MFDNN

Homework 11

24/05/30

(2/4)

$p_\theta(x) = \int p(x,z)dz = \int p_\theta(x|z)p_z(z)dz = \int \frac{p_\theta(z|x)p_z(z)}{q_\theta(z|x)} q_\theta(z|x)dz$

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

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

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

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

$\geq \mathbb{E}_{z_1,\dots,z_K\sim q_\phi(z|x)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} p_\theta(x|z_k)\frac{p_z(z_k)}{q_\phi(z_k|x)}\right)\right] = VLB_{\theta,\phi}^{(K)}(x)$ ■ ✓

(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 before. 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 $z_{i_k}\in\{z_1,\dots,z_K\}$

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

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

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

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

(c) Powerful enough refers to if the neural network, parameterized 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 ✓

$\exists \phi^*$ s.t. ... z_k z_k for $k=1,\dots,K$

In this case,

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

$$\overset{\text{closer to equality the more powerful } q_\phi \text{ is}}{\approx} \underset{\theta}{\text{maximize}} \sum_{i=1}^N \mathbb{E}_{z_1,\dots,z_K \sim p_\theta(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^K \frac{p_\theta(x|z_k)p_z(z_k)}{p_\theta(z_k|x)}\right] \checkmark$$

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

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

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

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

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

(b) $$\nabla VLB(X_i) = \left(\nabla_\theta VLB(X_i), \nabla_\phi VLB(X_i), \nabla_\lambda VLB(X_i)\right)$$

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

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

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

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

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

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

$$\nabla_{\theta,\phi,\lambda} VLB_{\theta,\phi,\lambda}(X_i) \approx \frac{1}{K}\sum_{k=1}^K \left(\nabla_\theta\left(\log(p_\theta(X_i|z_{i,k}))\right), \left(\nabla_\phi \log(z_{i,k})\right)\log\left(\frac{p_\theta(X_i|z_k)r_\lambda(z_k)}{q_\phi(z_k|X_i)}\right),\right.$$

$$\left.\nabla_\lambda\left(\log(r_\lambda(z_{i,k}))\right)\right) \checkmark \quad \text{where } z_k \sim q_\phi(z|X_i)$$

number of elements in a batch of SGA calculation

$z_{i,k}$ (sample for each X_i once)

2.(c) ∇_θ and ∇_λ same as above ✓ (Expectation distribution q_ϕ doesn't depend on θ or λ so no need for the trick)

∇_ϕ evaluation changes if we use reparameterization vs. log-derivative trick:

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

Redo 2.(c) $VLB_{\theta,\phi,\lambda}(X_i) = \underbrace{\mathbb{E}_{z\sim q_\phi(z|X_i)}[\log p_\theta(X_i|z)]}_{①} \underbrace{- D_{KL}(q_\phi(z|X_i)\|r_\lambda(z))}_{②}$ (slide 89)

$+\mathbb{E}[\log(\frac{r_\lambda}{q_\phi})]$

- $① = \mathbb{E}_{z\sim q_\phi(z|X_i)}\left[\log\left((2\pi\sigma^2)^{-d/2}\exp\left(-\frac{1}{2}(X_i - f_\theta(z))^T\left(\frac{1}{\sigma^2}I\right)(X_i - f_\theta(z))\right)\right)\right]$ $d = \dim(f_\theta)$

$$= -\frac{1}{2\sigma^2}\,\mathbb{E}_{z\sim\mathcal{N}(\mu_\phi(X_i),\Sigma_\phi(X_i))}\left[\|X_i - f_\theta(z)\|^2\right] - \frac{d}{2}\log(2\pi\sigma^2)$$

By reparameterization trick:

$$= -\frac{1}{2\sigma^2}\,\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\|X_i - f_\theta(\mu_\phi(X_i) + \Sigma_\phi^{1/2}(X_i)\varepsilon)\|^2\right] - \frac{d}{2}\log(2\pi\sigma^2)$$

↓

→ Apply ∇_ϕ : $-\frac{1}{2\sigma^2}\,\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\nabla_\phi\|\ldots\|^2\right]$ which can be evaluated + back-propagated

- $② = -\frac{1}{2}\left(tr\left([\mathrm{diag}(\lambda_2)]^{-1}\Sigma_\phi(X_i)\right) + (\lambda_1 - \mu_\phi(X_i))^T[\mathrm{diag}(\lambda_2)]^{-1}(\lambda_1 - \mu_\phi(X_i)) - k + \log\left(\frac{\det(\mathrm{diag}(\lambda_2))}{\det(\Sigma_\phi(X_i))}\right)\right)$

by HW9.5

→ Gradient ∇_ϕ can then be calculated + back propagated

Redo 4.(a) $\mathbb{E}_{p_A,p_B}[\text{points for B}] = p_A^T M p_B$ $\quad M := \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$ $\quad p_C = \begin{bmatrix} p_{C,\text{rock}} \\ p_{C,\text{paper}} \\ p_{C,\text{scissors}} \end{bmatrix}$

