MathDNN Homework 9

Department of Computer Science and Engineering
2021-16988 Jaewan Park

Problem 3

Consider Ω and $\Omega^\complement$ as ordered and sorted sets. Now define f and g as $f(x) = i|_{x \text{ is the } i\text{th element of } \Omega}$ and $g(y) = j|_{y \text{ is the } j\text{th element of } \Omega^\complement}$ for $x \in \Omega$ and $y \in \Omega^\complement$ each. We can calculate the Jacobian matrix between the layers in the form of

$$\frac{\partial z}{\partial x} = \left\{ \frac{\partial z_i}{\partial x_j} \right\}_{i,j}, \quad \frac{\partial z_i}{\partial x_j} = \begin{cases} 1 & \left(i \in \Omega,\ i = j\right) \\ \frac{\partial [s_\theta(x_\Omega)]_{g(i)}}{\partial x_j} e^{[s_\theta(x_\Omega)]_{g(i)}} x_i + \frac{\partial [t_\theta(x_\Omega)]_{g(i)}}{\partial x_j} & \left(i \in \Omega^\complement,\ j \in \Omega\right) \\ e^{[s_\theta(x_\Omega)]_{g(i)}} \left(= e^{[s_\theta(x_\Omega)]_{g(j)}}\right) & \left(i \in \Omega^\complement,\ j \in \Omega^\complement,\ i = j\right) \\ 0 & \left(\text{otherwise}\right) \end{cases}.$$

Selecting σ such that $\sigma^{-1}(i) = \begin{cases} f^{-1}(i) & (i \leq |\Omega|) \\ g^{-1}(i - |\Omega|) & (i > |\Omega|) \end{cases}$ gives

$$\begin{aligned} P_\sigma \frac{\partial z}{\partial x} P_{\sigma^{-1}} &= \begin{bmatrix} \partial z_{\sigma^{-1}(1)} / \partial x_{\sigma^{-1}(1)} & \partial z_{\sigma^{-1}(1)} / \partial x_{\sigma^{-1}(2)} & \cdots & \partial z_{\sigma^{-1}(1)} / \partial x_{\sigma^{-1}(n)} \\ \partial z_{\sigma^{-1}(2)} / \partial x_{\sigma^{-1}(1)} & \partial z_{\sigma^{-1}(2)} / \partial x_{\sigma^{-1}(2)} & \cdots & \partial z_{\sigma^{-1}(2)} / \partial x_{\sigma^{-1}(n)} \\ \vdots & \vdots & \ddots & \vdots \\ \partial z_{\sigma^{-1}(n)} / \partial x_{\sigma^{-1}(1)} & \partial z_{\sigma^{-1}(n)} / \partial x_{\sigma^{-1}(2)} & \cdots & \partial z_{\sigma^{-1}(n)} / \partial x_{\sigma^{-1}(n)} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ * & \operatorname{diag}\left(e^{s_\theta(x_\Omega)}\right) \end{bmatrix}. \end{aligned}$$

Therefore $\dfrac{\partial z}{\partial x}$ can be decomposed in the form of

$$\frac{\partial z}{\partial x} = P_{\sigma^{-1}} \begin{bmatrix} I & 0 \\ * & \operatorname{diag}\left(e^{s_\theta(x_\Omega)}\right) \end{bmatrix} P_\sigma,$$

and we can calculate the determinant as

$$\begin{aligned} \log \left| \frac{\partial z}{\partial x} \right| &= \log \begin{vmatrix} I & 0 \\ * & \operatorname{diag}\left(e^{s_\theta(x_\Omega)}\right) \end{vmatrix} \\ &= \log \prod_{i \in \Omega^\complement} e^{[s_\theta(x_\Omega)]_{g(i)}} = \sum_{i \in \Omega^\complement} [s_\theta(x_\Omega)]_{g(i)} \\ &= \mathbf{1}_{n-|\Omega|}^\intercal s_\theta(x_\Omega). \end{aligned}$$

Problem 4

(a) Since $-\log$ is a convex function, we can apply Jensen's inequality to $-\log$, which gives

$$\begin{aligned}
D_{\mathrm{KL}}(X||Y) &= \int_{\mathbb{R}^d} f(x)\log\left(\frac{f(x)}{g(x)}\right)dx = \mathbf{E}\left[\log\left(\frac{f(X)}{g(X)}\right)\right] = \mathbf{E}\left[-\log\left(\frac{g(X)}{f(X)}\right)\right] \\
&\geq -\log\left(\mathbf{E}\left[\frac{g(X)}{f(X)}\right]\right) = -\log\left(\int_{\mathbb{R}^d} f(x)\cdot\frac{g(x)}{f(x)}dx\right) = -\log 1 = 0.
\end{aligned}$$

