Problem 1 Theory

* a.

$$D_{KL}(q(x)\|p(x)) = -\int q(x)\log\left(\frac{p(x)}{q(x)}\right)dx > 0$$

$$= \left[\int q(x)\log(p(x))dx - \int q(x)\log(q(x))dx\right]$$

Cross Entropy — Entropy

$$\int q(x)\log(p(x))dx \qquad + -\int q(x)\log(q(x))dx$$

$$\int q(x)\left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x-\mu)^2\right]dx \qquad q(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu_q)^2}{2\sigma_q^2}}$$

$$= \frac{1}{2}\log(2\pi)\int q(x)dx + \frac{1}{2}\int x^2 p(x)dx \qquad \log(q(x)) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)$$

$$\frac{1}{2}\log(2\pi) + \frac{1}{2}(\mu_p^2+\sigma_p^2) \qquad \int q(x)\left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x-\mu_q)^2\right]dx$$

$$\frac{1}{2}\log 2\pi\sigma^2\int q(x)dx + \frac{1}{2\sigma}\int (x-\mu_q)^2 q(x)dx$$

$$\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_q^2) + \frac{1}{2}$$

$$D_{KL}(q(x)\|p(x)) = \frac{1}{2}\log(2\pi) + \frac{1}{2}(\mu_p^2+\sigma_p^2) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma_q^2) - \frac{1}{2}$$

$$D_{KL}(q(x)\|p(x)) = -\frac{1}{2}(1 + \log(\sigma_q^2) - \mu_p^2 - \sigma_p^2)$$

$$D_{KL}(q(x)\|p(x)) = \frac{1}{2}(1 - \log(\sigma_q^2) + \mu_p^2 + \sigma_p^2)$$

b. $\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \alpha \mathcal{L}_{prior}$

$\mathcal{L}_{recon} = \|\hat{x} - x\|^2$

$\mathcal{L}_{prior} = D_{KL}(q(z|x) \| p(z))$

What happens if α too high?

$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \alpha \mathcal{L}_{prior}$

$\mathcal{L}_{VAE} = \|\hat{x} - x\|^2 + \alpha D_{KL}(q(z|x) \| p(z))$

if α is too high $\mathcal{L}_{recon}$ gets washed out

$\mathcal{L}_{VAE}$ becomes dominated by $\mathcal{L}_{prior}$

This causes the $\mathcal{L}_{VAE}$ to output similar results as prior information dominates new information

c. Different Aspects between VAE & PCA

1. PCA can only linearly seperate data where VAE can seperate data non linearly

2. VAEs can have noise introduced to generate new outputs PCA cannot have a new noise input

Problem 2

a. False it is important to train the discriminator will improve much faster then the generator so the generator should be updated more frequently

b. Early in the training $D(G(z))$ is closer to 1 as the Generator produces fakes that are so bad that the discriminator can easily distinguish between real & fake

*c. I would rather use non-saturating cost $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m} \log(D(G(z^{(i)}))$ because there is higher gradient is higher early on resulting in faster early training

d. False the GAN is trained when $D(G(z))$ is close to 0.5, the generator should produce fakes that are so close to the real thing that the discriminator should have 50% accuracy