MFDNN

Homework 11

24/05/30

1.(a) $\log p_\theta(x) = \log\left(\mathbb{E}_{Z\sim p_z}\left[p_\theta(x|Z)\right]\right)$

$$= \log\left(\mathbb{E}_{Z\sim q_\phi(z|x)}\left[p_\theta(x|Z)\frac{p_z(Z)}{q_\phi(Z|x)}\right]\right)$$

$$= \log\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{Z_k\sim q_\phi(z|x)}\left[p_\theta(x|Z_k)\frac{p_z(Z_k)}{q_\phi(Z_k|x)}\right]\right)$$

$f(t) = \log\left(\frac{1}{K}\sum_{i=1}^{K} t\right) = \log(t)$ is concave so, by Jensen's inequality,

$$\geq \mathbb{E}_{Z_1,\dots,Z_K\sim q_\phi(z|x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} p_\theta(x|Z_k)\frac{p_z(Z_k)}{q_\phi(Z_k|x)}\right)\right] = VLB_{\theta,\phi}^{(K)}(x) \blacksquare$$

(b) Define $a_k = p_\theta(x|Z_k)\frac{p_z(Z_k)}{q_\phi(Z_k|x)}$

Then $VLB_{\theta,\phi}^{(K)} = \mathbb{E}_{Z_1,\dots,Z_K\sim q_\phi(z|x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} a_k\right)\right]$

and $VLB_{\theta,\phi}^{(M)} = \mathbb{E}_{Z_1',\dots,Z_M'\sim q_\phi(z|x)}\left[\log\left(\frac{1}{M}\sum_{k=1}^{M} a_k\right)\right]$

w.l.o.g. (without loss of generality) say that the $Z_1',\dots,Z_M'$ sampled for $VLB^{(M)}$ are a subset of the $Z_1,\dots,Z_K$ sampled for $VLB^{(K)}$ indicated by $I\subset\{1,\dots,K\}$ with $|I|=M\leq K$ as hinted. We can then rewrite $VLB^{(M)}$ as

$$VLB_{\theta,\phi}^{(M)} = \mathbb{E}_{Z_1,\dots,Z_K\sim q_\phi}\left[\mathbb{E}_{I=\{i_1,\dots,i_M\}}\left[\log\left(\frac{1}{M}\sum_{k=1}^{M} a_{i_k}\right)\right]\right]$$

i.e. for the outer expectation we sample the same Z_k as for $VLB_{\theta,\phi}^{(K)}$ and $a_{i_k} = p_\theta(x|Z_{i_k})\frac{p_z(Z_{i_k})}{q_\phi(Z_{i_k}|x)}$

where $Z_{i_k}\in\{Z_1,\dots,Z_K\}$

$$\leq \mathbb{E}_{Z_k\sim q_\phi}\left[\log\left(\mathbb{E}_I\left[\frac{1}{M}\sum_{k=1}^{M} a_{i_k}\right]\right)\right] \quad \text{by Jensen's inequality}$$

$$= \mathbb{E}_{Z_k\sim q_\phi}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} a_k\right)\right] \quad \text{by given hint}$$

$$= VLB_{\theta,\phi}^{(K)}(x)$$

So for $K\geq M$, $VLB_{\theta,\phi}^{(M)} \leq VLB_{\theta,\phi}^{(K)}$ $\blacksquare$

(c) Powerful enough refers to if the neural network, parameterised by ϕ, underlying q_ϕ can accurately represent the true posterior distribution. i.e. if $q_\phi(z|x) \approx p_\theta(z|x)$ sufficiently well

In this case,

$$\underset{\theta,\phi}{\text{maximise}} \sum_{i=1}^{N} VLB^{(K)}_{\theta,\phi}(X_i)$$

closer to equality the more powerful q_ϕ is

$$\approx \underset{\theta}{\text{maximise}} \sum_{i=1}^{N} \mathbb{E}_{z_1,\dots,z_k \sim p_\theta(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x|z_k)p_z(z_k)}{p_\theta(z_k|x)}\right]$$

$$= \underset{\theta}{\text{maximise}} \sum_{i=1}^{N} \log\left(\frac{1}{K}\sum_{k=1}^{K} p_\theta(x)\right) = \underset{\theta}{\text{maximise}} \sum_{i=1}^{N} \log p_\theta(x) \blacksquare$$

