PGM Course Notes: Mean Field Approximations

Karl Pichotta
pichotta@cs.utexas.edu

November 5, 2012

1 Motivation; restricting Ω

So we have

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

We looked at approximating this in general, but we can also approximate by looking at tractable subgraphs of G.

So we have

$$\begin{aligned} p(x, \theta, G) &= \exp\{\langle \theta, \phi(x) \rangle - A(\theta) & (1) \\ &= \sum_{(s,t) \in E} \theta_{st} \phi_{st}(x_s, x_t). & (2) \end{aligned}$$

What do we do?

$$\begin{aligned} A(\theta) &= \sum_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^{\star}(\mu)\} & (3) \\ &\geq \langle \theta, \mu \rangle - A^{\star}(\mu) \quad \forall \mu \in \mathcal{M} & (4) \end{aligned}$$

We have a graph $G = (V, E)$. Define $\Theta_{(s,t)}$ to be the parameter subvector of all parameters on edge (s, t).

For example, if we have an Ising model, then $\Theta_{(s,t)} = \theta_{st}$, that is, a single param.

As another example, if we have the canonical overcomplete model,

$$\Theta_{(s,t)} = (\theta_{st;jk}) \equiv \{\theta_{st;jk} : j, k \in \mathcal{X}\}$$

So parameters in general have the restriction

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

we can restrict our attention to a subset of Ω:

$$\Omega(F) \equiv \{\theta \in \Omega : \theta_\alpha = 0 \quad \forall \alpha \in I(G) \setminus I(F)\}$$

for a subgraph $F \subset G$, with $I(G)$ the index set of parameters given graph G. That is to say, we are requiring the parameters not in $I(F)$ to be 0. So, for example, if we just have edge parameters, then this would be equivalent to setting all edge params not in F to 0.

2 Some examples

Example, Consider $F_0 \equiv (V, \emptyset)$, the null graph of G with no edges. Then

$$\Omega(F_0) = \{\theta \in \Omega : \theta_\alpha = 0 \forall \alpha \in I(G) \setminus I(F_0)\}$$

Example, Consider the Ising model

$$p(x;\theta) = \exp\left\{ \sum_{(s,t)\in E} \theta_{st} x_s x_t + \sum_s \theta_s x_s - A(\theta) \right\}$$

Our set

$$\Theta = \{\{\theta_s\}_{s\in V}; \}\theta_{st}\}_{(s,t)\in E}\}$$

Ω is $R^{|V|+|E|}$. What is $\Omega(F_0)$? It is

$$\Omega(F_0) = \{\theta \in \Omega : \theta_{st} = 0 \forall (s,t) \in E\} \subseteq \mathbb{R}_{|V|+|E|}$$

So if $\theta \in \Omega(F_0)$, then this is a distribution respecting the graph F_0, and therefore must factor over F_0. This means

$$p(x;\theta) = \prod_{s\in V} p(x_s;\theta) \tag{5}$$

$$= \prod_{s\in V} \exp\{\theta_s x_s - \dots\} \tag{6}$$

Again, this is most clearly viewed as a graphical model distribution respecting graph F_0.

Example, If F is a tree T, then $\Omega(T)$ has 0 for the parameters not on the tree T, and so will factorize according to a graph structure.

3 Mean Field Approximation: Definition

So we have this correspondence between Ω and $\mathcal{M}$ parameters:

$$\Omega \to \mathcal{M}(G)$$
$$\Omega_F(G) \to \mathcal{M}_F(G)$$

What is $\mathcal{M}_F(G)$? Recall that

$$\nabla A(\Omega_F(G))$$

gives us the relative interior of $M_F(G)$. Thus, The closure $Cl(\nabla A(\Omega_F(G)))$ gives us the entirety of $\mathcal{M}_F(G)$. Take that as the definition of $\mathcal{M}_F(G)$.

$$A(\theta) \geq \langle \theta, \mu \rangle - A^\star(\mu) \tag{7}$$

for any μ; we saw this above.

We have Ω and $\mathcal{M}$. The set $\mathcal{M}$ is the image of Ω under ∇A (modulo boundaries, I believe). So the image $\nabla A[\Omega(F)] \subseteq \mathcal{M}$. We will later prove that $\mathcal{M}_F(G) \subseteq \mathcal{M}(G)$.

