# ANEEK DAS - CSC2515 - HW4.

Q1. Multilayer Perceptron .

i/p's $\longrightarrow$ $(x_1, x_2)$ ; $x_1, x_2$ are scalars .

o/p's $\longrightarrow$ $(y_1, y_2)$ ; $y_1 = \min(x_1, x_2)$.

$y_2 = \max(x_1, x_2)$.

All hidden units $\longrightarrow$ ReLU activation ; $\text{ReLU}(x) = \max(0, x)$.

All o/p units $\longrightarrow$ Linear/ no activation .

$$y_1 = \min(x_1, x_2)$$

Now, $\min(x_1, x_2) = -\max(0, x_1 - x_2) + x_1$

$= -\text{ReLU}(x_1 - x_2) + x_1$.

Now, we have to represent $x_1$ in terms of ReLU .

- If $x$ is '+ve' we can represent it as $\max(0, x_1) = \text{ReLU}(x_1)$
- If $x$ is '-ve' we can represent it as $-\max(0, -x_1) = -\text{ReLU}(-x_1)$

So,

$$\boxed{\min(x_1, x_2) = -\text{ReLU}(x_1 - x_2) + \text{ReLU}(x_1) - \text{ReLU}(-x_1)}$$

$y_1$

Now, $y_2 = \max(x_1, x_2)$

$\max(x_1, x_2) = \max(0, x_1 - x_2) + x_2$

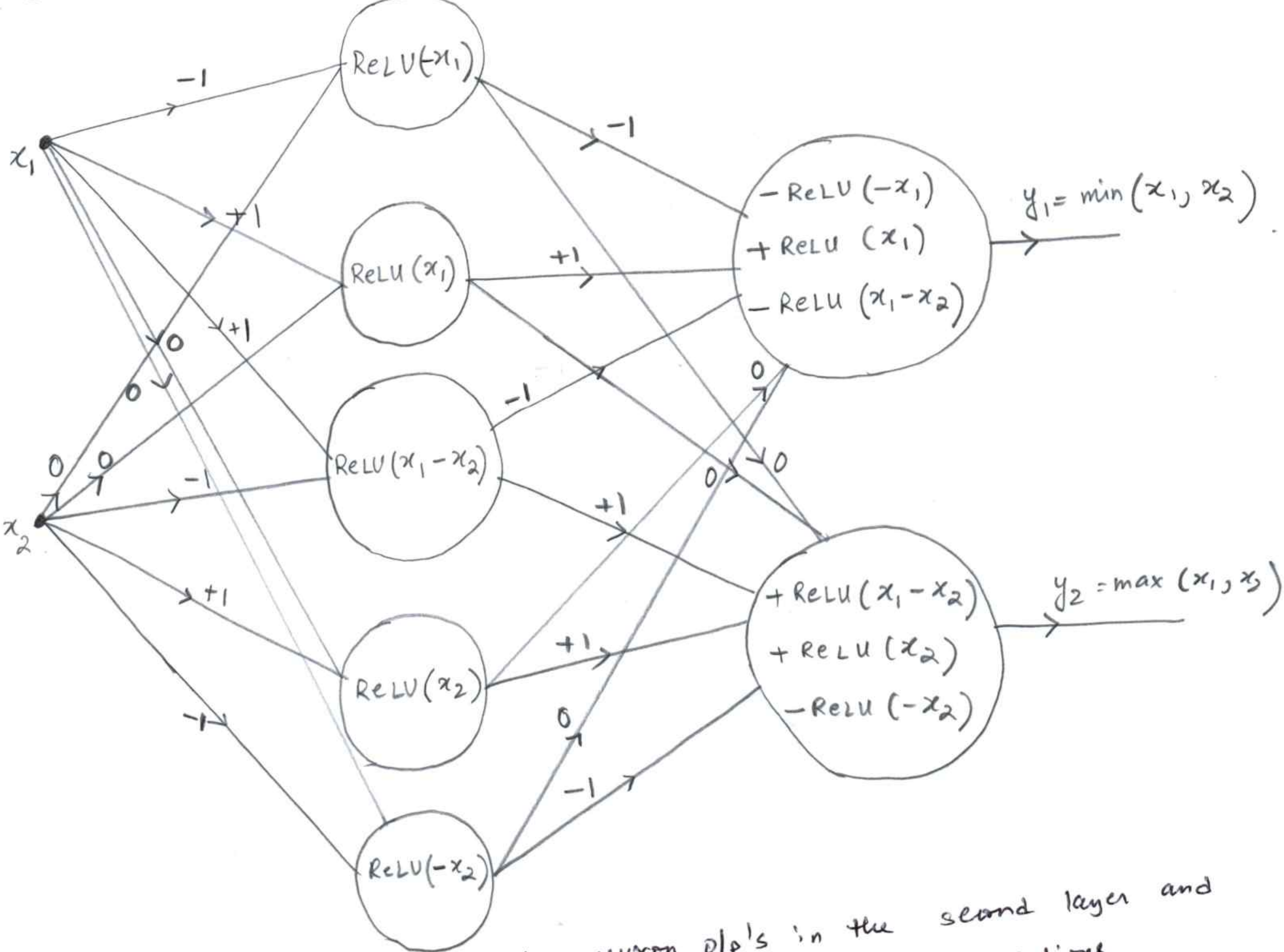
Now, similarly, $x_2 = \text{ReLU}(x_2) - \text{ReLU}(-x_2)$.

