CS395T: Learning Theory; Fall 2011

Lectures 4 and 5 ; Date: 09/19/2011 and 09/21/2011

Lecturer: Pradeep Ravikumar ; Scribe: Rashish Tandon

Keywords: graphical models, exponential families, inference, marginal polytope

Note: Scribed notes have only been lightly proofread.

1 Overview

In these lectures, we see how graphical models can be expressed as exponential families. Using this characterization of graphical models, we then look at some connections with convex analysis.

2 Exponential Family Via Maximum Entropy

We motivate exponential families for graphical models by interpreting it as a maximum entropy distribution. Suppose, we have a set $\mathcal{I}$ that indexes functions $\phi_\alpha : \mathcal{X} \to \mathbb{R} \quad \forall \alpha \in I$. For any distribution $p(X)$ over $\mathcal{X}$, we can compute the $|\mathcal{I}|$-dimensional vector of expectations

$$\mu = (\mu_\alpha \ : \ \alpha \in \mathcal{I}), \text{ where } \mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)]$$

Now, consider the reverse problem. We are given a vector of moments $\hat{\mu} = (\hat{\mu}_\alpha \ : \ \alpha \in \mathcal{I})$, and we would like to find a distribution $p(X)$ that is consistent with these moments i.e. $\mathbb{E}_p[\phi_\alpha(X)] = \hat{\mu}_\alpha$. There may be many distributions that are consistent with these moments. However, it turns out that if we choose the distribution with the maximum entropy, it belongs to the exponential family. More formally, let $\mathcal{P}$ be the set of all distributions on the random vector X. Then, we have the optimization problem :

$$\begin{aligned} \max_{p \in \mathcal{P}} H(p) &:= -\sum_{X \in \mathcal{X}} p(X) \log(p(X)) \\ \text{s.t.} \sum_{X \in \mathcal{X}} p(X) &= 1, \ p(X) \geq 0 \quad \forall X \in \mathcal{X} \\ \mathbb{E}_p[\phi_\alpha(X)] &= \hat{\mu}_\alpha \quad \forall \alpha \in \mathcal{I} \end{aligned}$$

If p^* is the optimal solution of the above, then p^* has the form

$$p^*(X) \propto \exp\left(\sum_\alpha \lambda_\alpha \phi_\alpha(X)\right)$$

where λ_α are the Lagrange multipliers associated with the moment constraints. *(Verify this as an exercise)*

3 Exponential Families

As described in earlier lectures, given a set of *sufficient statistics* $\phi = (\phi_\alpha \ : \ \alpha \in \mathcal{I})$, where $\phi_\alpha : \mathcal{X} \to \mathbb{R}$ and $\mathcal{I}$ is an index set ($|\mathcal{I}| = d$), we can associate an exponential family with ϕ by the distribution

$$p(X;\theta) \propto \exp\left(\sum_\alpha \theta_\alpha \phi_\alpha(X) \ - \ A(\theta)\right)$$

where $\theta = (\theta_\alpha \ : \ \alpha \in \mathcal{I})$ is an associated parameter vector. $A(\theta)$ is the normalizing term i.e. $\exp(A(\theta)) = \sum_{X \in \mathcal{X}} \exp\left(\sum_\alpha \theta_\alpha \phi_\alpha(X)\right)$, and is called the *log-partition function*. The set of valid parameters θ is

$$\Omega = \{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\}$$

A *minimal exponential family* is an exponential family where there is no non-zero vector $a \in \mathbb{R}^d$ such that

$$\sum_{\alpha \in \mathcal{I}} a_\alpha \phi(X) = b \text{ (a constant)}$$

The family is called *overcomplete* if such a non-zero vector $a \in \mathbb{R}^d$ exists. In case of overcomplete families, more than one set of parameters would lead to the same distribution. Given a set of parameters θ, consider a new set of parameters $\beta := \theta + a$. Then,

$$\begin{aligned} p(X;\beta) &\propto \exp\left(\sum_\alpha \beta_\alpha \phi_\alpha(X)\right) \\ &\propto \exp\left(\sum_\alpha \theta_\alpha \phi_\alpha(X)\right).e^b \\ &= p(X;\theta)\,, \end{aligned}$$

since e^b is just a constant and so it cancels with the normalizing term.

