Deep Learning 2024011303 Liu Hanzuo

HW2

True or False

P1 False; Whether q_θ collapse to all multimodal depends on whether the KL divergence is inclusive or exclusive.

P2 True; Randomness of sampling has been moved to the randomness of ϵ in the equation of

$$z = \mu + \sigma \cdot \epsilon, \epsilon \sim \mathcal{N}(0, 1)$$

P3 False; we need a nueral network $q_\phi(z)$ to approximate the posterior distribution $p(z|x)$.

P4 False; a larger β indicated the variables to be more independent.

QA

P5 We begin by show a equation

Lemma 1

$$KL(q(z)||p(z|x)) = \log p(x) - \sum_z q(z) \log \frac{p(z,x)}{q(z)}$$

Proof.

$$KL(q(z)||p(z|x)) = \sum_z q(z) \log \frac{q(z)}{p(z|x)} = \sum_z q(z) \log \frac{q(z)p(x)}{p(z,x)}$$

$$= \sum_z q(z) \log p(x) - \sum_z q(z) \log \frac{p(z,x)}{q(z)} = \log p(x) - \sum_z q(z) \log \frac{p(z,x)}{q(z)}$$

□

Thus, we have

$$F(\theta, q) = \sum_z q(z) \log p_\theta(x,z) - \sum_z q(z) \log q(z)$$

$$= \sum_z q(z) \log \frac{p(z,x)}{q(z)} = \log p(x) - KL(q(z)||p(z|x))$$

To maximize q, it's equivalent to minimize $KL(q(z)||p(z|x))$, thus $q(z) \leftarrow p(z|x)$ is equivalent to the E-step.
For the M-step

$$\arg\max_\theta F(\theta, q^t) = \mathbb{E}_{z\sim p_\theta^t(z|x)} [\log p_\theta(x,z)] + H(p_\theta^t(x|z)) = \arg\max_\theta Q(\theta|\theta^t) + H(p_\theta^t(x|z))$$

Since when maximizing θ part, $H(p_\theta^t(x|z))$ is a constant, thus

$$\arg\max_\theta F(\theta, q^t) = \arg\max_\theta Q(\theta|\theta^t)$$

Thus we prove the equivalent of two updating policy.

P6 1.

Proof.

$$KL(N_0||N_1) = \mathbb{E}_{N_0}\left[\log \frac{N_0(x)}{N_1(x)}\right]$$

Now we write down two PDFs:

$$N_0(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_0|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)$$

$$N_1(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)$$

The log ratio could be written as:

$$\log \frac{N_0(x)}{N_1(x)} = \log \frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_0|^{\frac{1}{2}}} - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)$$

$$KL(N_0||N_1) = \mathbb{E}_{N_0}\left[\log \frac{N_0(x)}{N_1(x)}\right]$$

$$= \log \frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_0|^{\frac{1}{2}}} - \frac{1}{2}\mathbb{E}_{N_0}\left[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right] + \frac{1}{2}\mathbb{E}_{N_0}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right] \quad (1)$$

For calculation, we note that:

$$\mathbb{E}_{N_0}\left[(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right] = tr(\Sigma_0^{-1}\Sigma_0) = d \quad (2)$$

let $\delta = \mu_0 - \mu_1$

$$(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) = (x-\mu_0+\mu_0-\mu_1)^T\Sigma_1^{-1}(x-\mu_0+\mu_0-\mu_1)$$

$$= \mathbb{E}_{N_0}\left[(x-\mu_0)^T\Sigma_1^{-1}(x-\mu_0)\right] + \delta^T\Sigma_1^{-1}\delta = tr(\Sigma_1^{-1}\Sigma_0) + \delta^T\Sigma_1^{-1}\delta \tag{3}$$

Bring these euqations ((2),(3)) back to the KL divergence ((1)), we have:

$$KL(N_0||N_1) = \frac{1}{2}\left[\log\frac{|\Sigma_1|}{|\Sigma_0|} - d + tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)\right]$$

□

2.