So, $\max(x_1, x_2) = \max(0, x_1 - x_2) + \max(0, x_2) - \max(0, -x_2)$.

$$\boxed{\max(x_1, x_2) = \text{ReLU}(x_1 - x_2) + \text{ReLU}(x_2) - \text{ReLU}(-x_2)}$$

$y_2$

Network diagram -




The weights of each of the neuron o/p's in the second layer and the weights of the i/p's are written along the connections.

<u>All biases are set to 0.</u> } Ans.

So, the wt. vector of $x_1 \longrightarrow [-1, +1, +1, 0, 0]$

the wt. vector of $x_2 \longrightarrow [0, 0, -1, +1, -1]$.

The learned quantity is written inside the neuron.

wts. of ReLU($-x_1$) $\longrightarrow [-1, 0]$

wts of ReLU ($x_1$) $\longrightarrow [+1, 0]$

wts of ReLU ($x_1 - x_2$) $\longrightarrow [-1, 1]$

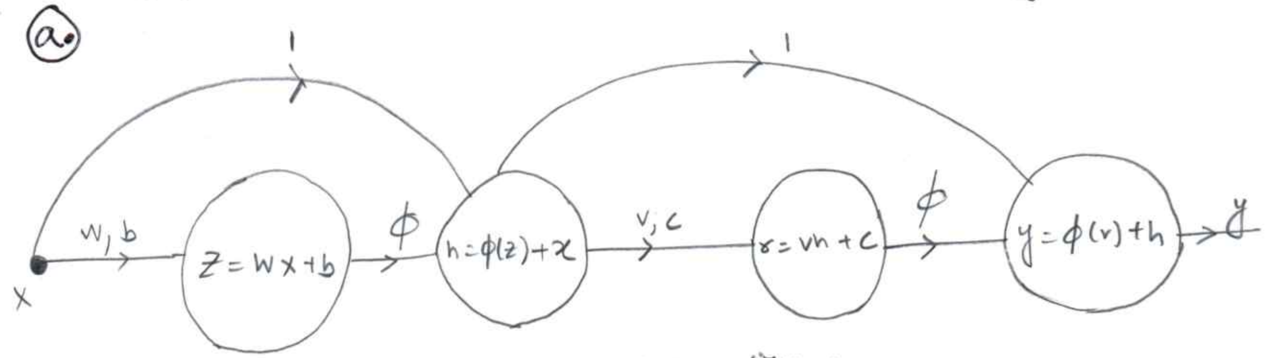
wts of ReLU ($x_2$) $\longrightarrow [0, +1]$