Returning to an equation above, we have

$$A(\theta) \geq \sum_{\mu \in \mathcal{M}_F(G)} \langle \theta, \mu \rangle - A^\star(\mu) \tag{8}$$

which follows from (7) and our observation that $\mathcal{M}_F(G)$ is in $\mathcal{M}(G)$. So mean field is, effectively, solving (8).

The **mean-field approximation** is:

$$A(\theta) \geq \sup_{\mu \in \mathcal{M}_F(G)} \{\langle \theta, \mu \rangle - A^\star(\mu)\} \tag{9}$$

For the mean-field $F = F_0 \equiv (V, \emptyset)$. On the otherhand, **Structured mean-field** is the more general case where F is a subgraph of G.

Note that as we take bigger and bigger subgraphs F, we get tighter lower bounds (i.e. better approximations). If we have a densely connected graph with lots of symmetry, mean-field performs surprisingly well.

4 Examples of Mean Field

Consider the Ising model:

$$p(x; \theta) = \exp\left\{ \sum_{(s,t)} \theta_{st} x_s x_t + \sum_s \theta_s x_s - A(\theta) \right\}$$

We have $\Omega = \mathbb{R}^{|V|+|E|}$ and $\Omega(F_0)$ (see above). Now, we have

$$\mathcal{M}_{F_0}(G) = \left\{ \{\mu_s\}_{s \in V}; \{\mu_{st}\}_{(s,t) \in E} \quad : \text{(some conditions defined below);} \right\}$$

such that we get these as expectations of distributions respecting F_0. Now, since there are no edges, we have the distribution of any variable being independent of all other variables. If we have independence, we also have X_s and X_t uncorrelated:

$$\mathbb{E}[X_s X_t] = \mathbb{E}[X_s]\mathbb{E}[X_t].$$

Note the LHS above is exactly μ_{st}. We therefore have a refinement of $\mathcal{M}_F(G)$:

$$\mathcal{M}_{F_0}(G) = \left\{ \{\mu_s\}_{s \in V}; \{\mu_{st}\}_{(s,t) \in E}; \quad : \mu_{st} = \mu_s \mu_t; \mu_s \in [0,1] \right\}$$

So this is the set of numbers between 0 and 1 (the μ_s's) and then products of them. So it's a unit hypercube.

Now, for $\mu \in \mathcal{M}_{F_0}(G)$, what is $A^\star(\mu)$? It is:

$$-A^\star(\mu) = \sum_{s \in V} H_S(\mu_s)$$

Because, since the distributions are all independent, to get the entropy of the joint (which is the negative of $A^\star(\mu)$), we just add the entropies—this is a basic property of entropy.

So cool, putting it together:

$$\sup_{\mu \in \mathcal{M}_{F_0}(G); \mu_s \in [0,1]} \left\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_s \left(\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s) \right) \right\} \tag{10}$$

again, the sup just ranges over the unit hypercube. Recall this is our lower bound for $A(\theta)$ as given by the equation above. This is nonconvex in general because of the second term, which is quadratic :(. If we maximize it we'll get a local maximum (which is better than nothing).

The standard mean-field updates are a way to solve this coordinate ascent problem.

Coordinate ascent is the problem where, supposing we have

$$\min_{\mu \in B} (f(u))$$

we iterate over $j \in \{1, \ldots, p\}$, one of our dimensions (assuming $\mu \in \mathbb{R}^p$). Fix all the variables other than μ_j, and then optimize over μ_j. Then repeat. That is, we iteratively optimize over a single coordinate by treating all the other coordinates as fixed. This is really simple (easy to code, and optimizing over a single variable is easy) and works pretty well a lot of the time. We apply this to our problem.

Suppose we use coordinate descent to optimize over μ_s. What's the gradient wrt μ_s? It is:

$$\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t - \log \mu_s - 1 + \log(1 - \mu_s) + 1$$

with $N(s)$ the set of neighbors of s. (recall we get this from (10)). Setting this to 0 gives us

$$\begin{aligned} \log \frac{\mu_s}{1 - \mu_s} &= \theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t \\ \frac{\mu_s}{1 - \mu_s} &= \exp(\ldots) \\ \mu_s = \frac{\exp(...)}{1 + \exp(...)} &= \sigma(\theta_s + \sum_{t \in N(s)} \theta_{st} \mu_t) \end{aligned}$$

with $\sigma(x)$ the logistic function. So we get our coordinate descent rule:

$$\mu_s \leftarrow \sigma(\theta_s + \sum_{t \in N(s)} \theta_{st \mu_t})$$

which is much simpler than sum-product.

