CS281A/Stat241A: Statistical Learning Theory

2002/12/02: Variational Methods for Approximate Interference

Lecturer: Martin Wainwright *Scribes: David Chinnery*

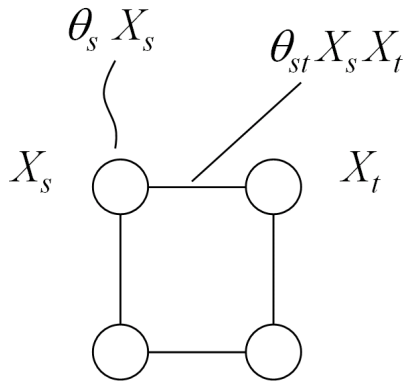
A. Mean Field Updates $x \in \{0,1\}^n$

Ising Model

$$p(x,\theta) \propto \exp(\sum_{s\in V} \theta_{s0} X_s + \sum_{(s,t)\in E} \theta_{st} X_s Y_t)$$

Mean parameters $\mu = \{\mu_s : s \in V\}$, with updates:

$$\mu_s \leftarrow [1 + \exp(-(\theta_{s0} + \sum_{t\in N(s)} \theta_{st}\mu_t))]^{-1}$$



B. Sum-Product Updates

Sum-product updates are an exact technique for trees.
It can also be applied to a graph with cycles, in which case it is approximate.

Exponential families: $x \in X^n, f_i : X^n \to \Re$
$f = \{f_i : i \in I\}$, for index set I, and we let $d = |I|$

$$p(x;\theta) = h(x)\exp(\sum_{i\in I} \theta_i f_i(x) - A(\theta)), \theta \in \Re^d$$

For the Ising model, $f = \{X_s : s \in V\} \cup \{X_s X_t : (s,t) \in E\}$

$$A(\theta) = \log(\int \exp(\sum_{i\in I} \theta_i f_i(x))dx)$$

The conjugate dual function of $A(\theta)$ with dual variable $\mu \in \Re^d$:

$$A^*(\mu) = \max_\theta(\sum_{i\in I} \theta_i \mu_i - A(\theta))$$

Example:
Bernoulli random variable $x \in \{0, 1\}$
$p(x; \theta) = \exp(\theta x - A(\theta))$
$A(\theta) = \log(\sum_{x=0}^{1} \exp(\theta x)) = \log(1 + \exp(\theta))$

Dual variable $\mu \in \Re$

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

Setting the derivative to zero to maximize: $\mu = \exp(\theta)/(1 + \exp(\theta)) = 1/(1 + \exp(-\theta))$
$\Rightarrow \theta = \log(\mu/(1 - \mu))$
$\Rightarrow A^*(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu)$.
This is the negative of the entropy function of the Bernoulli variable with mean μ.

An alternative parametrization for $p(x)$ is $p(x; \mu) = \mu^x (1 - \mu)^{1-x}$

Key points:
1. Derivatives of $A(\theta)$ give means (or marginals).

$$\frac{\partial A(\theta)}{\partial \theta_i} = \int p(x; \theta) f_i(x) dx$$

2. $A^*(\mu)$ is the negative entropy.

$$A^*(\mu) = \max_{\theta}(\sum_{i \in I} \theta_i \mu_i - A(\theta))$$

Take the gradient with respect to θ_i and set to zero:

$\mu_i = \int p(x; \hat{\theta}) f_i(x) dx$, where $\hat{\theta} = \hat{\theta}(\mu)$ is the optimum θ.

$$A^*(\mu) = \sum_{i \in I} \hat{\theta}_i \mu_i - A(\hat{\theta}) = \sum_{i \in I} \hat{\theta}_i (\int p(x; \hat{\theta}) f_i(x) dx) - A(\hat{\theta})$$

$\log p(x; \theta) = \sum_i \hat{\theta}_i f_i(x) - A(\hat{\theta})$
So, the integral is $\int p(x; \theta) \log p(x; \hat{\theta}) dx$, which is the negative entropy.
$A^*(\mu) = -H(\mu)$, where $H(\mu)$ is the entropy.

Realizable marginals:
For which $\mu \in \Re^d$ is $A^*(\mu)$ defined?
$MARG\{f_i\} = \{\mu \in \Re^d : \forall i \in I, \int p(x) f_i(x) dx = \mu_i\}$

e.g. Bernoulli $f(x) = x, \sum_{x=0}^{1} p(x) x = p(x = 1) \in [0, 1]$

If A is convex, then $(A^*)^* = A$. Hence

$$A(\theta) = \max_{\mu \in MARG(f_i)} (\sum_{i \in I} \mu_i \theta_i - A^*(\mu))$$

Comments:
1. This is a variational representation of $A(\theta)$.
2. The optimum is attained at unique $\hat{\mu}_i$.

Maximize entropy $H(p)$, subject to $\int p(x)f_i(x)dx = \hat{\mu}_i$.
We attach some Lagrangian multipliers to solve this,

$$L(p;\theta) = H(p) + \sum_i \theta_i(\int p(x)f_i(x)dx - \hat{\mu}_i)$$

gives an exponential form for the maximum entropy. The distribution that is recovered is of the form

$$p(x;\theta) \propto \exp(\sum_{i\in I} \theta_i f_i(x))$$

In terms of μ

$$L(\mu;\theta) = H(\mu) + \sum_{i\in I} \theta_i(\mu_i - \hat{\mu}_i) + \sum_{i\in I} \theta_i\hat{\mu}_i$$

and we get $\hat{\mu}_i = \int p(x;\theta)f_i(x)dx$.

