CS731 Spring 2011 Advanced Artificial Intelligence

Variational Methods

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

In this lecture we consider variational methods in inference (sum-product and mean field) and parameter learning (variational EM).

1 Inference: Variational Approximations

Recall that given θ, one can perform inference (equivalent to computing the mean parameters) by solving an optimization problem:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \mu^\top \theta - A^*(\mu), \tag{1}$$

using the fact that the solution is attained uniquely at the desired mean parameter

$$\mu = \mathbb{E}_\theta[\phi(\mathbf{x})]. \tag{2}$$

This is known as the variational principle, where a desired quantity (in this case μ) is defined as a solution to an optimization problem. However, in general (1) is difficult to solve even though it is a convex problem. *Variational approximation* aims to modify the optimization problem so that it is tractable, at the price of arriving at an approximate solution. We will interpret mean field and sum-product algorithms as different variational approximations to (1).

1.1 The Mean Field Method as Variational Approximation

In general, there are two difficulties with (1): (1) the marginal polytope $\mathcal{M}$, albeit convex, can be quite complex to describe and optimize over; (2) The dual function $A^*(\mu)$ does not admit an explicit form. The mean field method replaces $\mathcal{M}$ with a subset $\mathcal{M}(F)$ which is simple and on which $A^*(\mu)$ has a closed form.

Recall that the original exponential family is defined over a graph $G = (V, E)$. Now consider the fully disconnected subgraph $F = (V, \emptyset)$. This subgraph defines a sub-family

$$\Omega(F) = \{\theta \in \Omega \mid \theta_i = 0 \text{ if } \phi_i \text{ involves edges not in } F\}. \tag{3}$$

The densities in this sub-family are all fully factorized:

$$p_\theta(\mathbf{x}) = \prod_{s \in V} p(x_s; \theta_s). \tag{4}$$

F could also be a spanning tree of G or other tractable subgraphs, but we do not consider those cases here.

Clearly, $\Omega(F)$ maps to a subset of $\mathcal{M}$, call it $\mathcal{M}(F)$. Recall when $\{\mathbf{x}\}$ is finite, $\mathcal{M}$ is characterized by the convex hull of extreme points $\{\phi(\mathbf{x})\}$. Each particular extreme point $\phi(\mathbf{x})$ in $\mathcal{M}$ is realized by a distribution p that puts all mass on $\mathbf{x}$. Now we claim that these extreme points are also in $\mathcal{M}(F)$.

Example 1 *For the tiny Ising model $x_1, x_2 \in \{0, 1\}$ with $\phi = (x_1, x_2, x_1x_2)^\top$, the point mass probability $p(\mathbf{x} = (0, 1)^\top) = 1$ is realized as a limit to the series $p(\mathbf{x}) = \exp(\theta_1 x_1 + \theta_2 x_2 - A(\theta))$ where $\theta_1 \to -\infty$ and $\theta_2 \to \infty$. Note this series is in $\Omega(F)$ because $\theta_{12} = 0$. Therefore, the point mass probability on $\mathbf{x} = (0, 1)^\top$ is realizable by $\Omega(F)$ and hence the extreme point $\phi(\mathbf{x}) = (0, 1, 0)$ is in $\mathcal{M}(F)$. The same is true for the other three extreme points.*

Because the extreme points of $\mathcal{M}$ are in $\mathcal{M}(F)$, if the latter were convex, we would have $\mathcal{M} = \mathcal{M}(F)$. Therefore, whenever $\mathcal{M}(F)$ is a true subset of $\mathcal{M}$ (the general case), $\mathcal{M}(F)$ cannot be convex. Instead, $\mathcal{M}(F)$ is a nonconvex inner set of $\mathcal{M}$.