wts. of ReLU ($-x_2$) $\longrightarrow [0, -1]$

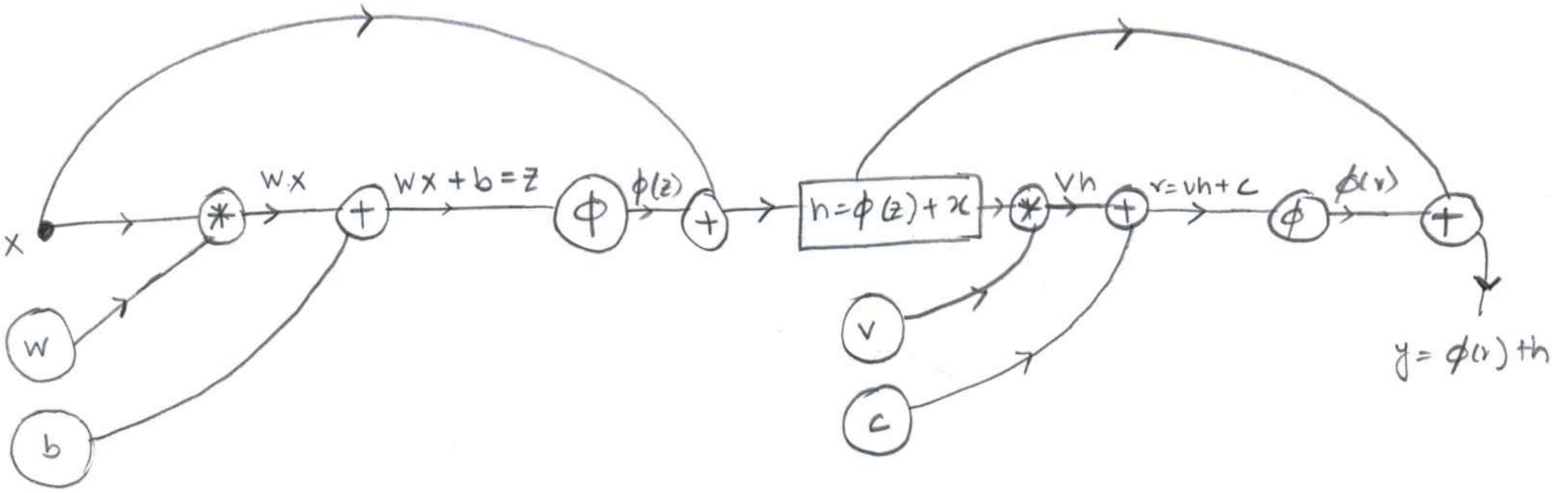
} Ans.

Q2. Back propagation through Residual Block.

(a)




So, the computational Graph looks like:




(b) Determine the backprop vales for gradient w.r.t. w, b, v & c.

Lets say the o/p (actual o/p) is $\bar{y}$, and the loss $l$ is defined as $l(y, \bar{y})$.

So, we have to find $\frac{\partial l}{\partial v}$, $\frac{dl}{dc}$, $\frac{dl}{dw}$, $\frac{dl}{\partial b}$.

first, $\frac{dl}{dv}$

$$\frac{dl}{dv} = \frac{dl}{dy} \cdot \frac{\partial y}{\partial r} \cdot \frac{\partial r}{\partial v}$$

putting the values,

$$= \frac{\partial l}{\partial y} \cdot \frac{\partial}{\partial r}\left(\phi(r) + h\right) \cdot \frac{\partial}{\partial v}(vh + c)$$

$$= \frac{\partial l}{\partial y} \cdot \left[\phi'(r) + \cancel{\frac{\partial h}{\partial r}}^{0}\right] \cdot h = \boxed{\frac{\partial l}{\partial y} \cdot \phi'(r) \cdot h = \frac{dl}{dv}}$$

Now, $\frac{\partial l}{\partial c} = \frac{\partial l}{\partial y} \cdot \frac{\partial y}{\partial r} \cdot \frac{\partial r}{\partial c}$

$$= \frac{\partial l}{\partial y} \cdot \frac{\partial}{\partial r}\left(\phi(r) + h\right) \cdot \frac{\partial}{\partial c}\left(Vh + c\right) = \boxed{\frac{\partial l}{\partial y} \cdot \phi'(r) = \frac{dl}{\partial c}}$$

Now, $\frac{dl}{dw}$ = gradient through (1) wt. layer $(V, c)$ + gradient through (2) skip connection.

