Andrew Elysee (aelys2)
CS446

HW3

1.

a.

$$p_\theta(x|z) = \prod_{j=1}^{G} \widehat{y_j}^{x_j}\left(1-\widehat{y_j}\right)^{1-x_j}$$

b. The output dimension of the encoder should be 2, since $z \in \mathbb{R}^2$.

c.

$$\log p_\theta(x) = \log \sum_z q_\phi(z|x)\frac{p_\theta(x,z)}{q_\phi(z|x)} \geq \sum_z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)}$$
$$= \sum_z q_\phi(z|x)\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}$$
$$= \sum_z q_\phi(z|x)\log p_\theta(x|z) + \sum_z q_\phi(z|x)\log\frac{p(z)}{q_\phi(z|x)}$$
$$= \mathbb{E}_{q_\phi}[\log p_\theta(x|z)] - D_{KL}\left(q_\phi(z|x), p(z)\right)$$

d. The KL divergence is non-symmetric and non-negative. Also $D_{KL}(q,p) = 0$ when q and p are the same distribution.

e. No, they will not be the same because the Equation 2 is more computationally expensive than Equation 1 but will be more accurate in the end.

f. No, we build the encoder because it allows us to compute a better Gaussian distribution for our data and slowly improve its accuracy in representing the data. This allows us to reduce our lower bound giving us an overall more accurate representation of our data's distribution.

g. The KL-divergence $D_{KL}\left(q_\phi(z|x), q_\phi(z|x)\right) = 0$ because $q_\phi(z|x) = q_\phi(z|x)$, This is a basic property of the KL-divergence. It makes sense intuitively because if the distributions are the same then we have a perfect representation of our data so the lower-bound between them would be 0.

h.

$$KL\left(q_\phi(z|x), p(z)\right) = \sum_z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)}$$
$$= \sum_z q_\phi(z|x)\log\frac{\frac{1}{\sqrt{2\pi\sigma_\phi^2}}\exp\left(-\frac{1}{2\sigma_\phi^2}\left(z-\mu_\phi\right)^2\right)}{\frac{1}{\sqrt{2\pi\sigma_p^2}}\exp\left(-\frac{1}{2\sigma_p^2}\left(z-\mu_p\right)^2\right)}$$

$$= \sum_z q_\phi(z|x) \log \frac{\exp\left(-\frac{1}{2\sigma_\phi^2}\left(z-\mu_\phi\right)^2\right)}{\exp\left(-\frac{1}{2\sigma_p^2}\left(z-\mu_p\right)^2\right)}$$
$$= \sum_z q_\phi(z|x) \left[-\frac{1}{2\sigma_\phi^2}\left(z-\mu_\phi\right)^2 + \frac{1}{2\sigma_p^2}\left(z-\mu_p\right)^2\right]$$
$$= \sum_z q_\phi(z|x) \frac{1}{2\sigma_\phi^2}\left[-\left(z-\mu_\phi\right)^2 + \left(z-\mu_p\right)^2\right]$$
$$= \sum_z q_\phi(z|x) \frac{1}{2\sigma_\phi^2}\left[-z^2 + 2z\mu_\phi - \mu_\phi^2 + z^2 - 2z\mu_p + \mu_p^2\right]$$

Since $var(x) = E[x^2] - E[x]^2$

$$= 1 + \sum_z \frac{1}{2\sigma_\phi^2}\left[2z\mu_\phi - \mu_\phi^2 - 2z\mu_p + \mu_p^2\right]$$

I am stuck here trying to do it without integrals but eventually we reach that.

$$KL\left(q_\phi(z|x), p(z)\right)$$
$$= \log\frac{\sigma}{\sigma} + \frac{\sigma^2 + \left(\mu_\phi - \mu_p\right)^2}{2\sigma^2} - \frac{1}{2} = \frac{1}{2} + \frac{\left(\mu_\phi - \mu_p\right)^2}{2\sigma^2} - \frac{1}{2}$$
$$KL\left(q_\phi(z|x), p(z)\right) = \frac{\left(\mu_\phi - \mu_p\right)^2}{2\sigma^2}$$

i. The Lagrangian is:

$$L\left(q_\phi(z|x), \lambda\right) = \sum_z q_\phi(z|x) \log(p_\theta) - \sum_z q_\phi(z|x) \log\frac{q_\phi(z|x)}{p(z)}$$
$$+ \lambda\left(\sum_z q_\phi(z|x) - 1\right)$$
$$0 = \frac{\partial L}{\partial q_\phi(z_i|x)} \Longrightarrow \log q_\phi(z_i|x) = \lambda - 1 + \log p_\theta(x|z_i)p(z_i)$$
$$1 = \sum_z q_\phi(z|x) = \sum_{z_i \in z} p_\theta(x|z_i)p(z_i)e^{\lambda-1} = e^{\lambda-1}\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)$$
$$e^{\lambda-1} = \frac{1}{\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)} \Longrightarrow \lambda = 1 - \log\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)$$

