COS513: Variational Inference

Scribes: Tina Lee, Zhizhen Zhao

January 16, 2011

1 Introduction

Variational Inference (VI) is a deterministic alternative to MCMC. Think of it this way: VI is a deterministic algorithm that can be randomized, whereas MCMC is a randomized algorithm. VI replaces sampling with optimization. Let's say the big circle in Figure 1 represents the space for all possible variables:

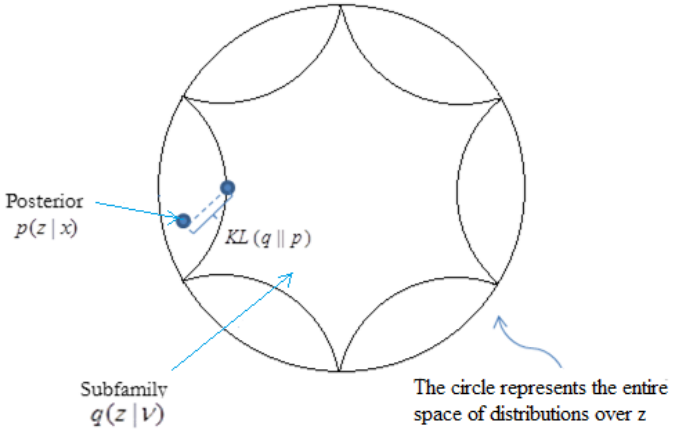
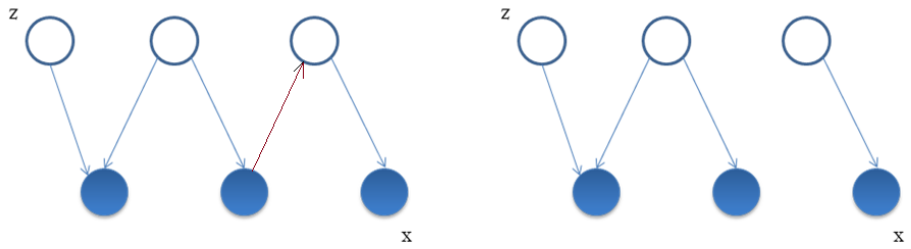


Figure 1: Space of distribution family p and the space of subfamily q. KL divergence measures the closeness of two distributions $p(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{z}|\boldsymbol{\nu})$.

In VI we define a subfamily of distribution over latent variables $q(\mathbf{z}|\boldsymbol{\nu})$. The objective here is to find a point in the subfamily distribution that is closest to our posterior $p(\mathbf{z}|\mathbf{x})$. KL divergence measures the closeness or distance between the two distributions (p and q). Note here that if you start with a graphical model as Figure 2a and take out one edge you would get a subfamily in Figure 2b.

Let's go back to a setting where we have a mixture model (Figure 3). The shaded $x_1, ..., x_4$ in Figure 3 represent observed data. Note here that the components β are in fact dependent on all these latent variables z.



(a) (b)

Figure 2: The connected graphical model (a) and the subfamily (b)

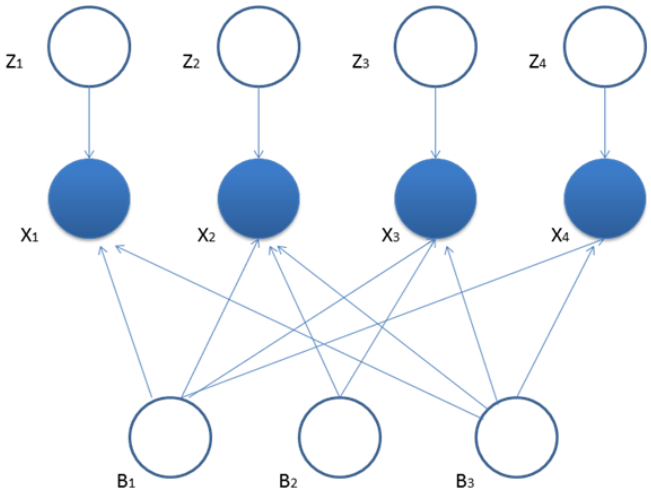


Figure 3: Bayesian Mixture Model

The VI method is an adaptation of ideas in statistical physics to more general statistical models. The VI method was pioneered by Michael Jordan's group at MIT in the mid 90s. The driving force behind this was a statistician by the name of Tommi Jaakkola. We are understanding more about how optimization can help us understand Bayesian Inference models because of VI. Other members of the group include Saul, Ghahramani, and Jordan himself. Jaakkola's dissertation is really interesting for understanding more about this. Others who have worked on optimization include Globerson, Wainwright, and Sontag.

Recall Jenson's bound of $p(\mathbf{x})$. Let $x_{1:N}$ be observations and $z_{1:M}$ be hidden variables.

$$\log p(\mathbf{x}) = \log \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \log \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \frac{q(\mathbf{z}|\boldsymbol{\nu})}{q(\mathbf{z}|\boldsymbol{\nu})} \tag{1}$$

$$\geq \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\nu}) \log p(\mathbf{x}, \mathbf{z}) - \int_{\mathbf{z}} q(\mathbf{z}|\boldsymbol{\nu}) \log q(\mathbf{z}|\boldsymbol{\nu}) \tag{2}$$

$$= \mathbb{E}_q \log p(\mathbf{x}, \mathbf{Z}) - \mathbb{E}_q \log q(\mathbf{Z}|\boldsymbol{\nu}). \tag{3}$$

The last line above defines our objective function for optimization.

