STA561: Probabilistic machine learning

Variational Inference (11/04/13)

Lecturer: Barbara Engelhardt *Scribes: Matt Dickenson, Alireza Samany, Tracy Schifeling*

1 Introduction

In this lecture we will further discuss variational inference and then proceed with *loopy belief propagation*, also known as LBP. In the last lecture we discussed *mean field variational inference* and we investigated these methods in the context of univariate Gaussians. In this lecture, we will talk further about variational inference in the context of the Ising model, which is a Markov random field (or an undirected graphical model), and then we will proceed with mean field and loopy belief propagation.

2 Variational Inference

Let us begin by reviewing variational inference. In variational inference, we have data set $\mathcal{D} = \{X_1, X_2, \ldots, X_n\}$ and we describe both our latent variables, $Z_{1:m}$, and the set of our model parameters, θ, by a single parameter Z such that $Z = \{Z_{1:m}, \theta\}$.

We are interested at estimating the posterior probability of latent variables given the data, which is $p(Z|X)$. By definition of the conditional distribution, we know that the posterior probability of latent variables given data is equal to the joint distribution of latent variables and data divided by the marginal probability of the data (or the observed variables). Therefore, we have:

$$p(Z|X) = \frac{p(Z,X)}{\int_Z p(Z,X)dZ}.$$

Variational inference is a method to compute an approximation to the posterior distribution. In general, computing the posterior distribution for several distributions, such as truncated Gaussians or Gaussians mixture models, can be very hard. This issue obliges us to resort to computing approximations for the posterior distribution. For example, consider the case of Gaussian Mixture Models (GMMs). Let us assume the standard mixture model where each class specific mean μ_k is distributed according to $\mu_k \sim \mathcal{N}(0, \tau^2)$, our latent variables Z_i have the distribution $Z_i \sim Mult(\pi)$ for a fixed hyperparameter π, the class-specific proportions, and a data point X_i is distributed as $X_i \sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})$. So our model will be:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \tau^2) \\ Z_i &\sim Mult(\pi) \\ X_i &\sim \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i}), \end{aligned}$$

which leads to the following expression for the posterior distribution using the chain rule:

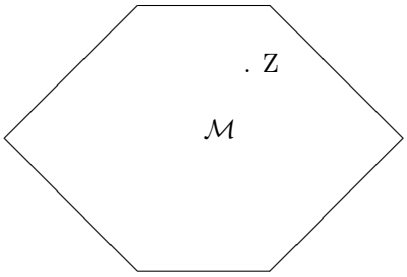

Figure 1: Convex polytope $\mathcal{M}$ containing all feasible parameter values Z

$$p(\mu, Z|X) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i|\pi) p(z_i|x_i, \mu_{1:k})}{\int_{\mu_{1:K}} \Sigma_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i|\pi) p(z_i|x_i, \mu_{1:k}) d\mu_{1:K}}$$

The numerator is simple to compute. However, computing the denominator is exponentially difficult (there are K^n possibilities for Z). In addition, the integral in the denominator is generally not straightforward. These problems serve as our motivation for finding approximate posterior distributions. We will need to tackle the same problem again when we discuss LDA models, which also lead to exponentially hard denominators.

Let us return to our previous discussion on the space of possible parameterizations of our original posterior distributions named $\mathcal{M}$. It turns out that if we consider all the values of Z that are consistent with the posterior distribution derived above, we have a theorem that for most models, including Gaussian models and other discrete models, for the given data, the set of all feasible Z parameters form a convex polytope. Consider each polytope boundary as defined by a constraint imposed by the statistical model. To ensure local and global consistency, the model must be parameterized in such a way as to ensure that the marginal probabilities of subsets of the variables in the model are proper probability distributions, according to the factorization of the joint probability defined by the graphical model structure.

We can try to take advantage of this fact to perform optimization over a convex space to find parameters Z. However, the posterior density on the space M might not have a form that is convenient for optimization.

2.1 Mean Field Approximation

Let us review mean field in order to understand our goal in variational inference. Our goal in mean field is to find the optimal parameters Z with respect to a fully factorized posterior distribution. We assume an approximation to the posterior in the form of $q(Z) = \prod_{i=1}^{m} q(Z_i)$ and optimize the reverse Kullback-Leibler divergence of this fully factorized approximation to the original posterior $p(Z|X)$:

$$\min_{q(Z)} KL(q(Z)||p(Z|X))$$

We should note that each variable in q is assumed to be independent of the others: $q(Z) = \prod_{m=1}^{M} q_m(Z_m)$. Based on the reverse KL divergence, we need the support of $q(Z)$ to lie entirely within the support of

$p(Z|X)$. The marginal independence assumption we imposed here leads us to a parameterization space which is no longer convex as shown in the figure below. However, this new space, called $\mathcal{M}_{MF}$, turns out to be straightforward to optimize over when we use computable expressions for each of the marginal probabilities $q_m(Z_m)$. However, it is possible that Z^*, the optimal set of parameters with respect to the full posterior distribution, does not lie in $\mathcal{M}_{MF}$.

