COS 513: FOUNDATIONS OF PROBABILISTIC MODELING
LECTURE 20

JIMIN SONG AND BANGPENG YAO

1. Models with Exponential Family Conditionals

Consider the distribution $p(\mathbf{X}) = p(x_1, \cdots, x_N)$ from which we wish to sample. In each step of the Gibbs sampling procedure, we replace x_i by a value sampled from the distribution $p(x_i|\mathbf{x}_{-i})$, where x_i denotes the i-th component of $\mathbf{X}$, and $\mathbf{x}_{-i}$ includes all components in $\mathbf{X}$ other than x_i. Gibbs sampling repeats this procedure by cycling through all the variables in some particular order.

Therefore, Gibbs sampling involves the computation of many conditional distributions. Whether those conditional distributions are easy to compute determines the feasibility of a Gibbs sampling approach. In most of the situations, the conditional distributions take the form of exponential family, from which the approach is very easy to sample. Later discussions will show that exponential family conditions are also helpful in variational methods.

In many graphical models, every conditional distribution $p(x_i|\mathbf{x}_{-i})$ is an exponential family, which makes the inference on these models very easy. For example, in the Gaussian mixture model, each conditional distribution takes the form of the following, respectively.

$$\texttt{uncollapsed} \begin{cases} p(\mu_k|\boldsymbol{\mu}_{-k}, \mathbf{z}_{1:N}) & \sim \texttt{Gaussian} \\ p(z_n|\mathbf{z}_{-n}, \boldsymbol{\mu}_{1:K}) & \sim \texttt{Multinomial} \end{cases} \tag{1}$$

$$\texttt{collapsed}: \quad p(z_n|\mathbf{z}_{-n}) \sim \texttt{Multinomial} \tag{2}$$

In general, conditional distributions in the exponential family are defined to be the set of distributions of the form

$$p(x_i|\mathbf{x}_{-i}) = \exp\{g(\mathbf{x}_{-i})^T t(x_i) - a(g(\mathbf{x}_{-i}))\} \tag{3}$$

where $t(x_i)$ is a function of x_i, $g(\mathbf{x}_{-i})$ are called the *natural parameters* of this distribution.

Besides the Gaussian mixture model mentioned above, other graphical models whose conditionals are exponential families include Kalman filters, Hidden Mixture Models, Mixtures of conjugate priors, etc.

2. VARIATIONAL METHODS

2.1. **Introduction.** There are two classes of approximate inference schemes, stochastic and deterministic approximates. Markov chain Monte Carlo belongs to stochastic techniques. Here we introduce variational inference, which is a deterministic technique.

Variational methods were firstly introduced in statistics physics, and then applied in machine learning problems. Recently, it has begun to be used in Bayesian inference. In this chapter, we care about using variational methods for inference. The key idea behind this is to use optimization to perform a difficult computation.

2.2. **Comparision to Markov Chain Monte Carlo.** The difference between variational methods and sampling (Markov chain monte carlo) are summarized as in Table 1. Besides what is shown in Table 1, one folk wisdom about sampling and variational methods is that variational methods are faster, but come with fewer theoretical guarantees.

Sampling (MCMC)	Variational
Key idea: Approximate the posterior with samples that are (hopefully) from it.	**Key idea:** Posit a family of distributions over the latent variables indexed by free variational parameters. Then fit these parameters to be (hopefully) close to the posterior.
Issue: What is the proposal distribution? Burn-in? Lag?	**Issue:** What is the family to use? How to optimize?
Computational bottleneck: Sampling, computing acceptance probability, assessing convergence.	**Computational bottleneck:** Optimization

TABLE 1. Comparison of variational and sampling methods.

2.3. **Variational Methods in Bayesian Models.** In this part, we consider how the variational optimization can be applied to the inference problem. Suppose we have a Bayesian model in which all parameters are given prior distributions. The set of all observed random variables are denoted by $\mathbf{X} = \{x_1, \cdots, x_N\}$. The model also have a set of latent variables which are denoted as $\mathbf{Z} = \{z_1, \cdots, z_M\}$. Note that N and M can be different. The probabilistic model specifies the joint distribution $p(\mathbf{X}, \mathbf{Z})$, and our goal is to find an

approximation for the posterior distribution $p(\mathbf{Z}|\mathbf{X})$. Because

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z})}{p(\mathbf{X})}, \tag{4}$$

in order to infer $p(\mathbf{Z}|\mathbf{X})$ we need to compute the normalizing constant $p(\mathbf{X})$, which is the marginal probability of the observations.

Using Jensen's inequality, we have

$$\begin{aligned} \log(p(\mathbf{X})) &= \log \int p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} \\ &= \log \int p(\mathbf{X}, \mathbf{Z}) \frac{q(\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &\geqslant \int q(\mathbf{Z}) \log(p(\mathbf{X}, \mathbf{Z})) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\ &= \mathbb{E}_q[\log(p(\mathbf{X}, \mathbf{Z}))] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &\equiv l(q, \mathbf{X}) \end{aligned} \tag{5}$$

where the bound is tight when $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$.

