10-708: Probabilistic Graphical Models, Spring 2015

13: Variational inference II

Lecturer: Eric P. Xing *Scribes: Ronghuo Zheng, Zhiting Hu, Yuntian Deng*

1 Introduction

We started to talk about variational inference which is sometimes referred to as the modern approaches in graphical model from last lecture. We covered loopy belief propagation in last class. Today we are going to cover variational (Bayesian) inference and mean field approximations. Before we start today's lecture, let's first review the goals and approaches for inference as well as the loopy belief propagation covered in last class.

There are four goals of inference in graphical models. The first goal is to compute the likelihood of observed data maybe in models with latent variables. The second goal is to compute the marginal distribution over a given subset of nodes in the model. The third one is to compute the conditional distribution over a subset of nodes. The last one is to compute a mode of the density.

There are two categories of approaches to inference: exact inference algorithms and approximate inference algorithms. The exact inference algorithms include brute force, the elimination algorithm, message passing (also called sum-product algorithm or belief propagation) and junction tree algorithm. Approximate inference algorithms include loopy belief propagation, which has been covered in last class, variational Bayesian inference and mean field approximations, which will be covered today, and stochastic simulation, sampling or MCMC, which will be covered in the future classes.

Now let's briefly review the loopy belief propagation. The key idea of the loopy belief propagation is to introduce message passing ("belief propagation") on loopy graphs or non-trees. Theoretically, messages may circulate indefinitely for loopy graphs. However, it seems that the loopy belief propagation often works empirically. To understand the puzzle, we can view loopy belief propagation as variational inference. First, we wrote down the KL-divergence between an approximate distribution Q and the distribution P we want to infer. Then, we define a similar value, i.e., the Gibbs "Free Energy", which consists of an entropy term and an expected log marginal term. However, computing the Free Energy is hard in general, so we instead use approximations, such as the Bethe approximation. We then can effectively minimize the Bethe Free Energy, i.e., the Free Energy with Bethe approximation.

2 Variational (Bayesian) Inference

Now let's start today's lecture on Variational (Bayesian) Inference and Mean Field Approximations (MFA). In modern machine learning, variational (Bayesian) inference, which we will refer to here as variational Bayes, is most often used to infer the conditional distribution over the latent variables given the observations (and parameters), which is also known as the posterior distribution over the latent variables given observations. We will start with introducing the notation used in the problem setup and motivating examples for using variational Bayes. The notation and examples used below are from David Blei's tutorial on Variational Inference.

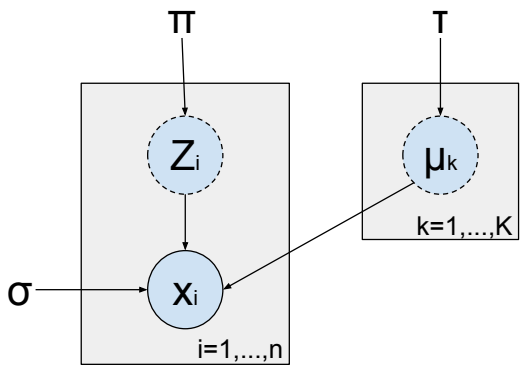

Figure 1: Graphical Model for Bayesian Mixture of Gaussians

2.1 Problem Setup

We use the following notation for the rest of the lecture. There are n observations, $x = x_{1:n}$, and m latent variables, $z = z_{1:m}$. The fixed parameters α could be for the distribution over the observations or over the hidden variables. The general framework $\{x_{1:n}, z_{1:m}, \alpha\}$ introduced here can be used to describe (just about) any graphical model. With the notation, the posterior distribution is written:

$$p(z|x, \alpha) = \frac{p(z, x|\alpha)}{\int_z p(z, x|\alpha)}$$

The posterior distribution is well defined given our notations. Why do we often need to use an approximate inference methods (such as variational Bayes) to compute the posterior distribution over nodes in our graphical model? The reason is that we cannot directly compute the posterior distribution for many interesting models, especially the integral term in the denominator. In other words, the posterior density is in an intractable form (often involving integrals) which cannot be easily analytically solved.

