10-708: Probabilistic Graphical Models, Spring 2015

13 : Theory of Variational Inference

Lecturer: Eric P. Xing *Scribes: Joey Robinson*

1 Introduction

In the previous lecture, we introduced the notion of variational inference and variational Bayes. In variational Bayes, we wish to approximate the joint distribution p of the latent variables with a family of distributions q that is easier to work with. We can then use standard optimization methods in order to find the parameter values for q that minimize the KL-divergence between p and q.

We then went on to introduce mean field variational inference, in which the variational distribution over latent variables is assumed to factor as a product of functions over each latent variable. In this sense, the joint approximation is made using the simplifying assumption that all latent variables are independent. Alternatively, we can use approximations in which only groups of variables are assumed to be independent provided that the necessary computational resources are available.

In this lecture, we examine the theory that underlies variational inference and mean field methods by taking a convex optimization approach. We also demonstrate connections between previous algorithms (belief propagation, Sum-Product) and variational inference.

2 Background

2.1 Exponential Families

In our study of variational inference, we first briefly return to the study of exponential families and their parameterization. Recall that an exponential family is any set of probability distributions that can be represented in the following form:

$$
\begin{aligned}
&p_\theta(x_1, \cdots, x_m) = \exp\{\theta^T \phi(x) - A(\theta)\}, \text{where} \\
&A(\theta) \text{ is the log normalization constant,} \\
&\theta \text{ is the vector of canonical parameters, and} \\
&\phi(x) \text{ is the vector of sufficient statistics}
\end{aligned}
$$

The above form is commonly referred to as the **canonical parameterization**. Note that the log normalization constant $A(\theta) = \log \int \exp\{\theta^T \phi(x)\} dx$ is a convex function.

Recall that the joint probability distribution over a Markov Random Field (MRF) can be written as the normalized product of clique potentials, i.e.

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

We can rewrite this using the canonical representation as described above:

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

From this canonical form, we see that computing the expectation of the sufficient statistics $\phi(x)$, given the canonical parameters θ, yields the following:

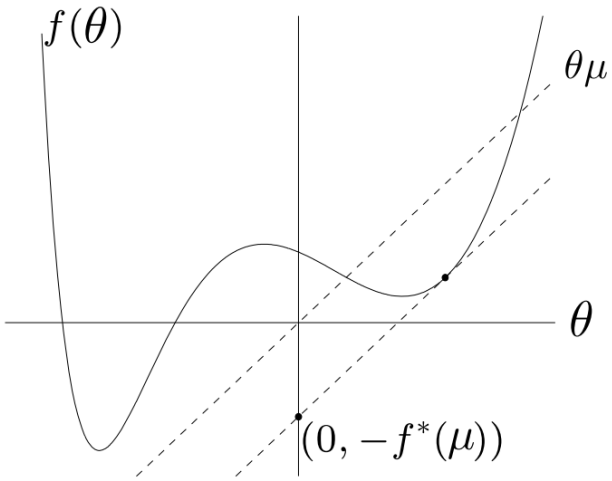
$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \ \forall j \in X_s$$
$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \ \forall (j,k) \in X_s \in X_t$$

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.

2.2 Conjugate Dual Function

Given any function $f(\theta)$, its **conjugate dual function** is defined as:

$$f^*(\mu) = \sup_\theta \{\langle \theta, \mu \rangle - f(\theta)\} \tag{1}$$



A convenient property of the dual function is that it is always convex. Additionally, when the original function f is both convex and lower semicontinuous, the dual of the dual is f.

We now step through a simple example of computing the mean parameters for a Bernoulli distribution, which takes on the following form:

$$p(x;\theta) = \exp\{\theta x - A(\theta)\}, \text{ where } A(\theta) = \log[1 + \exp(\theta)]$$

The conjugate dual function is then defined as:

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

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

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

Additionally, we can solve for θ:

$$\begin{aligned}\mu &= \frac{e^\theta}{1+e^\theta} \\ \mu &= \frac{1}{1+e^{-\theta}} \\ e^{-\theta} &= \frac{1}{\mu} - 1 \\ \theta &= \log[\frac{\mu}{1-\mu}]\end{aligned}$$

From this, we can see that if $\mu < 0$, $\theta = +\infty$ and so $A^*(\mu) = +\infty$. The same can easily be seen when $\mu > 1$ (logarithm of a negative number). Thus, we obtain the following formulation:

$$\begin{aligned}A^*(\mu) =& \mu \log \mu + (1-\mu)\log(1-\mu), \text{ if } \mu \in [0,1] \\ &+\infty \text{ otherwise}\end{aligned}$$

Using (**??**), $A(\theta)$ is then defined as:

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

To maximize this expression, we take the partial derivative with respect to μ and set the result equal to 0.

$$\begin{aligned}0 &= \theta - \log \mu - 1 + \log(1-\mu) + 1 \\ \log \mu - \log(1-\mu) &= \theta \\ \mu^{-1} - 1 &= e^{-\theta} \\ \mu &= \frac{e^\theta}{1+e^\theta}\end{aligned}$$

Here we see that this is, in fact, 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.