(2) $\frac{dl}{dw} = \frac{dl}{dy} \cdot \frac{dy}{dh} \cdot \frac{dh}{dz} \cdot \frac{dz}{dw}$

$$= \frac{dl}{dy} \cdot \frac{d}{dh}\left(\phi(r) + h\right) \cdot \frac{d}{dz}\left(\phi(z) + x\right) \cdot \frac{d}{dw}\left(wx + b\right)$$

$$= \frac{dl}{dy} \cdot \phi'(z) \cdot x$$

(1) $\frac{dl}{dw} = \frac{dl}{dy} \cdot \frac{dy}{dr} \cdot \frac{dr}{dh} \cdot \frac{dh}{dz} \cdot \frac{dz}{dw}$.

$$= \frac{dl}{dy} \cdot \frac{d}{dr}\left(\phi(r) + h\right) \cdot \frac{d}{dh}\left(Vh + c\right) \cdot \frac{d}{dz}\left(\phi(z) + x\right) \cdot \frac{d}{dw}\left(wx + b\right)$$

$$= \frac{dl}{dy} \cdot \phi'(r) \cdot V \cdot \phi'(z) \cdot x$$

So, $\frac{dl}{dw}$ = (1) + (2) = $\boxed{\frac{dl}{dy} \cdot \phi'(z) \cdot x\left[\phi'(r) \cdot V + I\right] = \frac{dl}{dw}}$

I is identity matrix.

Now, $\frac{dl}{db}$ = gradient through (1) wt. layer + gradient through (2) skip connection.

(1) $\frac{dl}{db} = \frac{dl}{dy} \cdot \frac{dy}{dr} \cdot \frac{dr}{dh} \cdot \frac{dh}{dz} \cdot \frac{dz}{db}$.

$$= \frac{dl}{dy} \cdot \frac{d}{dr}\left(\phi(r) + h\right) \cdot \frac{d}{dh}\left(Vh + c\right) \cdot \frac{d}{dz}\left(\phi(z) + x\right) \cdot \frac{d}{db}\left(wx + b\right)$$

$$= \frac{dl}{dy} \cdot \phi'(r) \cdot V \cdot \phi'(z) \cdot 1$$

② $\frac{dl}{db} = \frac{dl}{dy} \cdot \frac{dy}{dh} \cdot \frac{dh}{dz} \cdot \frac{dz}{db}$

$= \frac{dl}{dy} \cdot \phi'(z).$

So, $\frac{dl}{db} = ① + ② = \boxed{\frac{dl}{dy} \cdot \phi'(z) \cdot [\phi'(r) \cdot V + I] = \frac{dl}{db}}$

---

3. EM for Probabilistic PCA.

ⓐ

Z (latent code vector) drawn from Std. Gaussian $N(0, I)$

we will consider $z$ to be the 1-D projection.

$$\boxed{Z \sim N(0, 1)}$$

and, $\boxed{x | z \sim N(u.z, \sigma^2 I)}$

I changed the mean from $z.u$ to $u.z$ for the following reason.

Here is my approach.

$X \in \mathbb{R}^{J \times N}$ (there are 'N' samples of X, each of which has 'J' dimensions).

each sample $x_i \in \mathbb{R}^J$ where $i = 1$ to $N$.

$u$ is the principal component / parameter vector. i.e, when $x^{(i)}$ is projected on $u$ we get $z^{(i)}$ which is the latent representation of $x^{(i)}$ in 1-D space.

So, $u \in \mathbb{R}^{J \times 1}$ (transform J dimension to 1).

and, $z = u^T X$

So, $z \in \mathbb{R}^{1 \times N}$ ('N' samples each of which is 1-D).

So, $z^{(i)} \in \mathbb{R}$ (scalar).

So, when projecting back $z$ to a higher dimension to get approximation of X we get $X = u.z$.

So, $X = U.Z$

$(J \times 1) \cdot (1 \times N) = (J \times N)$ which is the dimension for $X$.

Hence, my reason to change $Z.U$ to $U.Z$.

