CS395T Graphical Models: Scribe Notes

Nevzat Onur Domaniç
onur@cs.utexas.edu

October 29, 2012

1 Properties of $A(\theta)$

Consider the distribution p_θ for some $\theta \in \Omega$. Remember $\mu = \mathbb{E}_{p_\theta}[\phi(X)] = \nabla A(\theta)$.

I. $\nabla A : \Omega \to \mathcal{M}$ is one-to-one iff the exponential family is minimal.

Proof. If the exponential family is not minimal then $\exists \theta_1, \theta_2$ such that $p_{\theta_1} = p_{\theta_2}$ and $\theta_1 \neq \theta_2$. Then $\nabla A(\theta_1) = \nabla A(\theta_2)$ hence ∇A is not one-to-one.

If the exponential family is minimal then A is strictly convex, $\nabla^2 A(\theta) \succ 0$, so ∇A is one-to-one. □

II. $\forall \mu \in Int(\mathcal{M})$, $\exists \theta(\mu) \in \Omega$ such that $\mathbb{E}_{p_{\theta(\mu)}}[\phi(x)] = \mu$, in other words $\nabla A(\theta(\mu)) = \mu$, so ∇A is onto $Int(\mathcal{M})$.

2 Fenchel-Conjugate of A

For $\mu \in \mathbb{R}^d$ the Fenchel-Conjugate function of A is defined as:

$$A^*(\mu) := \sup_{\theta \in \Omega} \{\langle \theta, \mu \rangle - A(\theta)\}$$

Theorem 1.

(I) $\forall \mu \in Int(\mathcal{M})$, $A^*(\mu) = -H(p_{\theta(\mu)})$ (negative entropy)

(II) $\forall \mu \notin Cl(\mathcal{M})$, $A^*(\mu) = +\infty$

(III) $\forall \mu \in \partial\mathcal{M}$, there exists a sequence $\{\mu_n\} \subset Int(\mathcal{M})$ converging to m, and $A^*(\mu) = \lim_{n \to \infty} A^*(\mu_n)$

Proof. We prove property (I). Since $\mu \in Int(\mathcal{M})$, $\exists \theta(\mu) \in \Omega$ such that $\nabla A(\theta(\mu)) = \mu$. Remember the definition of A^*:

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

The θ that achieves the supremum is $\theta(\mu)$, so $A^*(\mu) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu))$.

From the definition of the entropy:

$$\begin{aligned} -H(p_\theta) &= \mathbb{E}_{p_\theta} [\log p_\theta(x)] \\ &= \mathbb{E}_{p_\theta} [\langle \theta, \phi(x) \rangle - A(\theta)] \\ &= \langle \theta, \mu(\theta) \rangle - A(\theta) \end{aligned}$$

so $-H(p_{\theta(\mu)}) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) = A^*(\mu)$. □

3 Variational Principle

Theorem 2.

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

In other words, $A(\theta) = A^{**}(\theta)$, *i.e.*, A is the Fenchel-Conjugate of A^*.

Note that this is a convex optimization problem since both $\langle \theta, \mu \rangle$ and the negative entropy is convex. However specifying the objective function and the convex set takes exponential time.

So in Graphical Model Inference we

(a) Approximate $\mathcal{M}$

(b) Approximate entropy $-A^*(\mu)$

How to compute $A^*(\mu)$:

(i) Obtain $\theta(\mu)$ such that $\nabla A(\theta(\mu)) = \mu$.

(ii) Compute $-H(p_{\theta(\mu)})$

Let's look at an example:

3.1 Example: Bernoulli Distribution

Recall $p(x;\theta) = \exp\{x\theta - \log(1+\exp(\theta))\}$ for $x \in \{0,1\}$.

Let's check if $A^*(\mu) = \mu\log\mu + (1-\mu)\log(1-\mu)$. From the definition of the Fenchel-Conjugate we have:

$$A^*(\mu) = \sup_\theta \{\theta\mu - \log(1+\exp(\theta))\}$$

Taking derivative w.r.t. θ and setting it to zero:

$$\mu = \frac{\exp\theta}{1+\exp\theta} \implies \exp\theta = \frac{\mu}{1-\mu} \implies \theta = \log\frac{\mu}{1-\mu}$$

so

$$\begin{aligned} A^*(\mu) &= \theta\mu - \log(1+\exp(\theta(\mu))) \\ &= \mu\log\frac{\mu}{1-\mu} + \log(1-\mu) \\ &= \mu\log\mu + (1-\mu)\log(1-\mu) \end{aligned}$$

Now let's verify $A(\theta)$. We have:

$$A(\theta) = \sup_\mu \{\theta\mu - \mu\log\mu - (1-\mu)\log(1-\mu)\} \tag{1}$$

Taking derivative w.r.t. μ and setting it to zero:

$$\theta = \log\mu + 1 - 1 - \log(1-\mu) = \log\frac{\mu}{1-\mu} \implies \mu = \frac{\exp\theta}{1+\exp\theta}$$

Substituting μ in equation (1) gives $A(\theta) = \log(1+\exp\theta)$.