For **any** distribution $p(x)$ and a set of sufficient statistics $\phi(x)$, we define a vector of mean parameters:

$$\mu_i = \mathbb{E}_p[\phi_i(X)] = \int \phi_i(x) p(x) dx$$

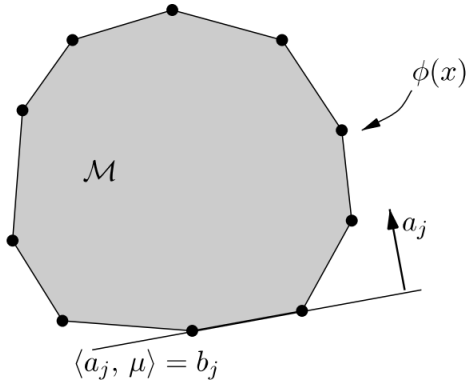
The set of all realizable mean parameters can then be described as:

$$\mathcal{M} := \{\mu \in \mathbb{R}^d \mid \exists p \; s.t. \; \mathbb{E}_p[\phi(X)] = \mu\}$$

Note that this is a convex set. For discrete exponential families such as the Bernoulli distribution in the previous example, this is called the **marginal polytype**. We now introduce a useful theorem from optimization theory:

Minkowski-Weyl Theorem: any non-empty convex polytype can be characterized by a finite collection of linear inequality constraints.

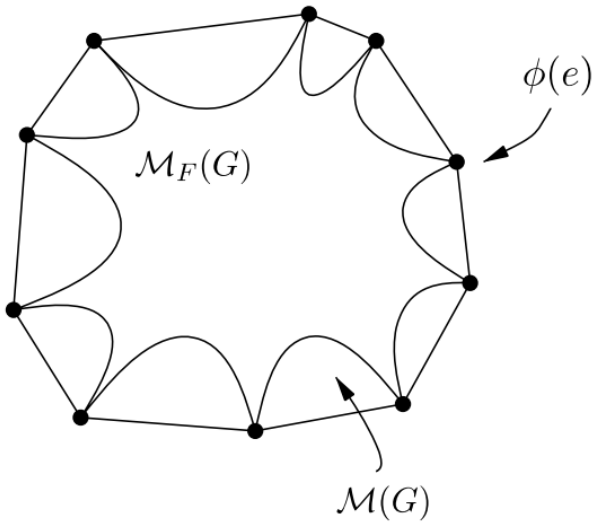
In essence, the theorem states that any non-empty convex polytype can be represented by its faces, each represented as a bounded hyperplane; this idea is summarized in the figure below. For tree graphical models, it can be shown that the number of faces is linear in the number of nodes in the graph. However, it becomes extremely difficult to characterize the shape of the marginal polytype for a general graph.



3 Variational Methods

3.1 Mean Field Methods

For a general graphical model, we require faster ways of estimating the mean parameter. Using the **variational principle**, we have shown that the conjugate dual function equals the negative entropy for all mean parameters within $\mathcal{M}$, and is $+\infty$ otherwise. Since $\mathcal{M}$ can be difficult to characterize, we instead approximate the marginal polytype with that of a simpler **tractable subgraph**. As an example, we might choose a tree approximation for a general graph. In the Ising model, we might break the full grid into disconnected "chunks". This is the essence of **mean field methods** – approximating the marginal polytype $\mathcal{M}$ with an inner approximation $\mathcal{M}_F(G)$ in order to make the problem tractable (see below).



3.2 Geometry of Mean Field

It can be shown that the approximated marginal polytype $\mathcal{M}$ used in mean field optimization is always non-convex for any exponential family in which $\mathcal{X}^m$ is finite, since $\mathcal{M}$ will be an inner approximation that touches all extremities at the intersection of faces in the exact marginal polytype.

3.3 Tree Graphical Models

Recall the Sum-Product algorithm used previously for inference over tree graphical models. It was stated that this algorithm is exact for trees, but only approximate for graphs containing cycles (often called loopy graphs). Here, we first show that belief propagation/Sum-Product can be restated as an exact optimization problem on trees. Then, in the next section, we show how variational inference can allow us to expand on these ideas to do approximate inference over arbitrary graphs using tractable subgraphs for arbitrary precision approximations (assuming sufficient computational resources exist).

Suppose we have a tree graphical model $T = (V, E)$ with vertices V and edges E, where our corresponding random variables X_s are discrete. The sufficient statistics are then the indicator variables for each possible value of every node, and similarly for each edge.

