

# NOTES ON VARIATIONAL INFERENCE

*He Li, NYU Stern*

Last update: October 6, 2018

## 1 Background

Consider a joint distribution  $p(x, z)$  where  $z$  is the latent variables. In the scenario of Bayesian statistics,  $z$  usually represents the model parameters. Given the prior on  $z$ :  $p(z)$  and the likelihood function we obtain from data,  $p(x|z)$ , we can construct the posterior distribution on  $z$ :

$$p(z|x) \propto p(z)p(x|z) \quad (1.1)$$

where  $z \in \mathbb{R}^m, x \in \mathbb{R}^n$ . However, it is sometimes very hard to sample or do calculation from this posterior distribution. Therefore, we may need some technique to approximate the posterior. One way is Markov chain Monte Carlo (MCMC), it is statistically convincing but computationally expensive. Another way is using variational inference, a method that is much faster.

In variational inference, our goal is to find the best alternative distribution within a family of densities  $\mathcal{Z}$  that is closed to the posterior under KL divergence.

$$q^*(z) = \underset{q(z) \in \mathcal{Z}}{\operatorname{argmin}} \operatorname{KL}(q(z) \| p(z|x)) \quad (1.2)$$

## 2 The Evidence Lower Bound

The minimization task (1.2) in last section is not computable, since it contains the term  $\log p(x)$ . Let's see it.

$$\begin{aligned} \operatorname{KL}(q(z) \| p(z|x)) &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\ &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, x)] + \mathbb{E}_q[\log p(x)] \\ &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, x)] + \log p(x) \end{aligned} \quad (2.1)$$

We call this term  $\log p(x) = \log \int p(z, x) dz$  the *evidence* and in some cases this term needs exponential time to compute (the integral).

Instead of minimizing the (1.2), we define the *evidence lower bound (ELBO)* as

$$\begin{aligned} \operatorname{ELBO}(q) &:= \mathbb{E}_q[\log p(z, x)] - \mathbb{E}_q[\log q(z)] \\ &= \mathbb{E}_q \log p(z) + \mathbb{E}_q p(x|z) - \mathbb{E}_q \log q(z) \\ &= \mathbb{E}_q p(x|z) - \operatorname{KL}(q(z) \| p(z)) \end{aligned} \quad (2.2)$$

Notice that

$$\text{ELBO}(q) = -\text{KL}(q(z)||p(z|x)) + \log p(x) \quad (2.3)$$

Obviously, maximizing ELBO is equivalent to minimizing the KL divergence. Given the fact that KL divergence is non-negative, we observe that

$$\text{ELBO}(q) \leq \log p(x) \quad (2.4)$$

which indicates the name of ELBO. This can also be derived from below,

$$\begin{aligned} \log p(x) &= \log \int p(x, z) \, dz \\ &= \log \int \frac{p(x|z)p(z)q(z)}{q(z)} \, dz \\ &= \log \mathbb{E}_q \left[ \frac{p(x|z)p(z)}{q(z)} \right] \\ &\geq \mathbb{E}_q \log \left[ \frac{p(x|z)p(z)}{q(z)} \right] \quad \text{Jensen's Inequality} \\ &= \mathbb{E}_q \log p(x|z) + \mathbb{E}_q \log p(z) - \mathbb{E}_q \log q(z) \\ &= \text{ELBO}(q) \end{aligned} \quad (2.5)$$

### 3 The Mean-field Variational Family and CAVI

We now focus on the mean-field variational family, where the latent variables are mutually independent and each governed by a distinct factor in the variational density, e.g.

$$q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j) \quad (3.1)$$

Regarding this mean-field variational family, we now introduce the most commonly used algorithm to solve this optimization problem: Coordinate ascent variational inference (CAVI). The CAVI is based on the following observation: consider a member  $q_j(z_j)$ , fix all the other variational factors, the optimal  $q_j(z_j)$  is given by

$$q_j^*(z_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(z_j | \mathbf{z}_{-j}, x)] \} \quad (3.2)$$

To see this result, according to chain rule in the probability,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}) \prod_{k=1}^m p(z_k | \mathbf{z}_{1:k-1}, \mathbf{x}) \quad (3.3)$$

For any given  $z_j$ , consider it as the last member in the product, therefore the terms that are related to  $q_j(z_j)$  in ELBO is

$$\begin{aligned} & \mathbb{E}_q[\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})] - \mathbb{E}_q[\log q_j(z_j)] \\ &= \int q_j(z_j) \mathbb{E}_{q_{-j}}[\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})] dz_j - \int q_j(z_j) \log q_j(z_j) dz_j \end{aligned} \quad (3.4)$$

To calculate the derivative of  $q_j(z_j)$  on ELBO, we first need to define the functional derivative.