So, we have

$$z^{(i)} \sim N(0, 1)$$

$$x^{(i)} | z^{(i)} \sim N(u.z^{(i)}, \sigma^2).$$

Now, from the relations given in the Appendix, we can infer the values of $p(x)$ and $p(z|x)$

$$p(x) \sim N(A\mu + b, A\Sigma A^T + S)$$

where, $A = U$ $\mu = 0$ $b = 0$ $\Sigma = I$ $S = \sigma^2 I$.

So, we get

$$\boxed{p(x^{(i)}) \sim N(0, u.u^T + \sigma^2 I)}.$$

and finally, $p(z|x) = N(C(A^T S^{-1}(x - b) + \Sigma^{-1}\mu), C)$.

where $C = (\Sigma^{-1} + A^T S^{-1} A)^{-1}$

Replacing the values we get. $p(z^{(i)} | x^{(i)}) =$

$$C(A^T S^{-1}(x - b) + \Sigma^{-1}\mu)$$

$$= C(U^T (\sigma^2 I)^{-1} x)$$

$$= C \cdot \left(\frac{U^T X}{\sigma^2}\right).$$

$$C = (1 + U^T (\sigma^2)^{-1} U)^{-1}$$

$$C = \left(1 + \frac{U^T U}{\sigma^2}\right)^{-1}$$

$$C = \left(\frac{\sigma^2 + U^T U}{\sigma^2}\right)^{-1}$$

$$C = \frac{\sigma^2}{U^T U + \sigma^2}.$$

The reason we can divide this is $U^T.U$ and $\sigma^2$ are both scalars.

So, $\mu = \frac{\sigma^2}{\sigma^2 + u^T.u} \cdot \frac{u^T x}{\sigma^2} = \frac{u^T x}{\sigma^2 + u^T.u}$.

$$c = \frac{\sigma^2}{\sigma^2 + u^T.u}.$$

$q(z^{(i)})$.

So, $\boxed{p(z^{(i)} \mid x^{(i)}) \sim N\left(\frac{u^T.x}{\sigma^2 + u^T.u}, \frac{\sigma^2}{\sigma^2 + u^T.u}\right)}$.

(*) $\sigma^2 + u^T.u$ is a scalar, so, we can divide the term by them, if, they were vectors we would multiply the numerators by $(\sigma^2 + u^T u)^{-1}$.

Now, $m = E[z \mid x] := \mu = (\sigma^2 I + u^T.u)^{-1} \cdot u^T.x$

and, $s = E[z^2 \mid x]$

we know, $var(z \mid x) = E[z^2 \mid x] - E[z \mid x]^2$.

substituting the values we get.

$$(\sigma^2 + u^T.u)^{-1} . \sigma^2 = E[z^2 \mid x] - ((\sigma^2 + u^T.u)^{-1} . u^T x)^2$$

$$s = (\sigma^2 + u^T.u)^{-1} . \sigma^2 I + ((\sigma^2 + u^T.u)^{-1} . u^T x)^2$$

$$\boxed{m^{(i)} = \frac{u^T x^{(i)}}{\sigma^2 + u^T.u}} \qquad \boxed{s^{(i)} = \frac{\sigma^2}{\sigma^2 + u^T.u} + \frac{(u^T x^{(i)})^2}{(\sigma^2 + u^T.u)^2}}$$

Since, $u^T x^{(i)} = z^{(i)}$ (scalar) & $\sigma^2 + u^T u$ is scalar, $m^{(i)}$ & $s^{(i)}$ are scalars

---

(b) M-Step.

$$u_{new} \leftarrow \underset{u}{\operatorname{argmax}} \sum_{i=1}^{N} E_{q(z^{(i)})} \left[\log p(z^{(i)}, x^{(i)})\right].$$

$E_{q(z^{(i)})}$ is the expected value of $q(z)$ for all values $z^{(i)}$.

