8 Exponential Families

8.1 Definition of Linear Exponential Family

Definition 1. *An* exponential family *parameterized by $\boldsymbol{\theta} \in \mathbb{R}^k$ is a family of distributions of the form*

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i=1}^{k} \theta_i f_i(\mathbf{x}) \right),$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is an N-dimensional vector.

The functions $f_i : \mathcal{X} \to \mathbb{R}\ \forall i$ are known as the *sufficient statistics*, or features, of the exponential family, while the parameters $\boldsymbol{\theta}$ are known as the *natural parameters.* We denote by

$$\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^k \mid Z(\boldsymbol{\theta}) < \infty\} \tag{1}$$

the space of natural parameters

Note that an exponential family consists of strictly positive distributions. There are a few additional technical definition that we may need later on:

- A *regular* exponential family is one where $\Theta \neq \varnothing$ and Θ is open.
- A *minimal* exponential family is one where $\nexists c \in \mathbb{R}^k \setminus \{0\}$ such that $\sum_{i=1}^{k} c_i f_i(\mathbf{x})$ is constant for all $\mathbf{x}$. Equivalently, there do not exist $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ such that $p(\mathbf{x}; \boldsymbol{\theta}_1) = p(\mathbf{x}; \boldsymbol{\theta}_2)$.

8.2 Examples of Exponential Families

1. In an discrete undirected graphical model, the distribution can be written as:

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C) \\ &= \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \ln \psi_C(\mathbf{x}_C) \right) \\ &= \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{C}} \sum_{\mathbf{x}'_C \in |\mathcal{X}|^{|C|}} \ln \psi_C(\mathbf{x}'_C) \mathbb{1}_{\mathbf{x}_C = \mathbf{x}'_C} \right). \end{aligned}$$

   This is an exponential family representation where the sufficient statistics correspond to indicator variables for each clique and each joint assignment to the variables in that clique:

$$f_{C,\mathbf{x}'_C}(\mathbf{x}) = \mathbb{1}_{\mathbf{x}_C = \mathbf{x}'_C},$$

   and the natural parameters $\boldsymbol{\theta}_{C,\mathbf{x}_C}$ correspond to the log potentials $\ln \psi_C(\mathbf{x}_C)$.

2. In a Gaussian graphical model, we can write the joint distribution in information form:

$$p(x) \propto \exp\left(\sum_{i=1}^{N} h_i x_i - \frac{1}{2}\sum_{ij} J_{ij} x_i x_j\right) \tag{2}$$

We observe that this is an exponential family with sufficient statistics

$$f(\mathbf{x}) = \begin{pmatrix} x_1 \\ x_1 \\ \vdots \\ x_N \\ x_1^2 \\ x_1 x_2 \\ \vdots \\ x_N^2 \end{pmatrix}.$$

and natural parameters

$$\boldsymbol{\theta} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \\ -\frac{1}{2}\mathbf{J}_{11} \\ -\frac{1}{2}\mathbf{J}_{12} \\ \vdots \\ -\frac{1}{2}\mathbf{J}_{NN} \end{pmatrix}.$$

Note that the sufficient statistics are first and second order moments.

3. Say that we have n i.i.d samples $\mathbf{x}^1, \ldots, \mathbf{x}^n$ from some distribution in an exponential family. Then the joint distribution of all the samples is

$$\begin{aligned} p(\mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta}) &= \prod_{j=1}^{n} \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i=1}^{k} \theta_i f_i(\mathbf{x}^j)\right) \\ &= \frac{1}{Z(\boldsymbol{\theta})^n} \exp\left(\sum_{i=1}^{k} \theta_i \sum_{j=1}^{n} f_i(\mathbf{x}^j)\right) \end{aligned}$$

This is an exponential family with natural parameter $\boldsymbol{\theta}$ and sufficient statistics $\sum_{j=1}^{n} f_i(\mathbf{x}^j)\ \forall\ i$.

4. Consider a Gaussian graphical model on the graph $\mathcal{G}$. This enforces the constraint that $\mathbf{J}_{ij} = 0$ for $(i, j) \notin \mathcal{E}$. This is an exponential family with sufficient statistics $f_i(\mathbf{x}) = x_i\ \forall\ i$ and $f_{ij}(\mathbf{x}) = x_i x_j$ for $(i, j) \in \mathcal{E}$ or $i = j$. Likewise, the natural parameters are $h_i\ \forall\ i$ and $-\frac{1}{2}\mathbf{J}_{ij}$ for $(i, j) \in \mathcal{E}$ or $i = j$.