2.2 Motivating Example

As a motivating example, we will try to compute the posterior for a (Bayesian) mixture of Gaussians. The likelihood or the generative process of Bayesian mixture of Gaussians is:

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1, \ldots, K$.
2. For $i = 1, \ldots, n$, draw $z_i \sim Cat(\pi)$ and draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$.

The graphical model for Bayesian mixture of Gaussians is as in Figure 1. Note that in Bayesian mixture of Gaussians, we have observed variables, $x_{1:n}$, latent variables $\mu_{1:k}$ and $z_{1:n}$, and parameters $\{\tau^2, \pi, \sigma^2\}$. We can rewrite the posterior distribution of Bayesian mixture of Gaussians as:

$$p\left(\mu_{1:K}, z_{1:n}|x_{1:n}\right) = \frac{\prod_{k=1}^{K} p\left(\mu_k\right) \prod_{i=1}^{n} p\left(z_i\right) p\left(x_i|z_i, \mu_{1:K}\right)}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p\left(\mu_k\right) \prod_{i=1}^{n} p\left(z_i\right) p\left(x_i|z_i, \mu_{1:K}\right)}$$

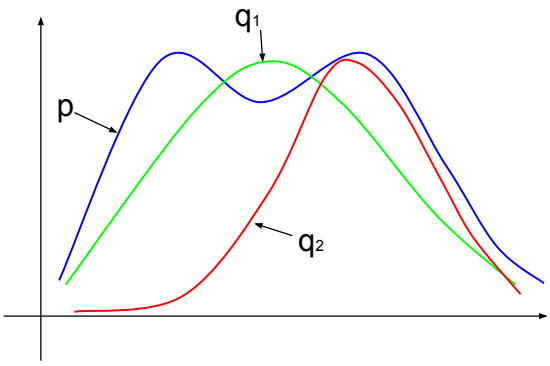

Figure 2: Example of the Limitation of KL Divergence

where we have suppressed writing the parameters for ease of notation. In the posterior distribution, the numerator can be computed for any choice of the latent variables. However, the integral in the denominator, which is the marginal probability of the observations, cannot easily be computed analytically. Moreover, even we can find some way to approximate the denominator numerically, the computational complexity increases exponentially on the domain of latent variables. Therefore, we need to use approximate inference methods (such as variational Bayes) to compute the posterior distribution over nodes in our graphical model.

2.3 Variational Bayes

The main idea behind variational Bayes is as following. First, we choose a family of distributions over the latent variables with its own set of variational parameters ν, i.e. $q(z_{1:m}|\nu)$. Then, we find the setting of the parameters that makes our approximation q closest to the posterior distribution. This is where optimization algorithms come in. After that, we can use q with the fitted parameters in place of the posterior. For example, we can use q with the fitted parameters to form predictions about future data, or to investigate the posterior distribution over the hidden variables, find modes, etc.

2.3.1 Kullback-Leibler Divergence

In order to find the setting of the parameters that makes our approximation q closest to the posterior distribution, we need find a way to measure the closeness of two distributions first. The Kullback-Leibler (KL) divergence is a widely used measurement of the closeness of two distribution. The KL divergence is defined to be:

$$KL\left(q||p\right) = \int_z q\left(z\right) \log \frac{q\left(z\right)}{p\left(z|x\right)} = \mathbb{E}_q\left[\log \frac{q\left(z\right)}{p\left(z|x\right)}\right]$$

By the definition of the KL divergence, there are three "cases" of importance intuitively. First, q and p are close according to KL divergence, i.e., KL divergence is low, if q is high and p is high. Second, q and p are not close according to KL divergence, i.e., KL divergence is high, if q is high and p is low. Third, q and p are close according to KL divergence, i.e., KL divergence is low, if q is low regardless of p. The third case may be problematic when we choose to minimize $KL\left(q||p\right)$ to select our closest approximation q for the posterior distribution p. We use an example in Figure 2 to illustrate the intuition. In Figure 2, intuitively, q_1 is a

better approximation to p than q_2, since the overall range and shape of q_1 is closer to p than q_2. However, if we use KL divergence, we may find that $KL(q_2||p) < KL(q_1||p)$. Intuitively, it might make more sense to consider $KL\left(p||q\right)$. However, we do not do this for computational reasons.

