10-708: Probabilistic Graphical Models 10-708, Spring 2016

14 : Theory of Variational Inference: Inner and Outer Approximation

Lecturer: Eric P. Xing *Scribes: Qi Guo, Chieh Lo, Wei-Chiu Ma*

1 Introduction

So far we have learned two families of approximate inference algorithms

1. Loopy belief propagation (sum-product)
2. Mean-field approximation

We will re-exam them together ,and form a unfied point of view based on the variational principle:

1. Loopy belief propagation: outer approximation
2. Mean-field approximation: inner approximation

2 Exponential Families

Before digging deeper into variational inference, we first slightly review the exponential families and their parameterization. Recall that exponential families refer to arbitrary set of probability distributions that can be represented in the following form:

$$p(x_1, \ldots, x_m; \theta) = \exp\{\theta^T \phi(x) - A(\theta)\}, \tag{1}$$

where $A(\theta)$ is the log partition function, θ is the canonical parameters, and $\phi(x)$ is the sufficient statistics. This form is well-known as the canonical parameterization. Note that θ and $\phi(x)$ can be either constants or vectors and the log partition function $A(\theta)$ is a always a convex function.

In addition, the joint probability distribution over Markov Random Fields (MRFs) can be expressed as the normalized product of clique potentials, i.e.

$$p(x; \theta) = \frac{1}{Z(\theta)} \prod_C \psi(x_C; \theta_C). \tag{2}$$

By re-formulating Equation 2 into the canonical representation defined in Equation 1, we will get:

$$p(x; \theta) = \exp\{\sum_C \log \psi(x_C; \theta_C) - \log Z(\theta)\} \tag{3}$$

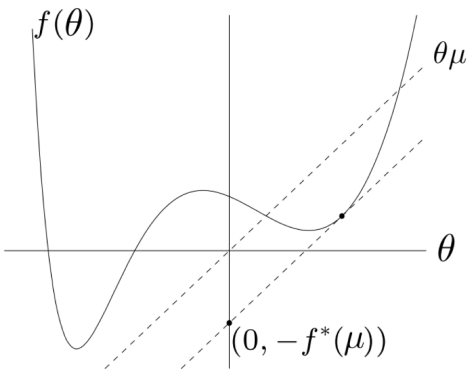

Figure 1: Conjugate Dual

From Equation 3, we can observe that computing the expectation of the sufficient statistics yields the following:

$$\mu_{s;j} = E_p[I_j(X_s)] = P[X_s = j], \forall j \in X_s \tag{4}$$
$$\mu_{st;jk} = E_p[I_{st;jk}(X_s, X_j)] = P[X_s = j, X_t = k], \forall (j,k) \in (X_s, X_t). \tag{5}$$

We can think of the first expectation here as the marginal mean value for a node, with the second expectation representing the marginal mean value for a pair of nodes.

3 Conjugate Dual

Given any function f(), its conjugate dual function is defined as:

$$f^*(\mu) = \sup\{\langle \theta, \mu \rangle - f(\theta)\} \tag{6}$$

A convenient property of the dual function is that it is **always convex**. Additionally, when the original function f is both convex and lower semi-continuous, the dual of the dual is f. We now step through a simple example of computing the mean parameters for a Bernoulli distribution.

3.1 Example: Bernoulli

The Bernoulli distribution takes the following form:

$$p(x;\theta) = \exp \theta x - A(\theta), \tag{7}$$

where $A(\theta) = \log[1 + \exp\{\theta\}]$.

The conjugate dual function is defined as:

$$A^*(\mu) = \sup_{\theta \in R}\{\mu\theta - \log[1 + \exp(\theta)]\} \tag{8}$$

Taking the partial with respect to θ, we obtain the following stationary point:

$$0 = \mu - \frac{e^\theta}{1+e^\theta} \tag{9}$$

$$\mu = \frac{e^\theta}{1+e^\theta} \tag{10}$$

From Equation 9, we can then solve θ:

$$\mu = \frac{e^\theta}{1+e^\theta} \tag{11}$$

$$= \frac{1}{1+e^\theta} \tag{12}$$

$$e^{-\theta} = \frac{1}{\mu} - 1 \tag{13}$$

$$\theta = \log[\frac{\mu}{1-\mu}] \tag{14}$$

From above, we can observe that if $\mu < 0$, $\theta = +\infty$, then $A^*(\mu) \to +\infty$, and vice versa. We can thus obtain the following formulation:

$$A^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu), \qquad \text{if } \mu \in [0,1] \tag{15}$$

$$= +\infty, \qquad \text{otherwise} \tag{16}$$

$A(\theta)$ can thus be defined as:

$$A(\theta) = \max_{\mu \in [0,1]} \{\mu^T \theta \ - A^*(\mu)\} \tag{17}$$

To maximize Equation 17, we take the partial derivative with respect to μ and set it to 0. The derivation is as follows:

$$\theta - \log \mu - 1 + \log(1-\mu) + 1 = 0 \tag{18}$$

$$\log \mu - \log(1-\mu) = \theta \tag{19}$$

$$\frac{1}{\mu} - 1 = e^{-\theta} \tag{20}$$

$$\mu = \frac{e^\theta}{1+e^\theta} \tag{21}$$

From above we can observe that this is the mean. In general, this will be true - the value of μ that maximizes the expression in our formulation of $A(\theta)$ will be the mean parameter. Additionally, just as our mean parameter was restricted to the range [0,1] above, our mean parameter in general will be restricted to some range of values. Note also that the dual function $A^*(\theta)$ is equal to the negative entropy of a Bernoulli distribution; the fact that the dual is equal to the negative entropy holds true in general and will be useful in the future.

We've shown that the mean computation of a Bernoulli distribution can be cast as an optimized problem on a restricted set of values. Does this methodology work in general? Unfortunately, computing the conjugate dual function over arbitrary graphs is intractable and the constraint set of possible mean values can be hard to determine. Thus, we turn to approximation methods.

3.2 Conjugate Dual for Exponential Family

Given

$$p(x_1, \ldots, x_m; \theta) = \exp\{\sum_{i=1}^{d} \theta_i \phi_i(x) - A(\theta)\}.$$

By definition, the dual function of $A(\theta)$ is

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

Stationary condition (one of the KKT conditions) is

$$\mu - \nabla A(\theta) = 0.$$

The derivatives of A yields *mean parameters*

$$\frac{\partial A}{\partial \theta_i}(\theta) = E_\theta[\phi_i(X)] = \int \phi_i(x) p(x; \theta) dx.$$

So the stationary condition becomes $\mu = E_\theta[\phi(X)]$.

If we have the dual solutions μ, can we get the primal solution $\theta(\mu)$ and how? Lets assume there is a solution $\theta(\mu)$ such that $\mu = E_{\theta(\mu)}[\phi(X)]$, then the dual of exponential family is

$$\begin{aligned} A^*(\mu) &= \langle \theta(\mu), \mu \rangle - A(\theta(\mu)) && (22) \\ &= E_{\theta(\mu)}[\langle \theta(\mu), \phi(X) \rangle - A(\theta(\mu))] && (23) \\ &= E_{\theta(\mu)}[\log p(X; \theta(\mu))] && (24) \end{aligned}$$

The entropy is defined as

$$H(p(x)) = -\int p(x) \log p(x) dx$$

, so we have that the dual

$$A^*(\mu) = -H(p(x; \theta(\mu)))$$

. It has this nice form not out of a coincident, but because of the property of exponential family.

The domain of $A^*(\mu)$ is a marginal polytope. We will explain it now.

First define a vector of mean parameters, which has been mentioned before. Given any distribution $p(x)$ and a set of sufficient statistics $\phi(x)$, **mean paramters** are

$$\mu_i = E_p[\phi_i(X)] = \int \phi_i(x) p(x) dx$$

, where $p(x)$ is not necessarily an exponential family.

