2. In pytorch,

Weights $(A_l)_{ij}$ are initialized with $U\left(-\frac{1}{\sqrt{fan\_in}}, \frac{1}{\sqrt{fan\_in}}\right)$ and

biases $(b_l)_i$ are initialized with $U\left(-\frac{1}{\sqrt{fan\_in}}, \frac{1}{\sqrt{fan\_in}}\right)$.

They are mutually independently initialized.

$$E((A_l)_{ij}) = \frac{1}{2}\left(-\sqrt{\frac{1}{fan\_in}} + \sqrt{\frac{1}{fan\_in}}\right) = 0$$

$$Var((A_l)_{ij}) = \frac{1}{12}\left(\sqrt{\frac{1}{fan\_in}} + \sqrt{\frac{1}{fan\_in}}\right)^2 = \frac{1}{3}\cdot\frac{1}{fan\_in} = \frac{1}{3}\cdot\frac{1}{n_{l-1}}$$

$$E((b_l)_i) = \frac{1}{2}\left(-\frac{1}{\sqrt{fan\_in}} + \frac{1}{\sqrt{fan\_in}}\right) = 0$$

$$Var((b_l)_i) = \frac{1}{12}\left(\frac{1}{\sqrt{fan\_in}} + \frac{1}{\sqrt{fan\_in}}\right)^2 = \frac{1}{3}\cdot\frac{1}{fan\_in} = \frac{1}{3}\cdot\frac{1}{n_{l-1}}$$

$$Var(y_L) = Var(A_L y_{L-1} + b_L) = var\left(\sum_j \{(A_L)_{ij}(y_{L-1})_j\} + (b_L)_i\right)$$

$$= n_{L-1}\, var\,(A_L)_{11}\, Var(y_{L-1})_1 + var(b_L)_1$$

$$= n_{L-1}\,\frac{1}{3}\cdot\frac{1}{n_{L-1}}\, Var(y_{L-1})_1 + \frac{1}{3}\cdot\frac{1}{n_{L-1}}$$

$$= \frac{1}{3}\, Var(y_{L-1})_1 + \frac{1}{3}\cdot\frac{1}{n_{L-1}}$$

$$Var(y_l)_1 = var\left(\sum_j \{(A_l)_{ij}(y_{l-1})_j\} + (b_l)_1\right)$$

$$= n_{l-1}\, var\,(A_l)_{11}\, var\,(y_{l-1})_1 + var\,(b_l)_1$$

$$= n_{l-1}\,\frac{1}{3}\cdot\frac{1}{n_{l-1}}\, var\,(y_{l-1})_1 + \frac{1}{3}\cdot\frac{1}{n_{l-1}}$$

$$= \frac{1}{3}\, Var\,(y_{l-1})_1 + \frac{1}{3}\cdot\frac{1}{n_{l-1}}, \qquad l = 1, 2, 3, \cdots, L-1, \qquad y_0 := x$$

$$\therefore \quad var\,(y_L) = \frac{1}{3^L}\, Var(y_0)_1 + \sum_{k=0}^{L-1}\frac{1}{3^{L-k}}\cdot\frac{1}{n_k}$$

$$= \frac{1}{3^L} + \sum_{k=0}^{L-1}\frac{1}{3^{L-k}}\cdot\frac{1}{n_k}$$

□

3.

$$\frac{\partial y_L}{\partial w_L} = \begin{bmatrix} (y_{L-1})_1 & (y_{L-1})_2 & \cdots & (y_{L-1})_{f_L} & 0 & \cdots & 0 \\ 0 & (y_{L-1})_2 & (y_{L-1})_3 & \cdots & (y_{L-1})_{f_L+1} & 0 \cdots & 0 \\ & & & \ddots & & & \\ 0 & \cdots & 0 & (y_{L-1})_{n_L} & (y_{L-1})_{n_L+1} & \cdots & (y_{L-1})_{\underbrace{f_L+n_L-1}_{n_{L-1}}} \end{bmatrix}$$

$y_l' := A_l y_{l-1} + b_l$, $l = 1, \cdots, L-1$.

$$\frac{\partial y_L}{\partial w_l} = \frac{\partial y_L}{\partial y_l'} \frac{\partial y_l'}{\partial w_l} = \frac{\partial y_L}{\partial y_l} \frac{\partial y_l}{\partial y_l'} \frac{\partial y_l'}{\partial w_l}, \quad l = 1, \cdots, L-1$$

$$= \frac{\partial y_L}{\partial y_l} \operatorname{diag}\left(\sigma'(y_l')\right) \begin{bmatrix} (y_{l-1})_1 & (y_{l-1})_2 & \cdots & (y_{l-1})_{f_l} \\ (y_{l-1})_2 & (y_{l-1})_3 & \cdots & (y_{l-1})_{f_l+1} \\ & & \ddots & \\ (y_{l-1})_{n_l} & \cdots & & (y_{l-1})_{\underbrace{f_l+n_l-1}_{n_{l-1}}} \end{bmatrix}$$