$\left(= p_{A1}(p_{B3} - p_{B2}) + p_{A2}(p_{B1} - p_{B3}) + p_{A3}(p_{B2} - p_{B1}) \quad \text{(as I did originally)} \right)$

- Clearly, $p_A^T M p_B \geq \underset{p_A \in \Delta^3}{\text{minimize}}\ p_A^T M p_B \Rightarrow \underset{p_B \in \Delta^3}{\text{maximize}}\ p_A^T M p_B \geq \underset{p_B}{\text{maximize}}\ \underset{p_A}{\text{minimize}}\ p_A^T M p_B$

$\Rightarrow \underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ p_A^T M p_B \geq \underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ \underset{p_A}{\text{minimize}}\ p_A^T M p_B$ by above (*)

$= \underset{p_B}{\text{maximize}}\ \underset{p_A}{\text{minimize}}\ p_A^T M p_B$

$\underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ p_A^T M p_B \leq \underset{p_B}{\text{maximize}}\ p_A^{*T} M p_B = 0$ (**)

$0 \geq \underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ p_A^T M p_B$ by (**)

$= \underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ -p_B^T M p_A$ $\because p_A^T M p_B = (p_A^T M p_B)^T = p_B^T M^T p_A = -p_B^T M p_A$

$= -\underset{p_A}{\text{maximize}}\ \underset{p_B}{\text{minimize}}\ p_B^T M p_A$

$= -\underset{p_B}{\text{maximize}}\ \underset{p_A}{\text{minimize}}\ p_A^T M p_B$

$\geq -\underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ p_A^T M p_B \geq 0$ by (*) and then (**)

$\Rightarrow \underset{p_A}{\text{minimize}}\ \underset{p_B}{\text{maximize}}\ p_A^T M p_B = \underset{p_B}{\text{maximize}}\ \underset{p_A}{\text{minimize}}\ p_A^T M p_B = 0 \Leftrightarrow$ existence of saddle pt solution p_A^*, p_B^*

- Show unique by proving $p_A^* M p_B \leq p_A^{*T} M p_B^* = 0 \leq p_A^T M p_B^* \quad \forall\, p_A, p_B \in \Delta^3$ only for given p_A^*, p_B^*

Suppose $p_B^* \neq [1/3, 1/3, 1/3]^T \Leftrightarrow M p_B^* = [p_{B_2} - p_{B_1},\ p_{B_3} - p_{B_2},\ p_{B_1} - p_{B_3}]^T \neq [0,0,0]^T$

$\Rightarrow$ At least one entry of $M p_B^*$ is negative

So pick p_A such that $p_A^T M p_B^* < 0$ to obtain a contradiction with $p_A^T M p_B^* \geq 0\ \forall\, p_A \in \Delta^3$ above.

$\Rightarrow p_B^* = [1/3, 1/3, 1/3]^T$

Analogous argument with $p_A^{*T} M p_B \leq 0$ to show $p_A^* = [1/3, 1/3, 1/3]^T$

So given p_A^*, p_B^* is indeed the unique solution.

(b) (I'd misunderstood the Q the first time - I assumed we had set that B plays p_B^*)

If A chooses $p_A \neq p_A^*$, then B can choose one p_B of $[1,0,0]^T, [0,1,0]^T, [0,0,1]^T$ to make $\mathbb{E}_{p_A,p_B}[\text{points for B}] > 0$! So no, not just any strategy is optimal for A. Only in the specific case where B plays p_B^* is A free to pick any strategy since $\mathbb{E}_{p_A,p_B^*}[\text{points for B}] = 0$.

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

$$\begin{aligned}\mathbb{E}[\text{points won by B in 1 game}] &= \big(\mathbb{P}(\text{B rock})\mathbb{P}(\text{A scissors}) + \mathbb{P}(\text{B paper})\mathbb{P}(\text{A rock}) + \\ &\quad \mathbb{P}(\text{B scissors})\mathbb{P}(\text{A paper})\big) - \big(\mathbb{P}(\text{B rock})\mathbb{P}(\text{A paper}) + \\ &\quad \mathbb{P}(\text{B paper})\mathbb{P}(\text{A scissors}) + \mathbb{P}(\text{B scissors})\mathbb{P}(\text{A rock})\big) + 0 \\ &= p_{B1}p_{A3} + p_{B2}p_{A1} + p_{B3}p_{A2} - p_{B1}p_{A2} - p_{B2}p_{A3} - p_{B3}p_{A1} \\ &= p_{B1}(p_{A3} - p_{A2}) + p_{B2}(p_{A1} - p_{A3}) + p_{B3}(p_{A2} - p_{A1})\end{aligned}$$

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

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

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

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

$\mathbb{E}_{p_A^*, p_B^*}[\text{points for B}] = 0 \Rightarrow \mathbb{E}_{p_A^*, p_B}[\text{points for B}] \le \mathbb{E}_{p_A^*, p_B^*}[\text{points for B}] \le \mathbb{E}_{p_A, p_B^*}[\text{points for B}]$ 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:

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

- 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.