If we substitute in λ:

$$\log q_\phi(z_i|x) = \lambda - 1 + \log p_\theta(x|z_i)p(z_i)$$
$$\log q_\phi(z_i|x) = \log p_\theta(x|z_i)p(z_i) - \log\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)$$

Thus,

$$q_\phi(z_i|x) = \frac{p_\theta(x|z_i)p(z_i)}{\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)}$$

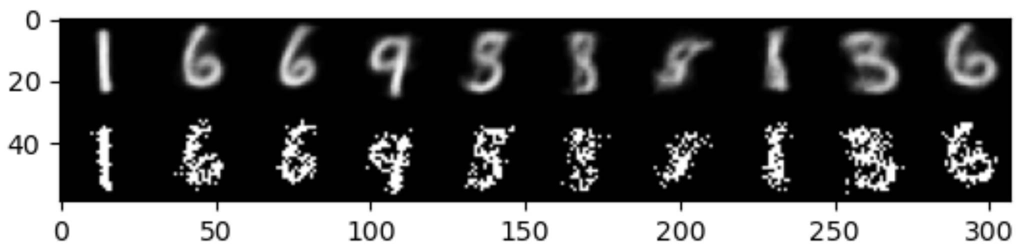
j. $q_\phi(z_i|x)$ should be (3) because according to Bayes Rule:

$$q_\phi(z_i|x) = \frac{p_\theta(x|z_i)p(z_i)}{\sum_{z_i \in z} p_\theta(x|z_i)p(z_i)} = p_\theta(z|x)$$

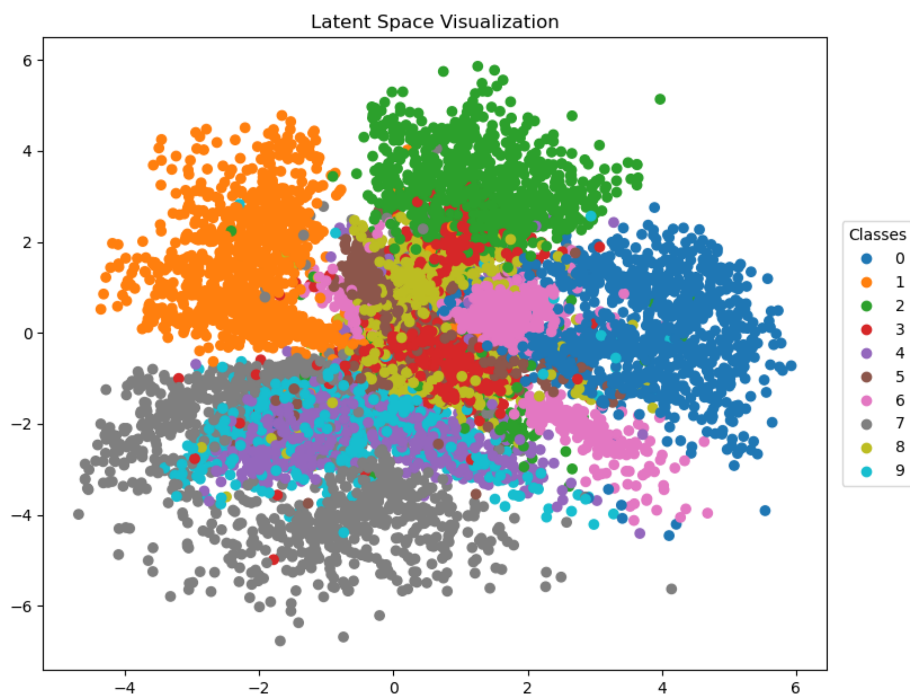
2.