4 Graphical Models as Exponential Families

Graphical Models have been described earlier as a product of functions. When representing these as exponential families however, the products become additive decompositions. We now have a few examples :

4.1 Ising Model

Given a graph $G = (V, E)$, we have a random variable $X_s \in \{0, 1\}$ associated with each node $s \in V$. We have potentials associated with each node and edge. The sufficient statistics for an Ising model are :

$$\{X_s \ \forall \ s \in V\} \cup \{X_s X_t \ \forall \ (s,t) \in E\}$$

This gives us the family :

$$p(X;\theta) = \exp\left(\sum_{s \in V} \theta_s X_s + \sum_{(s,t) \in E} \theta_{st} X_s X_t - A(\theta)\right)$$

Ising models can also be generalized to have k-order terms, for any $k \geq 2$. Such models are called k-spin Ising models. The above is a 2-spin Ising model.

4.2 Standard Overcomplete Model

This is a generalization of the Ising model where for each node $s \in V$, the random variable $X_s \in \{0, 1, \ldots, r-1\}$ (instead of just $\{0, 1\}$ as earlier). For each node $s \in V$, a particular label j defines the sufficient statistic :

$$\mathbb{I}_{s,j}(X) = \begin{cases} 1 & \text{if } X_s = j \\ 0 & \text{otherwise} \end{cases}$$

Also, for each edge $(s, t) \in E$, the tuple (i, j) of labels defines another sufficient statistic :

$$\mathbb{I}_{st,jk}(X) = \begin{cases} 1 & \text{if } (X_s, X_t) = (j, k) \\ 0 & \text{otherwise} \end{cases}$$

This gives us the distribution :

$$p(X; \theta) \propto \exp\left(\sum_{s \in V, j} \theta_{s,j} \mathbb{I}_{s,j}(X) + \sum_{(s,t) \in E, j, k} \theta_{st,jk} \mathbb{I}_{st,jk}(X) - A(\theta) \right)$$

Note that the set of sufficient statistics described above are overcomplete (*Why ?*), and such a representation using indicator functions is called the *standard overcomplete representation*.

This model is useful since any discrete graphical model can be expressed in this form.

5 Mean Parameterization

So far, an exponential family has been described by the set of parameters $\theta \in \Omega$. Any exponential family has an alternate parameterization in terms of the mean parameter vector $\mu = (\mu_\alpha \ : \ \alpha \in \mathcal{I})$, where $\mu_\alpha = \mathbb{E}_p[\phi_\alpha(X)]$. We define a marginal polytope $\mathcal{M}$ as the set of all possible mean parameter vectors, for sufficient statistics $\phi = \{\phi_\alpha \ : \ \alpha \in \mathcal{I}\}$. So,

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_p[\phi(X)], \text{for some } p : \mathcal{X} \to \mathbb{R} \text{ s.t. } \sum_{X \in \mathcal{X}} p(X) = 1 \text{ and } p(X) \geq 0 \ \forall X \in \mathcal{X}\}$$

We now describe the link between the log-partition function $A(\theta)$ and the mean vector μ. We know

$$A(\theta) = \log(\sum_{X \in \mathcal{X}} \exp(\theta^T \phi(X)))$$

Consider $\nabla A(\theta)$,

$$\begin{aligned} \nabla A(\theta) &= \frac{\phi(X). \sum_{X \in \mathcal{X}} \exp(\theta^T \phi(X))}{\sum_{X \in \mathcal{X}} \exp(\theta^T \phi(X))} \\ &= \sum_{X \in \mathcal{X}} \phi(X) \frac{\exp(\theta^T \phi(X))}{\sum_{X \in \mathcal{X}} \exp(\theta^T \phi(X))} \\ &= \mu \end{aligned}$$

Thus, $\nabla A(\theta) = \mu$. In fact, $\nabla A(\cdot)$ can be thought of as a map from the parameter vector θ to the mean parameter μ i.e. $\nabla A : \Omega \to \mathcal{M}$, and it is called the *forward mapping*. Some other properties of the log-partition function are :

- $\nabla^2 A(\theta) = \text{cov}(\phi, \phi)$ (*Verify as an exercise*)
- $A(\theta)$ is a convex function. This is because $\nabla^2 A(\theta) = \text{cov}(\phi, \phi) \succeq 0$. This also implies that Ω is convex.

