19 Learning Structure in Graphical Models

Thus far we have examined parameter estimation for a directed graphical model whose edge structure is given. We now turn to the problem of learning the edge structure itself in such models. In general, the degree to which such stucture can be learned, and efficient approaches to structure learning, are not yet well understood. As a result, our treatment will emphasize some broader principles, and some important special cases.

To begin, the problem of structure learning can be viewed as one of *model selection*. In particular, each possible configuration of edges in the graph represents one model, and thus our goal is to use training data to select an appropriate one. Conceptually, the model selection task can be broken down into two steps: 1) assigning a score to each of the candidate models (graphs) based on the data; and 2) finding the model with the highest score.

Example 1. The simplest example of a structure learning problem is an *independence test*. In particular, given samples of a pair of random variables (x, y), the goal of an independence test is to determine whether $\mathsf{x} \perp\!\!\!\perp \mathsf{y}$ or not. In the language of graphical models, this is equivalent to determining whether there is an edge or not between the nodes corresponding to x and y, respectively, in their graphical representation. This corresponds to a hypothesis test between the two graphical models $\mathcal{G}_0$ and $\mathcal{G}_1$ depicted in Fig. 1.

While at first glance, the general model selection procedure may appear straightforward (once the method of scoring has been chosen), it is important to appreciate the inherent complexity that is frequently encountered. Indeed, the number of possible graphs is typically enormous: for a directed graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = N$ random variables $\mathsf{x}_1, \ldots, \mathsf{x}_N$, the number of possible edges is

$$\binom{N}{2} = \frac{N(N-1)}{2},$$

and since an edge is either present or not in the graph, there are $2^{N(N-1)/2}$ possible graphs! However, as we will see, in some cases this apparent complexity can be avoided.

In the development that follows, we let each variable be from an alphabet $\mathcal{X}$ of size $|\mathcal{X}| = L$, and consider the case where our data $\mathcal{D}$ consists of K (complete) independent samples from the joint distribution, viz.,

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

(a) $\mathcal{G}_0 \colon \mathsf{x} \perp\!\!\!\perp \mathsf{y}$. (b) $\mathcal{G}_1 \colon \mathsf{x} \not\perp\!\!\!\perp \mathsf{y}$

Figure 1: Testing whether or not random variables x and y are independent based on training samples as a graphical model selection problem.

19.1 Likelihood Score

We start by noting that the method of maximum likelihood can be applied to the problem of learning structure in directed graphical models, where the model for the distribution is expressed in parameterized form as $p(\cdot; \mathcal{G}, \boldsymbol{\theta}^{\mathcal{G}})$, with $\mathcal{G}$ denoting the factorization structure in the distribution and $\boldsymbol{\theta}^{\mathcal{G}}$ denoting the parameters of the constituent conditional probability distributions.

With such a formulation, the maximum likelihood model structure is

$$\hat{\mathcal{G}} = \arg\max_{\mathcal{G}'} \underbrace{\max_{\boldsymbol{\theta}^{\mathcal{G}'}} \ell((\mathcal{G}', \boldsymbol{\theta}^{\mathcal{G}'}); \mathcal{D})}_{\triangleq \hat{\ell}(\mathcal{G}'; \mathcal{D})}. \tag{1}$$

Hence, the score of a candidate graph structure $\mathcal{G}'$ is $\hat{\ell}(\mathcal{G}'; \mathcal{D})$, which corresponds to the (log) likelihood for this model with its ML parameters. We emphasize that in this formulation the true graph $\mathcal{G}$ and its parameters $\boldsymbol{\theta}^{\mathcal{G}}$ are treated as deterministic but unknown.

19.1.1 Additional Information Measures

The log-likehood score has a useful interpretation in terms of information meaures. To develop this result we need some additional notation and definitions. First recall that the entropy in a random variable u is

$$H(\mathsf{u}) \triangleq H(p_{\mathsf{u}}) = -\sum_{u} p_{\mathsf{u}}(u) \ln p_{\mathsf{u}}(u) \, \mathrm{d}u. \tag{2}$$

In turn we can define *conditional* entropy as follows.

Definition 1 (Conditional Entropy)**.** *The (average) conditional entropy in u given v is*

$$H(\mathsf{u}|\mathsf{v}) \triangleq -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u, v) \ln p_{\mathsf{u}|\mathsf{v}}(u|v). \tag{3}$$

It is straightforward to verify that conditional entropy is nonnegative.