**Definition 3.1.** Given a manifold  $M$  representing (continuous/smooth) functions  $\rho$  (with certain boundary conditions etc.), and a functional  $F$  defined as

$$F : M \rightarrow \mathbb{R}$$

the functional derivative of  $F[\rho]$ , denoted  $\frac{\delta F}{\delta \rho}$ , is defined by

$$\int \frac{\delta F}{\delta \rho} \phi(x) dx = \lim_{\epsilon \rightarrow 0} \frac{F[\rho + \epsilon \phi] - F(\rho)}{\epsilon} = \left[ \frac{d}{d\epsilon} F[\rho + \epsilon \phi] \right]_{\epsilon=0} \quad (3.5)$$

where  $\phi$  is any nice function and  $\phi = 0$  on the boundary of the region of integration. The quantity  $\epsilon \phi$  is called the variation of  $\rho$ . In other words,

$$\phi \rightarrow \left[ \frac{d}{d\epsilon} F[\rho + \epsilon \phi] \right]_{\epsilon=0} \quad (3.6)$$

is a linear functional, so by the **Riesz–Markov–Kakutani representation theorem**, this functional is given by integration against some measure. Then  $\frac{\delta F}{\delta \rho}$  is defined to be the **Radon–Nikodym derivative** of this measure.

Given a functional

$$F[\rho] = \int f(x, \rho(x), \nabla \rho(x)) dx \quad (3.7)$$

and any function  $\phi$ , the functional derivative of  $F[\rho]$  is,

$$\begin{aligned} \int \frac{\delta F}{\delta \rho} \phi(x) dx &= \left[ \frac{d}{d\epsilon} \int f(x, \rho(x) + \epsilon \phi(x), \nabla \rho(x) + \epsilon \nabla \phi(x)) dx \right]_{\epsilon=0} \\ &= \int \left( \frac{\partial f}{\partial \rho} \phi + \frac{\partial f}{\partial \nabla \rho} \nabla \phi \right) dx \\ &= \int \left[ \frac{\partial f}{\partial \rho} \phi + \nabla \cdot \left( \frac{\partial f}{\partial \nabla \rho} \phi \right) - \left( \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \right) \phi \right] dx \\ &= \int \left[ \frac{\partial f}{\partial \rho} \phi - \left( \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \right) \phi \right] dx + \oint_S \left( \frac{\partial f}{\partial \nabla \rho} \phi \right) dx, \quad \text{Gauss Theorem} \\ &= \int \left[ \frac{\partial f}{\partial \rho} \phi - \left( \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \right) \phi \right] dx, \quad \phi = 0 \text{ on } S \\ &= \int \left[ \frac{\partial f}{\partial \rho} - \left( \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \right) \right] \phi dx \end{aligned} \quad (3.8)$$

Therefore, we have

$$\int \left[ \frac{\delta F}{\delta \rho} - \frac{\partial f}{\partial \rho} + \left( \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \right) \right] \phi \, dx = 0 \quad (3.9)$$

According to the fundamental lemma of calculus of variations below, we have

$$\frac{\delta F}{\delta \rho} = \frac{\partial f}{\partial \rho} - \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \quad (3.10)$$

which is called *Euler–Lagrange equation*. More properties of functional derivatives can be found [here](#).

**Theorem 3.2. Fundamental lemma of calculus of variations.** If a continuous multivariable function  $f$  on an open set  $\Omega \subset \mathbb{R}^d$  satisfies the equality

$$\int_{\Omega} f(x) h(x) \, dx = 0 \quad (3.11)$$

for all compactly supported smooth functions  $h$  on  $\Omega$ , then  $f \equiv 0$  on  $\Omega$ .