(b) Since $X_1, \cdots, X_d$ and $Y_1, \cdots, Y_d$ are each independent, when $f_1, \cdots, f_d$ and $g_1, \cdots, g_d$ are PDFs for $X_1, \cdots, X_d$ and $Y_1, \cdots, Y_d$ each, we can say

$$f(x) = f_1(x_1)\cdots f_d(x_d), \quad g(y) = g_1(y_1)\cdots g_d(y_d)$$

for any $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$. Therefore

$$\begin{aligned}
D_{\mathrm{KL}}(X||Y) &= \mathbf{E}\left[-\log\left(\frac{g(X)}{f(X)}\right)\right] = \mathbf{E}\left[-\log\left(\frac{g_1(X_1)}{f_1(X_1)}\right)\right] + \cdots + \mathbf{E}\left[-\log\left(\frac{g_d(X_d)}{f_d(X_d)}\right)\right] \\
&= D_{\mathrm{KL}}(X_1||Y_1) + \cdots + D_{\mathrm{KL}}(X_d||Y_d).
\end{aligned}$$

Problem 5

The PDF of a multivariate Gaussian random variable $X \sim \mathcal{N}(\mu, \Sigma)$ with dimension d is given by

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\right).$$

Let X_0, X_1 random variables that follow $\mathcal{N}(\mu_0, \Sigma_0)$, $\mathcal{N}(\mu_1, \Sigma_1)$ each. Also let their PDFs f_0, f_1. Then

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathcal{N}(\mu_0,\Sigma_0)\,||\mathcal{N}(\mu_1,\Sigma_1)) \\
&\quad= \mathbf{E}\left[-\log\left(\frac{f_1(X_0)}{f_0(X_0)}\right)\right] = \mathbf{E}\Big[\log f_0(X_0) - \log f_1(X_0)\Big] \\
&\quad= \mathbf{E}\left[\frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0} - \frac{1}{2}(X_0-\mu_0)^\intercal\Sigma_0^{-1}(X_0-\mu_0) + \frac{1}{2}(X_0-\mu_1)^\intercal\Sigma_1^{-1}(X_0-\mu_1)\right] \\
&\quad= \frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0} - \frac{1}{2}\mathbf{E}\big[\mathrm{tr}\big((X_0-\mu_0)^\intercal\Sigma_0^{-1}(X_0-\mu_0)\big)\big] + \frac{1}{2}\mathbf{E}\big[(X_0-\mu_1)^\intercal\Sigma_1^{-1}(X_0-\mu_1)\big] \\
&\quad= \frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0} - \frac{1}{2}\mathbf{E}\big[\mathrm{tr}\big((X_0-\mu_0)(X_0-\mu_0)^\intercal\Sigma_0^{-1}\big)\big] + \frac{1}{2}\big((\mu_0-\mu_1)^\intercal\Sigma_1^{-1}(\mu_0-\mu_1) + \mathrm{tr}\big(\Sigma_1^{-1}\Sigma_0\big)\big) \\
&\quad= \frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0} - \frac{1}{2}\mathrm{tr}\big(\mathbf{E}[(X_0-\mu_0)(X_0-\mu_0)^\intercal]\Sigma_0^{-1}\big) + \frac{1}{2}\big((\mu_1-\mu_0)^\intercal\Sigma_1^{-1}(\mu_1-\mu_0) + \mathrm{tr}\big(\Sigma_1^{-1}\Sigma_0\big)\big) \\
&\quad= \frac{1}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0} - \frac{1}{2}\mathrm{tr}\big(\Sigma_0\Sigma_0^{-1}\big) + \frac{1}{2}\big((\mu_1-\mu_0)^\intercal\Sigma_1^{-1}(\mu_1-\mu_0) + \mathrm{tr}\big(\Sigma_1^{-1}\Sigma_0\big)\big) \\
&\quad= \frac{1}{2}\left(\mathrm{tr}\big(\Sigma_1^{-1}\Sigma_0\big) + (\mu_1-\mu_0)^\intercal\Sigma_1^{-1}(\mu_1-\mu_0) - d + \log\left(\frac{\det\Sigma_1}{\det\Sigma_0}\right)\right).
\end{aligned}$$

Problem 6

For each θ, let $\phi_\theta \in \Phi$ the value of ϕ that makes $h(\theta, \phi) = 0$. Then we obtain

$$\begin{aligned}
\sup_{\theta,\phi} g(\theta,\phi) &= \sup_{\theta} \left(\sup_{\phi} g(\theta,\phi) \right) \\
&= \sup_{\theta} \left(\sup_{\phi} \Big(f(\theta) - h(\theta,\phi) \Big) \right) = \sup_{\theta} \Big(f(\theta) - \inf_{\phi} h(\theta,\phi) \Big) \\
&= \sup_{\theta} f(\theta)
\end{aligned}$$

since $\inf_\phi h(\theta, \phi) = 0$, more precisely $\min_\phi h(\theta, \phi) = 0$ when $\phi = \phi_\theta$. Therefore we can conclude that

$$\operatorname{argmax} f = \{\theta \mid (\theta, \phi) \in \operatorname{argmax} g\}$$

and the two given optimization problems are equivalent.