After introducing the KL divergence, to do variational Bayes, we want to minimize the KL divergence between our approximation q and our posterior p. However, we can not actually minimize this quantity (we will show why later), but we can minimize a function that is equal to it up to a constant. This function is known as the evidence lower bound (ELBO). Recall that the "evidence" is a term used for the marginal likelihood of observations (or the log of that).

2.3.2 Evidence Lower Bound

First, we derive the evidence lower bound by applying Jensen's inequality to the log (marginal) probability of the observations.

$$\begin{aligned}
\log p\left(x\right) &= \log \int_z p\left(x,z\right) \\
&= \log \int_z p\left(x,z\right) \frac{q\left(z\right)}{q\left(z\right)} \\
&= \log \left(\mathbb{E}_q \left[\frac{p\left(x,z\right)}{q\left(z\right)} \right] \right) \\
&\geq \mathbb{E}_q \left[\log p\left(x,z\right)\right] - \mathbb{E}_q \left[\log q\left(z\right)\right]
\end{aligned}$$

where the finial line is the ELBO, and the inequality comes from Jensen's inequality, i.e., when function f is concave, $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$. The inequality implies that ELBO is a lower bound for the evidence. In other words, ELBO is less than or equal to the evidence, i.e., the log marginal probability of the observations. All together, the Evidence Lower Bound (ELBO) for a probability model $p(x, z)$ and approximation $q(z)$ to the posterior is:

$$\mathbb{E}_q \left[\log p\left(x,z\right)\right] - \mathbb{E}_q \left[\log q\left(z\right)\right]$$

where the second expectation is the entropy, another quantity from information theory.

Now we will show that the KL divergence to the posterior is equal to the negative ELBO plus a constant. We can write the KL divergence as:

$$\begin{aligned}
KL\left(q||p\right) &= \mathbb{E}_q \left[\log \frac{q\left(z\right)}{p\left(z|x\right)} \right] \\
&= \mathbb{E}_q \left[\log q\left(z\right)\right] - \mathbb{E}_q \left[\log p\left(z|x\right)\right] \\
&= \mathbb{E}_q \left[\log q\left(z\right)\right] - \mathbb{E}_q \left[\log p\left(z,x\right)\right] + \log p\left(x\right) \\
&= -\left(\mathbb{E}_q \left[\log p\left(z,x\right)\right] - \mathbb{E}_q \left[\log q\left(z\right)\right]\right) + \log p\left(x\right)
\end{aligned}$$

where the final line is the negative ELBO plus a constant that does not depend on q. Therefore, finding an approximation that maximizes the ELBO is equivalent to finding the q that minimizes the KL divergence to the posterior. It is also interesting to note that the difference between the ELBO and the KL divergence is the log normalizer (i.e. the evidence), which is the quantity that the ELBO bounds. Choosing a proper family of variational distributions, the ELBO can be computed but the evidence itself cannot easily be computed as we have shown in the motivating example. Therefore, we cannot actually compute the KL divergence or minimize the KL divergence directly. Instead, we can maximize the ELBO over densities $q(z)$, which is equivalent to minimize KL divergence.

Overall, the basic procedure of variational Bayes is as following. First, we choose a family of variational distributions, i.e. a family of approximations, such that these two expectations in ELBO, i.e., $\mathbb{E}_q \left[\log p\left(x,z\right)\right]$

and $\mathbb{E}_q[\log q(z)]$, can be computed (we will discuss a specific family of approximations next). Then, we optimize ELBO over densities $q(z)$ in variational Bayes to find an "optimal approximation".

