10-708: Probabilistic Graphical Models 10-708, Spring 2014

14 : Theory of Variational Inference: Inner and Outer Approximation

Lecturer: Eric P. Xing *Scribes: Yu-Hsin Kuo, Amos Ng*

1 Introduction

Last lecture we learned about two families of approximate inference algorithms: loopy belief propagation and mean-field approximation. In this lecture, we are going to re-examine these two approaches and provide a unified view based on the variational principle, which is essentially a relaxation or redefinition of the objective and optimization space.

We can simply define the variational method as just an optimization-based formulation. Imagine you are interested in certain quantities, such as the likelihood of the data or the distribution of the marginal, and imagine these are difficult to compute. What the variational method does is to approximate the solution by relaxing the intractable optimization problem.

To demonstrate the idea of variational approaches, we present two examples below in Sections 1.1 and 1.2.

1.1 Largest Eigenvalue

Computing the largest eigenvalue of a matrix A can be viewed as solving the quadratic programming problem which maximizes

$$\lambda_{max}(A) = \max_{||x||_2=1} x^T A x$$

1.2 Solving a Linear System of Equations

It is obvious that the answer to this linear system of equations,

$$Ax = b,\ A \succ 0,$$

is $x^* = A^{-1}b$. However, to avoid computing the inverse of the matrix, the problem can also be formulated as

$$x^* = \underset{x}{argmin} \left\{ \frac{1}{2} x^T A x - b^T x \right\}$$

This is because for this convex optimization problem, the solution is at the stationary point which is defined by the original equation. To see why this is true, we can take the derivative of it and set it to zero, which is essentially $Ax = b$.

1.3 Next steps

Next, we focus on undirected GM but not directed GM because we can always transform the directed GM into an equivalent undirected GM. In undirected GM, we are interested in two quantities: the marginal

distributions and the partition function. However, calculating these quantities is exponentially difficult and we wish to convert them to an optimization problem. To come up with a generic formulation of this, we need to use the tools from exponential families and convex analysis.

2 Exponential Families

A probability density in the exponential family takes the following general form:

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

where θ represents the canonical/natural parameters, the function ϕ is a sufficient statistic, and A is a log partition function. Based on this definition, we can define the domain of the θ to be

$$\Omega := \left\{\theta \in R^d | A(\theta) < +\infty\right\}.$$

That is, θ should make sure that A is a finite quantity, because otherwise the whole probability will become zero.

The first example is Gaussian MRF and it has the form:

$$p(\mathbf{x}) = exp\left\{\frac{1}{2}\langle \mathbf{\Theta}, \mathbf{x}\mathbf{x}^T\rangle - A(\theta)\right\},$$

where $\theta = -\Sigma^{-1}$ and for sufficient statistics, we have a single term and a pairwise term. Another example is discrete MRF, where we have the indicator function of a particular node, single term potential and pairwise potential. The reason we use exponential families is that there is a simple mapping between mean parameterization and the natural inference problem.

3 Conjugate Duality

Finding the dual of functions was solved previously in closed form, but now we want to solve it in a variational way. Usually, if things go well, we find the stationary condition that maximizes the dual. We take the derivative with respect to μ and set the result to zero, obtaining the stationary condition with a relationship between μ and canonical θ. We have closed form solutions everything and things become very easy. So why do we want to solve the same thing with variational inference?

Because sometimes, life is not so good, and you are unable to get the solutions in closed form. Then you end up with symbolic distributions where nothing is in closed form and things get out of control. That is when you compute μ numerically and not from a closed form solution.

So to actually find the conjugate dual of any arbitrary function $f(\theta)$, we have

$$f^*(\mu) = \sup_\theta \{\langle \theta, \mu\rangle - f(\theta)\} \tag{2}$$

The linear term $\langle \theta, \mu\rangle$ has the free parameter μ, augmented on the dual function. So therefore this equation means that at every point of θ we are taking the maximum distance between the line $\langle \theta, \mu\rangle$ and the function $f(\theta)$. Readers can see this in Figure 1.