2.(a)

$$\log p_\theta(X_i) = \log\left(\mathbb{E}_{Z\sim r_\lambda(z)}\left[p_\theta(X_i|Z)\right]\right)$$

$$= \log\left(\mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right]\right)$$

$$\geq \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\log\left(\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right)\right] \quad \text{by Jensen's inequality}$$

$$= VLB_{\theta,\phi,\lambda}(X_i) \blacksquare$$

(b)

$$\underline{\nabla}\, VLB(X_i) = \left(\underline{\nabla}_\theta VLB(X_i), \underline{\nabla}_\phi VLB(X_i), \underline{\nabla}_\lambda VLB(X_i)\right)$$

$$\underline{\nabla}_\theta VLB(X_i) = \underline{\nabla}_\theta \int \log\left(\frac{p_\theta(X_i|z)r_\lambda(z)}{q_\phi(z|X_i)}\right) q_\phi(z|X_i)\,dz$$

$$= \int \underline{\nabla}_\theta\left(p_\theta(X_i|z)\right)\frac{1}{p_\theta(X_i|z)}\, q_\phi(z|X_i)\,dz + \underline{0}$$

$$= \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\underline{\nabla}_\theta\left(\log(p_\theta(X_i|Z))\right)\right]$$

$$\underline{\nabla}_\phi VLB(X_i) = \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\left(\underline{\nabla}_\phi \log q_\phi(Z)\right)\log\left(\frac{p_\theta(X_i|Z)r_\lambda(Z)}{q_\phi(Z|X_i)}\right)\right] \quad \text{by log-derivative trick for VAEs (HW10.1)}$$

$$\underline{\nabla}_\lambda VLB(X_i) = \mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\underline{\nabla}_\lambda\left(\log(r_\lambda(Z))\right)\right] \quad \text{by same logic as } \underline{\nabla}_\theta$$

So stochastic gradients of $VLB_{\theta,\phi,\lambda}(X_i)$ can be computed by:

$$\underline{\nabla}_{\theta,\phi,\lambda} VLB_{\theta,\phi,\lambda}(X_i) \approx \frac{1}{K}\sum_{k=1}^{K}\Bigg(\underline{\nabla}_\theta\left(\log(p_\theta(X_i|Z_k))\right), \left(\underline{\nabla}_\phi \log(Z_k)\right)\log\left(\frac{p_\theta(X_i|Z_k)r_\lambda(Z_k)}{q_\phi(Z_k|X_i)}\right),$$

$$\underline{\nabla}_\lambda\left(\log(r_\lambda(Z_k))\right)\Bigg) \quad \text{where } Z_k \sim q_\phi(z|X_i)$$

2.(c) $\nabla_{\underline{\theta}}$ and ∇_λ same as above (Expectation distribution q_ϕ doesn't depend on $\underline{\theta}$ or $\underline{\lambda}$ so no need for the trick)

∇_ϕ evaluation changes if we use reparametrization vs. log-derivative before:

$$\nabla_\phi \, VLB(X_i) = \nabla_\phi \, \mathbb{E}_{Z \sim q_\phi(Z|X_i)} \left[\log \left(\frac{p_\theta(X_i|Z) r_\lambda(Z)}{q_\phi(Z|X_i)} \right) \right]$$

(Reparameterization) $$= \nabla_\phi \, \mathbb{E}_{\underline{\varepsilon} \sim \mathcal{N}(0,1)} \left[\log \left(\frac{p_\theta(X_i|Y_\phi) r_\lambda(Y_\phi)}{q_\phi(Y_\phi|X_i)} \right) \right], \text{ where } Y_\phi(X_i, \varepsilon) = \mu_\phi(X_i) + \Sigma_\phi^{1/2}(X_i)\underline{\varepsilon}$$

$$= \mathbb{E}_{\underline{\varepsilon} \sim \mathcal{N}(0,1)} \left[\nabla_\phi \log \left(\text{''} \right) \right] = \mathbb{E}_{\underline{\varepsilon} \sim \mathcal{N}(0,1)} \left[\nabla_\phi \left(\log p_\theta + \log r_\lambda - \log q_\phi \right) \right]$$

- $p_\theta(X_i|Y_\phi) = (2\pi)^{-k/2} \sigma^{-1} \exp\left(-\frac{1}{2\sigma^2} \|X_i - f_\theta(Y_\phi)\|^2\right)$

$$\log p_\theta = \log\left((2\pi)^{-k/2}\sigma^{-1}\right) - \frac{1}{2\sigma^2} \|\underline{X}_i - \underline{f}_\theta(\underline{Y}_\phi)\|^2$$

$$\nabla_\phi \log p_\theta = \; ?$$

- $r_\lambda(Y_\phi) = (2\pi)^{-k/2} \|\underline{\lambda}_2\|^{-1} \exp\left(-\frac{1}{2}(Y_\phi - \lambda_1)^T \operatorname{diag}(\underline{\lambda}_2^{-1})(Y_\phi - \lambda_1)\right)$ element-wise power
  $= (2\pi)^{-k/2} \|\underline{\lambda}_2\|^{-1} \exp\left(-\frac{1}{2} \|\underline{\lambda}_2^{-1/2} \cdot (\underline{Y}_\phi - \underline{\lambda}_1)\|^2\right)$

$$\log r_\lambda = \log(\ldots) - \frac{1}{2} \|\underline{\lambda}_2^{-1/2} \cdot (\underline{Y}_\phi - \underline{\lambda}_1)\|^2$$

$$\nabla_\phi \log r_\lambda = \; ?$$

- $q_\phi(Y_\phi|X_i) = (2\pi)^{-k/2} |\Sigma_\phi|^{-1/2} \exp\left(-\frac{1}{2}(\Sigma_\phi^{1/2}(X_i)\varepsilon)^T \Sigma_\phi^{-1} (\Sigma_\phi^{1/2}(X_i)\varepsilon)\right)$
  $= (2\pi)^{-k/2} \|\underline{\Sigma}\|^{-1} \exp\left(-\frac{1}{2}\|\underline{\Sigma}^{-1/2} \cdot (\Sigma_\phi^{1/2}\varepsilon)\|^2\right)$, where $\underline{\Sigma}$ is the diag el. of Σ_ϕ

$$\log q_\phi = \log(\ldots) - \frac{1}{2} \|\underline{\Sigma}^{-1/2} \cdot (\Sigma_\phi^{1/2} \varepsilon)\|^2$$