3 Mean Field Variational Inference

We now describe a popular family of variational approximations called mean field approximations.

3.1 Mean Field Approximation

In order to make the posterior inference tractable, we assume the variational distribution over latent variables factorizes as:

$$q(z_1, \cdots, z_m) = \prod_{j=1}^{m} q(z_j)$$

The above variational approximation assumes complete factorization of the distribution over individual latent variables, which is often referred as "naive mean field". For a more general setting, we can assume the variational distribution factorizes into R groups $z_{G_1}, \cdots, z_{G_R}$, which is referred to as "generalized mean field":

$$q(z_1, \cdots, z_m) = q(z_{G_1}, \cdots, z_{G_R}) = \prod_{r=1}^{R} z_{G_r}$$

3.2 Mean Field Method

Under the above mean field approximation, we can optimize the ELBO using coordinate ascent optimization now. First, recall that the ELBO is defined as:

$$\mathcal{L} = \mathbb{E}_q[\log p(x, z)] - \mathbb{E}_q[\log q(z)]$$

Using chain rule, we can write $\log p(x, z)$ as:

$$p(z_{1:m}, x_{1:n}) = p(x_{1:n}) \prod_{j=1}^{m} p(z_j | z_{1:(j-1)}, x_{1:n})$$

Using mean field approximation, we can decompose $\mathbb{E}_q[\log q(z)]$ as:

$$\mathbb{E}_q[\log q(z_{1:m})] = \sum_{j=1}^{m} \mathbb{E}_{q_j}[\log q(z_j)]$$

Therefore, the ELBO can be written as:

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{j=1}^{m} (\mathbb{E}_q[\log p(z_j | z_{1:(j-1)}, x_{1:n})] - \mathbb{E}_{q_j}[\log q(z_j)])$$

Then we can derive the coordinate ascent update for a latent variable z_j while keeping all other latent variables fixed. First, we re-order the latent variables in the sum such that the j^{th} variable comes last, then

we take argmax of $\mathcal{L}$ with respect to $q(z_j)$ and remove the parts that do not depend on $q(z_j)$, we can write:

$$\begin{aligned}\text{argmax}_{q_j}\mathcal{L} &= \text{argmax}_{q_j}(\mathbb{E}_q[\log p(z_j|z_{1:(j-1)}, x)] - \mathbb{E}_{q_j}[\log q(z_j)]) \\ &= \text{argmax}_{q_j}(\mathbb{E}_q[\log p(z_j|z_{-j}, x)] - \mathbb{E}_{q_j}[\log q(z_j)]) \\ &= \text{argmax}_{q_j}(\int q(z_j)\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]dz_j - \int q(z_j)\log q(z_j)dz_j)\end{aligned}$$

Note that in the above derivation we have decomposed the expectation over q as an integral over z_j of an expectation over $q(z_{-j})$. Define the term inside the argmax on the last line to be $\mathcal{L}_j$, i.e.

$$\mathcal{L}_j = \int q(z_j)\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]dz_j - \int q(z_j)\log q(z_j)dz_j$$

Then we take the derivative of $\mathcal{L}_j$ with respect to $q(z_j)$ using Lagrange multipliers, and set the derivative to zero to find the argmax:

$$\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)] - \log q(z_j) - 1 = 0$$

Then we have obtained the coordinate ascent update of $q(z_j)$:

$$q^*(z_j) \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]\}$$

Since $p(z_j|z_{-j}, x(z_{-j}, x) = \frac{p(z_j, z_{-j}, x)}{p(z_{-j}, x)}$ and the denominator does not depend on z_j, we can equivalently write:

$$q^*(z_j) \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(z_j, z_{-j}, x)]\}$$

Note that there is generally no guarantee of convexity of ELBO, the coordinate ascent procedure converges to a local maximum.