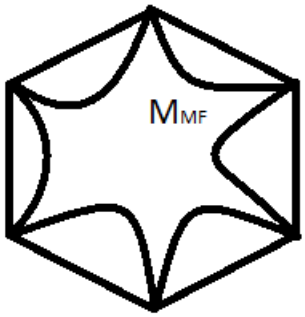

Figure 2: Feasible set of Z with mean field approximation.

Here we perform this optimization using an unnormalized $p(Z|X)$, as discussed last week in class. But let us look at the optimization above from a different perspective by considering a concept called the *Evidence Lower Bound*, or ELBO.

Let us recall Jensen's Inequality. We know that for a given concave function $f(\cdot)$ we have $E[f(x)] \leq f(E[x])$. This can be derived directly using the definition of expectation and applying the concavity property of $f(\cdot)$. As $\log(x)$ is a concave function, we can apply Jensen's Inequality to it. Let us consider $\log(p(X))$. We have:

$$\begin{aligned}
\log(p(x)) &= \log \int_Z p(x,z)dz \\
&= \log \int_Z p(x,z)\frac{q(z)}{q(z)}dz \\
&= \log E_q\left[\frac{p(x,z)}{q(z)}\right]
\end{aligned}$$

Using Jensen's Inequality, we have:

$$\log(p(z)) = \log E_q\left[\frac{p(x,z)}{q(z)}\right] \geq E_q\left[\log\left(\frac{p(x,z)}{q(z)}\right)\right] = E_q[\log(p(x,z))] - E_q[\log(q(z))] = \text{ELBO}$$

Let us recall a result we derived in the last lecture:

$$KL(q||\tilde{p}) = -E_q[\log(p(z,x)] + E_q[\log(q(z)] + \log(p(x)) \Rightarrow KL(q||\tilde{p}) = -\text{ELBO} + \log(p(x))$$

Since $p(x)$ is independent of $q(z)$, minimizing the KL divergence with respect to $q(z)$ is equivalent to maximizing the ELBO. There are two steps to minimize the KL divergence:

1. Choose a fully factorized approximation $q(z)$ to the true joint distribution, $p(z|x)$, such that the variational objective is computable,
2. Maximize the ELBO with respect to $q(z)$ in order to obtain the tightest possible approximation to $p(z|x)$.

3 Ising Model

Before proceeding with variational inference, it is helpful to review the Ising model, a canonical Markov random field, or undirected graphical model. The Ising model is a lattice of unobserved variables $(x_1, \ldots, x_n)$, each with its own (noisy) observation $(y_1, \ldots, y_n)$.

For example, suppose our goal is to reconstruct a denoised image given noisy observations of the pixels. We can think of the lattice as the pixels in a black and white image ($x_i \in \{-1, 1\}$), with a noisy grayscale observation of each pixel ($y_i \in \Re$). We wish to infer the unobserved lattice X from the observed values Y, taking into account the idea in images that there is often local consistency among neighboring pixels. Figure 3 illustrates an Ising model for $n = 9$, with the latent nodes colored white and the observed nodes shaded grey.

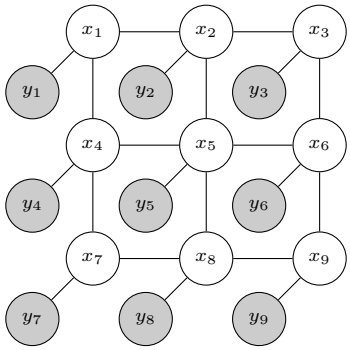

Figure 3: Schematic of an Ising Model

We now define the potential functions of the Ising model, which involve a function on each latent variable, and a function on each of the maximal cliques, or pairs of latent variables connected by edges in this model.

$$\begin{aligned} \psi_s(x_s) &= p(y_i|x_i) \equiv L_i(x_i) \\ \psi_{st}(x_s x_t) &= W_{st} x_s x_t \end{aligned}$$

Continuing with our image example above, we could set $W_{st} = 1$. In general, we set W to positive values if we want neighbors to tend to agree, and negative values if we want them to tend to differ.

Let $N(i)$ be a function that returns the first-degree (hidden) neighbors of node i. For example, in Figure 3, calling $N(x_1)$ would return nodes x_2 and x_4.