Proof. We only need to show that $\forall p_1, p_2, q_1, q_2$, we have:

$$(\lambda p_1 + (1-\lambda)p_2)\log\frac{(\lambda p_1 + (1-\lambda)p_2)}{\lambda q_1 + (1-\lambda)q_2} \le \lambda p_1 \log\frac{p_1}{q_1} + (1-\lambda)p_2\log\frac{p_2}{q_2}$$

Then taking the interval over all x, we could get the initial inequality proved.

$$F(p_1,p_2,q_1,q_2) := \lambda p_1\log\frac{p_1}{q_1} + (1-\lambda)p_2\log\frac{p_2}{q_2} - (\lambda p_1 + (1-\lambda)p_2)\log\frac{(\lambda p_1 + (1-\lambda)p_2)}{\lambda q_1 + (1-\lambda)q_2}$$

Note that

$$\frac{\partial}{\partial q_1}F(p_1,p_2,q_1,q_2) = -\lambda p_1\frac{1}{q_1} + (\lambda p_1 + (1-\lambda)p_2)\frac{1}{\lambda q_1 + (1-\lambda)q_2}$$

From Lagrange multiplier, we have:

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} := k$$

(Notation: this equation could also obtained from considering the minimum point for a singular variable q_1)

At this assumption, we have that:

$$LHS = (\lambda p_1 + (1-\lambda)p_2)\log k = RHS$$

Thus we have proved the inequality. □

3. Here we show the exclusive and inclusive KL divergence example.

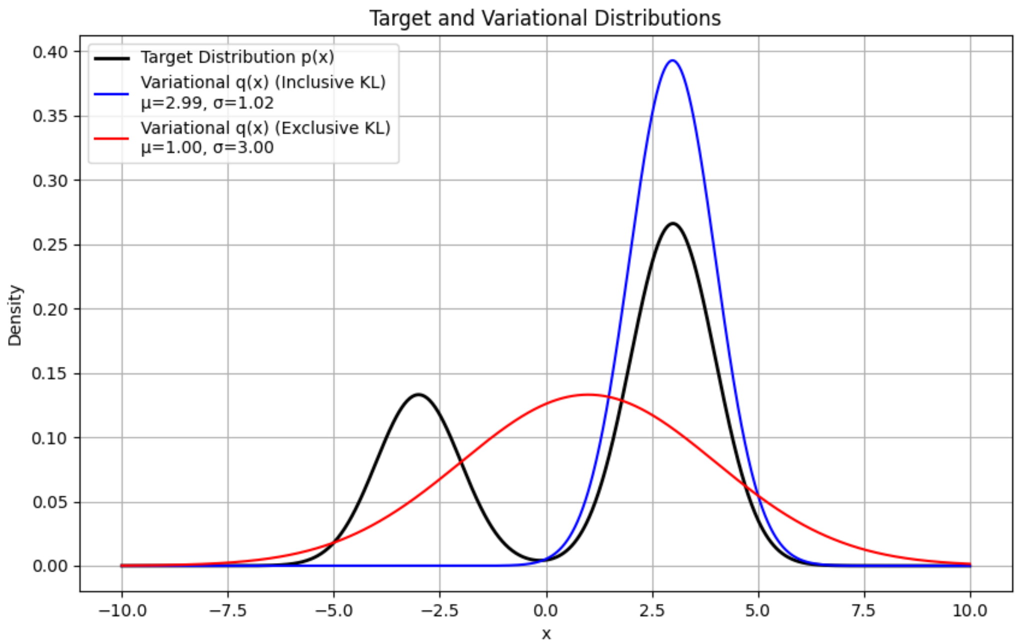

Figure 1: Figure 6-3: Exclusive and Inclusive KL divergence

4. The objective is:

$$\mathbb{E}_{z\sim p(z|x)}[\log \frac{p(z|x)}{q_\phi(z|x)}]$$

Note that though the optimal $q_\phi(z|x)$ is $p(z|x)$, there is pros and cons to take the inclusive KL divergence.

Pros:

1. $q_\phi(z|x)$ will tend to cover all possible areas that $p(z|x) > 0$, having a relative large penalty if $p(z|x) > 0$ but $q_\phi(z|x) = 0$. (Which indicate that $q_\phi(z|x)$ need to cover all modules and possible areas that $p(z|x) > 0$)

2. $q_\phi(z|x)$ will have a diverge result, cover larger areas

Cons:

1. $q_\phi(z|x)$ have a non-accurate probability distribution due to a diverge result, lower confidence over the high probability $p(z|x)$ result.