5. Multinomial undirected graphical models: We previously showed that we can reparameterize a binary undirected graphical model as

$$p_{x_{\mathcal{V}}}(\mathbf{x}_{\mathcal{V}}; \boldsymbol{\theta}) \propto \exp\left(\sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \theta_{\mathcal{C}} \prod_{c \in \mathcal{C}} x_c \right). \tag{3}$$

Note that this is indeed an exponential family, with natural parameters $\boldsymbol{\theta}$ and sufficient statistics $\prod_{c \in \mathcal{C}} x_c$. The joint distribution of n i.i.d samples is

$$p(\mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta}) \propto \exp\left(\sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \theta_{\mathcal{C}} \sum_{i=1}^{n} \prod_{c \in \mathcal{C}} x_c \right) = \exp\left(\sum_{\mathcal{C} \in \mathrm{cl}(\mathcal{G})} \theta_{\mathcal{C}} \cdot m(\mathbb{1}_{\mathcal{C}}) \right), \tag{4}$$

where $m(\mathbb{1}_{\mathcal{C}}) := \sum_{i=1}^{n} \prod_{c \in \mathcal{C}} x_c$ is the *marginal count* for the clique $\mathcal{C}$, which is just the number of samples where $x_c = 1$ for all $c \in \mathcal{C}$. In this case, we see that the sufficient statistics are just the marginal counts for each clique.

8.3 Maximum Likelihood Estimation in Exponential Families

Last lecture, we showed that maximum likelihood estimation is equivalent to the M-projection, i.e. minimizing KL divergence between the empirical distribution and the family of distributions we're parameterized by.

What then is the M-projection on to an exponential family? In turns out that the M projection has a very nice representation in terms of moment matching:

Theorem 1. *Let p be a distribution on $\mathsf{x}_1, \ldots, \mathsf{x}_N$ and Q an exponential family with sufficient statistics $f_i(x)$ for $i = 1, \ldots, k$ and natural parameters $\Theta \subset \mathbb{R}^k$. If there exists $\boldsymbol{\theta} \in \Theta$ such that $\mathbb{E}_{q_{\boldsymbol{\theta}}}[f_i(x)] = \mathbb{E}_p[f_i(x)]$ for all i, then the M-projection of p onto Q is*

$$q^M := \arg\min_{q \in Q} D(p||q) = q_\theta \tag{5}$$

Proof. Let $\boldsymbol{\theta}' \in \Theta$. Want to show $D(p||q_{\boldsymbol{\theta}'}) - D(p||q_{\boldsymbol{\theta}}) \geq 0$. We see that:

$$\begin{aligned}
D(p||q_{\boldsymbol{\theta}'}) - D(p||q_{\boldsymbol{\theta}}) &= -H(p) - \mathbb{E}_p(\log q_{\boldsymbol{\theta}'}) + H(p) + \mathbb{E}_p(\log q_{\boldsymbol{\theta}}) \\
&= -\mathbb{E}_p(\log q_{\boldsymbol{\theta}'}) + \mathbb{E}_p(\log q_{\boldsymbol{\theta}}) \\
&= -\mathbb{E}_p\left(\sum_{i=1}^{k} \theta_i' f_i(\mathbf{x}) - \log Z(\boldsymbol{\theta}') \right) + \mathbb{E}_p\left(\sum_{i=1}^{k} \theta_i f_i(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \right) \\
&= -\mathbb{E}_{q_{\boldsymbol{\theta}}}\left(\sum_{i=1}^{k} \theta_i' f_i(\mathbf{x}) - \log Z(\boldsymbol{\theta}') \right) + \mathbb{E}_{q_{\boldsymbol{\theta}}}\left(\sum_{i=1}^{k} \theta_i f_i(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \right) \\
&= -\mathbb{E}_{q_{\boldsymbol{\theta}}}(\log q_{\boldsymbol{\theta}'}) + \mathbb{E}_{q_{\boldsymbol{\theta}}}(\log q_{\boldsymbol{\theta}}) \\
&= D(q_{\boldsymbol{\theta}}||q_{\boldsymbol{\theta}'}) \\
&\geq 0,
\end{aligned}$$

where we used the condition that $\mathbb{E}_{q_{\boldsymbol{\theta}}}[f_i(x)] = \mathbb{E}_p[f_i(x)]$ to replace the expectation with respect to p to one with respect to $q_{\boldsymbol{\theta}}$. $\square$