In general, the constraint set $MARG\{f_i\}$ and the entropy are hard to deal with. Mean field avoids this.

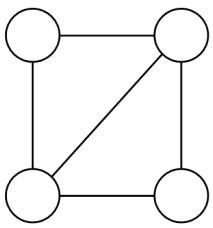
Mean field:
Choose some $\tilde{\mu} \in MARG\{f_i\}$
$A(\theta) \geq \sum \theta_i\tilde{\mu}_i + H(\tilde{\mu})$. We optimize over a tractable subset of μ.
(By applying Jensen's inequality to log, as log is concave, we can show that for any distribution q that $A(\theta) \geq \sum \theta_i(\int q(x)f_i(x)dx) + H(q)$.)

Mean field (Ising Model):
$p(x;\theta) \propto \exp(\sum_{s\in V} \theta_{s0}X_s + \sum_{(s,t)\in E} \theta_{st}X_sX_t)$
$\mu_s = E_\theta[x_s] = p(x_s = 1;\theta)$
$\mu_{st} = E_\theta[x_sx_t] = p(x_s = 1, x_t = 1;\theta)$

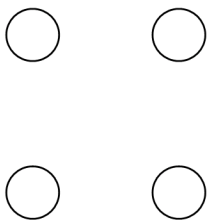
For a fully factorized graph,

$$q_\mu(x) = \prod_{s\in V} \mu_s^{x_s}(1-\mu_s)^{1-x_s}, \mu_s \in [0,1]$$

e.g. We represent



as a fully decoupled graph:



$$H(\mu) = -\sum_{s \in V} (\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s))$$

$\mu_{st} = E_q[x_s x_t] = \mu_s \mu_t$

So, the mean field is

$$A(\theta) \geq \max_{0 \leq \mu_s \leq 1} \{\sum_{s \in V} \theta_{s0} \mu_s + (\sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t) - (\sum_{s \in V} \mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s))\} \qquad (1)$$

where $A(\theta) = \log \sum_x \exp(\sum_s \theta_{s0} x_s + \sum_{(s,t)} \theta_{s,t} x_s x_t)$.

Alternate view: The following is equivalent to (1):

$$\min_{q_\mu \ fully \ factorized} D(q_\mu \parallel p(x; \theta))$$

By taking derivatives with respect to μ_s of (1) and setting derivatives to zero to maximize the expression we get:

$$\theta_{s0} + \sum_{t \in N(s)} \theta_{st} \mu_t - \log \mu_s + \log(1 - \mu_s) = 0$$

$$\Rightarrow \mu_s = 1/(1 + \exp(-\theta_{s0} + \sum_{t \in N(s)} \theta_{st} \mu_t))$$

This effectively is coordinate ascent on (1). It will at least find a local minimum.

Bethe Entropy Approximation and Sum-Product
Pairwise Markov random field

$$p(x) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t)$$

Mean parameters: single node marginal $\mu_s(x_s)$, and joint marginals $\mu_{st}(x_s, x_t)$

$$A(\theta) = \max_{\{\mu_s, \mu_{st}\} \in MARG} \{(\sum_{s \in V} \sum_{x_s} \theta_s(x_s) \mu_s(x_s)) + (\sum_{(s,t) \in E} \sum_{x_s, x_t} \theta_s(x_s, x_t) \mu_s(x_s, x_t)) + H(\mu)\}$$

Normalization conditions:

$$\sum_{x_s} \mu_s(x_s) = 1, \sum_{x_s, x_t} \mu_{st}(x_s, x_t) = 1$$

Marginalization constraints:

$$\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)$$

Define constraint set

$$TREE(G) = \{\mu : \sum_s \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\}$$

(i) If the graph is a tree, then $TREE(T) = MARG(T)$
(ii) In general, $MARG(G) \subset TREE(G)$

Recall for a tree:

$$p(x) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

$$-\sum_x p(x) \log p(x) = -\sum_{s \in V} \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s) - \sum_{(s,t) \in E} \sum_{x_s, x_t} \mu_s(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

$$= \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_s t)$$

where H_s is the entropy, and I_{st} is the mutual information.

This gives the Bethe approximation:

$$\tilde{H}_{Bethe}(\{\mu_s, \mu_t\}) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

$$\max_{\{\mu_s, \mu_{st}\} \in TREE(G)} \{(\sum_{s \in V} \sum_{x_s} \theta_s(x_s)\mu_s(x_s)) + (\sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\mu_{st}(x_s, x_t)) + \tilde{H}_{Bethe}(\mu)\}$$

The Bethe variational problem with Lagrangian:

$$L(\mu; \lambda) = G(\mu) + \sum_{(s,t)} \lambda_{st}(x_t)(\mu_t(x_t) - \sum_{x_s} \mu_(st)(x_s, x_t)) + \lambda_{ts}(x_s)[...]$$

$$\Rightarrow \mu_s(x_s) = \exp(\theta_s(x_s)) \prod_{t \in N(s)} \exp(\lambda_{ts}(x_s))$$