We can relate conditional entropy to entropy via

$$\begin{aligned}
H(\mathsf{u}|\mathsf{v}) &= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \ln p_{\mathsf{u}|\mathsf{v}}(u|v) \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \ln \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{v}}(v)} \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \ln \frac{p_{\mathsf{u}}(u)\, p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)\, p_{\mathsf{v}}(v)} \\
&= -\sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \left[\ln p_{\mathsf{u}}(u) + \ln \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)\, p_{\mathsf{v}}(v)} \right] \\
&= -\sum_{u} p_{\mathsf{u}}(u) \ln p_{\mathsf{u}}(u) - \sum_{u,v} p_{\mathsf{u},\mathsf{v}}(u,v) \ln \frac{p_{\mathsf{u},\mathsf{v}}(u,v)}{p_{\mathsf{u}}(u)\, p_{\mathsf{v}}(v)} \\
&= H(\mathsf{u}) - D(p_{\mathsf{u},\mathsf{v}} \| p_{\mathsf{u}} p_{\mathsf{v}}), \qquad (4)
\end{aligned}$$

where, as usual, $D(\cdot\|\cdot)$ denotes information divergence.

The second term in (4) is thus a measure of the degree of dependence between u and v, and is a sufficiently important quantity that it warrants its own special terminology.

Definition 2 (Mutual Information)**.** *The mutual information between* u *and* v *is*

$$I(\mathsf{u};\mathsf{v}) \triangleq D(p_{\mathsf{u},\mathsf{v}} \| p_{\mathsf{u}} p_{\mathsf{v}}). \qquad (5)$$

Evidentally, by the properties of divergence, mutual information is symmetric, nonnegative, and zero if and only if $\mathsf{u} \perp\!\!\!\perp \mathsf{v}$.

19.1.2 Empirical Information Measures

For a collection of independent samples $(u^1, v^1), \ldots, (u^K, v^K)$ from $p_{\mathsf{u},\mathsf{v}}$, we can obtain the empirical joint distribution

$$\hat{p}_{\mathsf{u},\mathsf{v}}(u,v) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}_{u=u^k} \mathbb{1}_{v=v^k}.$$

In turn, we can define the associated empirical entropy

$$\hat{H}(\mathsf{u}) \triangleq H(\hat{p}_{\mathsf{u}}),$$

where

$$\hat{p}_{\mathsf{u}}(u) = \sum_{v} \hat{p}_{\mathsf{u},\mathsf{v}}(u,v)$$

is the empirical marginal; and the empirical conditional entropy

$$\hat{H}(\mathsf{u}|\mathsf{v}) \triangleq -\sum_{u,v} \hat{p}_{\mathsf{u},\mathsf{v}}(u,v)\, \ln \hat{p}_{\mathsf{u}|\mathsf{v}}(u|v),$$

where

$$\hat{p}_{\mathsf{u}|\mathsf{v}}(u|v) = \frac{\hat{p}_{\mathsf{u},\mathsf{v}}(u,v)}{\sum_u \hat{p}_{\mathsf{u},\mathsf{v}}(u,v)} \tag{6}$$

is the empirical conditional. Finally, via (4) we can define the empirical mutual information

$$\hat{I}(\mathsf{u};\mathsf{v}) \triangleq \hat{H}(\mathsf{u}) - \hat{H}(\mathsf{u}|\mathsf{v}).$$

19.1.3 The Likelihood Score for Directed Graphical Models

For a directed graph model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = N$ nodes, the distribution factorizes according to

$$p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}; \boldsymbol{\theta}^\mathcal{G}) = \prod_{i=1}^{N} p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}; \boldsymbol{\theta}_i^\mathcal{G}),$$

where $\boldsymbol{\theta}^\mathcal{G} = (\boldsymbol{\theta}_1^\mathcal{G}, \ldots, \boldsymbol{\theta}_N^\mathcal{G})$ with $\boldsymbol{\theta}_i^\mathcal{G}$ denoting the parameters of conditional probability table $p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}$. We note that the parent relationship π_i for each node i is defined by the graph structure $\mathcal{G}$.