?

$$\nabla VLB_{\theta,\phi,\lambda}(X_i) \approx \sum_{k=1}^{k} \left(\nabla_\theta \left(\log \left(p_\theta(X_i|Y_k) \right) \right), \; \overset{?}{\ldots} \; , \nabla_\lambda \left(\log \left(r_\lambda(Y_k) \right) \right) \right) \text{ with } \varepsilon_k \sim \mathcal{N}(0,1)$$

4.(a) In this model, each game is independent so $\mathbb{E}[\text{points for } B] = \text{num games} \times \mathbb{E}[\text{points won by } B \text{ in } 1 \text{ game}]$

$$\mathbb{E}[\text{points won by } B \text{ in } 1 \text{ game}] = \big(\mathbb{P}(B \text{ rock})\,\mathbb{P}(A \text{ scissors}) + \mathbb{P}(B \text{ paper})\,\mathbb{P}(A \text{ rock}) + \mathbb{P}(B \text{ scissors})\,\mathbb{P}(A \text{ paper})\big) - \big(\mathbb{P}(B \text{ rock})\,\mathbb{P}(A \text{ paper}) + \mathbb{P}(B \text{ paper})\,\mathbb{P}(A \text{ scissors}) + \mathbb{P}(B \text{ scissors})\,\mathbb{P}(A \text{ rock})\big) + 0$$

Notation:
$(p_C)_i =: p_{Ci}$

$$= p_{B1} p_{A3} + p_{B2} p_{A1} + p_{B3} p_{A2} - p_{B1} p_{A2} - p_{B2} p_{A3} - p_{B3} p_{A1}$$
$$= p_{B1}(p_{A3} - p_{A2}) + p_{B2}(p_{A1} - p_{A3}) + p_{B3}(p_{A2} - p_{A1})$$

$$\Rightarrow \mathbb{E}_{p_A^*, p_B}[\text{points for } B] = p_{B1}\left(\tfrac{1}{3} - \tfrac{1}{3}\right) + p_{B2}\left(\tfrac{1}{3} - \tfrac{1}{3}\right) + p_{B3}\left(\tfrac{1}{3} - \tfrac{1}{3}\right) = 0 \quad \text{for all } p_B \in \Delta^3$$

is uniqueness clear from here as max over all p_A with any p_B will be 0 so the 0-setting solution (p_A^*, p_B^*) is unique?

$$\mathbb{E}_{p_A, p_B^*}[\text{points for } B] = \tfrac{1}{3}(p_{A3} - p_{A2}) + \tfrac{1}{3}(p_{A1} - p_{A3}) + \tfrac{1}{3}(p_{A2} - p_{A1}) = 0 \quad \text{for all } p_A \in \Delta^3$$

$$\mathbb{E}_{p_A^*, p_B^*}[\text{points for } B] = 0 \quad \Rightarrow \quad \mathbb{E}_{p_A^*, p_B}[\text{points for } B] \le \mathbb{E}_{p_A^*, p_B^*}[\text{points for } B] \le \mathbb{E}_{p_A, p_B^*}[\text{points for } B]$$
$$\text{for all } p_A, p_B \in \Delta^3$$

$\Rightarrow$ so p_A^*, p_B^* is a solution to minimax problem

Suppose there exists another solution, (p_A', p_B'), with at least one necessarily different from p_A^*, p_B^*.

- If only one is different (wlog assume $p_A' = p_A^*$ and $p_B' \neq p_B^*$),
  $\mathbb{E}_{p_A', p_B}[\text{pts for } B] = \mathbb{E}_{p_A', p_B'}[\text{pts for } B] = 0$ still and then $\mathbb{E}_{p_A, p_B'}[\text{pts for } B] \ge 0$ required:

  $$\mathbb{E}_{p_A, p_B'}[\text{pts for } B] = p_B' \cdot (p_A \times \underline{1}) \quad \ldots ?$$

- If both are different ($p_A' \neq p_A^*$ and $p_B' \neq p_B^*$) … ?

4.(b) Yes, since this is a symmetric game in the sense that a loss for one player is a win for the other. Therefore $\mathbb{E}_{p_A, p_B}[\text{pts for } B] = 0 \Rightarrow \mathbb{E}_{p_A, p_B}[\text{pts for } A] = 0$. So if B plays with $p_B = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ then no particular strategy can be better for A in terms of winning more points in expectation; i.e. any $p_A \in \Delta^3$ is optimal for A.