Variational Inference

David S. Hippocampus[*]
Department of Computer Science
Cranberry-Lemon University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Abstract

VI, VB, EM Summary

1 Summary

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

$$l(\theta; D) = \sum_z \log p(z|\theta)p(x|z,\theta) \tag{1}$$

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

$$l(\theta; D) = \log \sum_z q(z|x)\frac{p(x|\theta)}{q(z|x)} \tag{2}$$

$$\geq \sum_z q(z|x)\log\frac{p(x|\theta)}{q(z|x)} \tag{3}$$

The lower-bound is called free energy. The equality satisfies when $q(z|x) = p(z|x,\theta)$

1.1 Application 1: GMM

$$\log p(\theta; D) \geq \sum_z q(z|x)\log\frac{p(x|\theta)}{q(z|x)} \tag{4}$$

E-step:

$$q^{t+1} = \arg\max_q F(q, \theta^t) \tag{5}$$

M-step:

$$\theta^{t+1} = \arg\max_\theta F(q^{t+1}, \theta) \tag{6}$$

1.2 Application 2: VAE

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

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

$$E_{q(z|x)}\log p(x) = E_q \log p(x|z) - KL(q(z|x), p(z)) + KL(q(z|x), p(z|x)) \tag{8}$$

$$= ELBO + KL(q(z|x), p(z|x)) \tag{9}$$

and loss to optimize is:

$$l(\phi,\theta) = -E_{x\sim q(x|z)}\log p(x|z) + KL(q(z|x), p(z)) \tag{10}$$

[*]

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

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

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

2.3 Mutual Information

$$
I(x;y) := KL(p_{x,y}, p(x)\times p(y)) = h(x) - h(x|y) \geq 0 \tag{11}
$$