Setup

$$z \xrightarrow{G_\theta(z)} x \overbrace{\xrightarrow{h_\phi(x)} \text{h.val} \xrightarrow{\sigma} \text{prob.}}^{D_\phi(x)}$$

$$\mathbb{R}^k \qquad\qquad \mathbb{R}^n$$

2(a) $$L_G^{minimax}(\theta, \phi) = E_{z \sim N(0,1)} \left[\log \left(1 - D_\phi(G_\theta(z))\right) \right]$$

substituting $D_\phi(x)$ with $\sigma(h_\phi(x))$, we get

$$L_G^{minimax}(\theta, \phi) = E_{z \sim N(0,1)} \left[\log \left(1 - \sigma(h_\phi(G_\theta(z)))\right) \right]$$

$$\frac{\partial L_G^{minimax}}{\partial \theta} = E_{z \sim N(0,1)} \left[- \frac{\sigma'(h_\phi(G_\theta(z)))}{1 - \sigma(h_\phi(G_\theta(z)))} \frac{\partial}{\partial \theta} h_\phi(G_\theta(z)) \right]$$

we know that,

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

$$\therefore \frac{\partial L_G^{minimax}}{\partial \theta} = E_{z \sim N(0,1)} \left[- \frac{\sigma(h_\phi(G_\theta(z)))(1 - \sigma(h_\phi(G_\theta(z))))}{1 - \sigma(h_\phi(G_\theta(z)))} \frac{\partial}{\partial \theta} h_\phi(G_\theta(z)) \right]$$

$$= E_{z \sim N(0,1)} \left[- \underbrace{\sigma(h_\phi(G_\theta(z)))}_{\approx 0 \text{ when } D(G_\theta(z)) \approx 0} \frac{\partial}{\partial \theta} h_\phi(G_\theta(z)) \right]$$

$$\therefore \frac{\partial L_G^{minimax}}{\partial \theta} \approx 0$$

because we are multiplying $\frac{\partial}{\partial \theta} h_\phi(G_\theta(z))$ with a term ≈ 0. So, overall value ≈ 0.

- Why is this a problem for generator?

Ans: It is problematic because if the gradients are very small (≈ 0), it would be hard to train the generator as the updates to the weights θ would be negligible. The training would be essentially stalled.

3 a) $L_D(\phi, \theta) = -E_{x \sim p_{data}(x)}[\log D_\phi(x)] - E_{x \sim p_\theta(x)}[\log(1 - D_\phi(x))]$

$$= -\int p_{data}(x) \log D_\phi(x)\, dx$$

$$- \int p_\theta(x) \log(1 - D_\phi(x))\, dx$$

$$= -\int \left(p_{data}(x) \log D_\phi(x) + p_\theta(x) \log(1 - D_\phi(x)\right) dx$$

$$= \int f(D_\phi(x))\, dx$$

where $f(t) = -p_{data}(x) \log t - p_\theta(x) \log(1-t)$

and $t = D_\phi(x)$ ——— (1)

The point where $f(t)$ minimizes (and in turn $L_D(\phi, \theta)$) is where

$$f'(t) = 0$$

$$\Rightarrow \quad \frac{-p_{data}(x)}{t} + \frac{p_\theta(x)}{1-t} = 0$$

$$\Rightarrow \quad \frac{P_\theta(x)}{1-t} = \frac{P_{data}(x)}{t}$$

$$\Rightarrow \quad \frac{P_\theta(x)}{P_{data}(x)} = \frac{1-t}{t}$$

$$\Rightarrow \quad \frac{P_\theta(x)}{P_{data}(x)} = \frac{1}{t} - 1$$

$$\Rightarrow \quad \frac{P_\theta(x)}{P_{data}(x)} + 1 = \frac{1}{t}$$

$$\Rightarrow \quad \frac{P_\theta(x) + P_{data}(x)}{P_{data}(x)} = \frac{1}{t}$$

$$\Rightarrow \quad t = \frac{P_{data}(x)}{P_\theta(x) + P_{data}(x)}$$

Substitution based on ①

$\Rightarrow$

$$D_\phi(x) = \frac{P_{data}(x)}{P_\theta(x) + P_{data}(x)}$$

$$\therefore \quad D^*(x) = \frac{P_{data}(x)}{P_\theta(x) + P_{data}(x)}$$

$\Uparrow$

point where $f(D_\phi(x))$

minimizes

3(b) Given,

$$D_\phi(x) = \sigma(h_\phi(x))$$

$$= \frac{1}{1+e^{-h_\phi(x)}} \quad \text{—— (1)}$$

If $D_\phi = D^*$, then

$$D_\phi(x) = \frac{P_{data}(x)}{P_\theta(x) + P_{data}(x)} \longrightarrow \text{from (3a)}$$