3.3 Simple Example: Multinomial Conditionals

If we choose a model whose conditional distribution is multinomial, i.e.

$$p(z_j|z_{-j}, x) = \pi(z_{-j}, x)$$

Then the coordinate update for $q(z_j)$ is:

$$q^*(z_j) \propto \exp\{\mathbb{E}[\log \pi(z_{-j}, x)]\}$$

Which is also multinomial. Below we will see a more general conclusion that for models whose conditional probabilities are in exponential family, the updates are also in the same exponential family.

4 Exponential Family Conditionals

4.1 Exponential Family Conditional Models

Exponential-family-conditional models, also known as conditionally conjugate models are models whose conditional densities that are in exponential family, i.e. of the form:

$$p(z_j|z_{-j}, x) = h(z_j)\exp\{\eta(z_{-j}, x)^T t(z_j) - a(\eta(z_{-j}, x))\}$$

where h, η, t and a are functions that parameterize the exponential family. Different choices of these parameters lead to popular densities such as normal, gamma, exponential, Bernoulli, Dirichlet, categorical, beta, Poisson, geometric, etc. Many popular models such as Bayesian mixtures of exponential family models with conjugate priors, hierarchical hidden Markov models, Bayesian linear regression, to name a few, fall into this category, while some popular models such as topic model do not fall into this category.

4.2 Variational Inference for Exponential Family Conditionals

We derive a general formula for the coordinate ascent update for exponential-family-conditional models in this section. First, we choose the form of local variational approximation $q(z_j)$ to be the same as the conditional distribution (i.e. in an exponential family). Then we will show that the coordinate ascent update for $q(z_j)$ yields an optimal $q^*(z_j)$ in the same family.

Recall from Section 3.2 the coordinate ascent update for optimizing the ELBO with respect to $q(z_j)$ is:

$$q^*(z_j) \propto \exp\{\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)]\}$$

In exponential-family-conditional models, the log of the conditional is:

$$\log p(z_j|z_{-j}, x) = \log h(z_j) + \eta(z_{-j}, x)^T t(z_j) - a(\eta(z_{-j}, x))$$

Take the expectation with respect to $q(z_{-j})$, we have:

$$\mathbb{E}_{q_{-j}}[\log p(z_j|z_{-j}, x)] = \log h(z_j) + \mathbb{E}_{q_{-j}}[\eta(z_{-j}, x)]^T t(z_j) - \mathbb{E}_{q_{-j}}[a(\eta(z_{-j}, x))]$$

The last term does not depend on $q(z_j)$, so we have the update for $q(z_j)$:

$$q^*(z_j) \propto h(z_j) \exp\{\mathbb{E}_{q_{-j}}[\eta(z_{-j}, x)]^T t(z_j)\}$$

From the above optimal form, we can see that $q^*(z_j)$ is in the same exponential family as the conditional.

We can also write the update formula in terms of variational parameters ν: Assume the factorization of joint distribution takes the following form:

$$q(z_{1:m}|\nu) = \prod_{j=1}^{m} q(z_j|\nu_j)$$

where each local variational approximation $q(z_j|\nu_j)$ has an exponential family form. Then the coordinate ascent update for $q(z_j)$ can be written as update in its parameter:

$$\nu_j^* = \mathbb{E}_{q_{-j}}[\eta(z_{-j}, x)]$$

5 Example: Generalized Mean Field

5.1 Markov Random Fields

If we assume the joint distribution completely factorizes by variables in Markov Random Fields, i.e.

$$q(x) = \prod_{s \in V} q(x_s)$$

then we are making the assumption that the posterior distribution of variables are independent of each other, i.e. dropping all edges between variables, as shown in the Ising model in Figure 3.

The above completely factorization assumption is fairly general, which is called naive mean field approximation. We can also apply more general forms of the mean field approximations, i.e. clusters of disjoint latent variables are independent, while the dependencies of latent variables in each clusters are preserved. An illustration example is shown in Figure 4, where the left figure is the original Ising model, the middle figure visualizes the independency assumptions of generalized mean field with 2×2 clusters and the right figure visualizes the independency assumptions of generalized mean field with 4×4 clusters.