In our development, we restrict attention to the case where there is no coupling of parameters between different conditional probability tables. As a result, we can solve for each the ML estimates of $\boldsymbol{\theta}_i^\mathcal{G}$ independently for each i, whence

$$\hat{\ell}(\mathcal{G}; \mathcal{D}) = \sum_{i=1}^{N} \hat{\ell}_i(\mathcal{G}; \mathcal{D}_i) \tag{7}$$

where $\mathcal{D}_i$ here refers to the data corresponding to random variables $\{\mathsf{x}_i, \mathsf{x}_{\pi_i}\}$

In addition, we focus on the fully parameterized case, so $\boldsymbol{\theta}_i^\mathcal{G}$ are the full set of (L) entries in the conditional probability table $p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(\cdot|\cdot)$, i.e.,

$$p_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}; \boldsymbol{\theta}_i^\mathcal{G}) = \left[\boldsymbol{\theta}_i^\mathcal{G}\right]_{x_i, x_{\pi_i}}.$$

Hence, the ML parameter estimates are the empirical conditional distributions themselves, i.e.,

$$\left[\hat{\boldsymbol{\theta}}_i^\mathcal{G}\right]_{x_i, x_{\pi_i}} = \hat{p}_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i})$$

as we developed earlier.

In turn, from our development in the last installment of the notes it follows that we can express the ith term in likelihood score (7) in the form

$$\begin{aligned}\hat{\ell}_i(\mathcal{G};\mathcal{D}_i) &= \sum_{x_i,x_{\pi_i}} \hat{p}_{\mathsf{x}_i,\mathsf{x}_{\pi_i}}(x_i,x_{\pi_i}) \ln \hat{p}_{\mathsf{x}_i|\mathsf{x}_{\pi_i}}(x_i|x_{\pi_i}) \\ &= -\hat{H}(\mathsf{x}_i|\mathsf{x}_{\pi_i}) \\ &= \hat{I}(\mathsf{x}_i;\mathsf{x}_{\pi_i}) - \hat{H}(\mathsf{x}_i), \end{aligned} \tag{8}$$

where to obtain the last equality we have used (4) with (5).

Using (8) in (7) we obtain our likelihood score in the form

$$\hat{\ell}(\mathcal{G};\mathcal{D}) = \sum_{i=1}^{N} \Big[\hat{I}(\mathsf{x}_i;\mathsf{x}_{\pi_i}) - \hat{H}(\mathsf{x}_i)\Big] = \sum_{i=1}^{N} \hat{I}(\mathsf{x}_i;\mathsf{x}_{\pi_i}) - \underbrace{\sum_{i=1}^{N} \hat{H}(\mathsf{x}_i)}_{\text{constant}},$$

where we note that the second term on the right-hand side does not depend on the graph structure, i.e., locations of edges. As a result, an equivalent score is

$$\hat{\ell}(\mathcal{G};\mathcal{D}) = \sum_{i=1}^{N} \hat{I}(\mathsf{x}_i;\mathsf{x}_{\pi_i}). \tag{9}$$

To obtain different directed graph structures, we vary the topological ordering and parent assignments, each of which give rise to different combinations of empirical mutual information in the likelihood score.

19.1.4 Generalized Likelihood Ratio

19.1.5 Learning Trees

One immediate application of the likelihood score is learning tree-structured graphs. In particular, it is straightforward to find the tree model over N nodes with the maximum likelihood score. First, recall Cayley's formula that there are N^{N-2} possible trees. However, since the likelihood score can be expressed in the form (9), solving for the optimum tree is equivalent to finding a maximum-weight spanning tree within a fully-connected graph with edge weights $\hat{I}(\mathsf{x}_i,\mathsf{x}_j)$ between nodes i and j, for all $i,j \in \mathcal{V}$.

As we discussed earlier, finding a maximum-weight spanning tree can be solved using a greedy procedure, an example of which is Kruskal's algorithm. In particular, we create a list of empirical mutual informations between all pairs of variables, and order the list from largest to smallest. We then build an undirected graph from an empty one as follows. We add edges to the graph, in order, starting from the beginning of this list, bypassing any entry that would break the tree property (i.e., introduce a

cycle). We continue this process until $N-1$ edges have been added and the tree is complete. This is referred to as the *Chow-Liu algorithm*, after its inventors.[1]

The resulting undirected tree can of course be converted to a directed one in the usual way. In particular, we choose any node as the root, and propagate edge orientations away from the root node all the way to the leaf nodes.

19.1.6 Limitations of the Likelihood Score

While the likelihood score is effective in learning trees, this is because all trees equally complex. When choosing among models of differing complexities, the likelihood score is difficult to use directly, as we illustrate in this section.