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_q \log p(\mathbf{x}, \mathbf{Z}) - \mathbb{E}_q \log q(\mathbf{Z}|\boldsymbol{\nu}). \tag{4}$$

This is called "ELBO": Evidence Lower BOund.

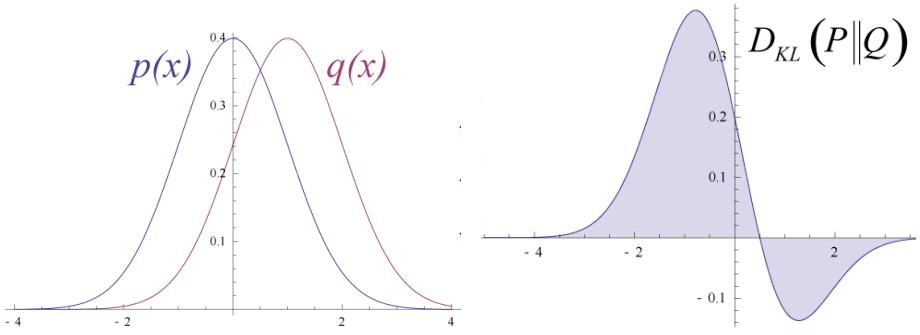
2 Kullback-Leibler divergence

First we define what KL divergence is.

$$KL(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) = \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \tag{5}$$

$$= \mathbb{E}_q \log \frac{q(\mathbf{z}|\boldsymbol{\nu})}{p(\mathbf{z}|\mathbf{x})} \tag{6}$$

$$= \mathbb{E}_q \log q(\mathbf{z}|\boldsymbol{\nu}) - \mathbb{E}_q \log p(\mathbf{z}|\mathbf{x}). \tag{7}$$



(a) (b)

Figure 4: An illustration of KL divergence. a. Blue curve is the distribution $p(x)$ and red curve is the distribution $q(x)$. b. The shade region is to be integrated to get $D_{KL}(p\|q)$. Note KL divergence is not symmetric.

Recall $p(\mathbf{z}|\mathbf{x}) = p(\mathbf{x}, \mathbf{z})/p(\mathbf{x})$. Then equation (7) is equal to

$$\mathbb{E}_q \log q(\mathbf{z}|\boldsymbol{\nu}) - \mathbb{E}_q \log p(\mathbf{x}, \mathbf{z}) + \log p(\mathbf{x}). \tag{8}$$

Compare equation (4) and equation (8), we see here that minimizing KL is equivalent to maximizing ELBO.

3 Variational inference

Assume $p(\mathbf{x}, \mathbf{z})$ is in the exponential family,

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\eta}) = \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x}, \mathbf{z}) - a(\boldsymbol{\eta})\right\}. \tag{9}$$

This implies that

$$p(z_i|\mathbf{z}_{-i}, \mathbf{x}) = \exp\left\{\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})^T u(z_i) - a(\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x}))\right\}. \tag{10}$$

Now define the variational family

$$q(\mathbf{z}|\boldsymbol{\nu}) = \prod_{i=1}^{m} q(z_i|\nu_i). \tag{11}$$

This is called the mean-field family (or the fully factorized family). We can denote entropy here as

$$H(q) = \sum_{i=1}^{m} H(q(z_i|\nu_i)) \tag{12}$$

Further assume each $q(z_i|\nu_i)$ is in the same exponential family as $p(z_i|\mathbf{z}_{-i}, \mathbf{x})$

$$q(z_i|\nu_i) = \exp\left\{\nu_i^T u(z_i) - a(\nu_i)\right\}. \tag{13}$$

Notice that a is the same as in equation (10).

$$\begin{aligned} H(q(z_i|\nu_i)) &= -\mathbb{E}\log q(z_i|\nu_i) && (14)\\ &= -\nu_i^T \mathbb{E}[u(z_i)] + a(\nu_i) && (15)\\ &= -\nu_i^T a'(\nu_i) + a(\nu_i), && (16) \end{aligned}$$

where $a'(\nu_i) = \frac{\partial a(\nu_i)}{\partial \nu_i}$.

Here we use coordinate ascent with ν_i. The partial derivative of $\mathbb{E}[\log p(\mathbf{x}, \mathbf{z})]$ with respect to ν_i equals $\frac{\partial}{\partial \nu_i}\mathbb{E}[\log p(z_i|\mathbf{z}_{-i}, \mathbf{x})]$.

$$\begin{aligned} \frac{\partial}{\partial \nu_i}\mathcal{L}(\boldsymbol{\nu}) &= \frac{\partial}{\partial \nu_i}\left(\mathbb{E}\left[\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})^T u(z_i)\right] - \mathbb{E}\left[a(\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x}))\right] - \nu_i^T a'(\nu_i) + a(\nu_i)\right) && (17)\\ &= \frac{\partial}{\partial \nu_i}\left(\mathbb{E}\left[\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})\right]^T a'(\nu_i) - \mathbb{E}\left[a(\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x}))\right] - \nu_i^T a'(\nu_i) + a(\nu_i)\right) && (18)\\ &= a''(\nu_i)\mathbb{E}\left[\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})\right]^T - a'(\nu_i) - \nu_i^T a''(\nu_i) + a'(\nu_i) && (19)\\ &= a''(\nu_i)(\mathbb{E}\left[\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})\right]^T - \nu_i^T) && (20) \end{aligned}$$

where starting from (17) to (18), we used the mean-field approximation that $u(z_i)$ and $\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})$ are independent. Notice here that, $\nu_i = \mathbb{E}\left[\boldsymbol{\eta}(\mathbf{z}_{-i}, \mathbf{x})\right]$ is our coordinate update and it does not depend on ν_i because of the mean-field approximation.