Lecture 17 — March 9, 2017

Lecturer: David Tse *Scribe: J. Zhang, F. Farnia*

17.1 Maximum entropy and graphical models

Last time, we introduced the maximum entropy principle and briefly mentioned its relationship to graphical models. Given $(X_1, \ldots, X_n)$ and a graph $G = (V, E)$, we let $\mathbb{E}[X_i X_j] = K_{ij}$ if $(i, j) \in E$, $\mathbb{E}[X_i^2] = K_{ii}$, and $\mathbb{E}[X_i X_j]$ unspecified if $(i, j) \notin E$. We showed last time that the maximum entropy distribution is $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ for a covariance matrix $\mathbf{K}$ satisfying $(\mathbf{K}^{-1})_{ij} = 0$ if $(i, j) \notin E$. That implies that if we have two nodes that are not connected in the graph, then their corresponding random variables are conditionally independent given all other variables in the graph.

Now suppose we have n discrete random variables $X_1, \ldots, X_n$. We are given pairwise marginal distributions $p(x_1, x_2), p(x_2, x_3), \ldots, p(x_{n-1}, x_n)$. In other words, though we are dealing with a high-dimensional joint distribution, we only access to some low-dimensional statistics to infer that high-dimensional distribution. In this case, the graph where we connect every two nodes with given pairwise statistic looks as follows

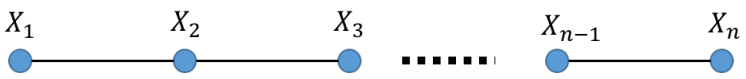

Figure 17.1: Graph of given pairwise marginals

What does the entropy-maximizing joint distribution look like? Observe that the graph looks like a Markov chain; hence, we can expect that the solution satisfies Markov properties. Here the entropy maximization problem is

$$\begin{aligned} \max_{p_{\mathbf{X}}} \quad & H(X_1, \ldots, X_n) \\ \text{s.t.} \quad & p(x_1, x_2), \ldots, p(x_{n-1}, x_n) \text{ fixed} \end{aligned}$$

Using the chain rule for entropy,

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \ldots, X_{n-1}) \\ &\leq H(X_1) + H(X_2|X_1) + H(X_3|X_2) + \cdots + H(X_n|X_{n-1}) \end{aligned}$$

Thus, for any distribution with those pairwise marginals, $H(X_1, \ldots X_n)$ is upper bounded by the entropy of a Markov chain with the same pairwise marginals, because conditioning reduces entropy. Therefore, a simple Markov chain maximizes entropy in this case.

Example. Consider a Mickey Mouse Markov chain $X_i \in \{0, 1\}$. We fix $p(x_1, x_2)$, $p(x_2, x_3)$, $\ldots, p(x_{n-1}, x_n)$. Note that this is equivalent to fixing $P(X_i = 1) = \alpha_i$, $i = 1, \ldots, n$ and $P(X_i = 1, X_{i+1} = 1) = \beta_{i,i+1}$, $i = 1, \ldots, n-1$.
In general, as shown in the last lecture if we have $\Lambda = \{p : \mathbb{E}[r_i(X)] = \alpha_i, i = 1, \ldots, I\}$, then the entropy-maximizing distribution is

$$p^*(\mathbf{x}) \propto \exp\left(\lambda_1 r_1(x) + \lambda_2 r_2(x) + \cdots + \lambda_I r_I(x)\right) \text{ (exponential family)}$$

In the Mickey Mouse Markov chain example, what are the r_i's? $P(X_i = 1) = \mathbb{E}[X_i] = \alpha_i$, and $P(X_i = 1, X_{i+1} = 1) = \mathbb{E}[X_i X_{i+1}] = \beta_{i,i+1}$. Therefore

$$\begin{aligned} p^*(\mathbf{x}) &\propto \exp\left(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n + \lambda_{1,2} x_1 x_2 + \lambda_{2,3} x_2 x_3 + \cdots + \lambda_{n-1,n} x_{n-1} x_n\right) \\ &= \exp\left(\sum_{i=1}^{n} \lambda_i x_i + \sum_{i=1}^{n-1} \lambda_{i,i+1} x_i x_{i+1}\right) \end{aligned}$$

For example, if $n = 3$

$$\begin{aligned} p^*(\mathbf{x}) &= \exp\left(\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_{1,2} x_1 x_2 + \lambda_{2,3} x_2 x_3\right) \\ &= g_1(x_1, x_2) g_2(x_2, x_3) \end{aligned}$$

which you will prove in the homework to be the joint distribution of a Markov chain.