Before we begin, we must first determine the negative entropy of a distribution over our tree. We have:

$$\begin{aligned} H(p(x;\mu)) &= -\sum_x p(x;\mu)\log p(x;\mu) \\ &= \sum_{s\in V}[-\sum_{x_s}\mu_s(x_s)\log\mu_s(x_s)] + \sum_{(s,t)\in E}[\sum_{x_s,x_t}\mu_{st}(x_s,x_t)\log\frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_x(x_t)}] \\ &= \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}) \end{aligned}$$

In order to compute the mean parameters, we use the variational formulation from earlier:

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

Adding in our joint and marginal normalization constraints, we can solve the above maximization problem:

$\mathcal{L}(\mu,\lambda) = \langle\theta,\mu\rangle + \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}) + \sum_{s\in V}\lambda_{ss}C_s s(\mu) + \sum_{(s,t)\in E}[\sum_{x_t}\lambda_{st}C_{st}(x_t;\mu) + \sum_{x_s}\lambda_{ts}C_{ts}(x_s;\mu)]$, where

$C_s s(\mu) = 1 - \sum_{x_s}\mu_s(x_s)$, and

$C_{ts}(x_s;\mu) = \mu_s(x_s) - \sum_{x_t}\mu_{st}(x_s,x_t)$

By taking the partial derivatives of this function with respect to each mean parameter μ_s and μ_{st} and setting the result equal to 0, it can be shown that optimization problem decomposes exactly into the message passing algorithm for belief propagation. Thus, message passing is a Lagrangian method for computing the stationary condition of the above variational formulation.

4 Belief Propagation on Arbitrary Graphs

Suppose, instead, that we wish to perform inference on arbitrary graphs. In the previous section, we showed how variational inference can be used to estimate mean parameters in trees. Additionally, we showed that this Lagrangian optimization procedure is equivalent to that of the message passing algorithm. However, mean field methods provide a much more general framework than message passing in that we can perform optimization over arbitrary tractable subgraphs – not just trees.

4.1 Bethe Variational Problem

The two main difficulties that arise when using the variational formulation are the determination of the marginal polytype $\mathcal{M}$ and the computation of the exact entropy. Suppose that, for an arbitrary connected graph $\mathcal{G}$, we use the tree-based approximation for $\mathcal{M}$. Since $\mathcal{G}$ contains additional constraints (via additional edges) compared to any tree formed with its edges, this tree approximation is an outer bound. We can formalize this set in the following way:

$$\mathbb{L}(G) = \{\tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\}$$

In the equation above, we call the locally consistent vectors τ **pseudo-marginals**. Suppose that we also use the tree approximation for entropy, such that we have:

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

$H_{\text{Bethe}}(\tau)$ is referred to as the **Bethe approximation**. We can combine these two approximations – that of the marginal polytype and the entropy – to formulate what is known as the **Bethe Variational Problem** (BVP), which we state below:

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

This problem is differentiable, with each constraint set being a simple convex polytype. It can be shown that loopy belief propagation is equivalent to an iteration method for solving the BVP, with message passing on trees as an analytical solution. Another result is that the mean parameter estimates obtained by loopy belief propagation are a fixed point iff they are local stationary points of the Bethe Variational Problem.

One might ask whether the solution of the BVP ever results in a non-realizable mean parameters. That is, do we even need to worry about our tree-based marginal polytope approximation. Unfortunately, yes. In 2003, Wainwright et. a. proved that for any element of our tree-based outer bound $\mathbb{L}(G)$, it is possible to construct a distribution p such that the solution to the BVP falls within this gap.

5 Summary

In the first section, we provided a brief recap of exponential families, and showed how the joint distribution specified by a Markov Random Field (MRF) can be recast as an exponential family. We then defined the conjugate dual function, and described the conditions (convexity and lower semi-continuity) under which the dual of the dual is the original function itself. We then described how, in the general case, the mean parameter will be restricted.

Since the range of possible mean parameters can be described by a convex set, and our joint distribution to maximize is of an exponential family, we recast the problem as a Lagrangian optimization problem. In the Mean Field Approximation, we utilize an inner approximation of the realizable space. This allows us to use the exact formulation for entropy, however may result in situations where our true maximum would fall out of our approximated space.

Alternatively, we can use an outer approximation to the marginal polytype $\mathcal{M}$. However, $\mathcal{M}$ and our entropy H are often intractable to compute exactly, so we use a tree-based approximation for both. In the case of negative entropy, this approximation is called the Bethe Approximation (H_{Bethe}). The optimization problem that results from both approximations is called the Bethe Variational Problem (BVP). Interestingly, solving the Lagrangian in the Bethe Variational Problem is equivalent to the message-passing/Sum-Product algorithm discussed previously. This not only explains why Sum-Product is [only] exact on trees, but also provides further insights into the conditions under which a stationary and reasonably accurate solution can be found for general (non-tree) graphs.

The significance of the results here lie both in the improve understanding of existing methods (e.g. loopy belief propagation, message passing, Sum-Product), and in the flexibility of this new framework and ability to work with existing convex optimization tools.