Homework 1

(1) $G_\theta : \mathbb{R}^k \to \mathbb{R}^n$ Get samples by sampling $z \in \mathbb{R}^k$
$x_G \sim p_{G_\theta}(z)$

$D_\phi : \mathbb{R}^n \to [0,1]$ Higher if detects real image
Last layer is usually a sigmoid fcn.

$\sigma(z) = \frac{1}{1+e^{-z}}$

Perform gradient descent on $L_D(\phi;\theta) = -\mathbb{E}_{x\sim p_{data}}[\log D_\phi(x)] - \mathbb{E}_{z\sim N(0,1)}[\log[1-D_\phi(G_\theta(z))]]$

For real x, we want the discriminator to output $\boxed{1}$. For fake x, $D_\phi(x) \to \boxed{0}$

And gradient descent on $L_G(\theta;\phi) = \mathbb{E}_{z\sim N(0,1)}[\log[1-D_\phi(G_\theta(z))]]$

1. Fix vanishing gradient problem on L_G
   Show that:

$$\frac{d}{d\theta} L_G(\theta;\phi) \approx 0 \text{ if } D_\phi(G_\theta(z)) \approx 0 \Leftrightarrow h_\phi(G_\theta(z)) \ll 0$$

Recall that: $\sigma'(z) = \sigma(z)(1-\sigma(z))$

//

$$L_G(\theta;\phi) = \mathbb{E}_{z\sim N(0,1)}[\log[1-\sigma(h_\phi(G_\theta(z)))]] \qquad \frac{d}{dx}\log x = \frac{1}{x} \qquad \frac{d}{du}\log x(u) = \frac{1}{x(u)}x'(u)$$

$$\frac{d}{d\theta} L_G(\theta;\phi) = \mathbb{E}_{z\sim N(0,1)}\left[\frac{d}{d\theta}\log[1-\sigma(h_\phi(G_\theta(z)))]\right]$$: Both expectation and derivative are linear operators

$$= \mathbb{E}_{z\sim N(0,1)}\left[\frac{1}{1-\sigma(h_\phi(G_\theta(z)))}\cdot \sigma(h_\phi(G_\theta(z)))(\sigma(h_\phi(G_\theta(z)))-1)\cdot \frac{d}{d\theta}[h_\phi(G_\theta(z))]\right]$$

If $h_\phi(G_\theta(z)) \to -\infty \Leftrightarrow \sigma(h_\phi(G_\theta(z))) \to 0$

$$= \mathbb{E}_{z\sim N(0,1)}\left[\frac{1}{1-0}\cdot 0(0-1)\cdot\frac{d}{d\theta}[h_\phi(G_\theta(z))]\right] = 0 \quad //$$

This is bad for the generator because, when $h_\phi(G_\theta(z)) \to -\infty$ (discriminator successfully identifies fake image), the derivative of $L_G(\theta;\phi)$ w.r.t. θ will be very small. Since learning and improving relies on gradient descent, the generator won't be able to improve and won't escape that region.

Alternatively, the non-saturating loss shows large derivatives when $h_\phi(G_\theta(z)) \to -\infty$, allowing the generator to improve and escape such regions faster.

② Let $\mu : \mathcal{X} \to \mathbb{R}^K$, $\sigma : \mathcal{X} \to \mathbb{R}^K$

$$q(z|x) = \mathcal{N}(z; \mu(x), \sigma^2(x))$$
$$p(z) = \mathcal{N}(z; 0, I)$$

Show that:

$$D_{KL}(q(z|x) \,\|\, p(z)) = \frac{1}{2}\sum_{i=1}^{K} \sigma_i(x)^2 + \mu_i(x)^2 - 1 - \log \sigma_i(x)$$
$$D_{KL}(q(z|x) \,\|\, p(z)) = \int q(z|x) \cdot \log \frac{q(z|x)}{p(z)}\, dz$$