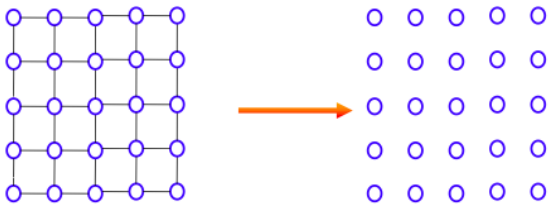

Figure 3: Naive Mean Field for MRF

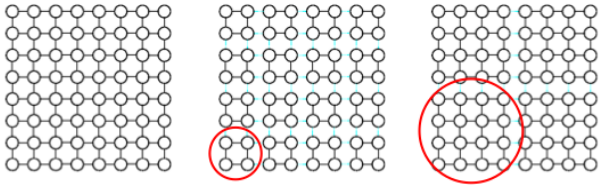

Figure 4: Generalized Mean Field for MRF

5.2 Factorial HMMs

Let's take a look at another example: factorial HMM. We can make naive mean field approximations such that all variables are assumed to be independent in posterior. We can also make generalized mean field approximations such that each disjoint cluster contains one hidden Markov chain, two hidden Markov chains or three hidden Markov chains. The original factorial HMM and a generalized mean field approximation based on clusters with two hidden Markov chains are shown in Figure 5.

Figure 6 shows the singleton marginal error and CPU time for naive mean field, generalized mean field with clusters of different number of Markov chains and exact inference BP algorithm. From the Singleton marginal error histogram, we can see that as expected, naive mean field drops all edges in the posterior, thus has the highest error, while generalized mean field with clusters of more chains generally behave better in terms of singleton marginal error. From the CPU time histogram, we can see that the exact inference method takes the most time, while mean field approximations generally take less time.

From the above example, we can draw the conclusion that mean field approximation makes inference tractable or cheap compared with exact inference methods, at the cost of additional independency assumptions, thus leads to a bias in the inference result.

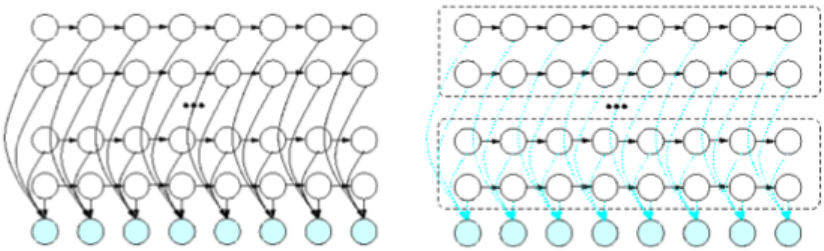

Figure 5: Mean Field Approximation for fHMM

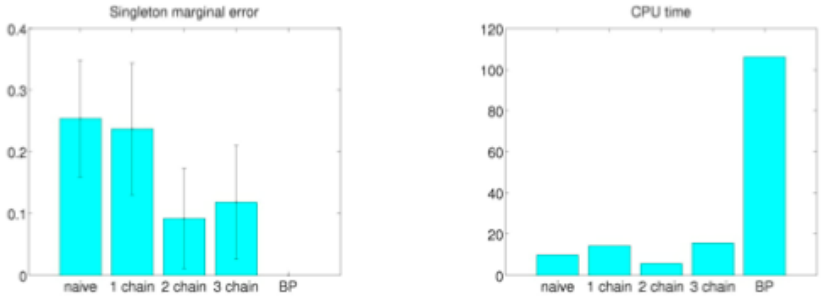

Figure 6: Error and inference cost for fHMM

Note: Some figures in this note are adapted from `http://www.cs.cmu.edu/~epxing/Class/10708-15/slides/lecture13-VI.pdf`.