It suffices to consider the simple scenario of Example 1. In particular, our goal is to choose between the graphs $\mathcal{G}_0$ and $\mathcal{G}_1$ depicted in Fig. 1, for which the (equivalent) likelihood scores are, respectively,

$$\hat{\ell}(\mathcal{G}_0; \mathcal{D}) = 0 \tag{10a}$$

$$\hat{\ell}(\mathcal{G}_1; \mathcal{D}) = \hat{I}(\mathsf{x}; \mathsf{y}). \tag{10b}$$

Note that since

$$\hat{\ell}(\mathcal{G}_1; \mathcal{D}) - \hat{\ell}(\mathcal{G}_0; \mathcal{D}) = \hat{I}(\mathsf{x}; \mathsf{y}) \geq 0,$$

the more complicated model $\mathcal{G}_1$ will always be at least as good as $\mathcal{G}_0$. Thus, likelihood scoring will always prefer $\mathcal{G}_1$. This phenomenon extends to larger models as well, where the likelihood score will always favor more complex models.[2]

There are a variety of ways to address this shortcoming by appropriately penalizing models that are more complex than the amount of available data can justify. In the next section, we show how, as an example, a Bayesian formulation can achieve this effect.

19.2 Bayesian Score

Here we develop a score by treating both the graph $\mathcal{G}$ and its parameters $\boldsymbol{\theta}^{\mathcal{G}}$ as random, placing prior distributions $p(\mathcal{G})$ and $p(\boldsymbol{\theta}^{\mathcal{G}}|\mathcal{G})$ on them. In this case, a reasonable criterion is to see the maximum a posteriori (MAP) graphical model, i.e., that for which the posterior

$$p(\mathcal{G}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{G})\, p(\mathcal{G})}{p(\mathcal{D})}. \tag{11}$$

[1]The paper in which the algorithm was introduced is C. K. Chow and C. N. Liu, *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[2]Even before advent of learning theory, there was an appreciation of the liabilities of choosing overly complex models. Indeed, it is embodied in the principle of Occam's razor: that among all possible models one should pick the simplest one consistent with the evidence.

is largest. Since the denominator in (11) does not depend on the graph structure, we can use as our score the logarithm of the numerator, i.e.,.

$$\ell^{\mathrm{B}}(\mathcal{G};\mathcal{D}) = \ln p(\mathcal{D}|\mathcal{G}) + \ln p(\mathcal{G}), \tag{12}$$

which we refer to as the *Bayesian* score. In (12) we refer to

$$p(\mathcal{D}|\mathcal{G}) = \int p(\mathcal{D}|\mathcal{G},\boldsymbol{\theta}^{\mathcal{G}})\, p(\boldsymbol{\theta}^{\mathcal{G}}|\mathcal{G})\, \mathrm{d}\boldsymbol{\theta}^{\mathcal{G}} \tag{13}$$

as the *marginal likelihood*, since it marginalizes out random parameters $\boldsymbol{\theta}^{\mathcal{G}}$, which are not of direct interest.[3]

19.2.1 Marginal Likelihoods for Independence Testing

To develop additional insight into the marginal likelihoods that constitute the Bayesian score, let us return to the problem of independence testing of Example 1.

To simplify our exposition, let us restrict attention to the binary alphabet $\mathcal{X} = \{0,1\}$, fully-parameterized (Bernoulli) distributions, and uniform priors. More specifically, For $\mathcal{G}_0$ the unknown (random) parameters are $\boldsymbol{\Theta}_{\mathcal{G}_0} = (\Theta_{\mathsf{x}}, \Theta_{\mathsf{y}}) = (p_{\mathsf{x}}(1), p_{\mathsf{y}}(1))$, which are independent and each uniformly distributed on $[0,1]$. For $\mathcal{G}_1$, the unknown (random) parameters are $\boldsymbol{\Theta}_{\mathcal{G}_1} = (\Theta_{\mathsf{x}}, \Theta_{\mathsf{y}|0}, \Theta_{\mathsf{y}|1}) = (p_{\mathsf{x}}(1), p_{\mathsf{y}|\mathsf{x}}(1|0), p_{\mathsf{y}|\mathsf{x}}(1|1))$, which are mutually independent and each uniformly distributed on $[0,1]$.

As convenient notation, we let $K_{\mathsf{x},\mathsf{y}}(i,j)$ denote the number of times the pair $(x,y) = (i,j)$ occurs in the data $\mathcal{D}$, and so $K_{\mathsf{x}}(i) = \sum_j K_{\mathsf{x},\mathsf{y}}(i,j)$ is the number of times $x = i$ occurs in the sample; $K_{\mathsf{y}}(j)$ is defined similarly. Hence, $\hat{p}_{\mathsf{x},\mathsf{y}}(i,j) = K_{\mathsf{x},\mathsf{y}}(i,j)/K$, $\hat{p}_{\mathsf{x}}(i) = K_{\mathsf{x}}(i)/K$, and $\hat{p}_{\mathsf{y}}(j) = K_{\mathsf{y}}(j)/K$ are the associated empirical joint and marginal distributions for x and y based on the data $\mathcal{D}$.