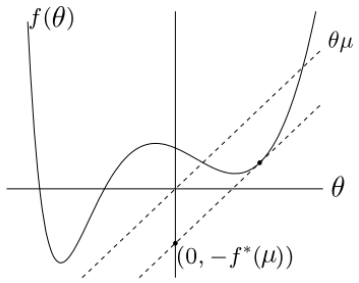

Figure 1: A plot of $\langle \theta, \mu \rangle$ with some function $f(\theta)$

3.1 Properties

The dual found in this way has some nice properties. It is always convex, because it is really just a pointwise bundle of supremums over $\langle \theta, \mu \rangle$ and $f(\theta)$. That, by definition, is convex no matter what $f(\theta)$ may be.

If the original function $f(\theta)$ were convex, then the dual of the dual (which is also convex, as all duals must be) is simply the original function. You can have an inference task which expresses the partition function as the supremum of $\langle \theta, \mu \rangle$ minus another function, as well.

Generally for members of the exponential family, the dual is the negative of the entropy, as we will show in Section 3.3. μ is restricted, and solving the optimization problem gives you both the mean parameter and the log partition function.

Even more generally, actually computing the conjugate dual is an intractable problem, and setting the constraints in the real world can be non-trivial. As such, approximations are often needed.

3.2 Bernoulli Mean Parameter

For example, to obtain the mean parameter of a Bernoulli distribution, we have

$$A^*(\mu) := \sup_{\theta \in R} \left\{ \mu\theta - \log \left[1 + e^{\theta} \right] \right\}$$

The stationary condition is found to be

$$\mu = \frac{e^{\theta}}{1 + e^{\theta}} \tag{3}$$

Now we can plug (3) back into $A^*(\mu)$ to get $A(\theta)$ expressed in terms of θ and $A^*(\mu)$, and optimize that instead. From this, we find the variational form to be

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

From which we get that the optimum (which is also the mean) is

$$\mu(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}$$

3.3 Entropy

Given an exponential family defined according to Equation 1

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

Using Equation 2, we can give the dual as

$$A^*(\mu) := \sup_{\theta \in \Omega} \{ \langle \mu, \theta \rangle - A(\theta) \} \tag{4}$$

The stationary condition would be

$$\mu - \nabla A(\theta) = 0 \tag{5}$$

We take the derivative of $A(\theta)$ with respect to θ_i to obtain

$$E_\theta \left[\phi_i(X) \right] = \int \phi_i(x) p(x; \theta) dx$$

Plugging back into Equation 5 we get

$$\mu = E_\theta \left[\phi(X) \right] \tag{6}$$

If there does exist a solution $\theta(\mu)$ such that Equation 6 is satisfied for this value of θ, then you can pretty much plug it back in to Equation 4 and get the μ replaced by ϕ:

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

This derivation shows that if everything goes well and μ is defined, then your dual function is now the entropy of the original distribution, since entropy is defined as:

$$\begin{aligned} H(p(x)) &= -\int p(x) \log p(x) dx \\ &= -A^*(\mu) \end{aligned}$$

4 Polytope

Obtaining the inverse mapping given by Equation 6, for the canonical parameters $\theta(\mu)$, is a non-trivial task. So what exactly is the feasible space over which $\mu \in R^d$ has a solution $\theta(\mu)$?

For discrete exponential families, this would be the *marginal polytope*, defined as

$$M = \left\{ \mu \in R^d \mid \exists p \text{ s.t. } E_p \left[\phi(X) \right] = \mu \right\} \tag{7}$$

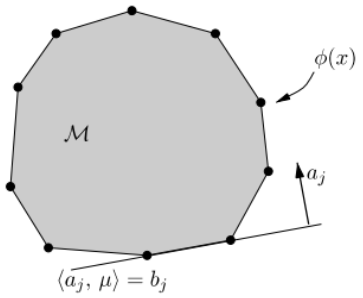

Figure 2: A demonstration of how a collection of linear inequality constraints can characterize a convex polytope.