The mean field method is defined simply by replacing $\mathcal{M}$ with $\mathcal{M}(F)$ in (1):

$$\mathcal{L}(\theta) = \sup_{\mu \in \mathcal{M}(F)} \mu^\top \theta - A^*(\mu), \tag{5}$$

Obvious $\mathcal{L}(\theta) \leq A(\theta)$. The solution that achieves $\mathcal{L}(\theta)$ may not be the mean parameter μ (2), depending on whether that $\mu \in \mathcal{M}(F)$ or not. Furthermore, even when that $\mu \in \mathcal{M}(F)$, because $\mathcal{M}(F)$ is nonconvex, in practice we may not be able to find it (instead we might get stuck in a local maximum). Therefore, the mean field problem (5) is fraught with difficulties. Then, why would one want to use the mean field method? The key lies in the fact that $A^*(\mu) = -H(p_\theta(\mu))$ has a very simple form for $\mu \in \mathcal{M}(F)$, as the following example shows.

Example 2 (Mean Field for Ising Model) *Recall that the Ising model has mean parameters which are the node and edge marginals: $\mu_s = p(x_x = 1), \mu_{st} = p(x_s = 1, x_t = 1)$. Since $\mathcal{M}(F)$ corresponds to the fully factorized product distributions (4), its mean parameters are simply defined by the μ_s's, with the edge marginals begin $\mu_{st} = \mu_s \mu_t$. For such μ's, the dual function $A^*(\mu) = -H(p_\theta(\mu))$ has the simple form*

$$A^*(\mu) = \sum_{s \in V} -H(\mu_s) = \sum_{s \in V} \mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s). \tag{6}$$

Thus the mean field problem (5) can be written as

$$\begin{aligned} \mathcal{L}(\theta) &= \sup_{\mu \in \mathcal{M}(F)} \mu^\top \theta - \sum_{s \in V} \left(\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)\right) & (7) \\ &= \max_{(\mu_1 \ldots \mu_m) \in [0,1]^m} \left(\sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H(\mu_s) \right) & (8) \end{aligned}$$

This is a concave problem in a single dimension μ_s. An iterative coordinate-wise maximization (fixing μ_t for $t \neq s$ and optimizing μ_s) procedure can be derived by setting the partial derivative w.r.t. μ_s to 0. This yields the update

$$\mu_s = \frac{1}{1 + \exp\left(-(\theta_s + \sum_{(s,t) \in E} \theta_{st} \mu_t)\right)}. \tag{9}$$

We therefore derived the mean field algorithm for Ising model in a previous lecture.

However, (8) is not jointly concave in $\mu_1 \ldots \mu_m$. Therefore, the iterative procedure will converge to a local maximum of (8) depending on the initialization of $\mu_1 \ldots \mu_m$. It may not reach the lower bound $\mathcal{L}(\theta)$ (though it is guaranteed to produce a lousier lower bound).

★ *To see how a function that is concave in each dimension may not be concave jointly, consider $f(x, y) = xy$.*

1.2 The Sum-Product Algorithm as Variational Approximation

The sum-product algorithm makes two approximations to the variational problem (1): it relaxes $\mathcal{M}$ to an *outer* set, and replaces the dual A^* with an approximation.

Recall that for standard overcomplete exponential families on discrete nodes, the mean parameter is $\mu = (\ldots \mu_{sj} \ldots \mu_{stjk} \ldots) \in \mathbb{R}^d_+$ where $\mu_{sj} = p(x_s = j), \mu_{stjk} = p(x_s = j, x_t = k)$. The marginal polytope is

$\mathcal{M} = \{\mu \mid \exists p$ with node and edge marginals $\mu\}$. Now consider non-negative vectors $\tau \in \mathbb{R}^d_+$ satisfying the following conditions:

$$\sum_{j=0}^{r-1} \tau_{sj} = 1 \qquad \forall s \in V \tag{10}$$

$$\sum_{k=0}^{r-1} \tau_{stjk} = \tau_{sj} \qquad \forall s,t \in V, j = 0 \ldots r-1 \tag{11}$$

$$\sum_{j=0}^{r-1} \tau_{stjk} = \tau_{tk} \qquad \forall s,t \in V, k = 0 \ldots r-1. \tag{12}$$

These can be understood as node normalization and edge-node marginal consistency conditions, respectively. Now define $L = \{\tau$ satisfying the above conditions$\}$. Clearly $\mathcal{M} \subseteq L$. It turns out if the graph has a tree structure, then $\mathcal{M} = L$. But if the graph has cycles then $\mathcal{M} \subset L$ (i.e., L is too lax to satisfy some other constraints that true marginals need to satisfy; see example 4.1 in Wainwright & Jordan). However, L is a much simpler set than $\mathcal{M}$. The first approximation in sum-product is to replace $\mathcal{M}$ with L in the variational problem (1).

