Contents

1 VI: Variational Inference

TL;DR	
Use simple distributions that we know how to sample from to approximate complicated distributions	
Related Topics: *EM, VAE*	**Date:** *Apr 2nd, 2021*

In the typical Bayesian inference setting, we are interested in the posterior distribution

$$p_\theta(z \mid x) = \frac{p(x, z)}{p(x)}. \tag{1.1}$$

where $p_\theta(z \mid x)$ is the posterior distribution of latent variables given data x parametrized by θ; $p(x, z)$ is the joint distribution of x and z; $p(x)$ is the data distribution.

Usually, we won't able to derive an analytic solution to the posterior due to the intractability of data distribution $p(x)$. Variational inference (VI) is a set of algorithm that uses a known distribution $q_\phi(x)$ to approximate other unknown distribution $p(x)$ (this refers to a general distribution not the data distribution), and for Bayesian inference, we are approximating the posterior distribution $p_\theta(z \mid x)$ given data x.

1.1 Measuring the Difference between Distributions

By saying we are approximating a target p with a source q, we need to evaluate how good a given approximation is. The Kullback-Leibler divergence, or KL divergence for short, is a common choice for measuring the "distance" between two distributions. The KL divergence is defined as:

$$D_{KL}(p(x) \parallel q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{x \sim p} \log \frac{p(x)}{q(x)}. \tag{1.2}$$

KL divergence has a few decent properties:

- $D_{KL}(p \parallel q) \geq 0$.
- $D_{KL}(p \parallel q) = 0$ iff p and q are exactly the same.

KL divergence is called a divergence rather than a distance because KL divergence doesn't satisfy the symmetry property of a distance metric, i.e., $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$ in general. Since they are different, which one shall we use in the VI to evaluate our approximation? To answer this question, let's take a close look at these two types of KL divergence in Table 1.1.

	$D_{KL}(p \parallel q)$	$D_{KL}(q \parallel p)$
$\mathbb{E}$	$\mathbb{E}_{x\sim p} \log \frac{p(x)}{q(x)}$	$\mathbb{E}_{x\sim q} \log \frac{q(x)}{p(x)}$
Naming	forward KL moment projection	reverse KL information projection
Sampling	$x \sim p$	$x \sim q$
Value $p(x) >> q(x)$	$+\infty$	$-\infty$
Value $p(x) \approx q(x)$	≈ 0	≈ 0
Value $p(x) << q(x)$	$-\infty$	$+\infty$

Table 1.1: Comparison between $D_{KL}(p \parallel q)$ and $D_{KL}(q \parallel p)$

We observe from its values that forward and reverse KL behaves differently on approximation overshoots and undershoots. Forward KL has better sensitive when $p(x) >> q(x)$ and reverse KL penalizes when $q(x) >> p(x)$. What if we average over them? This in turns leads to the Jensen-Shannon divergence, we will not introduce it for now.

The major difference between such two KL divergences is on the sampling perspective where forward KL requires sampling from p and reverse KL requires sampling from q. In the typical VI setting, we usually don't have access to the PDF of p, or at least, it is not normalized. Therefore, sampling from p is generally not plausible. While, we can freely choose a variational distribution q_ϕ such that sampling from it is simple and straightforward, for example, a fully factorized Gaussian distribution. Therefore, we will use the reverse KL divergence $D_{KL}(q \parallel p)$ in VI.

1.2 VI: The Main Algorithm

We are now clear that VI finds the optimal parameters ϕ of the variational distribution $q_\phi(z)$ by minimizing the reverse KL divergence $D_{KL}(q \parallel p)$. Let's derive the VI algorithm for Bayesian inference, in other words, approximating the posterior distribution.

$$\phi^* = \arg\min_\phi D_{KL}(q_\phi(z) \parallel p_\theta(z \mid x)) \tag{1.3}$$

$$= \arg\min_\phi \mathbb{E}_{z\sim q} \log \frac{q_\phi(z)}{p_\theta(z \mid x)} \tag{1.4}$$

$$\mathbb{E}_{z\sim q} \log \frac{q_\phi(z)}{p_\theta(z \mid x)} = \mathbb{E}_{z\sim q} \left[\log q_\phi(z) - \log p_\theta(z \mid x)\right] \tag{1.5}$$

$$= \mathbb{E}_{z\sim q} \left[\log q_\phi(z) - \log p_\theta(z \mid x)\right] \tag{1.6}$$

$$= \underbrace{\mathbb{E}_{z\sim q} \left[\log q_\phi(z) - \log p_\theta(x, z)\right]}_{-\text{Evidence Lower BOund } (-\text{ELBO})} + \underbrace{\log p(x)}_{\text{evidence}}. \tag{1.7}$$

By Eqn (1.7), we know that the KL divergence between the variational distribution and the posterior can be re-written as the sum of negative evidence lower bound (ELBO) and evidence. In the meantime, given data x, $p(x)$ is a constant. Therefore, minimizing the KL divengence is equivalent to minimizing the negative ELBO. So, let's further simplify the ELBO.

$$
\begin{aligned}
-\text{ELBO} &\triangleq \mathcal{J}(q) && (1.8)\\
&= \mathbb{E}_{z\sim q}\left[\log q_\phi(z) - \log p_\theta(x,z)\right] && (1.9)\\
&= \mathbb{E}_{z\sim q}\log q_\phi(z) - \mathbb{E}_{z\sim q}\left[\log q_\phi(z) - \log p_\theta(x,z)\right] && (1.10)\\
&= \underbrace{\mathbb{E}_{z\sim q}\log q_\phi(z)}_{\text{variational}} - \underbrace{\mathbb{E}_{z\sim q}\log p(x \mid z)}_{\text{likelihood}} - \underbrace{\mathbb{E}_{z\sim q}\log p(z)}_{\text{prior}}. && (1.11)
\end{aligned}
$$

1.3 Summary

The variation inference finds the set of parameter ϕ that minimizes the reverse KL divergence of variational distribution and the target distribution. This is equivalent to minimize negative ELBO, i.e.,

$$
\phi^* = \arg\min_\phi \mathbb{E}_{z\sim q}\log q_\phi(z) - \mathbb{E}_{z\sim q}\log p(x \mid z) - \mathbb{E}_{z\sim q}\log p(z). \tag{1.12}
$$

All of these terms in Eqn (1.12) can be easily evaluated from the variational distribution and the latent variable. In the expectation-maximization (EM) algorithm, we will visit ELBO again in the optimization process.