NOTES ON VARTIATIONAL INFERENCE

He Li, NYU Stern

Last update: September 20, 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:

$$p(z|x) \propto p(z)p(x|z) \tag{1.1}$$

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.

$$q*(z) = \operatorname*{argmin}_{q(z)\in\mathcal{Z}} \mathrm{KL}\left(q(z)\|p(z|x)\right) \tag{1.2}$$

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.

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

We call this term $\log p(x) = \log \int p(z,x)\,\mathrm{d}z$ 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

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

Notice that

$$\mathrm{ELBO}(q) = -\mathrm{KL}\left(q(z)\|p(z|x)\right) + \log p(x) \tag{2.3}$$

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

$$\mathrm{ELBO}(q) \leq \log p(x) \tag{2.4}$$

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

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

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.

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

Regarding this mean-field variational family, we now introduce the most commonly used alogrithm 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

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

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

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

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

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

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 ρ (with certain boundary conditions etc.), and a functional F defined as

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

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

$$
\int \frac{\delta F}{\delta \rho}\phi(x)\,\mathrm{d}x = \lim_{\epsilon \to 0} \frac{F[\rho + \epsilon\phi] - F(\rho)}{\epsilon} = \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}F[\rho + \epsilon\phi]\right]_{\epsilon=0} \tag{3.5}
$$

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

$$
\phi \to \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}F[\rho + \epsilon\phi]\right]_{\epsilon=0} \tag{3.6}
$$

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

$$
F[\rho] = \int f(x, \rho(x), \nabla\rho(x))\,\mathrm{d}x \tag{3.7}
$$

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

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

Therefore, we have

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

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

$$\frac{\delta F}{\delta \rho} = \frac{\partial f}{\partial \rho} - \nabla \cdot \frac{\partial f}{\partial \nabla \rho} \tag{3.10}$$

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

$$\int_\Omega f(x) h(x) \, \mathrm{d}x = 0 \tag{3.11}$$

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

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

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

$$\int_\Omega f(x) h(x) \, \mathrm{d}x = 0 \tag{3.12}$$

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

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

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

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,

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

which is equivalent to

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