Now we can specify functions for our prior and data likelihood:

$$p(x) = \frac{1}{Z_0} \exp\{-\sum_{i=1}^{n} \sum_{j \in N(i)} x_i x_j\}$$

$$p(y|x) \quad = \quad \prod_{i=1}^{n} \exp\{-L_i(x_i)\}$$

where $\frac{1}{Z_0}$ is the normalizing constant. From this, we have the posterior:

$$p(x|y) \quad = \quad \frac{1}{Z} \exp\left\{ -\sum_{i=1}^{n} \sum_{j \in N(i)} x_i x_j - \sum_{i=1}^{n} L_i(x_i) \right\}.$$

3.1 Mean Field Approximation of the Ising Model

Having seen an example of the Ising model for image denoting, we now turn our attention to how we can infer the latent values of this model using a mean field approximation. We do this by "breaking" the edges between the latent variables. We add a mean value (or variational parameter) μ to each x, such that $\mu_i = \mathbb{E}[x_i]$. The new structure is illustrated in Figure 4, with the same color coding as above.

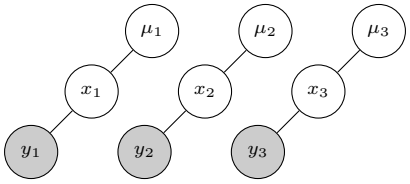

Figure 4: Mean Field Approximation of an Ising Model

If we approximate each local distribution as Bernoulli with parameter μ_i:

$$q(x) = \prod_{i=1}^{n} q_i(x_i),$$

then the general coordinate ascent method we derived to optimize each element $\mu_i = E_{q_i}[q_i(x_i)]$ in the factorized distribution looks like this:

$$\log(q_i(x_i)) = \mathbb{E}_{\neg q_i}[\log \tilde{p}(x)],$$

or the expected value under all of the other q_i except for i of the log of the unnormalized local probability $p(x)$.

We can also approach this optimization by rewriting the ELBO in terms of the Ising model, and optimizing this function with respect to $q_i(x_i)$:

$$\begin{aligned} \log(q_i(x_i)) \quad &= \quad \mathbb{E}_{\neg q_i}[-x_i \sum_{j \in N(i)} x_j - L_i(x_i) + c] \\ q_i(x_i) \quad &\propto \quad \exp\{-x_i \sum_{j \in N(i)} \mu_j - L_i(x_i)\} \end{aligned}$$

where in the second line we push in the expectation (does not impact equality), and we remove the constant term. We find that this term is equivalent to the term we derived using the general coordinate ascent method above. At the end of the day, we will have estimates throughout the full Ising model for each μ_i that represent a point estimate of the denoised image.

Thus,

$$q_i(x_i = +1) \quad = \quad \frac{\exp\{-\sum_{j \in N(i)} \mu_j - L_i(+1)\}}{\sum_{x_i' \in \{+1,-1\}} \exp\{-\sum_{j \in N(i)} \mu_j - L_i(x_i')\}},$$

which can loosely be recognized as a logistic function. We will also write out this term for $x_i = -1$ and take the average of these two terms to find the update for μ_i. Iteratively update:

$$\begin{aligned} \mu_i &= +1(q_i(x_i = +1)) + -1(q_i(x_i = -1)) \\ q_i(x_i) &\propto \exp[\mathbb{E}_{\neg q_i}[\log p(x|y)]]. \end{aligned}$$

Effectively, this q_i is an approximation of the marginalized posterior, the equivalent of a Gibbs step (although we sample from this distribution for Gibbs sampling; for mean field inference we instead take the expected value of this conditional probability).

In the model, we have three parameters of interest: μ_1, μ_2, and μ_{12}, where $\mu_{12} = \mathbb{E}[\psi_{12}(x_1 x_2)]$. We have the constraint $0 \leq \mu_{12} \leq \mu_1 \mu_2$. We limit acceptable values to those within the portion of the simplex that satisfies this constraint. We can visualize this in Figure 5, where the acceptable values occupy the space under the shaded face.

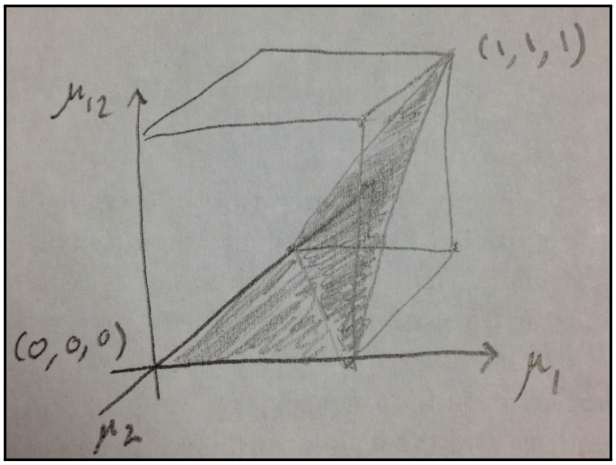

Figure 5: Visualizing Constraints on μ.

