1) a)

$$h_{t+1} = \cancel{Wh}\ \phi(z_{t+1})$$

$$z_{t+1} = W h_t$$

$$\Rightarrow \nabla_W h_{t+1} = \nabla_W \phi(z_{t+1}) = \left(\nabla_{z_{t+1}} \phi\right)\left(\nabla_W z_{t+1}\right)$$

$$\nabla_W h_{t+1} = \phi'(z_{t+1}) \left(h_t + W \nabla_W h_t\right)$$

$$\cancel{\Rightarrow \nabla h_t = \frac{\nabla_W h_{t+1}}{}}$$

~~Also~~ Also, $\phi(z) = \sigma(z) - 0.5$

$$\Rightarrow \phi'(z) = \sigma'(z)$$

$$\Rightarrow \nabla_W h_{t+1} = \sigma'(z_{t+1}) \left(h_t + W \nabla_W h_t\right)$$

$$\Rightarrow \boxed{\nabla_W h_t = \left(\frac{\nabla_W h_{t+1}}{\sigma'(z_{t+1})} - h_t\right) \frac{1}{W}}$$

b) ~~for vanishing gradient,~~ for

b) for $h_t = 0 \quad \forall t$

$$\Rightarrow \nabla_w h_t = \frac{1}{w}\left(\frac{\nabla_w h_{t+1}}{\sigma'(z_{t+1})}\right)$$

now ~~if~~ $z_{t+1} = w h_t = 0 \quad \forall t$

$$\Rightarrow \nabla_w h_t = \frac{1}{w}\left(\frac{\nabla_w h_{t+1}}{\sigma'(0)}\right) = \left(\frac{4}{w}\right)\left(\nabla_w h_{t+1}\right)$$

for vanishing gradient,

$$\nabla_w h_t < \nabla_w h_{t+1}$$

$$\Rightarrow \frac{4}{w} < 1$$

$$\Rightarrow \boxed{w > 4}$$

for <u>exploding gradient</u>,

$$\nabla_w h_t > \nabla_w h_{t+1}$$

$$\Rightarrow \frac{4}{w} > 1 \quad \Rightarrow \boxed{w < 4}$$

$$\Rightarrow \boxed{\alpha = 4}$$

2) a) False. $h_t \neq h_{t-1}$ due to the bias term.

~~Counter~~ Eg:-

Let $h_0 = 0$, $x_1 = 0$, $c_0 = 0$

$\Rightarrow$ $f_1 = \sigma(b_f)$

$i_1 = \sigma(b_i)$

$\tilde{c}_1 = \cancel{\tanh(b_i} \tanh(b_c)$

$c_1 = f_1 \odot c_0 + i_1 \odot \tilde{c}_1$

$c_1 = \cancel{0+} \sigma(b_i) \odot \tanh(b_c)$

$o_1 = \sigma(b_o)$

$h_1 = o_1 \odot \tanh(c_1)$

$h_1 = \sigma(b_o) \odot \tanh(\sigma(b_i) \odot \tanh(b_c)) \neq 0$

$\Rightarrow \boxed{h_1 \neq h_0}$

Scanned with CamScanner

b) False. Even though due to f_t being 0 the gradient flowing through the forget gate will be zero but ~~th~~ there will still be ~~gr~~ gradients flowing through the input and output gate to the previous time steps.

c) True. f_t, o_t, i_t are all sigmoid functions ~~th~~ and sigmoid outputs a value between 0 an 1.

d) False. ~~Alth~~ Although each entry of f_t, i_t, o_t will be between 0 and 1, but the sum of all entries will ~~to~~ not be equal 1.

~~eg~~ say f_t is a n-dimensional vector,

$\Rightarrow$ ~~$f_{t_j} \leq 1$~~

$$0 < f_{t_j} < 1$$

$$\Rightarrow \sum_{j=1}^{n} f_{t_j} \leq n$$

for f_t to be a probability distribution

$$\sum_{j=1}^{n} f_{t_j} = 1$$

3)a) Dimension of f_t : 1×1
Dimension of i_t : 1×1
Dimension of o_t : 1×1
Dimension of h_t : 1×1

b) ~~a)~~ $h_0 = 0, \quad c_0 = 0$

$x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x_1 = \begin{bmatrix} 0.5 \\ -1 \end{bmatrix}$

$y_0 = 0.5, \quad y_1 = 0.8$

~~$f_1 = \sigma(1$~~ $W_f = [1, 2] \quad U_f = [0.5] \quad b_f = [0.2]$
$W_i = [-1, 0] \quad U_i = [2] \quad b_i = [-0.1]$

$W_c = [1, 2] \quad U_c = [1.5], \quad b_c = [0.5]$

$W_o = [3, 0], \quad U_o = [-1], \quad b_o = [0.8]$

for $t = 1$

$f_1 = \sigma(1 + 0 + 0.2) = \sigma(1.2) = 0.769$
$i_1 = \sigma(-1 + 0 - 0.1) = 0.249$

$\hat{c}_1 = \tanh(1 + 0 + 0.5) = 0.905$
$c_1 = f_1 \odot c_0 + i_1 \odot \tilde{c}_1 = 0.226$