4.1 Representations of Convex Polytopes

Two representations of a convex polytope are the convex hull representation,

$$M = \left\{ \mu \in R^d | \sum_{x \in X^m} \phi(x) p(x) = \mu, \text{for some } p(x) \geq 0, \sum_{x \in X^m} p(x) = 1 \right\} \triangleq \text{conv}\left\{ \phi(x), x \in X^m \right\}, \quad (8)$$

and the half-plane representation using the Minkowski-Weyl Theorem, which says that for any convex polytope that is non-empty, we can characterize it using a finite collection of linear equality constraints. This is defined formally by Equation 9, and demonstrated graphically by Figure 2.

$$M = \left\{ \mu \in R^d | a_j^T \mu \geq b_j, \forall j \in J \right\} \text{ where } |J| \text{ is finite} \quad (9)$$

Note that for tree graphical models, the number of half planes only grows linearly with the graph size. For graphical models in general, however, it is very hard to generalize the marginal polytope.

4.2 Variational Principle

The dual function here takes the form of

$$A^*(\mu) = \begin{cases} -H(p_\theta(\mu)) & \text{if } \mu \in M^\circ \\ +\infty & \text{if } \mu \notin \bar{M} \end{cases}$$

where $\theta(\mu)$ once again satisfies Equation 6.

From Equation 2 we see that the log partition function is in the form

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

where M represents the marginal polytope and $A^*(\mu)$ represents the negative entropy function (with no explicit form). The optimization problem has a unique solution for all $\theta \in \Omega$, when $\mu(\theta) \in M^\circ$ for $\mu(\theta)$ that satisfies Equation 6.

The mean field method produces a non-convex inner bound and an exact form of entropy. Bethe approximation (discussed in Section 5) and loopy belief propagation on the other hand produce a polyhedral outer bound and a non-convex Bethe approximation.

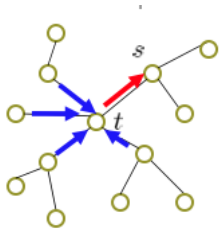

Figure 3: An illustration of message passing

5 Bethe Approximation and Sum-Product

To recap, we have learned that the message passing rule for Fig 3 is:

$$M_{ts} \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t)\psi_t(x'_t) \prod_{\mu \in N(t)/s} M_{ut}(x'_t) \right\},$$

where $\kappa > 0$ denotes a normalization constant. This algorithm is an exact inference for trees, but approximate for loopy graphs.

5.1 Variational Inference and Sum-Product Algorithm

In a **tree graphical model** of discrete variables $X_S \in \{0, 1, ..., m_s - 1\}$, the sufficient statistics are the indicator function given by:

$$\begin{cases} \mathbb{I}_j(x_s) & \text{for } s = 1, ..., n, \quad j \in \chi_s \\ \mathbb{I}_j(x_s, x_t) & \text{for } (s,t) \in E, \quad (j,k) \in \chi_s \times \chi_t \end{cases}$$

Using these sufficient statistics, we define an exponential family of the form

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

where we have introduced the convenient shorthand notation

$$\theta_s(x_s) := \sum_j \theta_{s;j} \mathbb{I}_{s;j}(x_s), \text{ and}$$

$$\theta_{st}(x_s, x_t) := \sum_{(j,k)} \theta_{st;jk} \mathbb{I}_{st;jk}(x_s, x_t).$$

The mean parameters correspond to singleton and pairwise marginal probabilities, which are given below:

$$\mu_s(x_s) = \sum_{j \in \chi_s} \mu_{s:j} \mathbb{I}_j(x_s) = \mathbb{P}(X_s = x_s)$$

$$\mu_{st}(x_s, x_t) = \sum_{(j,k) \in \chi_s \times \chi_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t) = \mathbb{P}(X_s = x_s, X_t = x_t),$$

Note that μ_s is a $|\chi_s|$-dimensional marginal distribution over X_s, whereas μ_{st} is a $|\chi_x| \times |\chi_t|$ matrix, representing a joint marginal over (X_s, X_t)

With the mean parameters defined above, and the junction tree theorem for a tree T we have:

$$\mathbb{M}(T) = \left\{ \mu \geq 0 | \sum_{x_s} \mu_s(x_s) = 1, \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \right\},$$

which maintains the normalization condition and marginalization constraints.

