A short note on Variational Inference

Tong Zhou

November 24, 2020

In this note, the emphasis will be placed on the topic of Bayes learning. That is, how a posterior density is estimated and its pros and cons with the MCMC method.

Before plunging into the Bayesian setting, let's first review two basic rules in probability theory: *Product Rule* and *Sum Rule*:

- **Product Rule:** $p(x, y, z) = p(x|y, z)p(y|z)p(z)$.
- **Sum Rule:** $p(y) = \int p(x, y)\,\mathrm{d}x$.

Note that the two rules apply to any probability distributions. They are useful in that when their roles are not symmetric, such manipulations can convert intractable quantities to tractable ones. One such application is the *Bayes Rule:*

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)\,\mathrm{d}y}$$

In the Bayes setting, this relation can be reiterated as:

$$\text{Posterior} = \frac{\text{Likelihood}\times\text{Prior}}{\text{Evidence}}$$

Or more specifically,

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}$$

where the prior is $\pi(\theta)$, i.e. $\theta \sim \pi(\theta)$, and $p(x|\theta)$ is the likelihood, i.e. $X|\theta \sim p(x|\theta)$.

The evidence $p(x) = \int p(x|\theta)\pi(\theta)\,\mathrm{d}\theta$ is usually intractable.

Since the numerator is easy to compute, MCMC can be used to draw a Markov chain from $p(x|\theta)\pi(\theta)$. Then a equivalent random sample extracted from the Markov chain can be used to approximate the posterior density $\pi(\theta|x)$. There are various sampling algorithms to draw such Markov chains, i.e. Metropolis-Hasting algorithm, Gibbs sampling, Hybrid Monte Carlo and NUTS (No-U-Turn-Sampling) and etc.

Another idea to approximate the posterior density is via VI. Instead of using a random sample by MCMC, the VI replaces an intractable posterior density with a tractable density belonging to a known class of probability densities. In other words, we hope to seek a density $q(\theta) \in \mathcal{Q}$, such that $\pi(\theta|x) \approx q(\theta)$, where $\mathcal{Q}$ is some known class of densities and easily computable.

A few questions need addressing:

1. *How to measure the distance between $\pi(\theta|x)$ and $q(\theta)$.* We use the Kullback-Leibler divergence, i.e. our goal is

$$\min_{q(\theta)\in Q} \mathsf{KL}\left(q(\theta)\,||\,\pi(\theta|x)\right) = \min_{q(\theta)\in Q} \int q(\theta)\log\frac{q(\theta)}{\pi(\theta|x)}\,\mathrm{d}\theta$$

2. *How to evaluate the KL divergence ?* Since the target still involves the unknown density $\pi(\theta|x)$, the objective function is still infeasible to optimize.

   We use the following observations:

$$\begin{aligned}
\log p(x) &= \int q(\theta)\log p(x)\,\mathrm{d}\theta = \int q(\theta)\log\frac{p(x,\theta)}{\pi(\theta|x)}\,\mathrm{d}\theta \\
&= \int q(\theta)\log\frac{p(x,\theta)q(\theta)}{\pi(\theta|x)q(\theta)}\,\mathrm{d}\theta \\
&= \int q(\theta)\log\frac{p(x,\theta)}{q(\theta)}\,\mathrm{d}\theta + \int q(\theta)\log\frac{q(\theta)}{\pi(\theta|x)}\,\mathrm{d}\theta \\
&= \underbrace{\mathcal{L}(q(\theta))}_{\text{Evidence Lower Bound}} + \underbrace{\mathsf{KL}\left(q(\theta)\,||\,\pi(\theta|x)\right)}_{\text{KL-divergence}}
\end{aligned}$$

   where $\mathcal{L}(q(\theta)) = \mathbb{E}_{\theta\sim q(\theta)}[\log p(x,\theta)] - \mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta)]$.

   In other words,

$$\log p(x) = \mathsf{Constant} = \mathcal{L}(q(\theta)) + \mathsf{KL}$$

   Using it, the minimization problem of the KL divergence is equivalent to maximize $\mathcal{L}(q(\theta))$, i.e.

$$\mathsf{KL}(q(\theta)\,||\,\pi(\theta|x)) \to \min_{q(\theta)\in Q} \quad\Longleftrightarrow\quad \mathcal{L}(q(\theta)) \to \max_{q(\theta)\in Q}$$