It will be convenient to use that the uniform distribution is a member of of the beta family of distributions, which are conjugate priors for Bernoulli distributions. The beta distribution with parameters α, β, denoted $\mathtt{Beta}(\alpha,\beta)$, takes the form

$$p_{\Theta}(\theta;\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1], \quad \alpha,\beta > 0. \tag{14}$$

The uniform distribution evidently corresponds to the case $\alpha = \beta = 1$.

The following identity will be useful in our analysis

[3]While beyond the scope of this subject, it is possible to approximate the marginal likelihood using a *Laplace approximation*, which involves approximating the integrand of (13) as Gaussian. The results in

$$p(\mathcal{D}|\mathcal{G}) \cong p(\mathcal{D}|\hat{\theta}^{\mathrm{ML}}_{\mathcal{G}},\mathcal{G})\, \underbrace{p(\hat{\theta}^{\mathrm{ML}}_{\mathcal{G}}|\mathcal{G})\, \sigma_{\theta|\mathcal{D}}}_{\text{Occam factor}},$$

where $\sigma_{\theta|\mathcal{D}}$ is a standard deviation related to the Gaussian approximation. The indicated Occam factor has the effect of favoring simpler models.

Fact 1. *For all nonnegative integers s and t,*

$$\int_0^1 w^s(1-w)^t\,dw = \left[(s+t+1)\binom{s+t}{s}\right]^{-1}. \tag{15}$$

Likewise, we will find the following version of Stirling's approximation convenient.

Fact 2 (Stirling's Approximation)**.**

$$\binom{n}{m} = e^{nH(\mathtt{B}(m/n))}\sqrt{\frac{n}{2\pi m(n-m)}}(1+o(1)), \tag{16}$$

where $\mathtt{B}(m/n)$ is a Bernoulli distribution with parameter m/n, and where $o(1)$ is a term that vanishes as $m\to\infty$ and $n\to\infty$.

We first evaluate the score for the model $\mathcal{G}_0$. The marginal likelihood for structure $\mathcal{G}_0$ can be expressed in the form

$$p(\mathcal{D}|\mathcal{G}_0) = p(\mathcal{D}_x|\mathcal{G}_0)\,p(\mathcal{D}_y|\mathcal{G}_0),$$

where $\mathcal{D}_x = \{x_k,\ k=1,\ldots,K\}$ and $\mathcal{D}_y = \{y_k,\ k=1,\ldots,K\}$.

Proceeding, we have

$$\begin{aligned}
p(\mathcal{D}_y|\mathcal{G}_0) &= \int p(\mathcal{D}_y|\theta_y,\mathcal{G}_0)\,p(\theta_y|\mathcal{G}_0)\,\mathrm{d}\theta_y \\
&= \int_0^1 \theta_y^{K_y(1)}(1-\theta_y)^{K_y(0)}\,\mathrm{d}\theta_y \\
&= \left[(K_y(0)+K_y(1)+1)\binom{K_y(0)+K_y(1)}{K_y(1)}\right]^{-1} \\
&= \frac{1}{K+1}e^{-KH(\hat{p}_y)}\sqrt{\frac{2\pi K_y(0)K_y(1)}{K}}(1+o(1)) \\
&= \frac{1}{K+1}e^{-KH(\hat{p}_y)}\sqrt{2\pi K\hat{p}_y(0)\,\hat{p}_y(1)}(1+o(1)) \\
&= \exp\left\{-KH(\hat{p}_y) - \ln K + \frac{1}{2}\ln K + O(1)\right\},
\end{aligned}$$

where $O(1)$ denotes a term that does not grow with K, whence

$$\ln p(\mathcal{D}_y|\mathcal{G}_0) = -KH(\hat{p}_y) - \frac{\ln K}{2} + O(1). \tag{17}$$

Moreover, by symmetry we also have

$$\ln p(\mathcal{D}_x|\mathcal{G}_0) = -KH(\hat{p}_x) - \frac{\ln K}{2} + O(1). \tag{18}$$

Thus, the Bayesian score is

$$\ell^{\mathrm{B}}(\mathcal{G}_0;\mathcal{D}) = -K[H(\hat{p}_x) + H(\hat{p}_y)] - \ln K + O(1).$$