- If the exponential family is minimal, A is a strictly convex function. This can be seen as follows,

$$\begin{aligned}\text{For any } a \in \mathbb{R}^d, \text{ we have } a^T\phi(X) \neq \text{ a constant}\\ \Rightarrow \text{Var}(a^T\phi(X)) \neq 0\\ \Rightarrow \text{Var}(a^T\phi(X)) > 0\\ \Rightarrow a^T\text{cov}(\phi,\phi)a > 0\\ \Rightarrow a^T\nabla^2 A(\theta)a > 0\end{aligned}$$

  Thus, $\nabla^2 A(\theta) \succ 0 \Rightarrow A$ is strictly convex. Also, this implies that ∇A is a one-one function.

- ∇A is onto $Int(\mathcal{M})$ ($Int(\cdot)$ represents the interior of a set). This means that for any $\mu \in Int(\mathcal{M})$, $\exists\, \theta \in \Omega$ such that $\mathbb{E}_{p_\theta}[\phi(X)] = \mu$.

Some properties of the marginal polytope $\mathcal{M}$ are :

- $\mathcal{M}$ is a convex set (*Verify as an exercise*)
- $\mathcal{M} = \text{conv}\{\phi(X) \mid \forall\ X \in \mathcal{X}\}$, which means that $\mathcal{M}$ is the convex hull of all points $\{\phi(X)\}\ \forall\ X \in \mathcal{X}$. This implies that $\mathcal{M}$ is a polytope (and hence its name). This also implies that $\mathcal{M}$ can be written as an intersection of halfspaces i.e.

$$\mathcal{M} = \cap_j \{u \mid a_j^T u \geq b_j\}$$

The mean parameters have an important role in marginal inference. Consider the following examples :

- **Ising Model** For the Ising Model, we have the sufficient statistics $\{X_s\}_{s\in V} \cup \{X_sX_t\}_{(s,t)\in E}$. Then,

$$\mu_s = \mathbb{E}_p[X_s] = P(X_s = 1) \text{ and } \mu_{st} = \mathbb{E}_p[X_sX_t] = P(X_s = 1, X_t = 1)$$

  The mean vector μ is the vector of all μ_s and μ_{st}.

- **Potts Model** For the Potts Model, we have the sufficient statistics $\{\mathbb{I}_{s,j}(X)\}_{s\in V,j} \cup \{\mathbb{I}_{st,jk}(X)\}_{(s,t)\in E,j,k}$. Then,

$$\mu_{s,j} = \mathbb{E}_p[\mathbb{I}_{s,j}(X)] = P(X_s = j) \text{ and } \mu_{st,jk} = \mathbb{E}_p[\mathbb{I}_{st,jk}(X)] = P(X_s = j, X_t = k)$$

Thus, clearly, in both the ising model and the potts model, the mean parameters correspond to marginals. In general, the problem of computing the forward mapping ∇A can be considered an important inference problem for exponential families.

The problem of computing the reverse mapping $(\nabla A)^{-1}$, also has an interesting interpretation in terms of the max log-likelihood estimate of θ for a given set of samples. Suppose we have a set of p samples $D = \{X^1, X^2, \ldots, X^p\}$ that have been independently drawn from an exponential family for which θ is unknown. Suppose that we choose to obtain an estimate $\hat{\theta}$ by maximizing the likelihood of the data. Then,

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta\in\Omega} \sum_{i=1}^{p} \log\left(P(X^i;\theta)\right)\frac{1}{p}\\ &= \arg\max_{\theta\in\Omega} \sum_{i=1}^{p} \frac{\theta^T\phi(X^i) - A(\theta)}{p}\\ &= \arg\max_{\theta\in\Omega}\ \theta^T\hat{\mu} - A(\theta), \text{ where } \hat{\mu} = \sum_{i=1}^{p}\frac{\phi(X^i)}{p}\end{aligned}$$

Thus, $\hat{\theta}$ satisfies

$$\begin{aligned}\nabla A(\hat{\theta}) &= \hat{\mu}\\ \Rightarrow \hat{\theta} &= (\nabla A)^{-1}(\hat{\mu})\end{aligned}$$

Thus, finding the MLE solution $\hat{\theta}$, is equivalent to computing the backward mapping (under suitable conditions when it is unique).

Let us now look at the conjugate dual of the log partition function A. This helps in providing a variational representation to A and can be used to approximate the log-partition function, as we shall see.