$o_1 = \sigma(3 + 0 + 0.8) = 0.978$

$\boxed{h_1 = o_1 \odot \tanh(c_1) = 0.217}$

Scanned with CamScanner

for $\underline{t = 2}$

$$f_2 = \sigma\left(-1.5 + \frac{0.216}{2} + 0.2\right) = 0.233$$

$$i_2 = \sigma(-0.5 + 0.108 \cancel{+} - 0.1) = 0.458$$

$$\tilde{c}_2 = \tanh(-1.5 + 0.324 + 0.5) = -0.587$$

$$c_2 = f_2 \odot c_1 + i_2 \odot \tilde{c}_2 = -0.216$$

$$o_2 = \sigma(1.5 \cancel{+} - 0.216 + 0.8) = 0.889$$

$$\boxed{h_2 = o_2 \odot \tanh(c_2) = -0.189}$$

c) for $t = 1$

$$MSE = \|h_1 - y_0\|_2^2 = (0.217 - 0.5)^2 = \boxed{0.0801}$$

for $t = 2$

$$MSE = \|h_2 - y_1\|_2^2 = (-0.189 - 0.8)^2 = \boxed{0.978}$$

$$\boxed{\text{overall MSE Loss} = \frac{1}{2}(0.0801 + 0.978) = 0.529}$$

4) a) $D_{KL}(q(z|x) \,||\, p(z))$

$$= -\int_{z \sim \mathcal{N}(0,1)} q(z|x) \log \frac{p(z)}{q(z|x)} dz$$

$$\cancel{= -\int q(z|x) \log p(z)}$$

$$= -\int q(z|x) \log p(z) dz + \int q(z|x) \log(q(z|x)) dz \quad -(1)$$

now $$p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad -(2)$$

$$q(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \quad -(3)$$

$$\Rightarrow \log p(z) = \log\left(\frac{1}{\sqrt{2\pi}}\right) - \left(\frac{z^2}{2}\right)$$

$$\log q(z|x) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(z-\mu)^2}{2\sigma^2}$$

$$\Rightarrow D_{KL}(q(z|x) \,||\, p(z)) = -\int q(z|x) \log\left(\frac{1}{\sqrt{2\pi}}\right) dz + \int q(z|x) \frac{z^2}{2} dz$$

$$+ \int q(z|x) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) dz - \int q(z|x) \frac{(z-\mu)^2}{2\sigma^2} dz$$

$$= \frac{1}{2}\log(2\pi) + \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q}[z^2] \cancel{+} \frac{1}{2}\log(2\pi\sigma^2)$$

$$-\frac{1}{2\sigma^2}\int q(z|x)\, z^2\, d\cancel{z} - \frac{1}{2\sigma^2}\int q(z|x)\,\mu^2\, d\cancel{z}\, z$$

$$+\frac{1}{2\sigma^2}\int q(z|x)(2z\mu)\, dz$$

$$= \frac{1}{2}\log\left(\frac{1}{\sigma^2}\right) + \frac{1}{2}\mathop{\mathbb{E}}_{z\sim q}[z^2] - \frac{1}{2\sigma^2}\cancel{\mathbb{E}}\mathop{\mathbb{E}}_{z\sim q}[z^2] - \frac{\mu^2}{2\sigma^2}$$

$$+ \frac{\mu}{\cancel{2}\sigma^2}\mathop{\mathbb{E}}_{z\sim q}[z]$$

now, $\mathop{\mathbb{E}}_{z\sim q}[z] = \mu$

and $\mathop{\mathbb{E}}_{z\sim q}[z^2] = \sigma^2 + \mu^2$

$$\Rightarrow D_{KL}(q(z|x)\,||\,p(z)) = \frac{1}{2}\log\left(\frac{1}{\sigma^2}\right) + \frac{(\sigma^2+\mu^2)}{2}\left(1 - \frac{1}{\sigma^2}\right) + \frac{\mu^2}{2\sigma^2}$$

$$\boxed{D_{KL}(q(z|x)\,||\,p(z)) = \frac{1}{2}\log\left(\frac{1}{\sigma^2}\right) + \frac{\sigma^2+\mu^2}{2} - \frac{1}{2}}$$

Scanned with CamScanner

b) $L_{VAE} = L_{recon} + \alpha L_{prior}$

If α is too high, ~~the VAE w~~ all the inputs will be mapped to the sampe standand normal ~~dis~~ distribution. This mean irrespective of the input similar encoded vector z will be selected thus all the reconstructed results will look the same.

c) Differences between VAE an ~~a~~ PCA :-

i) PCA is a linear mapping between the input space to the encoded space, whears VAE maps the inputs ~~zi~~ using nonlinear transformation to ~~t~~ a probability distribution.

ii) The ~~eato~~ basis vectors ~~as~~ ~~learned~~ obtained using PCA are orthogonal whereas the features learnt using VAE might be correlated.

Scanned with CamScanner

5) a) False, because in case of a multimodal ~~data~~ data, over training the generator for a fixed discriminator can make the generator produce only one single mode of the data which fools the discriminator the best irrespective of the sampled noise. In other ~~words~~ words it can lead to mode collapse.

b) Early in the training $D(G(z))$ will be close to 0. This happens because initially the ~~dis~~ generator's output and the real data will be very be very differ-ent and it will be easier for the discriminator to ~~figu~~ learn to discriminate between them.

c) Non-saturating cost should be used to train the GAN, because initially when the discriminator is easily able to identify ~~fake~~ generated data, $D(G(z))$ will be close to 0. In this case the non-saturating loss will give higher magnitude ~~the~~ gradients which will allow the generator to run faster.

d) False. The GAN will be trained when the generator is able to generate data as close as possible to real data. In this case the discriminator will not be able to distinguish between real & fake data. In this case the minima of Loss is achieved at $D(G(z)) = 0.5$