural choice for a single model would be the MAP parameter estimate

$$\hat{\boldsymbol{\theta}}^{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p_{\Theta|\mathcal{D}}(\boldsymbol{\theta}|\mathcal{D}). \tag{37}$$

In the case of our categorical models with Dirichlet priors, it is straightforward to solve (37) explicitly, yielding

$$\hat{\theta}_l^{\text{MAP}} = \frac{K\hat{p}(a_l) + \alpha_l - 1}{K - L + \sum_{l=1}^{L} \alpha_l}.$$

Note that while, like the ML estimate, $\hat{\boldsymbol{\theta}}^{\text{MAP}}$ converges to the correct parameters with probability one as $K \to \infty$, it does so in a somewhat different manner. In particular, unlike the ML estimate, the MAP estimate relies less on the empirical distribution as an estimate of the unknown distribution when K is small, when the empirical distribution is less reliable. Instead, in this regime it relies more on the prior to estimate the model. In this sense, Bayesian parameter estimation can be seen as one way of penalizing models that are more complex that the available data can justify.