For an exponential family, the set of all realizable mean parameters of

$$\mathcal{M} := \{\mu \in R^d \mid \exists p \text{ s.t. } E_p[\phi(X)] = \mu\} \quad (25)$$

The above is a convex set. And for discrete exponential families, this is called marginal polytope.

$$\mathcal{M} = \text{conv}\{\phi(x), x \in \mathcal{X}^m\}$$

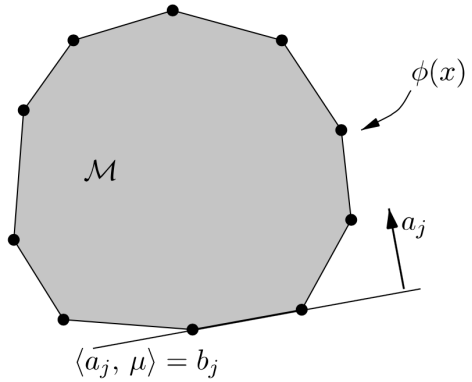

Figure 2: Marginal Polytope (Half-plane Representation)

According to *Minkowski-Weyl Theorem*, any non-empty convex polytope can be characterized by a finite collection of linear inequality constraints.

So we have a half-plane representation:

$$\mathcal{M} = \{\mu \in R^d \mid a_j^T \mu \geq b_j, \forall j \in \mathcal{J}\}$$

, where $|\mathcal{J}|$ is finite (Figure 2).

4 Variational Method

The exact variational formulation of the log partition function is

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

, where

$$\mathcal{M} := \{\mu \in R^d \mid \exists p \text{ s.t. } E_p[\phi(X)] = \mu\}$$
$$A^*(\mu) = -H(p_\theta(\mu)) \text{ if } \mu \in \mathcal{M}^o(\text{ else } +\infty)$$

We have taken a long way to get here, combining all former efforts involving convex optimization and exponential family. This is **THE** optimization problem we aim to solve. Two difficulties to optimize it are

- $\mathcal{M}$: the marginal polytope, difficult to characterize
- A^*: the negative entropy function, no explicit form, involving an integral

5 Mean Field Method

Mean field tackles the hard optimization problem by *non-convex inner bound* and *exact form of entropy.*

For a general graph G, the marginal polytope $\mathcal{M}(G; \phi)$ (Equation 25) is hard to characterize. We find a subgraph F that is tractable to approximate it. For example, a tree or or a graph with no edge at all. This is the essence of *mean-field approximation.*

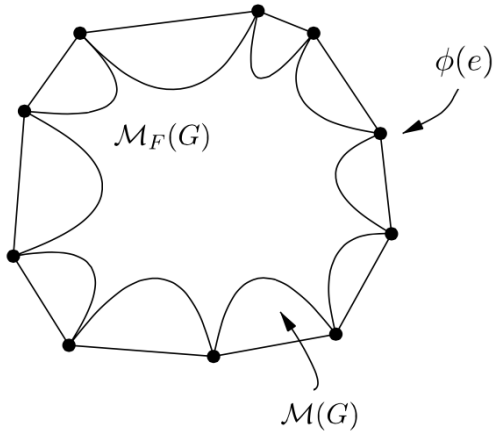

Figure 3: Mean field optimization is always non-convex for any exponential family in which the state space $\mathcal{X}$ is finite. $\mathcal{M}_F(G)$ contains all the extreme points.

For a tractable subgraph F,

$$\mathcal{M}(F;\phi)^o := \{\tau \in R^d \mid \exists \theta \in \Omega(F) \text{ s.t. } E_\theta[\phi(X)] = \tau\} \tag{26}$$

This is an inner approximation $\mathcal{M}(F;\theta)^o \subseteq \mathcal{M}(G;\theta)^o$ (Figure 3). And mean field solve the relaxed problem

$$\max_{\tau \in \mathcal{M}_F(G)} \{\langle \tau, \theta \rangle - A_F^*(\tau)\}$$

. Where A_F^* is the exact dual function w.r.t. $\mathcal{M}_F(G)$.