Thus to find the MLE, we just need to find the natural parameters which match the moments of the empirical distribution. This is also why the "M" in "M-projection" stands for moment! Note that this theorem does not give us a way to compute the MLE yet, but rather a set of equations which the MLE needs to satisfy.

8.3.1 Existence and Uniqueness

There are a couple questions which arise from this theorem. When does such a $\boldsymbol{\theta}$ exist, and if it does, is it unique?

Proposition 1. *Define the function $g : \Theta \to \mathbb{R}^k$ where $g(\boldsymbol{\theta}) \mapsto \mathbb{E}_{q_{\theta}}[f(\mathbf{x})]$. Then, if $\mathbb{E}_{\hat{p}}[f(\mathbf{x})] \in im(g)$ and g is one-to-one, the MLE is unique. It is the unique $\hat{\boldsymbol{\theta}} \in \Theta$ that satisfies $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[f(\mathbf{x})] = \mathbb{E}_{\hat{p}}[f(\mathbf{x})]$.*

It turns out that g is one-to-one for minimal exponential families. The first condition, that $\mathbb{E}_{\hat{p}}[f(\mathbf{x})] \in \mathrm{im}(g)$, is true for regular exponential families for almost all $\hat{p}$ (i.e. for all but a measure zero set of empirical distributions). Thus for minimal, regular, exponential famililes, the MLE almost always exists and is unique.

8.3.2 Computing the MLE: Examples

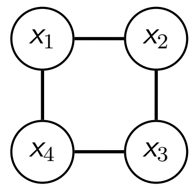
- In the discrete setting, the sufficient statistics are simply the marginal counts $m(x_{\mathcal{C}})$. Therefore we require the observed marginal counts to be equal to the expected marginal counts, or equivalently, the observed marginal probabilities over each clique $\frac{m(x_{\mathcal{C}})}{n}$ to be equal to $\hat{p}(x_{\mathcal{C}})$, the probability of setting a clique equal to 1.

- Let's consider estimating the MLE of the covariance $\mathbf{\Sigma}$ for a mean zero undirected Gaussian graphical model with graph $\mathcal{G}$. This means that we require $(\hat{\mathbf{\Sigma}}^{-1})_{ij} = 0$ for all $(i,j) \notin \mathcal{E}, i \neq j$.

  Let $\mathbf{S}$ be the sample covariance matrix. By the above theorem, the MLE must match moments with the empirical distribution. Therefore we must have our MLE $\hat{\mathbf{\Sigma}}$ be such that
  $$\hat{\mathbf{\Sigma}}_{ij} = \mathbb{E}_{\hat{\mathbf{\Sigma}}}[x_i x_j] = \mathbf{S}_{ij}$$
  for all $(i,j) \in \mathcal{E}$ or $i = j$.

  To fill in the missing values $\hat{\mathbf{\Sigma}}_{ij}$ for $(i,j) \notin E$, we use the constraint that $(\hat{\mathbf{\Sigma}}^{-1})_{ij} = 0$. This gives a system of equations which can then be solved.

  As an example, consider the 4 cycle:



Then our MLE must be of the form:

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} & a & \mathbf{S}_{14} \\ \mathbf{S}_{21} & \mathbf{S}_{22} & \mathbf{S}_{23} & b \\ a & \mathbf{S}_{32} & \mathbf{S}_{33} & \mathbf{S}_{34} \\ \mathbf{S}_{41} & b & \mathbf{S}_{43} & \mathbf{S}_{44} \end{bmatrix} \tag{6}$$

To solve for the unknowns a and b, we use the sparsity constraints $(\hat{\mathbf{\Sigma}}^{-1})_{13} = 0, (\hat{\mathbf{\Sigma}}^{-1})_{24} = 0$. This gives a system of 2 equations in the 2 unknowns a, b.