3. *How to compute ELBO?*

   Observe that

$$\begin{aligned}
\mathcal{L}(q(\theta)) &= \int q(\theta)\log\frac{p(x,\theta)}{q(\theta)}\,\mathrm{d}\theta = \int q(\theta)\log\frac{p(x|\theta)\pi(\theta)}{q(\theta)}\,\mathrm{d}\theta \\
&= \int q(\theta)\log p(x|\theta)\,\mathrm{d}\theta + \int q(\theta)\log\frac{\pi(\theta)}{q(\theta)}\,\mathrm{d}\theta \\
&= \underbrace{\mathbb{E}_{\theta\sim q(\theta)}[\log p(x|\theta)]}_{\text{Data term}} - \underbrace{\mathrm{KL}\left(q(\theta)\,||\,\pi(\theta)\right)}_{\text{Regularizer}}
\end{aligned}$$

4. *How to find $q(\theta)\in Q$? Δ_1, Σ* We have several methods:

   - **Mean field approximation:**
   - **Parametric approximation:**
   - **Stochastic VI:**

The intuition is that we seek $q(\theta)$ to maximize the log-likelihood while minimize the divergence between $q(\theta)$ and $\pi(\theta)$.

1 Relation to EM algorithm

The above discussion on VI can be modified so that a close relation to the EM algorithm will unfold that the EM algorithm can see seen as a special case of VI.

Suppose z is a hidden variable and x is observed. Then $p(x,z;\theta) = p(x|z;\theta)p(z;\theta)$. The goal is to recover the posterior density $p(z|x;\theta)$ that can be associated with the following relation:

$$p(z|x;\theta) = \frac{p(x,z;\theta)}{\int p(x,z;\theta)\,\mathrm{d}z}$$

VI then can be employed to use $q(z)$ to approximate $p(z|x;\theta)$.

Similar to above discussions about the generic VI model, the KL divergence and ELBO can be easily obtained:

$$\log p(x;\theta) = \int q(z)\log\frac{p(x,z;\theta)}{q(z)}\,\mathrm{d}z + \mathsf{KL}(q(z)\,||\,p(z|x;\theta))$$

and the ELBO can be written as:

$$\begin{aligned}\int q(z)\log\frac{p(x,z;\theta)}{q(z)}\,\mathrm{d}z &= \int q(z)\log p(x,z;\theta)\,\mathrm{d}z - \int q(z)\log q(z)\,\mathrm{d}z \\ &= \mathbb{E}_{q\sim q(\lambda)}[\log p(x,z;\theta)] + S(q(z))\end{aligned}$$

where $S(q(z)) = -\int q(z)\log q(z)\,\mathrm{d}z$ is called the *Gibbs entropy*.

Now we can see why EM is special case of VI model. Recall that in the *E-step*, given an initial value $\theta^{(s)}$, compute $\mathbb{E}\left[\log p(x,z;\theta)|x,\theta^{(s)}\right]$. Note that the expectation is taken over $z|x,\theta^{(s)}$. So when $p(z|x_i;\theta)$ is simple, the classical EM algorithm can be directly used. In such trivial case, substituting $q_i(z) = p(z|x_i;\theta)$ facilitates a VI algorithm.

The EM algorithm can be summarized as follows:

EM Algorithm

- **Step -1:** given $x_i, i = 1,\cdots,N$, both $p(x,z;\theta)$ and $p(z|x;\theta)$ have closed forms.
- **Step 0:** Given initial value $\theta^{(0)}$.
- **Step 1: E-step** In the t-th iteration, compute $p(z|x_i,\theta^{(t)})$.
- **Step 2: M-step** Obtain θ_{t+1} by

$$\theta_{t+1} = \arg\max_{\theta}\sum_{i=1}^{N}\int p(z|x_i;\theta^{(t)})\log p(x_i,z|\theta)\,\mathrm{d}z$$

slightly

However, if we substitute $q_i(z) = p(z|x_i;\theta)$, the VI algorithm will work as the following:

VI Algorithm

- **Step -1:** same.
- **Step 0:** same.
- **Step 1:** same.
- **Step 2:**

$$\theta_{t+1} = \arg\max_{\theta} \sum_{i=1}^{N} \left[\int p(z|x_i; \theta) \log p(x_i, z; \theta) - \int p(z|x_i; \theta) \log p(z|x_i; \theta) \, \mathrm{d}z \right].$$

Therefore, their difference lies in step-2, the EM-algorithm simplifies the computation by fixing $\theta = \theta^{(t)}$. This also circumvents computing the second therm $-\int p(z|x_i; \theta^{(t)} \log p(z|x_i; \theta^{(t)})) \, \mathrm{d}z$, because it is just a constant in the M-step and hence can be dropped. Therefore, it is not hard to see the EM algorithm can only attain local maximum.