Note that if we fix more pairwise marginals as well, like $p(x_1, x_5)$, we should consider the new corresponding terms in p^* (for example given $p(x_1, x_5)$ we should add a term $\lambda_{1,5} x_1 x_5$). For a general setting, we can get the **Ising model** which extends the above discussion for a general graph of pairwise marginals specified by an edge set E. Then we have

$$\begin{aligned} p^*(\mathbf{x}) &\propto \exp\left(\sum_i \lambda_i x_i + \sum_{(i,j) \in E} \lambda_{i,j} x_i x_j\right) \\ \Rightarrow p^*(\mathbf{x}) &= \frac{1}{A(\{\lambda_i\}, \{\lambda_{i,j}\})} \exp\left(\sum_i \lambda_i x_i + \sum_{(i,j) \in E} \lambda_{i,j} x_i x_j\right) \end{aligned}$$

where $A(\{\lambda_i\}, \{\lambda_{i,j}\})$ is a normalization constant to ensure that p^* is a valid probability distribution. As an example, consider a 2D version of Markov chain in Figure 17.2. The Ising model coming from this multi-grid network has been well-studied in statistical physics. Note that in this case we seek to find the distribution maximizing entropy given pairwise marginals specified by this graph.

To compute the joint distribution, we need to find the value of the normalization constant $A(\{\lambda_i\}, \{\lambda_{i,j}\})$ as well. In general, this is a hard problem from a computational point of view. A technique for approximating the solution is called the **variational method**, which is based on approximating the maximum entropy problem. Further discussion on this technique is beyond the scope of this course.

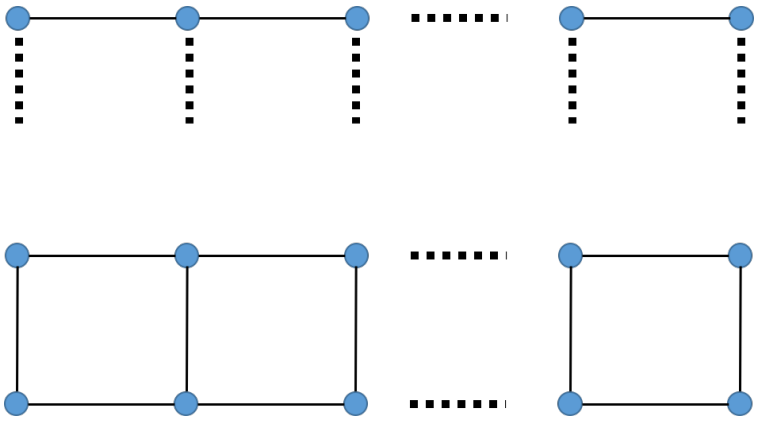

Figure 17.2: A 2D version of Markov Chains

17.2 Supervised setting

In a supervised setting, we have access to a feature vector $\mathbf{X} = (X_1, \ldots, X_p)$ from which we want to predict a target variable Y. For example, the feature vector may represent the pixel intensities in an image. Y can be either real valued (e.g. price of a house) or categorical (e.g. cat or dog). For prediction, the main goal is to find the conditional distribution of Y given the feature vector $\mathbf{X}$. However, $P_{\mathbf{X},Y}$ is usually too high-dimensional to estimate. We may reliably estimate, however, lower-dimensional statistics, such as cross-moment $\mathbb{E}[X_iY]$'s. We now explore how we could use the lower-order statistics to find a predictive conditional distribution $P_{Y|\mathbf{X}}$.

Notice that the problem formulation for maximizing $H(X_1, \ldots, X_p)$ is symmetric with respect to X_i's. For this reason, we cannot simply append a Y, which we know has a different role in prediction problem than X_i's. A suggested formulation given that the underlying joint distribution $P_{\mathbf{X},Y}$ is in a set Γ would be

$$\max_{P_{\mathbf{X},Y} \in \Gamma} H(Y|\mathbf{X})$$

Here, we are looking for a conditional distribution that satisfies the lower order constraints but which yields the maximum uncertainty of Y given $\mathbf{X}$. To find the structure of the solution to this optimization problem, we will again use the non-negativity property of relative entropy. Suppose we are searching over all distributions satisfying some pairwise moments:

$$\Gamma = \{f_{\mathbf{X},Y} : \mathbb{E}[X_iX_j] = K_{ij},\ \mathbb{E}[X_iY] = \alpha_i,\ \mathbb{E}[Y^2] = \beta,\ \forall i, j \in \{1, \ldots, p\}\}$$

An idea to solve this problem is to reduce it to the unconditional case by using the chain rule

$$h(Y|\mathbf{X}) = h(\mathbf{X}, Y) - h(\mathbf{X})$$

The first term $h(\mathbf{X}, Y)$ becomes maximized if $(\mathbf{X}, Y)$ has a jointly Gaussian distribution. However, the second term $h(\mathbf{X})$ also gets maximized given the jointly Gaussian distribution. So it is not clear if choosing $(\mathbf{X}, Y)$ to be jointly Gaussian will maximize $h(\mathbf{X}, Y) - h(\mathbf{X})$.

To solve, this problem, we use the fact that

$$D(f(\cdot|\mathbf{X}=\mathbf{x})\|f^*(\cdot|\mathbf{X}=\mathbf{x})) \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^p$$

We will conclude this discussion next lecture.