5 Framing in terms of KL divergence

KL divergence is a typical way to look at the difference between distributions p and q. The KL Divergence between them is

$$
\begin{aligned}
D(p\|q) &\equiv \mathbb{E}_p\left[\log\frac{p(x)}{q(x)}\right] && (11)\\
&= \sum_x p(x)\log\frac{p(x)}{q(x)} && (12)
\end{aligned}
$$

Note this looks a lot like entropy (it's sometimes called the relative entropy of p and q). Note also it's not symmetric, so not a distance metric.

So ok, let $p \equiv P_{\theta_1}$ and $q \equiv P_{\theta_2}$, the distributions of params θ_1 and θ_2. We have

$$
\begin{aligned}
D(p\|q) &= \log\frac{P_{\theta_1}(x)}{P_{\theta_2}(x)} && (13)\\
&= \langle\theta_1, \phi(x)\rangle - A(\theta_1) - \langle\theta_2, \phi(x)\rangle - A(\theta_2)\rangle && (14)\\
&= -A(\theta_1) - A(\theta_2) - E_{P_{\theta_1}}\left[\langle\phi(x), \theta_2 - \theta_1\rangle\right] && (15)\\
&= -A(\theta_1) - A(\theta_2) - \langle\mu_1, \theta_2 - \theta_1\rangle && (16)\\
&= -A(\theta_1) - A(\theta_2) - \langle\nabla A(\theta_1), \theta_2 - \theta_1\rangle && (17)
\end{aligned}
$$

where (16) hides an application of linearity of expectation, and then uses the definition of $E[\phi(x)] = \mu$, and (17) uses the fact that $\nabla A(\theta_1) = \mu_1$ (this is mentioned in one of the previous notes).

Note that the first-order Taylor approximation of $A(\theta)$ at θ_1 is

$$A(\theta_1) + \nabla A(\theta_1)(\theta - \theta_1)$$

At a new point θ_2, we look at how far apart the above Taylor approximation is from $A(\theta_2)$, and we get exactly the value (17) (this is sometimes called the Bregman divergence).

So OK, another way of looking at this: suppose we have θ from some graph G. Suppose we don't like that because it's not a tree or something. We ask, then, what is

$$\min_{\overline{\theta}\in\Omega(F)} D(P_{\overline{\theta}}\|P\theta)$$

for some F that we like. That is, give us the distribution that's closest (under KL) such that it factorizes according to F. Turns out this is exactly the variational principle we derived above.

We have:

$$
\begin{aligned}
&\min_{\overline{\theta}\in\Omega(F)} D(P_{\overline{\theta}}\|P_\theta) && (18)\\
&= \min_{\overline{\theta}\in\Omega(F)} \left\{-A(\overline{\theta}) - A(\theta) - \langle\nabla A(\overline{\theta}), \theta - \overline{\theta}\rangle\right\} && (19)\\
&= \min_{\overline{\theta}\in\Omega(F)} \left\{-A(\overline{\theta}) - \langle\nabla A(\overline{\theta}), \theta - \overline{\theta}\rangle\right\} && (20)
\end{aligned}
$$

So we've expressed everything in terms of $\overline{\theta}$. However, we already saw the connection between μ's and θ's:

$$A(\overline{\theta}) = \langle \overline{\mu}, \overline{\theta} \rangle - A^\star(\overline{\mu})$$

And so we can write (20) as:

$$\begin{aligned}
(20) &= \min_{\overline{\theta} \in \Omega(F)} \left\{ -\langle \overline{\mu}, \overline{\theta} \rangle + A^\star(\overline{\mu}) - \langle \nabla A(\overline{\theta}), \theta - \overline{\theta} \rangle \right\} && (21) \\
&= \min_{\overline{\theta} \in \Omega(F)} \left\{ A^\star(\overline{\mu}) - \langle \nabla A(\overline{\theta}), \theta \rangle \right\} && (22) \\
&= \min_{\overline{\theta} \in \Omega(F)} \{ A^\star(\overline{\mu}) - \langle \overline{\mu}, \theta \rangle \} && (23) \\
&= \sup_{\mu \in \mathcal{M}_F(G)} \{ \langle \theta, \mu \rangle - A^\star(\mu) \} && (24)
\end{aligned}$$

where in (23) we use the fact that $\nabla A(\overline{\theta}) = \overline{\mu}$ (we used this above too, in (17)). The astute reader will notice that this is simply (9), the mean field approximation.