By comparison, the likelihood score for this model is readily calculated to be

$$\hat{\ell}(\mathcal{G}_0;\mathcal{D}) = -K[H(\hat{p}_x) + H(\hat{p}_y)], \tag{19}$$

so we have

$$\ell^{\mathrm{B}}(\mathcal{G}_0;\mathcal{D}) = \hat{\ell}(\mathcal{G}_0;\mathcal{D}) - \ln K + O(1). \tag{20}$$

Next we evaluate the score for the alternative model $\mathcal{G}_1$. The marginal likelihood in this case takes the form

$$p(\mathcal{D}|\mathcal{G}_1) = p(\mathcal{D}_x|\mathcal{G}_1)\, p(\mathcal{D}_y|\mathcal{D}_x, \mathcal{G}_1) \tag{21}$$

Since $p(\mathcal{D}_x|\mathcal{G}_1) = p(\mathcal{D}_x|\mathcal{G}_0)$, which we have already calculated, we can focus on the second term.

Proceeding, we have

$$\begin{aligned}
p(\mathcal{D}_y|\mathcal{D}_x, \mathcal{G}_1) &= \int_0^1 \theta_{y|0}^{K(0,1)}(1-\theta_{y|0})^{K(0,0)}\,\mathrm{d}\theta_{y|0} \int_0^1 \theta_{y|1}^{K(1,1)}(1-\theta_{y|1})^{K(1,0)}\,\mathrm{d}\theta_{y|1} \\
&= \frac{1}{K_x(0)+1} e^{-K_x(0)H(\hat{p}_{y|x}(\cdot|0))} \sqrt{2\pi K_x(0)\,\hat{p}_{y|x}(0|0)\,\hat{p}_{y|x}(1|0)} \\
&\qquad \cdot \frac{1}{K_x(1)+1} e^{-K_x(1)H(\hat{p}_{y|x}(\cdot|1))} \sqrt{2\pi K_x(1)\,\hat{p}_{y|x}(0|1)\,\hat{p}_{y|x}(1|1)} \\
&\qquad\qquad \cdot (1+o(1)), \\
&= \frac{1}{K\hat{p}_x(0)+1} e^{-K\hat{p}_x(0)H(\hat{p}_{y|x}(\cdot|0))} \sqrt{2\pi K\hat{p}_x(0)\,\hat{p}_{y|x}(0|0)\,\hat{p}_{y|x}(1|0)} \\
&\qquad \cdot \frac{1}{K\hat{p}_x(1)+1} e^{-K\hat{p}_x(1)H(\hat{p}_{y|x}(\cdot|1))} \sqrt{2\pi K\hat{p}_x(1)\,\hat{p}_{y|x}(0|1)\,\hat{p}_{y|x}(1|1)} \\
&\qquad\qquad \cdot (1+o(1)),
\end{aligned}$$

so

$$\begin{aligned}
\ln p(\mathcal{D}_y|\mathcal{D}_x, \mathcal{G}_1) &= -K\big[\hat{p}_x(0)H(\hat{p}_{y|x}(\cdot|0)) + \hat{p}_x(1)H(\hat{p}_{y|x}(\cdot|1))\big] - \ln K + O(1) \\
&= -KH(\hat{p}_{y|x}) - \ln K + O(1).
\end{aligned} \tag{22}$$

As a result, using (18) and (22) in (21), we obtain

$$\begin{aligned}
\ell^{\mathrm{B}}(\mathcal{G}_1;\mathcal{D}) &= -K\big[H(\hat{p}_x) + H(\hat{p}_{y|x})\big] - \frac{3\ln K}{2} + O(1) \\
&= -K\big[H(\hat{p}_x) + H(\hat{p}_y) - \hat{I}(x;y)\big] - \frac{3\ln K}{2} + O(1)
\end{aligned}$$

By comparison, the likelihood score for this model is readily calculated to be

$$\hat{\ell}(\mathcal{G}_1; \mathcal{D}) = -K[H(\hat{p}_x) + H(\hat{p}_y) - \hat{I}(x; y)], \tag{23}$$

so we have

$$\ell^{\mathrm{B}}(\mathcal{G}_1; \mathcal{D}) = \hat{\ell}(\mathcal{G}_1; \mathcal{D}) - \frac{3}{2} \ln K + O(1). \tag{24}$$