Similarly, one may consider a continuous function  $f$  on the closure of  $\Omega$ , assuming that  $h$  vanishes on the boundary of  $\Omega$  (rather than compactly supported).

Also, for discontinuous multivariable functions, Let  $\Omega \subset \mathbb{R}^d$  be an open set, and  $f \in L^2(\Omega)$  satisfy the equality

$$\int_{\Omega} f(x) h(x) \, dx = 0 \quad (3.12)$$

for all compactly supported smooth functions  $h$  on  $\Omega$ . Then  $f \equiv 0$ .

Now let's head back to (3.2), denote  $\text{ELBO}(q) = \mathcal{L}$ , we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial q_j(z_j)} &= \frac{\partial}{\partial q_j(z_j)} \left[ \int q_j(z_j) \mathbb{E}_{q_{-j}} [\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})] \, dz_j - \int q_j(z_j) \log q_j(z_j) \, dz_j \right] \\ &= \mathbb{E}_{q_{-j}} [\log p(z_j | \mathbf{z}_{-j}, \mathbf{x})] - \log q_j(z_j) - 1 \end{aligned} \quad (3.13)$$

where we use the Euler–Lagrange equation (3.10) (note that this functional does not have  $\nabla \rho$  term).

Therefore, the optimal  $q_j(z_j)$  is,

$$q_j^*(z_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(z_j | \mathbf{z}_{-j}, x)] \} \quad (3.14)$$

which is equivalent to

$$q_j^*(z_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(z_j, \mathbf{z}_{-j}, x)] \} \quad (3.15)$$

---

**Algorithm 1:** Coordinate Ascent Variational Inference (CAVI)

---

**Data:** A model  $p(\mathbf{x}, \mathbf{z})$ , a data set  $\mathbf{x}$   
**Result:** A variational density  $q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j)$   
**Initialize:** Variational factors  $q_j(z_j)$ ;  
**while** the ELBO has not converged **do**  
    **for**  $j \in 1, \dots, m$  **do**  
        Set  $q_j^*(z_j) \propto \exp \{ \mathbb{E}_{-j} [\log p(z_j | \mathbf{z}_{-j}, x)] \}$ ;  
    **end**  
**end**

---

## 4 Black Box Variational Inference

### 4.1 Stochastic Optimization

Under Mean-field assumptions, consider the ELBO

$$\mathcal{L}(\lambda) \triangleq \mathbb{E}_{q_\lambda(z)} [\log p(x, z) - \log q(z)] \quad (4.1)$$

Let's consider **stochastic optimization** method. Define  $f(x)$  to be the target function we are going to maximize, and  $h(t)$  be the realization of a random variable whose expectation is the gradient of  $f(x)$ . Then, we can optimize the target iterally with

$$x_{t+1} \leftarrow x_t + \rho_t h_t(x) \quad (4.2)$$

where  $\rho_t$  is the learning rate. This converges to a maximum of  $f(x)$  when the learning rate schedule follows the Robbins-Monro conditions

$$\sum_{t=1}^{\infty} \rho_t = \infty, \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty \quad (4.3)$$

As derived in the appendix of [2], the gradient of ELBO can be written as

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_q [\nabla_\lambda \log q(z|\lambda) (\log p(x, z) - \log q(z|\lambda))] \quad (4.4)$$

Therefore, we can use Monte-Carlo to do the stochastic optimization with

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^S \nabla_\lambda \log q(z_s|\lambda) (\log p(x, z_s) - \log q(z_s|\lambda)) \quad (4.5)$$

where  $z_s \sim q(z|\lambda)$ .