6 Bethe Approximation and Sum-Product

6.1 Recap: Sum-Product/Belief Propagation Algorithm

The update for each node in sum-product algorithm is given by:

$$M_{ts}(x_s) \to k \sum_{x_t'} \{\phi_{st}(x_s, x_t')\phi_t(x_t') \prod_{u \in N(t)\setminus s} M_{ut}(x_t')\}.$$

where ϕ is the potential function. And the marginal for node s is given by:

$$\mu_s(x_s) = k\phi_s(x_s) \prod_{t \in N(s)} M_{ts}^*(x_s).$$

6.2 Variational Inference for Sum-Product/Belief Propagation Algorithm

The sum-product algorithm can do exact inference on trees, but can only approximate for loopy graphs. Hence, by using variation inference methods, we can formulate an optimization problem to estimate mean parameters in trees. We explicitly describe the algorithm in the following paragraphs.

6.3 Exact Variational Principle for Trees

Let's begin by computing the mean parameters for tree graphical models. For a discrete tree with variables $X_s \in \{0, 1, \ldots, m_s - 1\}$, the sufficeint statistics of a tree is given by:

$$I_j(x_s) \text{ and } I_{jk}(x_s, x_t) \text{ for } s = 1, \ldots n \text{ and } (s,t) \in E$$

The mean parameters are

$$\mu_s(x_s) = P(X_s = x_s) \text{ and } \mu_{st}(x_s, x_t) = P(X_s = x_s, X_t = x_t).$$

We can construct the marginal polytope for the tree and by the junction tree theorem

$$\mathcal{M}(T) = \{\mu \geq | \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)\}$$

If $\mu \in \mathcal{M}(T)$, then

$$p_\mu(x) := \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, X_t)}{\mu_s(x_s)\mu_t(x_t)}.$$

For trees, the entropy can be decomposes as:

$$H(p_\mu(x)) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}\mu_{st}.$$

Hence, the variational function is given by:

$$A(\theta) = \max_{\mu \in \mathcal{M}(T)} \{< \theta, \mu > + \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})\}.$$

Next, by utilizing the Lagrangian method, we can get the general variational principal for belief propagation algorithm for tree:

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

$$\mu_s(x_s, x_t) \propto \exp\left(\theta_s(x_s) + \theta_t(x_t) + \theta_{st}(x_s, x_t)\right) \prod_{u \in N(s) \setminus t} \exp\left(\lambda_{us}(x_s)\right) \prod_{v \in N(t) \setminus s} \exp\left(\lambda_{vt}(x_t)\right)$$

where λ_{ss} and $\lambda_{ts}(x_s)$ are the Lagrange multiplier. This yields:

$$M_{ts}(x_s) \to \sum_{x_t} \exp \theta_t(x_t) + \theta_{st}(x_s, x_t) \prod_{u \in N(t) \setminus s} M_{ut}(x_t).$$

6.4 Belief Propagation on Arbitrary Graphs

There are two main difficulties for belief propagation on arbitrary graphs: (1) determination of the marginal polytype $\mathcal{M}$ and (2) computation of the exact entropy. Suppose that, for an arbitrary connected graph G, we use the tree-based approximation for $\mathcal{M}$. Since G contains additional constraints (via additional edges) compared to any tree formed with its edges, this tree approximation is an outer bound. We formulate as follows: The outer bound:

$$\mathbb{L}(G) = \{\tau | \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\}$$

where τ is the pseudo-marginals. Hence, we can approximate the exact entropy as:

$$-A^*(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}).$$

The Bethe Variational Problem is given by:

$$\max_{\tau \in \mathbb{L}(\mathbb{G})} \{< \theta, \tau > + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st})\}.$$

This problem is differentiable, with each constraint set being a simple convex polytype with message passing on trees as an analytical solution. However, there is no guarantees on the convergence of the algorithm on loopy graphs. Also, the Bethe variational problem is usually non-convex. Therefore, there are no guarantees on the global optimum.

7 Summary

In this scribes, we discussed the following things: (1) recap of the exponential family (2) mean-field approximation (3) belief propagation approximation.

Variational methods can be thought of as turning inference problem into an optimization problem by using exponential families and convex duality. In the mean field approximation, we use an inner approximation of the realizable space. This allows us to use the exact formulation for entropy, however may cause the true maximum fall out of our approximated space. In the Bethe approximation for belief propagation, we use an outer approximation to the marginal polytype $\mathcal{M}$. By solving the Lagrangian in the Bethe Variational Problem, we are equivalent solve the message-passing/Sum-Product algorithm.