Hence, to implement our independence test, we compare (20) and (24). Since $\hat{\ell}(\mathcal{G}_0; \mathcal{D})$ and $\hat{\ell}(\mathcal{G}_1; \mathcal{D})$ grow like $O(K)$, we can neglect the $O(1)$ terms in these expressions and express the test as comparing the likehood score to a threshold that depends on the amount of data used, viz.,

$$\hat{\ell}(\mathcal{G}_1; \mathcal{D}) - \hat{\ell}(\mathcal{G}_0; \mathcal{D}) \underset{\text{select } \mathcal{G}_0}{\overset{\text{select } \mathcal{G}_1}{\gtrless}} \frac{\ln K}{2}.$$

Equivalently, using (19) and (23) we can express this test as comparing the empirical mutual information $\hat{I}(x; y)$ to a threshold, viz.,

$$\hat{I}(x; y) \underset{\text{select } \mathcal{G}_0}{\overset{\text{select } \mathcal{G}_1}{\gtrless}} \frac{\ln K}{2K}. \tag{25}$$

Note that this threshold eventually decays to zero with K, but quite slowly.

As a final comment, throughout our development in this section we have neglected the term $\ln p(\mathcal{G})$ in (12) in evaluating our models since this term does not depend on K and thus can be combined with other $O(1)$ terms. This observation also reveals why the choice of prior is comparatively arbitrary.

19.2.2 Large-Sample Approximations

The approximation results we used in the preceding section for independence testing are special cases of a more general approximation that can be used for choosing the graph structure maximizing the Bayesian score. In particular, for large K, we have

$$\ell^{\mathrm{B}}(\mathcal{G}; \mathcal{D}) = \hat{\ell}(\mathcal{G}; \mathcal{D}) - \frac{\ln K}{2} \dim \mathcal{G} + O(1), \tag{26}$$

where $\dim \mathcal{G}$ is the number of independent parameters in model $\mathcal{G}$. When we neglect the $O(1)$ term in (26), the resulting score corresponds to what is referred to as the *Bayesian Information Criterion (BIC)* for model selection. We emphasize that the first term (likelihood score) is $O(K)$ and that the second term, which is our "Occam factor," penalizes models based on their complexity relative to the amount of training data available.

Note, for example, that in our independent testing development, $\dim \mathcal{G}_0 = 2$ and $\dim \mathcal{G}_1 = 3$, which our test (25) reflects.

19.3 Relationship Between Inference and Structure Learning

In this last section, we will draw a connection between inference and learning by making the following statement: if you can solve an inference problem (i.e marginalization) for any graphical model, then you can learn any graphical model from data, and vice versa. We will show how this applies to a very specific example setup.

For our specific setup, assume we have n random variables $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_N \end{bmatrix}$. We will restrict the joint distribution of these variable to the class of graphical models that follow the form:

$$P(\mathbf{x}) \propto \exp \sum_{(i,j)} x_i \theta_{ij} x_j = \exp \mathbf{x}^T \theta \mathbf{x} \tag{27}$$

where $\theta = [\theta_{ij}]$ $\forall i, j = 1, \dots, N$.

When each x_i takes on binary values, this is known as an Ising model. Let us assume an Ising model for our specific setup, although what we will show for this model extends more generally to any exponential family that we have seen up until now. We leave it as an exercise for you to extend our results to the case when $P(x) \propto \exp\left(\sum_{(k)} \theta_k f_k(x)\right)$ for any subset $k \in x_1, \dots, x_N$ and function $f_k(x)$.

In our example, say we have observed data $x^{(1)}, \dots, x^{(n)}$ and we want to find estimate $\hat{\theta}$ of the parameters. We define the sufficient statistics

$$\hat{\mu}_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_i^{(k)} x_j^{(k)} \quad \text{for all } i, j = 1, \dots, N \tag{28}$$

We can view $\hat{\mu}_{ij}$ as a proxy for estimating $\mathbb{E}[x_i x_j]$, i.e. $\mu_{ij} \cong \mathbb{E}[x_i x_j]$. Although $\mathbb{E}[x_i x_j]$ is unknown, we can calculate the $\hat{\mu}_{ij}$ from the data.

Now, we propose to do something which may seem out of the blue, but will soon connect back to our goal. Let's propose to find a distribution P_θ over $\{0,1\}^N$ that

(a) maximizes entropy

(b) is subject to being consistent with the sufficient statistics $\hat{\mu}_{ij}$.

