Quick Reflesher on Variational Inference and Mean Field Approximation

Shoichiro Yamanishi

April 3, 2020

1 Introduction

This is a quick refresher on variational inference and mean field approximation. The purpose of this document is as a self-contained document to enhance the course notes [3] by D. Blei with my own annotations to fill the gaps between deductions in order for my future-self to refresh this topic quickly without a pencil and paper.

The variational inference and the mean field approximation are explained in PRML[2] 10.1 and Barber[1] 28.3, 28.4, but I like the course notes [3] by D. Blei available online at Princeton is best, as it gives the flow of explanation from the problem setting down to the optimality of the factroized approximator for the exponential families. However, it is a bit too terse to me.

2 Variational Inference

Variational Inference is applicable for the following case.

- you have samples $\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}$, that are i.i.d.
- you want to model a probability distribution over $\mathbf{x}$ with latent variables and parameters like $\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}$.
- you want the posterior $p(\mathbf{z}|, \mathbf{x}, \boldsymbol{\theta})$ but it is intractable. Especially $p(\mathbf{z}|, \mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{z},\mathbf{x}|\boldsymbol{\theta})}{\int p(\mathbf{z},\mathbf{x}|\boldsymbol{\theta})d\mathbf{z}}$ and the denominator is intractable due to huge dimensionality etc.
- you want to approximate $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$ by some probability distribution $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})$. $\boldsymbol{\nu}$ is called variational parameter.

3 How to find $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})$? KL divergeance and ELBO

Consider the following.

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \ln(\int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z})$$

Assume $0 < q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) < \infty$.

$$\ln p(\mathbf{x}|\boldsymbol{\theta}) = \ln(\int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})} d\mathbf{z})$$

By Jensen's inequality ($\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$ for any concave function $f(x)$),

$$\begin{aligned}
\ln p(\mathbf{x}|\boldsymbol{\theta}) &\geq \int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) \ln(\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})} d\mathbf{z} = \mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta}) \\
&= \int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) \ln(\frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})} d\mathbf{z} \\
&= \int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) \ln p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{z} + \int q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) \ln(\frac{p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{q(\mathbf{z})}) d\mathbf{z} \\
&= \ln p(\mathbf{x}|\boldsymbol{\theta}) - D_{KL}(q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) || p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))
\end{aligned} \tag{1}$$

If ($q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$) then the KL divergence above vanishes, and hence the ELBO is maximal. Therfore for a fixed $\mathbf{x}_i$ and unknown $\boldsymbol{\theta}^*$,

$$\begin{aligned}
\underset{q(\mathbf{z})}{\operatorname{argmax}}(ELBO) &= \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta}^*) \\
\max_{q(\mathbf{z})}(ELBO) &= \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})}[\ln p(\mathbf{x}_i, \mathbf{z}|\boldsymbol{\theta}) - \ln p(\mathbf{z}|\mathbf{x}_i, \boldsymbol{\theta})] \\
&= \ln p(\mathbf{x}_i|\boldsymbol{\theta}^*)
\end{aligned} \tag{2}$$

$\mathcal{L}(q(\mathbf{z}), \boldsymbol{\theta})$ is ELBO (Evidence Lower Bound).

In the variational inference, we try to maximize ELBO. In the following argument we omit $\boldsymbol{\theta}$ as it is not relevant to the argument. Maximization of ELBO w.r.t $\boldsymbol{\theta}$ can be handled by the M-step of the variational EM algorithm.

4 Mean Field Approximation

One way to find a good $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu})$ to increase ELBO is to use mean field approximation. In the mean field approximation, you break some of the dependencies among $\mathbf{z}$ conditioned on $\mathbf{x}$, and factrize.

$$q(\mathbf{z}|\mathbf{x}, \boldsymbol{\nu}) = \prod_i q_i(\mathbf{z}_i) \tag{3}$$

where $\mathbf{z} = \cup \mathbf{z}_i$.

In the following argument we denote $\prod_{i \neq j} q_i(\mathbf{z}_i) (= \frac{\prod_i q_i(\mathbf{z}_i)}{q_j(\mathbf{z}_j)})$ by $q_{\setminus j}(\mathbf{z}_{\setminus j})$. Then we try to improve ELBO with each $q_i(\mathbf{z}_i)$ separately, fixing $q_{\setminus i}(\mathbf{z}_{\setminus i})$.

$$\begin{aligned} q_i^*(\mathbf{z}_i) &= \underset{q(\mathbf{z}_i)}{\operatorname{argmax}}(\mathcal{L}(q(\mathbf{z}_i), \boldsymbol{\theta})) \\ &= \underset{q(\mathbf{z}_i)}{\operatorname{argmax}}(-D_{KL}(q(\mathbf{z}_i)||p(\mathbf{z}, \mathbf{x}))) \\ &= \underset{q(\mathbf{z}_i)}{\operatorname{argmax}}(\int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln p(\mathbf{z}, \mathbf{x}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i - \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln q(\mathbf{z}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i) \end{aligned} \tag{4}$$