From (1), (3a)

$$\frac{1}{1+e^{-h_\phi(x)}} = \frac{P_{data}(x)}{P_\theta(x) + P_{data}(x)}$$

$$1+e^{-h_\phi(x)} = \frac{P_\theta(x) + P_{data}(x)}{P_{data}(x)}$$

$$1 + e^{-h_\phi(x)} = \frac{P_\theta(x)}{P_{data}(x)} + 1$$

cancelling "1" on each side and taking Log

$$-h_\phi(x) = \log \frac{P_\theta(x)}{P_{data}(x)}$$

$$h_\phi(x) = -\log \frac{P_\theta(x)}{P_{data}(x)}$$

$$= \log \frac{P_{data}(x)}{P_\theta(x)}$$

3(c) Given,

$$L_G(\theta, \phi) = E_{x \sim p_\theta(x)}\left[\log\left(1 - D_\phi(x)\right)\right] - E_{x \sim p_\theta(x)}\left[\log D_\phi(x)\right]$$

$$= E_{x \sim p_\theta(x)}\left[\log \frac{1 - D_\phi(x)}{D_\phi(x)}\right]$$

if $D_\phi = D^*$, then

$$L_G(\theta, \phi) = E_{x \sim p_\theta(x)}\left[\log \frac{1 - \frac{p_{data}(x)}{p_{data}(x) + p_\theta(x)}}{\frac{p_{data}(x)}{p_{data}(x) + p_\theta(x)}}\right]$$

$$= E_{x \sim p_\theta(x)}\left[\log \frac{\frac{p_\theta(x)}{p_{data}(x) + p_\theta(x)}}{\frac{p_{data}(x)}{p_{data}(x) + p_\theta(x)}}\right]$$

$$= \mathbb{E}_{x \sim p_\theta(x)} \left[\log \frac{p_\theta(x)}{p_{data}(x)} \right]$$

$$= KL\left(p_\theta(x) \,||\, p_{data}(x) \right)$$

3(d) $$nLL = -E_{x \sim p_{data}(x)} \left[\log p_\theta(x) \right]$$

$$= -E_{x \sim p_{data}(x)} \left[\log \frac{p_\theta(x)}{p_{data}(x)} p_{data}(x) \right]$$

$$= -E_{x \sim p_{data}(x)} \left[\log \frac{p_\theta(x)}{p_{data}(x)} \right] - E_{x \sim p_{data}(x)} \left[\log p_{data}(x) \right]$$

A term which is independent of 'θ' and hence can be ignored for the purpose of training the generator as that is tuning 'θ'.

$$= -E_{x \sim p_{data}(x)} \left\{ \log \frac{p_\theta(x)}{p_{data}(x)} \right\}$$

$$= E_{x \sim p_{data}(x)} \left\{ \log \frac{p_{data}(x)}{p_\theta(x)} \right\}$$

$$= KL\left(p_{data}(x) \,||\, p_\theta(x)\right)$$

So, VAE is in essence trying to minimize $KL\left(p_{data}(x), p_\theta(x)\right)$ and L_G is trying to minimize $KL\left(p_\theta(x) \,||\, p_{data}(x)\right)$.

Since, KL divergence is not symmetric, I don't think we can say that these objectives are same.

But I would note that these are similar in nature and are varying because of the mechanics of the learning process. In VAE, we start with the actual samples (p_{data}) & we try to see whether model (p_θ) can generate similar sample. In GAN, the process starts with generation i.e p_θ. I feel this is cause of different KL divergences.