### 4.2 Rao-Blackwellization

In practice, people usually use the **Rao-Blackwellization** method to reduce the variance of Monte-Carlo sampling. If we want to compute the expectation of a function  $\mathbb{E}f(x, y)$ , Rao-Blackwellization can be simplified as using conditional expectation  $\hat{f}(x, y) = \mathbb{E}(f(x, y)|x)$  as an estimator, given the fact that

$$\mathbb{E}[\mathbb{E}(f(x, y)|x)] = \mathbb{E}(f(x, y)) \quad (4.1)$$

and

$$\text{Var}(f(x, y)) = \mathbb{E}(\text{Var}(f(x, y)|x)) + \text{Var}(\mathbb{E}(f(x, y)|x)) \geq \text{Var}(\mathbb{E}(f(x, y)|x)) \quad (4.2)$$

This means that  $\hat{f}(x, y)$  is a lower variance estimator than  $f(x, y)$ . Due to the mean-field assumption,

$$\begin{aligned} \mathbb{E}(f(x, Y)|x) &= \int \frac{f(x, Y)p(x, Y)}{p(x)} dY \\ &= \int \frac{f(x, Y)p(x)p(Y)}{p(x)} dY \\ &= \int f(x, Y)p(Y) dY \\ &= \mathbb{E}_Y f(x, Y) \end{aligned} \quad (4.3)$$

Therefore, for each component of the gradient, we should compute expectations with respect to the other factors. Derived from the supplement material of [2], we conclude that

$$\nabla_{\lambda_i} \mathcal{L} = \mathbb{E}_{q(i)}[\nabla_{\lambda_i} \log q(z_i|\lambda_i)(\log p_i(x, z_{(i)}) - \log q(z_i|\lambda_i))] \quad (4.4)$$

where  $q(i)$  be the distribution of variables in the model that depend on the  $i$ th variable, i.e., the Markov blanket of  $z_i$ ; and  $p_i(x, z_{(i)})$  be the terms in the joint that depend on those variables.

### 4.3 Control Variables

A control variate is a family of functions with equivalent expectation. Consider a function  $h$ , which has a finite first moment, and a scalar  $a$ . Define  $\hat{f}$  to be

$$\hat{f}(z) \triangleq f(z) - a[h(z) - \mathbb{E}(h(z))] \quad (4.1)$$

Therefore, we can choose  $a$  to minimize the variance of  $\hat{f}$ :

$$\text{Var}(\hat{f}) = \text{Var}(f) + a^2 \text{Var}(h) - 2a \text{Cov}(f, h) \quad (4.2)$$

where the minimal is obtained at  $a^* = \frac{\text{Cov}(f, h)}{\text{Var}(h)}$ . Notice that when  $a = 0$ ,  $\hat{f} = f$ , therefore the minimal variance is no greater than  $f$ 's variance.

Back to our algorithm, we can define

$$\begin{aligned} f_i(z) &= \nabla_{\lambda_i} \log q(z_i|\lambda_i)(\log p_i(x, z) - \log q_i(z|\lambda_i)) \\ h_i(z) &= \nabla_{\lambda_i} \log q(z_i|\lambda_i) \end{aligned} \quad (4.3)$$

Note that

$$\begin{aligned}
\mathbb{E}(\nabla_{\lambda_i} \log q(z_i|\lambda_i)) &= \mathbb{E}\left(\frac{\nabla_{\lambda_i} q(z_i|\lambda_i)}{q(z_i|\lambda_i)}\right) \\
&= \int \nabla_{\lambda_i} q(z_i|\lambda_i) \mathrm{d}z_i \\
&= \nabla_{\lambda_i} \int q(z_i|\lambda_i) \mathrm{d}z_i \\
&= \nabla_{\lambda_i} 1 = 0
\end{aligned} \tag{4.4}$$

Therefore  $\mathbb{E}(\nabla_{\lambda_i} \log q(z|\lambda_i)) = 0$ . Moreover, we can use sampling value to estimate  $\hat{a}_i^* = \frac{\hat{\text{Cov}}(f_i, h_i)}{\hat{\text{Var}}(h_i)}$ . Finally, we can use the estimated gradient:

$$\nabla_{\lambda_i} \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^S \nabla_{\lambda_i} \log q_i(z_s|\lambda_i) (\log p_i(x, z_s) - \log q_i(z_s|\lambda_i) - \hat{a}_i^*), \quad z_s \sim q_{(i)}(z|\lambda) \tag{4.5}$$

The logic of black box variational inference is as follows:

*find a maximum value  $\implies$  use stochastic optimization  $\implies$  monte-carlo sampling to estimate gradient  $\implies$  use Rao-Blackwellization and control variables to lower the sampling variance.*

## References

- [1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- [2] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822, 2014.