$$= A_L \left[\operatorname{diag}\left(\sigma'(y_{L-1})\right) A_{L-1}\right] \cdots \left[\operatorname{diag}\left(\sigma'(y_{l+1})\right) A_{l+1}\right] \operatorname{diag}\left(\sigma'(y_l)\right) \begin{bmatrix} (y_{l-1})_1 & (y_{l-1})_2 & \cdots & (y_{l-1})_{f_l} \\ (y_{l-1})_2 & (y_{l-1})_3 & \cdots & (y_{l-1})_{f_l+1} \\ & & \ddots & \\ (y_{l-1})_{n_l} & \cdots & & (y_{l-1})_{\underbrace{f_l+n_l-1}_{n_{l-1}}} \end{bmatrix}$$

$$\frac{\partial y_L}{\partial b_L} = I_{n_L}$$

Using result of Homework 4 problem 6,

$$\frac{\partial y_L}{\partial b_l} = \frac{\partial y_L}{\partial y_l} \frac{\partial y_l}{\partial \beta_l} \frac{\partial \beta_l}{\partial b_l} = \frac{\partial y_L}{\partial y_{L-1}} \cdot \frac{\partial y_{L-1}}{\partial y_{L-2}} \cdots \frac{\partial y_{l+1}}{\partial y_l} \cdot \frac{\partial y_l}{\partial \beta_l} \frac{\partial \beta_l}{\partial b_l}$$

$$= A_L \cdot \left[\operatorname{diag}\left(\sigma'(y_{L-1}')\right) A_{L-1}\right] \cdots \left[\operatorname{diag}\left(\sigma'(y_{l+1}')\right) A_{l+1}\right] \operatorname{diag}\left(\sigma'(y_l')\right) \cdot \mathbb{1}_{n_l}$$

$l = 1, \cdots, L-1$

Each $\operatorname{diag} \cdot A$ is computed by $(A^T \cdot \operatorname{diag})^T$ with transposed convolution.

By storing output of each layer during forward propogation,

back propogation can be done efficiently.

□

4. $Y = \varphi(x) = A^{-1}(x-b)$

$$P_X(x) = P_Y(\varphi(x)) \left| \det \frac{\partial \varphi}{\partial x}(x) \right|$$

$$= P_Y\left(A^{-1}(x-b)\right) \left| \det A^{-1} \right|$$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}(x-b)^T (A^{-1})^T A^{-1}(x-b)} |A|^{-1}, \qquad |\Sigma| = |AA^T| = |A||A^T| = |A|^2$$

$$= \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-b)^T \Sigma^{-1}(x-b)}$$

□

5.

(a) $$D_{KL}(X \| Y) = \int_{\mathbb{R}^d} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$$

$$= \mathbb{E}\left[-\log \frac{g(x)}{f(x)} \right]$$

$$\geq -\log \mathbb{E}\left[\frac{g(x)}{f(x)}\right] \quad \text{(Since } -\log \text{ convex, Jensen's inequality)}$$

$$= -\log \int_{\mathbb{R}^d} \frac{g(x)}{\cancel{f(x)}} \cancel{f(x)}\, dx$$

$$= -\log 1$$

$$= 0.$$

□

(b) $$f(\underline{x}) = pdf_{X_1}(x_1) \cdots pdf_{X_d}(x_d)$$

$$g(\underline{y}) = pdf_{Y_1}(y_1) \cdots pdf_{Y_d}(y_d)$$

$$D_{KL}(X|Y) = \int_{\mathbb{R}^d} f(\underline{x}) \log\left(\frac{f(\underline{x})}{g(\underline{x})}\right) d\underline{x}$$

$$= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} pdf_{X_1}(x_1) \cdots pdf_{X_d}(x_d) \log \frac{pdf_{X_1}(x_1) \cdots pdf_{X_d}(x_d)}{pdf_{Y_1}(x_1) \cdots pdf_{Y_d}(x_d)} dx_1 \cdots dx_d$$

$$= \sum_{i=1}^{d} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} pdf_{X_1}(x_1) \cdots pdf_{X_d}(x_d) \log \frac{pdf_{X_i}(x_i)}{pdf_{Y_i}(x_i)} dx_1 \cdots dx_d$$

$$= \sum_{i=1}^{d} \int_{\mathbb{R}} pdf_{X_i}(x_i) \log \frac{pdf_{X_i}(x_i)}{pdf_{Y_i}(x_i)} dx_i$$

$$= \sum_{i=1}^{d} D_{KL}(X_i \| Y_i)$$

□

6. $X \sim N(\mu_0, \Sigma_0)$

$Y \sim N(\mu_1, \Sigma_1)$

$$D_{KL}(X \| Y) = \mathbb{E}\left[-\log \frac{pdf_Y(x)}{pdf_X(x)}\right]$$

$$= \mathbb{E}\left[-\log \frac{\frac{1}{\sqrt{(2\pi)^d |\Sigma_1|}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right)}{\frac{1}{\sqrt{(2\pi)^d |\Sigma_0|}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)\right)}\right]$$

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

Use the fact that $\mathbb{E}(y^T A y) = \mathrm{tr}(AV) + \mu^T A \mu$, where $y \sim N(\mu, V)$

$X - \mu_0 \sim N(0, \Sigma_0)$, $X - \mu_1 \sim N(\mu_0 - \mu_1, \Sigma_0)$

Hence, $D_{KL}(X \| Y) = -\frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}\Sigma_0) + \frac{1}{2}\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + \frac{1}{2}(\mu_0-\mu_1)^T \Sigma_1^{-1}(\mu_0-\mu_1) + \frac{1}{2}\log\frac{|\Sigma_1|}{|\Sigma_0|}$

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

□