3.1.1 Approximating Polytope in Discrete Models

$X = (X_1, \ldots, X_p)$ where $X_s \in \mathcal{X} = \{0, \ldots, r-1\}$. We have:

$$p_\theta(X) = \exp\left\{\sum_{s,j} \mathbb{1}_{s,j}(X)\theta_{sj} + \sum_{\substack{(s,t)\in E \\ j,k}} \mathbb{1}_{st,jk}(X)\theta_{st,jk} - A(\theta)\right\}$$

We introduce the following shorthand notations:

$$\begin{aligned} \theta_s(X_s) &= \sum_j \mathbb{1}_{s,j}(X_s)\theta_{s,j} \\ \theta_{st}(X_s, X_t) &= \sum_{j,k} \mathbb{1}_{st,jk}(X_s, X_t)\theta_{st,jk} \end{aligned}$$

so $p_\theta(X)$ becomes $= \exp\left\{\sum_s \theta_s(X_s) + \sum_{(s,t)\in E} \theta_{st}(X_s, X_t) - A(\theta)\right\}$.

Let's have

$$\phi_s(X_s) = \sum_{j\in\mathcal{X}} \mathbb{1}(X_s = j)\phi_s(j)$$

and

$$\phi_{st}(X_s, X_t) = \sum_{j,k\in\mathcal{X}} \mathbb{1}(X_s = j, X_t = k)\phi_{st}(j,k)$$

so any sufficient statistics over discrete random variables can be expressed as a linear combination of indicator functions, thus the overcomplete representation is general.

Now let's look at the mean parameters. Remember that the Marginal Polytope $\mathbb{M}(G)$ is defined as:

$$\mathbb{M}(G) = \{\mu : \exists p, \mu = \mathbb{E}_p\left[\phi(x)\right]\}$$

The mean parameters are:

$$\mu_{s,j} = \mathbb{E}\left[\mathbb{1}_{s,j}(X_s)\right] = \mathbb{P}(X_s = j)$$
$$\mu_{st,jk} = \mathbb{E}\left[\mathbb{1}_{st,jk}(X_s, X_t)\right] = \mathbb{P}(X_s = j, X_t = k)$$

So we define:

$$\mu_s(X_s) = \sum_j \mathbb{1}_{s,j}(X_s)\mu_{s,j}$$
$$\mu_{st}(X_s, X_t) = \sum_{jk} \mathbb{1}_{st,jk}(X_s, X_t)\mu_{st,jk}$$

Then the Marginal Polytope is:

$$\mathbb{M}(G) = \left\{\mu_s(X_s), \mu_{st}(X_s, X_t) : \begin{array}{l}\text{node-wise and pair-wise}\\ \text{marginals of some distribution } p\end{array}\right\}$$

Note that we have the following set of constraints:

- $$\mu_{s,j} \in [0,1]$$
  $$\mu_s(X_s) \in [0,1]$$
  $$\mu_{st}(X_s, X_t) \in [0,1]$$

- Normalization Constraints

$$\sum_{X_s} \mu_s(X_s) = 1$$
$$\sum_{X_s, X_t} \mu_{st}(X_s, X_t) = 1$$

- Marginalization Constraints

$$\sum_{X_t} \mu_{st}(X_s, X_t) = \mu_s(X_s)$$
$$\sum_{X_s} \mu_{st}(X_s, X_t) = \mu_t(X_t)$$

These are called the "local" constraints and the number of constraints is $\mathcal{O}(|V| + |E|)$.

We define:

$$\mathbb{L}(G) = \left\{ \begin{array}{l} T_s(X_s) \\ T_{st}(X_s, X_t) \end{array} : T_s, T_{st} \text{ satisfy "local" constraints} \right\}$$

where T_s and T_{st} are called the pseudo-marginals.