If we take a slice of this convex polytope in Figure 5 at $\mu_1 = \mu_2$, it is a triangle like the one in Figure 6. The x-axis in the figure is μ_1. The quadratic curve indicates where $\mu_1^2 = \mu_{1,2}$ – below this curve is the $\mathcal{M}_{MF}$ space. This quadratic function is due to the fully factorized form of the approximation, leading to the marginal independence of μ_1 and μ_2.

4 Loopy Belief Propagation (LBP)

We can now switch to loopy belief propagation. In this variational approximation, q (our approximation to the true distribution) is not necessarily a valid joint probability. Instead, q has "local consistency" or "pairwise consistency," meaning q satisfies the following conditions:

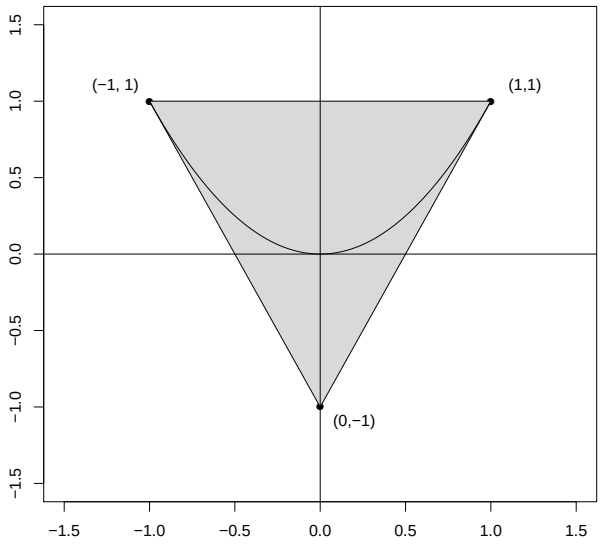

Figure 6: Visualizing μ_{12} when $\mu_1 = \mu_2$

- $\sum_{x_i} q_i(x_i) = 1$
- $\sum_{x_i} q_{ij}(x_i, x_j) = q_j(x_j)$ for all pairs i, j

But we do not guarantee global consistency, e.g., it may not be true that $\sum_{x_i, x_j, x_k} q_{ijk}(x_i, x_j, x_k) = 1$.

Recall that belief propagation in a tree is exact; we get our final answer after one forward and backward pass. This is not true in a graphical model with loops. We apply belief propagation to a loopy graph as follows:

1. Initialize the messages $m_{s \to t}(x_t) = 1$. Initialize the beliefs $\mu_s = 1$.
2. Send messages; $m_{s \to t}(x_t) = \sum_{x_s} L_s(x_s) x_s x_t \prod_{u \in N(s) \setminus t} m_{u \to s}(x_s)$. Update beliefs of each node $\mu_s \propto L_s(x_s) \prod_{t \in N(s)} m_{t \to s}(x_s)$.
3. Repeat until convergence (i.e., until the beliefs do not change substantially).

We can consider this message passing method on our example Ising model above, where $\Psi_s(x_s) = L_s(x_s)$ and $\Psi_{st}(x_s x_t) = x_s x_t$. This represents a factorization that is locally consistent. Then when we draw out this local graph, we see that the message from x_s to x_t is the product over all incoming messages to node x_s except for x_t, scaled by the marginal likelihood of x_s and x_s, x_t, and marginalizing out x_s.

LBP does not always converge for loopy graphs. There are some methods that encourage convergence, including:

- Dampening to avoid oscillation: often these messages oscillate in their estimates
- Asynchronous updates: compute messages along every edge in parallel (synchronous updates) versus compute messages in a sequential way
- Scheduling the messages (update the beliefs in an order that makes sense) using, for example, a tree-based reparameterization

The local consistency requirement means there are fewer constraints on the space $\mathcal{M}$, and so $\mathcal{M}_{LBP}$ is an outer-polytope that contains the polytope of $\mathcal{M}$. In particular, the local consistency constraints means that there will be a total of $O(|V|+|E|)$ constraints (represented by every marginal and every pairwise marginal, captured by the number of edges). See Figure 22.7 (c) in the textbook.

Comparing mean field and LBP, note that

- Mean field is not exact in trees, and LBP is exact for trees.
- Mean field optimizes over node marginals, and LBP optimizes over node and edge marginals.
- Mean field has more local optima than the LBP