Variational Inference

Zhuoyuan Chen*
Facebook AI Research
Menlo Park, CA 94025
chenzhuoyuan07@gmail.com

Abstract

VI, VB, EM Summary

1 Summary

Given θ as parameter, x observed, z latent variables, we have

$$\begin{aligned} l(\theta; D) = \log p(x) &= \log \sum_z p(z|\theta) p(x|z, \theta) && (1) \\ &= \log \sum_z p(z|\theta) p(x|z, \theta) \frac{q(z)}{q(z)} && (2) \\ &= \log \mathbb{E}_z \frac{p(x|z)p(z)}{q(z)} && (3) \end{aligned}$$

According to Jensen's Inequality (log is concave), we have

$$\begin{aligned} l(\theta; D) &= \log p(z) = \log \sum_z q(z) \frac{p(x|\theta)}{q(z)} && (4) \\ &\geq \sum_z q(z) \log \frac{p(x|\theta)}{q(z)} && (5) \\ &= \mathbb{E}_{z \sim q(z)} [\log p(x|z) + \log p(z)] - H(q) = \mathbb{E}_q \log p(x|z) - KL(q||p) && (6) \end{aligned}$$

The evidence lower-bound (**ELBO**) is called free energy. The equality satisfies when $q(z|x) = p(z|x, \theta)$. The difference between the gap is:

$$\log p(x) - ELBO = KL(q(z), p(z|x)) \tag{7}$$

1.1 Application 1: GMM

$$\log p(\theta; D) \geq \sum_z q(z|x) \log \frac{p(x|\theta)}{q(z|x)} \tag{8}$$

E-step:

$$q^{t+1} = \arg\max_q F(q, \theta^t) \tag{9}$$

M-step:

$$\theta^{t+1} = \arg\max_\theta F(q^{t+1}, \theta) \tag{10}$$

*

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

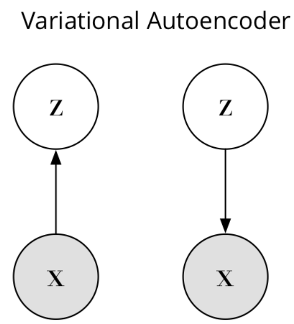

Figure 1: VAE.

1.2 Application 2: VAE

Graphical model can be shown in Figure 1. Traditional VI process:

1. Calculate $\nabla_\theta L_i(p, q_i)$ by
   (a) Sample $z \sim q_i(x_i)$
   (b) $\nabla_\theta L_i(p, q_i) \approx \nabla_\theta \log p_\theta(x_i|z)$
2. $\theta \leftarrow \theta + \alpha \nabla_\theta L_i(p, q_i)$
3. update q_i to maximize $L_i(p, q_i)$

Each q_i is different. So, the number of parameters is $|\theta| + (|\mu_i| + |\sigma_i|) \times N$, and step 3 is intractable. Using a neural network for $q(z_i) = q_\phi(x_i)$ makes the number of parameters independent of sample points. (This step is called **Amortized Variational Inference**). Then step 3 becomes:

$$\phi \leftarrow \phi + \alpha \nabla_\phi L_i(p, q_i)$$

Learn by PG versus reparametrization trick:

$$J(\phi) \approx \frac{1}{M} \nabla_\phi \log q_\phi(z|x) r(x_i, z_i) \tag{11}$$

$$\frac{1}{M} \nabla_\phi r(x_i, \mu_i + \sigma_i * \epsilon) \tag{12}$$

The second one has lower variance since it makes use of derivative of $r(x, z)$.

Encoder: $z = q(\phi, x)$, decoder: $x = p(\theta, z)$. We have MLE:

$$KL(q(z|x), p(z|x)) = E_{q(z|x)} \log q(z|x) - E_{q(z|x)}(\log P(x|z) + \log p(z) - \log p(x)) \tag{13}$$

$$E_{q(z|x)} \log p(x) = E_q \log p(x|z) - KL(q(z|x), p(z)) + KL(q(z|x), p(z|x)) \tag{14}$$

$$= ELBO + KL(q(z|x), p(z|x)) \tag{15}$$

and loss to optimize is:

$$l(\phi, \theta) = -E_{x \sim q(x|z)} \log p(x|z) + KL(q(z|x), p(z)) \tag{16}$$

1.3 Semi-Supervised VAE

Graphical model can be shown in Figure 2.