a. Code
b. Code
c. Code
d. Code
e. Next page

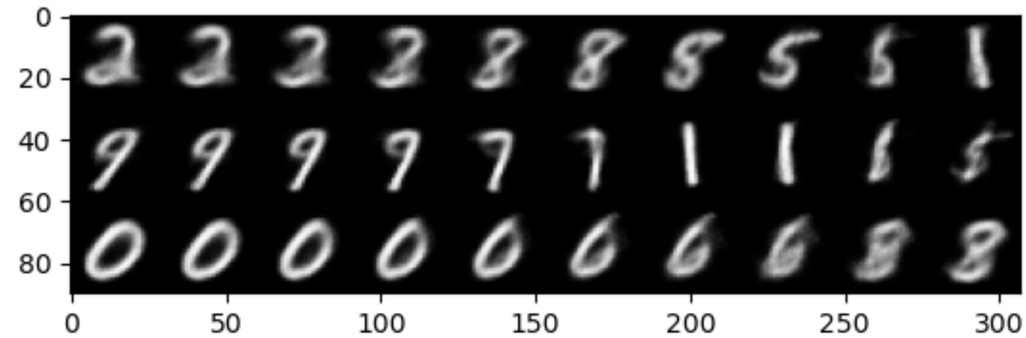
i.



ii.



iii.



3.

a. The cost function for GANs is:

$$\max_{\theta} \min_{\omega} -\sum_{x} \log D_{\omega}(x) - \sum_{z} \log(1 - D_{\omega}(G_{\theta}(z)))$$

b. Source: lecture slides

$$\min_{D} -\int_{x} p_{data} \log D(x) dx - \int_{z} p_{z}(z) \log\Big(1 - D\big(G_{\theta}(z)\big)\Big)\, dz =$$

$$\min_{D} -\int_{x} p_{data} \log D(x) + p_{G}(x) \log(1 - D(x)) dx$$

c. Using Euler-Lagrange formalism:

$$S(D) = \int_{x} L\big(x, D, \dot{D}\big) dx$$

Since there are no $\dot{D}$:

$$0 = \frac{\partial L\big(x, D, \dot{D}\big)}{\partial D} - \frac{d}{dx} \frac{\partial L\big(x, D, \dot{D}\big)}{\partial \dot{D}} = \frac{\partial L\big(x, D, \dot{D}\big)}{\partial D}$$

$$\frac{\partial L\big(x, D, \dot{D}\big)}{\partial D} = \frac{p_{data}(x)}{D(x)} + \frac{p_{G}(x)}{1 - D(x)} = 0$$

So,

$$D^{*} = \frac{p_{data}(x)}{p_{data}(x) + p_{G}(x)}$$

d.

$$-\int_{x} p_{data} \log D(x) + p_{G}(x) \log\big(1 - D(x)\big) dx =$$

$$-\int_{x} p_{data} \log \frac{p_{data}(x)}{p_{data}(x) + p_{G}(x)} + p_{G}(x) \log \frac{p_{G}(x)}{p_{data}(x) + p_{G}(x)} dx$$

Using JSD,

$$-\int_{x} p_{data} \log \frac{p_{data}(x)}{p_{data}(x) + p_{G}(x)} + p_{G}(x) \log \frac{p_{G}(x)}{p_{data}(x) + p_{G}(x)} dx$$
$$= -2\, JSD(p_{data}, p_{G}) + \log(4)$$

and

$JSD(p_{data}, p_{G}) = \frac{1}{2} KL(p_{data}, M) + \frac{1}{2} KL(p_{G}, M)$ with $M = \frac{1}{2}(p_{data} + p_{G})$

Thus, the optimal Generator $G^*(x)$ generates the distribution p_G^*, such that $p_G^* = p_{data}$.

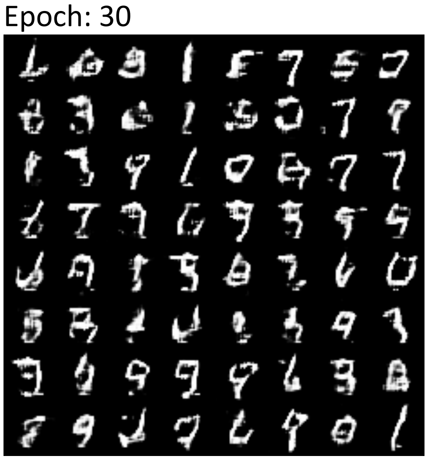
e. $KL(P_1, P_2) = Undefined,\ KL(P_1, P_3) = Undefined$. The KL divergence would be undefined for both since there are some values of x for which the probability in P_1 is 0 or P_2 is 0 and same for P_1 and P_3, while the other is nonzero. This would lead to a division by 0 in the KL divergence equation.

$$W_1(P_1, P_2) = 0.5,\ \ W_1(P_1, P_3) = 1$$

4.

a. Code
b. Code
c. Code
d. Code
e. Code
f. Epoch: 10

Epoch: 30



Epoch: 50

Epoch: 90