For $A : \Omega \to \mathbb{R}$, we have its fenchel conjugate defined as :

$$A^*(\mu) = \sup_{\theta \in \Omega} \theta^T \mu - A(\theta)$$

Note that $\theta^T \mu - A(\theta)$ is simply the negative entropy for the distribution $p(X;\theta)$. This is because :

$$\begin{aligned} \log p(X;\theta) &= \theta^T \phi(X) - A(\theta) \\ \Rightarrow \mathbb{E}[\log p(X;\theta)] &= \theta^T \mu - A(\theta) \\ \Rightarrow \ -H(p(X;\theta)) &= \theta^T \mu - A(\theta) \end{aligned}$$

Now,

$$\begin{aligned} A^*(\mu) &= \sup_{\theta \in \Omega} \theta^T \mu - A(\theta) \\ &= \theta_\mu^T \mu - A(\theta_\mu) \text{ for some } \theta_\mu \text{ such that } \nabla A(\theta_\mu) = \mu, \text{ i.e. } \theta_\mu = (\nabla A)^{-1}(\mu) \\ &= -H(p(X;\theta_\mu)) \end{aligned}$$

However, such a θ_μ may not always exist. It turns out that,

$$A^*(\mu) = \begin{cases} -H(p(X;\theta_\mu)) & \text{if } \mu \in \mathcal{M} \\ +\infty & \text{if } \mu \notin Cl(\mathcal{M}) \end{cases}$$

where $Cl(\cdot)$ represents the closure of a set, i.e. the set with all its limit points. For any $\mu \in Cl(\mathcal{M}) \setminus \mathcal{M}$ (i.e. μ on boundary of $\mathcal{M}$), we have

$$A^*(\mu) = \lim_{k \to +\infty} A^*(\mu_k) \text{ for any sequence } \{\mu_k\}_{k=1}^{\infty} \to \mu \text{ such that } \mu_k \in \mathcal{M} \ \forall \ k \in \mathbb{N}$$

Then, the conjugate of A^* is,

$$\begin{aligned} A^{**}(\theta) &= \sup_{\mu \in \mathbb{R}^d} \theta^T \mu - A^*(\mu) \\ &= \sup_{\mu \in \mathcal{M}} \theta^T \mu - A^*(\mu) \end{aligned}$$

Now, since A is known to be convex, we must have $A(\theta) = A^{**}(\theta)$. Thus,

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \theta^T \mu - A^*(\mu)$$

The above is a convex optimization problem. But, computing $A(\theta)$ is still intractable since we do not have a closed form for $A^*(\mu)$ or $\mathcal{M}$ and describing $\mathcal{M}$ explicitly can require up to an exponential number of constraints. However, the above formulation does give a direction in which to proceed to approximate $A(\theta)$ viz. by approximating the set $\mathcal{M}$ and the function $A^*(\mu)$.

Recall that in the Potts model, we had the mean parameters $\{\mu_{s,j} \ \forall \ s \in V, j\} \cup \{\mu_{st,jk} \ \forall \ (s,t) \in E, j, k\}$. Using these, we define the set satisfying the obvious constraints on μ,

$$L = \{\mu \in \mathbb{R}^d \mid \mu_{s,j} \geq 0, \mu_{st,jk} \geq 0, \sum_j \mu_{s,j} = 1, \sum_k \mu_{st,jk} = \mu_{s,j}\}$$

Then, it can be shown that

- $\mathcal{M} \subseteq L$
- $\mathcal{M} = L$ if the underlying graph for the distribution is a tree.

Note that L has been described by only a polynomial number of constraints.

To approximate $A^*(u)$, one can use the Bethe function (which is convex), defined as :

$$B^*(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

where $H_s(\mu_s) = -\sum_{X_s \in \mathcal{X}_s} \mu_{s,X_s} \log(\mu_{s,X_s})$ and

$$I_{st}(\mu_{st}) = \sum_{X_s \in \mathcal{X}_s, X_t \in \mathcal{X}_t} \mu_{st,X_s X_t} \log \left(\frac{\mu_{st,X_s X_t}}{\sum_t \mu_{st,X_s X_t} + \sum_s \mu_{st,X_s X_t}} \right)$$

Therefore, the optimization problem becomes :

$$\sup_{u \in L} \theta^T \mu - B^*(\mu)$$

It can be shown that the Sum-Product algorithm solves the above optimization problem.