The second approximation is on $A^* = -H(p)$. First we point out that if the graph is a tree, one can exactly reconstruct the *joint* probability p_μ from μ (which only specifies node and edge marginals) as follows:

$$p_\mu(\mathbf{x}) = \prod_{s \in V} \mu_{sx_s} \prod_{(s,t) \in E} \frac{\mu_{stx_sx_t}}{\mu_{sx_s}\mu_{tx_t}}. \tag{13}$$

And when the graph is a tree, the entropy of the joint distribution above is easy to compute:

$$\begin{aligned} H(p_\mu) &= -A^*(\mu) && (14)\\ &= \sum_{s \in V} H(\mu_s) - \sum_{(s,t) \in E} I(\mu_{st}) && (15)\\ &= -\sum_{s \in V} \sum_{j=0}^{r-1} \mu_{sj} \log \mu_{sj} - \sum_{(s,t) \in E} \sum_{j,k} \mu_{stjk} \log \frac{\mu_{stjk}}{\mu_{sj}\mu_{tk}}. && (16) \end{aligned}$$

Note neither (13) nor (16) holds for graph with cycles. Nonetheless, we define the *Bethe entropy* for $\tau \in L$ on loopy graphs in the same way:

$$H_{Bethe}(p_\tau) = -\sum_{s \in V} \sum_{j=0}^{r-1} \tau_{sj} \log \tau_{sj} - \sum_{(s,t) \in E} \sum_{j,k} \tau_{stjk} \log \frac{\tau_{stjk}}{\tau_{sj}\tau_{tk}}. \tag{17}$$

Recall that τ is not a true marginal, and H_{Bethe} is not a true entropy. The second approximation in sum-product is to replace $A^*(\tau)$ with $-H_{Bethe}(p_\tau)$.

With these two approximations, we arrive at a different variational problem than (1):

$$A_{sum-product}(\theta) = \sup_{\tau \in L} \tau^\top \theta + H_{Bethe}(p_\tau). \tag{18}$$

This is a constrained optimization problem with constraints $\tau \in L$. Optimality conditions require that the gradients vanish w.r.t. both the primal variables τ and the Lagrangian multipliers on those constraints. The sum-product algorithm can be derived as an iterative fixed point procedure to achieve optimality. Details can be found in section 4.1.3 in Wainwright & Jordan. At the solution, $A_{sum-product}(\theta)$ is not guaranteed to be either an upper or a lower bound of $A(\theta)$, and τ may not correspond to a true marginal distribution. They are approximations.

2 Parameter Learning: Variational Interpretation of EM for Exponential Families

So far, we have focused on the inference problem where the parameter θ is fixed. In what follows, we address the *learning* problem where the parameter is unknown and must be estimated from iid data $\mathbf{x}_1 \ldots \mathbf{x}_n$. The underlying principle will be maximum likelihood. We distinguish the case where we have *fully observed data* where all dimensions of $\mathbf{x}$ are observed, from the case where we have *partially observed data* where some dimensions of $\mathbf{x}$ are unobserved.

2.1 Fully Observed Data

We consider exponential family $p_\theta(\mathbf{x}) = \exp\left(\theta^\top \phi(\mathbf{x}) - A(\theta)\right)$. Given iid data $\mathbf{x}_1 \ldots \mathbf{x}_n$, the log likelihood is

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(\mathbf{x}_i) = \theta^\top \left(\frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)\right) - A(\theta) = \theta^\top \hat{\mu} - A(\theta), \tag{19}$$

where $\hat{\mu} \equiv \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)$ is the mean parameter of the empirical distribution on $\mathbf{x}_1 \ldots \mathbf{x}_n$. Clearly $\hat{\mu} \in \mathcal{M}$. The maximum likelihood principle seeks

$$\theta^{ML} = \arg\sup_{\theta \in \Omega} \theta^\top \hat{\mu} - A(\theta). \tag{20}$$