This gives us the following maximization problem:

$$\max_{P \in \mathcal{P}} H(P) \quad \text{s.t.} \qquad P(x) \geq 0, \tag{29}$$

$$\sum_{x \in \{0,1\}^N} P(x) = 1 \tag{30}$$

$$\mathbb{E}\left[x_i x_j\right] = \hat{\mu}_{ij} \quad \forall i, j = 1, \dots, N \tag{31}$$

We can solve this maximization problem using Lagrange multipliers λ for (30) and θ_{ij} for (31) to get the Lagrangian

$$\mathcal{L}(P, \theta, \lambda) = H(P) + \lambda(\sum_x P(x) - 1) + \sum_{i,j} \theta_{ij}(\mathbb{E}[x_i x_j] - \mu_{ij}) \tag{32}$$

Taking the derivative with respect to $P(x)$, we get

$$\frac{\partial \mathcal{L}}{\partial P(x)} = -1 + \log P(x) + \lambda + \sum_{i,j} \theta_{ij} x_i x_j \tag{33}$$

Setting (33) to zero gives

$$P_\theta(x) \propto \exp(\sum_{i,j} \theta_{ij} x_i x_j) \tag{34}$$

We see that the optimizing distribution has precisely the form we are looking for. Replacing P(x) by the optimizing $P_\theta(x)$ in the Lagrangian, we compute the Lagrangian as a function $g(\theta)$ of θ:

$$g(\theta) \triangleq \mathcal{L}(P, \theta, \lambda) = \sum_x P_\theta(x) \log \frac{1}{P_\theta(x)} + \sum_{i,j} \theta_{ij}(\mathbb{E}_{P_\theta}[x_i x_j] - \hat{\mu}_{ij}) \tag{35}$$

$$= \sum_x P_\theta(x)(\log Z(\theta) - \sum_{i,j} \theta_{ij} x_i x_j) + \sum_{i,j} \theta_{ij}(\mathbb{E}_{P_\theta}[x_i x_j] - \hat{\mu}_{ij}) \tag{36}$$

$$= \log Z(\theta) - \sum_{i,j} \theta_{ij} \mathbb{E}_{P_\theta}[x_i x_j] + \sum_{i,j} \theta_{ij}(\mathbb{E}_{P_\theta}[x_i x_j] - \hat{\mu}_{ij}) \tag{37}$$

$$= \log Z(\theta) - \sum_{i,j} \hat{\mu}_{ij} \theta_{ij} \tag{38}$$

Therefore, solving the maximum entropy problem through Lagrangian multipliers boils down to minimizing $g(\theta)$ over the choice of θ,

$$\min_\theta g(\theta) = \log Z(\theta) - \sum_{i,j} \hat{\mu}_{ij} \theta_{ij} \tag{39}$$

The $\hat{\mu}_{ij}$ comes from the data, and we are looking for the parameter θ. The simplest algorithm for evaluating θ is gradient descent. We set up the gradient descent algorithm with the following initialization and iterative update:

$$\theta^{(0)} = 0 \tag{40}$$

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)} \frac{\partial g}{\partial \theta_{ij}}\bigg|_{\theta=\theta^{(t)}} \tag{41}$$

Solving for the gradient, we get the nice expression

$$\frac{\partial g}{\partial \theta_{ij}} = \frac{\partial}{\partial \theta_{ij}}(\log Z(\theta) - \sum_{i',j'} \hat{\mu}_{i'j'}\theta_{i'j'}) \tag{42}$$

$$= \frac{1}{Z(\theta)} \sum_{x} \exp\left(x^T \theta x\right) x_i x_j - \hat{\mu}_{ij} \tag{43}$$

$$= \mathbb{E}[x_i x_j] - \hat{\mu}_{ij} \tag{44}$$

which gives update equation

$$\theta^{(t+1)} = \theta^{(t)} - \alpha^{(t)}[\mathbb{E}[x_i x_j] - \hat{\mu}_{ij}] \tag{45}$$

In the gradient update, we compare the expectation of the marginal of $x_i x_j$ over P_θ, and compare it to the observed mean $\hat{\mu}_{ij}$ of the observed data. The gradient is only nonzero when these two values do not match, and the algorithm stops changing exactly when the parameters produce the same mean as the observed mean for all $x_i x_j$. This procedure is referred to as moment-matching. Moment-matching describes the process of starting with a guess of an unknown parameter, using that guess to compute the moments of sufficient statistics that we are interested in, and comparing it with the observed sufficient statistics. If they are different, we update our guess, otherwise we stop.

Computing the moments with respect to the parameters that we are interested in is an inference problem. If we can solve this inference problem, then we can do moment-matching. And if we can do moment-matching, we can learn the parameters of the model. Because we have started with the assumption of an Ising model, knowing the parameters gives us the graph structure of the model. With this example, we have shown the dual relationship between inference and learning, and how solving one of these problems allows us to solve the other.