Then variational methods try to compute $\log(p(\mathbf{X}))$ through optimization by finding q that maximizes $l(q, X)$,

$$\log(p(\mathbf{X})) = \max_{q \in \mathcal{M}} (\mathbb{E}_q[\log(p(\mathbf{X}, \mathbf{Z}))] - \mathbb{E}_q[\log q(\mathbf{Z})]) \tag{6}$$

where $\mathcal{M}$ is a family of distributions including $p(\mathbf{Z}|\mathbf{X})$ that assert no more conditional independences than those in $p(\mathbf{Z}|\mathbf{X})$. For instance, $\mathcal{M}$ can be all joint probability distributions on $\mathbf{Z}$ that contains no conditional independence.

In general, we cannot do this optimization over $\mathcal{M}$. So, in order to compute $l(q, \mathbf{X})$ and attempt optimization, we need to restrict $\mathcal{M}$ to a simpler family $\mathcal{M}_{tract}$, which is tractable, so that

$$\log(p(\mathbf{X})) \geq \max_{q \in \mathcal{M}_{tract}} (l(q, \mathbf{X})). \tag{7}$$

3. Mean-Field Variational Methods

3.1. **Introduction.** Two fundamental problems with variational methods are both $\mathcal{M}$ and the conjugate dual function of A, A^* are hard to characterize in explicit form. Mean-field variational methods are one type of variational methods where $\mathcal{M}$ is restricted to tractable subfamiles of distributions $\mathcal{M}_{tract}$ in such a way that the distributions are fully factorized. That is, the distributions in $\mathcal{M}_{tract}$ have the form

$$q(\mathbf{Z}) = q(z_1|\nu_1) q(z_2|\nu_2) \cdots q(z_M|\nu_M) \tag{8}$$

where $\nu_{1:M}$ are *variational parameters* and each z_i is parameterized by ν_i as in Figure 1.

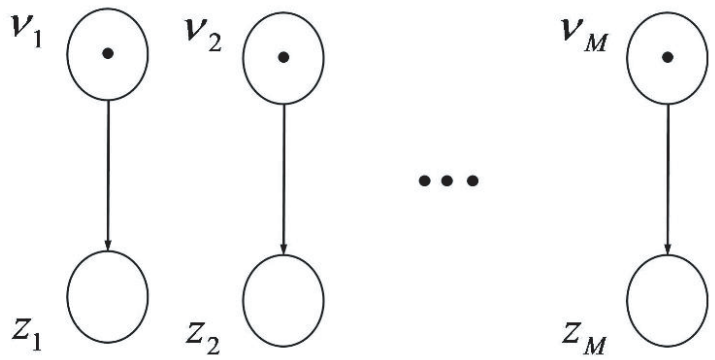

FIGURE 1. A graphical model of the distributions in $\mathcal{M}_{tract}$

Since each z_i can be any distribution governed by a variational parameter ν_i, there are M different distributions and $z_1, z_2, \ldots, z_M$ are independent. This is called a naive mean field approach. For more complicated approach, we can consider dependences between z_i's by putting some edges between the nodes in the graphical model. This is called a structured mean field approach.

We can find the approximating distribution q that maximizes the objective function $l(q, \mathbf{X})$ over $\mathcal{M}_{tract}$. Actually, that q is the approximate posterior distribution. We do not need to go through computations of equation 7 and equation 4 by incorporating q into them.

3.2. **Mean-field and Kullback-Leibler divergence.** An alternative interpretation to explain mean-field variational methods is to minimize the difference between the approximating distribution and the target distribution using KL divergence (Kullback-Leibler divergence). KL divergence is an information theoretic measure of the "distance" between two distributions, defined as for $p_1(\mathbf{X})$ and $p_2(\mathbf{X})$,

$$KL(p_1(\mathbf{X})\|p_2(\mathbf{X})) = \mathbb{E}_{p_1}\left[\log\left(\frac{p_1(\mathbf{X})}{p_2(\mathbf{X})}\right)\right]. \tag{9}$$

Maximizing the objective function $l(q, \mathbf{X})$ with respect to q is equivalent to finding the $q \in \mathcal{M}_{tract}$ that minimizes $KL(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X}))$. I.e.,

$$
\begin{aligned}
(10) \qquad q^* &= \arg\max_{q \in \mathcal{M}_{tract}} l(q, X) \\
&= \arg\max_{q \in \mathcal{M}_{tract}} \left(\mathbb{E}_q[\log(p(\mathbf{X}, \mathbf{Z}))] - \mathbb{E}_q[\log q(\mathbf{Z})]\right) \\
&= \arg\min_{q \in \mathcal{M}_{tract}} \left(\mathbb{E}_q[\log q(\mathbf{Z})] - \mathbb{E}_q[\log(p(\mathbf{Z}|\mathbf{X}))]\right) \\
&= \arg\min_{q \in \mathcal{M}_{tract}} KL(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X})).
\end{aligned}
$$