As stated earlier, the solution is

$$\theta^{ML} = \theta(\hat{\mu}), \tag{21}$$

i.e., the exponential family density whose mean parameter matches $\hat{\mu}$. When $\hat{\mu} \in \mathcal{M}^0$ and ϕ minimal, there is a unique maximum likelihood solution θ^{ML}. The value of the log likelihood function $\ell(\theta^{ML}) = A^*(\hat{\mu}) = -H(p_{\theta^{ML}})$.

2.2 Partially Observed Data

We assume that the value of some nodes in the graphical model are unobserved. We denote each input item as $(\mathbf{x}, \mathbf{z})$ where $\mathbf{x}$ is the observed part and $\mathbf{z}$ the unobserved part. That is, the full data would be $(\mathbf{x}_1, \mathbf{z}_1) \ldots (\mathbf{x}_n, \mathbf{z}_n)$ but we only observe partial data $\mathbf{x}_1 \ldots \mathbf{x}_n$[1]. One can still learn parameters using the maximum likelihood principle on

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(\mathbf{x}_i). \tag{22}$$

However, the difficulty stems from the fact that $p_\theta(\mathbf{x}_i)$ is now the marginal over observed variables (note $\phi(\mathbf{x}, \mathbf{z})$ is defined over the complete data):

$$p_\theta(\mathbf{x}_i) = \int p_\theta(\mathbf{x}_i, \mathbf{z}) d\mathbf{z} = \int \exp\left(\theta^\top \phi(\mathbf{x}, \mathbf{z}) - A(\theta)\right) d\mathbf{z}. \tag{23}$$

In this case, we call $\ell(\theta)$ the incomplete log likelihood:

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \int \exp\left(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}) - A(\theta)\right) d\mathbf{z} = \left(\frac{1}{n}\sum_{i=1}^{n} \log \int \exp\left(\theta^\top \phi(\mathbf{x}_i, \mathbf{z})\right) d\mathbf{z}\right) - A(\theta) \tag{24}$$

EM maximizes a lower bound of the incomplete log likelihood. First consider the conditional probability

$$p_\theta(\mathbf{z} \mid \mathbf{x}_i) = \frac{\exp(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}) - A(\theta))}{\int \exp(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}') - A(\theta)) d\mathbf{z}'}. \tag{25}$$

[1] Each item can have different missing variables and everything follows exactly the same. For notational simplicity we do not consider that here.

Note this is (of course) an exponential family too, since it can be written as

$$p_\theta(\mathbf{z} \mid \mathbf{x}_i) = \exp\left(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}) - \log \int \exp(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}'))d\mathbf{z}'\right) \equiv \exp\left(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}) - A_{\mathbf{x}_i}(\theta)\right), \tag{26}$$

where we defined a new log partition function for this conditional probability conditioned on $\mathbf{x}_i$:

$$A_{\mathbf{x}_i}(\theta) = \log \int \exp(\theta^\top \phi(\mathbf{x}_i, \mathbf{z}'))d\mathbf{z}'. \tag{27}$$

With this, (24) can be written as

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} A_{\mathbf{x}_i}(\theta) - A(\theta) \tag{28}$$

We now lower-bound each $A_{\mathbf{x}_i}(\theta)$ using variational principle. Consider the mean parameter realizable by any distribution on $\mathbf{z}$ while holding $\mathbf{x}_i$ fixed:

$$\mathcal{M}_{\mathbf{x}_i} = \{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_p[\phi(\mathbf{x}_i, \mathbf{z})] \text{ for some } p\}. \tag{29}$$

Recall that the variational definition of $A_{\mathbf{x}_i}(\theta)$ is

$$A_{\mathbf{x}_i}(\theta) = \sup_{\mu \in \mathcal{M}_{\mathbf{x}_i}} \theta^\top \mu - A^*_{\mathbf{x}_i}(\mu). \tag{30}$$