2. Hard to sample from the posterior distribution $z \sim p(z|x)$, larger variance for gradients.

3. Waste probability in low probability areas.

Theoretically, we could also gain the similar expression:

$$KL(p(z|x)||q_\phi(z)) = \sum_z p(z|x) \log \frac{p(z|x)}{q_\phi(z)} = \sum_z p(z|x) \log \frac{p(z,x)}{q_\phi(z)p(x)}$$

$$= -\sum_z p(z|x) \log \frac{q_\phi(z)}{p(z,x)} - \log p(x) = -\frac{1}{p(x)} \sum_z p(z,x) \log \frac{q_\phi(z)}{p(z,x)} - \log p(x)$$

Thus we have:

$$-\log p(x) = KL(p(z|x)||q_\phi(z)) + \frac{1}{p(x)} ELBO$$

While, does not have the beautiful formula that the exclusive KL divergence has.

P7 1.

$$\log p_{\mu,\sigma,\theta}(x) \geq \mathbb{E}_{w,z\sim q_{\psi,\phi}(w,z|x)} \left[\log \frac{p(w,z,x)}{q_{\psi,\phi}(w,z|x)} \right] = ELBO$$

Thus we have:

$$ELBO = \sum q_\psi(w|x) q_\phi(z|w,x) \log \frac{p(w,z,x)}{q_\psi(w|x) q_\phi(z|w,x)}$$

$$= \mathbb{E}_{q(w,z|x)} [\log p_\theta(x|w)] + \mathbb{E}_{q(w,z|x)} [\log p_{\mu,\sigma}(w|z)] + \mathbb{E}_{q(w,z|x)} [\log p(z)]$$

$$-\mathbb{E}_{q(w,z|x)} [\log q_\psi(w|x)] - \mathbb{E}_{q(w,z|x)} [\log q_\phi(z|w,x)]$$

$$= \mathbb{E}_{w\sim q_\psi(w|x)} [\log p_\theta(x|w)] - \mathbb{E}_{w\sim q_\psi(w|x)} [KL(q_\phi(z|w,x)\|p(z))] - KL(q_\psi(w|x)\|\mathbb{E}_{z\sim p(z)}[p_{\mu,\sigma}(w|z)])$$

2.

1. Here we calculate the result of different term in the loss function.

   $$\mathbb{E}_{q(w,z|x)} [\log p_\theta(x|w)] = \mathbb{E}_{q(w,z|x)} \left[-\log\det \sigma_\theta(w) + \frac{(x-\mu_\theta(w))^T * \sigma_\theta(w)^{-2}(x-\mu_\theta(w))}{2} \right] + Const$$

2. If we have the closed form easy-calculating $q(w,z|x)$ (such as Gaussian), we could use the reparameterization trick to calculate the expectation and expectation.

   $$\mathbb{E}_{w\sim q_\psi(w|x)} [KL(q_\phi(z|w,x)\|p(z))] = \mathbb{E}_{w\sim q_\psi(w|x)} \sum_{k=1}^{K} q_\phi(k|w,x) \log \frac{q_\phi(k|w,x)}{\pi_k}$$

   Could use reparameterization trick:

   $$x = \mu_\theta(w) + \sigma_\theta(w) \cdot \epsilon, \epsilon \sim \mathcal{N}(0,I)$$

   $$w = \mu_z + \sigma_z \cdot \epsilon, \epsilon \sim \mathcal{N}(0,I)$$

   to back propogate the gradients.

3. the third term could be written as:

   $$KL(q_\psi(w|x)\|\mathbb{E}_{z\sim p(z)}[p_{\mu,\sigma}(w|z)]) = \mathbb{E}_{w\sim q_\psi(w|x)} \left[\log q_\psi(w|x) - \log \mathbb{E}_{z\sim p(z)}[p_{\mu,\sigma}(w|z)] \right]$$

$$= \mathbb{E}_{w \sim q_\psi(w|x)} \left[\log q_\psi(w|x) - \log \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \right]$$

Could also be back-propogated by reparameterization trick!

To summarize, we show that all the terms in the loss function could be back-propogated by reparameterization trick. For the training procedure, we divide it into two parts:

- back-propogate the gradients of $\log p_{\mu,\sigma,\theta}(x|w)$ to μ, σ, θ
- back-propogate the gradients of $\log p_{\mu,\sigma,\theta}(x|w)$ to ψ, ϕ by setting $q(w, x|x) = q_\psi(w|x) \cdot q_\phi(z|w, x) \leftarrow p(w, z|x)$ use the loss of ELBO and back-propogate the gradients using reparameterization trick.