4(a)

$$h_\phi(x,y) = \log \frac{P_{data}(x,y)}{P_\theta(x,y)}$$

$$= \log \frac{P_{data}(x|y)}{P_\theta(x|y)} + \log \frac{P_{data}(y)}{P_\theta(y)}$$

$$= \log \frac{P_{data}(x|y)}{P_\theta(x|y)}$$

$$= \log \frac{N(\varphi(x) \mid \mu_y, I)}{N(\varphi(x) \mid \hat{\mu}_y, I)}$$

$$= \log \frac{\cancel{\frac{1}{\sqrt{2\pi}}}\, e^{-1/2(\varphi(x) - \mu_y)^2}}{\cancel{\frac{1}{\sqrt{2\pi}}}\, e^{-1/2(\varphi(x) - \hat{\mu}_y)^2}}$$

$$= \log\left(e^{-\frac{1}{2}\left(\left(\varphi(x)-\mu_y\right)^2-\left(\varphi(x)-\hat{\mu}_y\right)^2\right)}\right)$$

$$= -\frac{1}{2}\left(\left(\varphi(x)-\mu_y\right)^2-\left(\varphi(x)-\hat{\mu}_y\right)^2\right)$$

$$= -\frac{1}{2}\left(\mu_y^2 - 2\varphi(x)\mu_y - \hat{\mu}_y^2 + 2\varphi(x)\hat{\mu}_y\right)$$

$$= -\frac{1}{2}\left(-2\varphi(x)\left(\mu_y-\hat{\mu}_y\right) + \mu_y^2 - \mu_y^2\right)$$

$$= \varphi(x)\left(\mu_y-\hat{\mu}_y\right) - \frac{\mu_y^2-\hat{\mu}_y^2}{2}$$

<To Be Completed>

5(a)

$$KL\left(p_\theta(x) \,\|\, p_{data}(x)\right) = E_{x\sim N(\theta,\epsilon^2)}\left[\log \frac{\exp\left(-\frac{1}{2\epsilon^2}(x-\theta)^2\right)}{\exp\left(-\frac{1}{2\epsilon^2}(x-\theta_0)^2\right)}\right]$$

$$= E_{x\sim N(\theta,\epsilon^2)}\left[\log\left(\exp\left(-\frac{1}{2\epsilon^2}\left((x-\theta)^2-(x-\theta_0)^2\right)\right)\right)\right]$$

$$= E_{x\sim N(\theta,\epsilon^2)}\left[-\frac{1}{2\epsilon^2}\left((x-\theta)^2-(x-\theta_0)^2\right)\right]$$

$$= E_{x\sim N(\theta,\epsilon^2)}\left[\frac{1}{2\epsilon^2}\left((x-\theta_0)^2-(x-\theta)^2\right)\right]$$

$$= \frac{1}{2\epsilon^2} E_{x\sim N(\theta,\epsilon^2)}\left[-2x\theta_0+\theta_0^2+2x\theta-\theta^2\right]$$

$$= \left(\frac{\theta_0^2-\theta^2}{2\epsilon^2}\right) + \frac{1}{2\epsilon^2} E_{x\sim N(\theta,\epsilon^2)}\left[-2x\theta_0+2x\theta\right]$$

$$= \left(\frac{\theta_0^2-\theta^2}{2\epsilon^2}\right) + \left(\frac{2\theta-2\theta_0}{2\epsilon^2}\right) E_{x\sim N(\theta,\epsilon^2)}[x]$$

$$= \left(\frac{\theta_0^2 - \theta^2}{2\epsilon^2}\right) + \left[\frac{2\theta - 2\theta_0}{2\epsilon^2}\right]\theta$$

$$= \frac{\theta_0^2 - \theta^2 + 2\theta^2 - 2\theta_0\theta}{2\epsilon^2}$$

$$= \frac{\theta_0^2 + \theta^2 - 2\theta_0\theta}{2\epsilon^2}$$

$$= \frac{(\theta - \theta_0)^2}{2\epsilon^2}$$

Hence proved that

$$KL(P_\theta(x) \,||\, P_{data}(x)) = \frac{(\theta - \theta_0)^2}{2\epsilon^2}$$

(b) In the above expression, if $\epsilon \to 0$ the $KL \to \infty$ i.e. KL divergence would be a very high number.

$$\frac{\partial KL}{\partial \theta} = \frac{1}{2\epsilon^2}\left(2(\theta - \theta_0)\right)$$

$$= \frac{\theta - \theta_0}{\epsilon^2}$$

This derivative value also be very high if $\theta \neq \theta_0$. The issue that it will cause during trainings would be the updates to weights are very large and the training will not be stable. The weights will keep moving in different directions without converging in a stable manner.

(c) All the loss functions we have discussed so far have either a very high value or zero. Essentially, the gradient descent process which navigates the loss curve becomes a process of random search which is randomly looking for the point where loss = 0. There is no smooth path which leads to that point. So, we need a function where the loss curve is smooth and there is a path to minima from all the points.