Now, we proceed to the loss function where the entropy is decomposed as

$$H(p(x:\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}) = -A^*(\mu),$$

which is a closed form definition. With these definitions, we now have the variational formulation,

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

To solve the optimization problem, we can apply two Lagrange multipliers λ_{ss} and λ_{ts} for the normalization constraint and for each marginalization constraint, respectively. Then we take the derivative and set them to zero to get the final answer. To see the detailed derivation, readers can refer to pages 36-37 in the slide.

5.2 Belief Propagation on arbitrary graphs

Next, we discuss the belief propagation on **arbitrary graphs** where we cannot compute the entropy and marginal polytope. The marginal polytope M is hard to characterize and hence it is approximated by an outer bound:

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

Also, exact entropy $-A^*(\mu)$ lacks an explicit form, hence we approximate it to the expression of a tree:

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

Combining the two approximations in the original optimization problem leads to the Bethe variational problem (BVP):

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

One can use loopy BP for solving a Lagrangian formulation of the BVP.

To understand why it is called outer bound for $\mathbb{L}(G)$, we can look at the geometry of BP. From Figure 4, we can see clearly that $\mathbb{M}(G) \subseteq \mathbb{L}(G)$. This is because the way we define $\mathbb{L}(G)$ is based on local consistency, which might lead to more distributions that don't exist in the original graph and the equality holds if and only if the graph is a tree. When searching for the solution to the BVP, one should be aware that it's possible for the solution to appear inside the gap.

The above formulation provides a principled basis for applying the sum-product algorithm to loopy graphs. However, there are no guarantees on the convergence of the algorithms on loopy graphs. Also, because the BVP is usually non-convex, there are no guarantees on the global optimum. This connection offers us multiple ways to improve the usual sum-product algorithm. We could produce progressively better approximations to our entropy function, as well as better approximations to the marginal polytope.

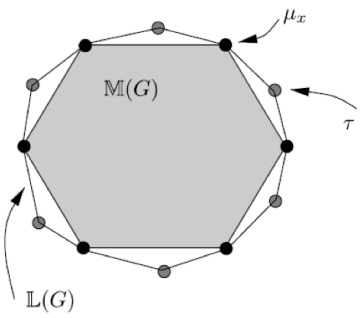

Figure 4: The geometry of the belief propagation

6 Mean Field Approximation

For exponential families that have sufficient statistics ϕ defined for a graph G, Equation 7 gives us the set of feasible mean parameters. A generic graph G may not necessarily be *tractable* (as in you can perform exact inference on it), though. Mean field methods help us by simplifying the production of node marginals to the production of mean parameters for the graph.

We restrict p so that it is only in a subset of distributions associated with a tractable subgraph F in G. The distributions associated with F would be a subset of the canonical ones:

$$M(F;\phi) = \left\{\tau \in R^d | \tau = E_\theta\left[\phi(X)\right] \text{ for some } \theta \in \Omega(F)\right\}$$

An inner approximation of G would have

$$M(F;\phi)^\circ \subseteq M(G;\phi)^\circ$$

The mean field approach would solved a relaxed version of the problem, producing pseudo-marginals of the form

$$\max_{\tau \in M_F(G)} \left\{\langle \tau, \theta\rangle - A_F^*(\tau)\right\}$$

We essentially have

$$A_F^* = A_F^*|_{M_F(G)}$$

as the exact dual function, but restricted to $M_F(G)$.