1. Label y is known:
   $\log p_\theta(x, y) \geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\phi(z|x, y)] = -L(x, y)$
2. Label y is unknown:

$$\log p_\theta(x) = \log \sum_y \int_z q(z, y|x) \frac{p(x, y, z)}{q(z, y|x)} dz \geq \sum_y q(y|x) \int_z q(z|x, y) \log \frac{p(x, y, z)}{q(z, y|x)} dz$$

$$= \sum_y q(y|x)(-L(x, y)) + H(q(y|x))$$

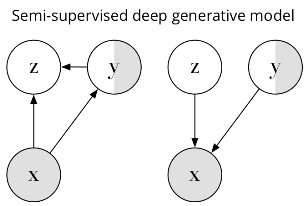

Figure 2: Semi-supervised VAE.

1.4 DVIB

Graphical model: Y - X - Z, with cost function:

$$\arg\max_{\theta} I(Z,Y;\theta) - \beta I(Z,X;\theta) \tag{17}$$

Then, the graphical model is:

$$p(X,Y,Z) = p(X)p(Z|X)p(Y|X)$$

1. Lower bound of $I(Z;Y)$, with approximation $q_1(y|z)$:
$$I(Z,Y) \geq \int p(y,z)\log\frac{q(y|z)}{p(y)}dydz = \int p(y,z)\log q_1(y|z)dydz + H(Y)$$
where we can drop $H(Y)$, using graphical model $p(x,y,z) = p(x)p(y|x)p(z|x)$, then we have:
$$I(Z,Y) \geq \int p(x)p(y|x)p(z|x)\log q_1(y|z)dxdydz$$
2. Upper bound of $I(X;Y)$, with approximation $q_2(z)$:
$$I(Z,X) = \int p(x,z)\log\frac{p(z|x)}{p(z)}dzdx = \int p(x,z)\log p(z|x)dzdx - \int p(x,z)\log p(z)dzdx$$
Then, we have
$$I(Z,X) \leq \int p(x)p(z|x)\log\frac{p(z|x)}{q_2(z)}$$

2 Fisher Information Matrix, Natural Gradient

2.1 KL-Divergence

$$\begin{aligned}
& KL(p_w(x), p_{w+\triangle w}(x)) \\
= \; & E_{x\sim p_w(x)}\log p_w(x) - \log p_{w+\triangle w}(x) \\
= \; & E_{x\sim p_w(x)}\{\log p_w(x) - [\log p_w(x) + \nabla_w \log p_w(x)\triangle w + \tfrac{1}{2}\triangle w^T \nabla_w^2 \log p_w(x)\triangle w)]\} \\
= \; & [E_{x\sim p_w(x)}\nabla_w \log p_w(x)]\triangle w - \tfrac{1}{2}\triangle w^T [E_{x\sim p(x)}\nabla_w^2 \log p_w(x)]\triangle w \\
= \; & \tfrac{1}{2}\triangle w[E_{x\sim p(x)}\nabla_w \log p_w(x)\nabla_w \log p_w(x)^T]\triangle w^T
\end{aligned}$$

where

$$\begin{aligned}
\nabla_w^2 \log p_w(x) &= \frac{\nabla_w^2 p_w(x)}{p_w(x)} - \frac{\nabla_w p_w(x)\nabla_w p_w(x)^T}{p_w^2(x)} \\
&= \frac{\nabla_w^2 p_w(x)}{p_w(x)} - \nabla_w \log p_w(x)\nabla_w \log p_w(x)^T
\end{aligned}$$

Also, we use the following property:

$$\begin{aligned}
E_{x\sim p_w(x)}\nabla_w \log p_w(x) &= \int_x p_w(x)\nabla_w \log p_w(x)dx = \int_x \nabla_w p_w(x)dx \\
&= \nabla_w(\int_x p_w(x)dx) = 0 \\
E_{x\sim p_w(x)}\nabla_w^2 \log p_w(x) &= 0
\end{aligned}$$

2.2 Fisher-Information Matrix

$$E_{x\sim p(x)}\nabla_w \log p_w(x)\nabla_w \log p_w(x)^T$$

3 Mutual Information

In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables:

$$I(x;y) := KL(p_{x,y}, p_x \otimes p_y) = h(x) - h(x|y) \geq 0 \tag{18}$$

Intuitively, mutual information measures the information that X and Y share: It measures how much knowing one of these variables reduces uncertainty about the other. Properties:

$$I(X;Y) \geq 0 \tag{19}$$
$$I(X;Y) = I(Y;X) \tag{20}$$
$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \tag{21}$$
$$H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X|Y) - H(Y|X) \tag{22}$$
$$I(X;Y) = \mathbb{E}_Y[KL(p_{x|y}, p_x)] \tag{23}$$