Lets start with the $\log P(z^{(i)}, x^{(i)})$ term.

$$\log\left[p(x^{(i)}, z^{(i)})\right] = \log\left[p(x^{(i)} \mid z^{(i)}) \cdot p(z^{(i)})\right]$$

$$= \log p(x^{(i)} \mid z^{(i)}) + \log p(z^{(i)}).$$

Now, we know, $x^{(i)} \mid z^{(i)}$ is multivariate Gaussian as $x^{(i)}$ has $J$ dims.

$$x^{(i)} \mid z^{(i)} \sim N(u \cdot z^{(i)}, \sigma^2)$$

we know, for multivariate Gaussian $p(x; \mu, \Sigma) =$

$$\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1}(x-\mu)}.$$

putting the values we get.

$$p(x^{(i)} \mid z^{(i)}) = \frac{1}{(2\pi)^{J/2} \cdot \sigma} \cdot e^{-\frac{1}{2\sigma^2}(x^{(i)} - u \cdot z^{(i)})^T \cdot (x^{(i)} - u \cdot z^{(i)})}.$$

$\sigma^2$ is scalar.

So, $$\log p(x^{(i)} \mid z^{(i)}) = \log \frac{1}{(2\pi)^{J/2} \cdot \sigma} - \frac{1}{2\sigma^2}(x^{(i)} - u \cdot z^{(i)})^T \cdot (x^{(i)} - u \cdot z^{(i)}).$$

Now, $p(z^{(i)})$ is a univariate dist. as $z^{(i)}$ is a scalar.

$$p(z^{(i)}) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(z^{(i)})^2}.$$

So, $$\log p(z^{(i)}) = \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(z^{(i)})^2.$$

So, $$\log\left[p(x^{(i)}, z^{(i)})\right] = \log \frac{1}{(2\pi)^{J/2}\sigma} - \frac{1}{2\sigma^2}(x^{(i)} - u z^{(i)})^T \cdot (x^{(i)} - u \cdot z^{(i)})$$

$$+ \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(z^{(i)})^2$$

Now, since, we will optimize w.r.t. $u$, we can remove the terms that are not dependent on $u$.

So, we have

$$-\frac{1}{2\sigma^2}(x^{(i)} - Uz^{(i)})^T \cdot (x^{(i)} - U \cdot z^{(i)})$$

All other terms would yield '0' if differentiated w.r.t. U, so, we drop them.

So, we have after expanding

$$-\frac{1}{2\sigma^2}\left\{x^{(i)T} \cdot x^{(i)} - 2x^{(i)T} \cdot U \cdot z^{(i)} + U^T \cdot U (z^{(i)})^2\right\}$$

Now, putting back in our original equation and equating to 0

$$\frac{\partial}{\partial U}\sum_{i=1}^{N} E_{q(z^{(i)})}\left[-\frac{1}{2\sigma^2}\left(x^{(i)T} \cdot x^{(i)} - 2x^{(i)T} \cdot U z^{(i)} + U^T \cdot U (z^{(i)})^2\right)\right] = 0$$

$$-\frac{1}{2\sigma^2}\frac{\partial}{\partial U}\left[\sum_{i=1}^{N} E_{q(z^{(i)})}\left(x^{(i)T} \cdot x^{(i)} - 2x^{(i)T} \cdot U \cdot z^{(i)} + U^T \cdot U (z^{(i)})^2\right)\right] = 0$$

Now, using linearity of Expectation.

$$\frac{\partial}{\partial U}\sum_{i=1}^{N} x^{(i)T} \cdot x^{(i)} - 2x^{(i)T} \cdot U \cdot E[z^{(i)}] + U^T \cdot U \; E[z^{(i)2}] = 0$$

Now, $E[z^{(i)}] = m^{(i)}$ & $E[(z^{(i)})^2] = S^{(i)}$

$$\sum_{i=1}^{N}\frac{\partial}{\partial U}\left(x^{(i)T} \cdot x^{(i)} - 2x^{(i)T} \cdot U \cdot m^{(i)} + U^T \cdot U \cdot S^{(i)}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{N} -2x^{(i)T} \cdot m^{(i)} + 2U \cdot S^{(i)} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \cancel{2}x^{(i)T} \cdot m^{(i)} = \sum_{i=1}^{N} \cancel{2} \cdot U \cdot S^{(i)}$$

$$\Rightarrow \boxed{U = \frac{\sum_{i=1}^{N} x^{(i)T} \cdot m^{(i)}}{\sum_{i=1}^{N} S^{(i)}}} \longleftarrow \text{Ans}$$