Therefore, for any $\mu^i \in \mathcal{M}_{\mathbf{x}_i}$ we have the trivial variational lower bound

$$A_{\mathbf{x}_i}(\theta) \geq \theta^\top \mu^i - A^*_{\mathbf{x}_i}(\mu^i). \tag{31}$$

This translates to a lower bound $\mathcal{L}$ on the incomplete log likelihood:

$$\ell(\theta) \geq \frac{1}{n}\sum_{i=1}^{n}\left(\theta^\top \mu^i - A^*_{\mathbf{x}_i}(\mu^i)\right) - A(\theta) \equiv \mathcal{L}(\mu^1, \ldots, \mu^n, \theta). \tag{32}$$

2.2.1 Exact EM

The EM algorithm is coordinate ascent on $\mathcal{L}$. In the E step, it optimizes each μ^i in turn for $i = 1 \ldots n$, fixing all other variables:

$$\mu^i \leftarrow \arg\max_{\mu^i \in \mathcal{M}_{\mathbf{x}_i}} \mathcal{L}(\mu^1, \ldots, \mu^n, \theta). \tag{33}$$

The maximization problem on the RHS is equivalent to

$$\text{argmax}_{\mu^i \in \mathcal{M}_{\mathbf{x}_i}} \theta^\top \mu^i - A^*_{\mathbf{x}_i}(\mu^i). \tag{34}$$

We recognize the argmax as the variational representation of the mean parameter

$$\mu^i(\theta) = \mathbb{E}_\theta[\phi(\mathbf{x}_i, \mathbf{z})]. \tag{35}$$

It is this $\mathbb{E}_\theta[]$ operation, under the current parameters θ, that earned it the name "E step." In the M step, it optimizes θ, holding the μ's fixed:

$$\theta \leftarrow \arg\max_{\theta \in \Omega} \mathcal{L}(\mu^1, \ldots, \mu^n, \theta) = \arg\max_{\theta \in \Omega} \theta^\top \hat{\mu} - A(\theta), \tag{36}$$

where we define

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu^i. \tag{37}$$

We recognize this as the standard fully observed maximum likelihood problem, hence the name "M step." The solution is attained at $\theta(\hat{\mu})$ which satisfies the condition

$$\mathbb{E}_{\theta(\hat{\mu})}[\phi(\mathbf{x})] = \hat{\mu}. \tag{38}$$

Furthermore, at the end of E step (35) these μ^i_{new} achieve equality in the variational lower bound (31). Hence the lower bound $\mathcal{L}$ in (32) is tight at this moment:

$$\ell(\theta_{old}) = \mathcal{L}(\mu^1_{new}, \ldots, \mu^n_{new}, \theta_{old}). \tag{39}$$

Therefore, if a subsequent solution θ_{new} to the M step improves upon $\mathcal{L}(\mu^1_{new}, \ldots, \mu^n_{new}, \theta_{old})$, it also improves the incomplete log likelihood:

$$\ell(\theta_{new}) \geq \mathcal{L}(\mu^1_{new}, \ldots, \mu^n_{new}, \theta_{new}) \geq \mathcal{L}(\mu^1_{new}, \ldots, \mu^n_{new}, \theta_{old}) = \ell(\theta_{old}). \tag{40}$$

2.2.2 Variational EM

Recall that for loopy graphs, computing the mean parameter (34) is often intractable, which renders exact EM impossible. One solution is to use an approximate variational inference algorithm that improves, but not necessarily maximizes, the quantity in (34). One such algorithm is the mean field algorithm, which attempts (up to local maximum) to solve

$$\text{argmax}_{\mu^i \in \mathcal{M}_{\mathbf{x}_i}(F)} \theta^\top \mu^i - A^*_{\mathbf{x}_i}(\mu^i). \tag{41}$$

Recall the set $\mathcal{M}_{\mathbf{x}_i}(F)$ is an inner approximation to $\mathcal{M}_{\mathbf{x}_i}$, using an appropriate tractable subgraph F. Such "mean field E step" guarantees that the whole procedure is still coordinate ascent on $\mathcal{L}$.

It should be noted that the sum-product algorithm does not enjoy the coordinate ascent property.