Let:

$$Z \triangleq \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_K \end{bmatrix} \qquad Z \sim \mathcal{N}(\mu(x), \sigma^2(x)) \qquad Z_i \sim \mathcal{N}(\mu_i(x), \sigma_i^2(x))$$

The PDF of Z is given by: $f_Z(\vec{z}) \triangleq q(\vec{z}|x)$

The KL divergence can be rewritten as:

$$D_{KL}(q(\cdot|x) \,\|\, p(\cdot)) = E_{Z\sim\mathcal{N}(\mu(x), \sigma^2(x))}\left[\log \frac{q(z|x)}{p(z)}\right]$$

$$p \sim \mathcal{N}(0, I_K)$$
$$p(z) = \frac{1}{(2\pi)^{K/2}} e^{-\frac{1}{2}z^T z}$$
$$\log p(z) = -\frac{K}{2}\log 2\pi - \frac{1}{2}z^T z$$

$$q(\cdot|x) \sim \mathcal{N}(\mu, \sigma^2)$$
$$q(z|x) = \frac{1}{(2\pi)^{K/2}\det(\sigma^2 I_K)^{1/2}} \cdot e^{-\frac{1}{2}(z-\mu)^T(\sigma^2 I_K)^{-1}(z-\mu)}$$
$$= \frac{1}{(2\pi)^{K/2} \cdot \prod_{i=1}^{K}\sigma_i} e^{-\frac{1}{2}(z-\mu)^T\left(\frac{1}{\sigma^2}I_K\right)(z-\mu)}$$
$$\log q(z|x) = -\frac{K}{2}\log 2\pi - \sum_{i=1}^{K}\log \sigma_i - \frac{1}{2}(z-\mu)^T\left(\frac{1}{\sigma^2}I_K\right)(z-\mu)$$

$$= E_{z\sim\dots}\left[-\cancel{\frac{K}{2}}\log 2\pi - \sum_{i=1}^{K}\log\sigma_i - \frac{1}{2}(z-\mu)^T\left(\frac{1}{\sigma^2}I_K\right)(z-\mu) + \frac{K}{2}\cancel{\log} 2\pi + \frac{1}{2}z^T z\right]$$

$$= \underbrace{E_{z\sim\dots}\left[-\sum_{i=1}^{K}\log\sigma_i\right]}_{\text{Constant independent of } z} - \frac{1}{2}\cdot E_{Z\sim\mathcal{N}(\mu(x),\sigma^2(x))}\left[(z-\mu)^T\left(\frac{1}{\sigma^2}I_K\right)(z-\mu)\right] + \frac{1}{2}E_{Z\sim\mathcal{N}(\mu(x),\sigma^2(x))}\left[Z^T Z\right]$$

Let $S \sim \mathcal{N}(0, I_K)$

Rewrite: $Z = \mu + \sigma \cdot S$ — element-wise product

Thus: $(Z-\mu)^T\left(\frac{1}{\sigma^2}I_K\right)(Z-\mu) = S^T S$

$$= S_1^2 + S_2^2 + S_3^2 + \dots + S_K^2$$
$$E[S^T S] = \sum_{i=1}^{K} E[S_i^2] = \sum_{i=1}^{K} Var[S_i] + E[S_i]^2$$
$$= \sum_{i=1}^{K} 1 = K$$

$$Z^T Z = Z_1^2 + Z_2^2 + \dots + Z_K^2$$
$$E[Z^T Z] = \sum_{i=1}^{K} E[Z_i^2] = \sum_{i=1}^{K} Var[Z_i] + E[Z_i]^2$$
$$= \sum_{i=1}^{K} \sigma_i^2 + \mu_i^2$$

$$= -\sum_{i=1}^{K}\log\sigma_i - \frac{1}{2}\underbrace{K}_{\sum_{i=1}^{K} 1} + \frac{1}{2}\sum_{i=1}^{K}\sigma_i^2 + \mu_i^2$$

$$= \frac{1}{2}\sum_{i=1}^{K}\left[-\log\sigma_i^2 - 1 + \sigma_i^2 + \mu_i^2\right] \quad //$$