The first term inside the argmax in the RHS will be:

$$\begin{aligned} Term1 &= \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln p(\mathbf{z}_i, \mathbf{z}_{\setminus i}, \mathbf{x}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i \\ &= \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln p(\mathbf{z}_i, \mathbf{x} | \mathbf{z}_{\setminus i}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i - \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln p(\mathbf{z}_{\setminus i}, \mathbf{x}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i \\ &= \int q(\mathbf{z}_i) \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}_{\setminus i})} [\ln p(\mathbf{z}_i, \mathbf{x} | \mathbf{z}_{\setminus i})] d\mathbf{z}_i - const \end{aligned} \tag{5}$$

The second term will be:

$$\begin{aligned} Term2 &= - \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln(q(\mathbf{z}_i) q(\mathbf{z}_{\setminus i})) d\mathbf{z}_{\setminus i} d\mathbf{z}_i \\ &= - \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln q(\mathbf{z}_i) d\mathbf{z}_{\setminus i} d\mathbf{z}_i - \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln q(\mathbf{z}_{\setminus i})) d\mathbf{z}_{\setminus i} d\mathbf{z}_i \\ &= - \int q(\mathbf{z}_{\setminus i}) d\mathbf{z}_{\setminus i} \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i - \int q(\mathbf{z}_i) d\mathbf{z}_i \int q(\mathbf{z}_{\setminus i}) \ln q(\mathbf{z}_{\setminus i})) d\mathbf{z}_{\setminus i} \\ &= - \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i - \int q(\mathbf{z}_{\setminus i}) \ln q(\mathbf{z}_{\setminus i})) d\mathbf{z}_{\setminus i} \\ &= - \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i - const \end{aligned} \tag{6}$$

Therefore, equation 4 will be reorganized as follows:

$$q_i^*(\mathbf{z}_i) = \underset{q(\mathbf{z}_i)}{\operatorname{argmax}}(\int q(\mathbf{z}_i) \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}_{\setminus i})} [\ln p(\mathbf{z}_i, \mathbf{x} | \mathbf{z}_{\setminus i})] d\mathbf{z}_i - \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i) \tag{7}$$

Now we form a Lagrangian multiplier together with an equality constraint $\int q(\mathbf{z}_i) d\mathbf{z}_i = 1$.

$$\begin{aligned} \mathcal{L}(q(\mathbf{z}_i), \lambda) = &\int q(\mathbf{z}_i) \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}_{\setminus i})} [\ln p(\mathbf{z}_i, \mathbf{x} | \mathbf{z}_{\setminus i})] d\mathbf{z}_i \\ &- \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i - \lambda \left(\int q(\mathbf{z}_i) d\mathbf{z}_i - 1 \right) \end{aligned} \tag{8}$$

$$\begin{aligned}\frac{\partial \mathcal{L}(q(\mathbf{z}_i), \lambda)}{\partial q(\mathbf{z}_i)} &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}_{\setminus i})}[\ln p(\mathbf{z}_i, \mathbf{x}|\mathbf{z}_{\setminus i})] - \ln q(\mathbf{z}_i) - 1 - \lambda \\ &= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}_{\setminus i})}[\ln p(\mathbf{z}_i, \mathbf{x}|\mathbf{z}_{\setminus i})] - \ln q(\mathbf{z}_i) + const = 0\end{aligned} \tag{9}$$

Here we used $\frac{d}{dq}\ln q = \ln q + 1$. This suggests the following form.

$$q_i^*(\mathbf{z}_i) \propto exp\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}_{\setminus i})}[\ln p(\mathbf{z}_i, \mathbf{x}|\mathbf{z}_{\setminus i})]\right) \tag{10}$$

In fact, equation 5 can be expressed by:

$$\begin{aligned}Term1 &= \int q(\mathbf{z}_i) \int q(\mathbf{z}_{\setminus i}) \ln p(\mathbf{z}_i, \mathbf{z}_{\setminus i}, \mathbf{x}) d\mathbf{z}_{\setminus i} d\mathbf{z}_i \\ &= \int q(\mathbf{z}_i) \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}_{\setminus i})}[\ln p(\mathbf{z}_i, \mathbf{z}_{\setminus i}, \mathbf{x})] d\mathbf{z}_i\end{aligned} \tag{11}$$

and $q_i^*(\mathbf{z}_i)$ can be also expressed by:

$$q_i^*(\mathbf{z}_i) \propto exp\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}_{\setminus i})}[\ln p(\mathbf{z}_i, \mathbf{z}_{\setminus i}, \mathbf{x})]\right) \tag{12}$$

The equations 10 and 12 suggest the form (family) of the approximation $q_i^*(\mathbf{z}_i)$ is determined by the model joint distribution or conditional distribution. It also suggests if the model is in the exponential family, then not only the form but also the optimum approximation $q_i^*(\mathbf{z}_i)$ itself is determined.

References

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[3] David M. Blei. Variational inference. https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf. Accessed: 2020-03-30.