605 796 227

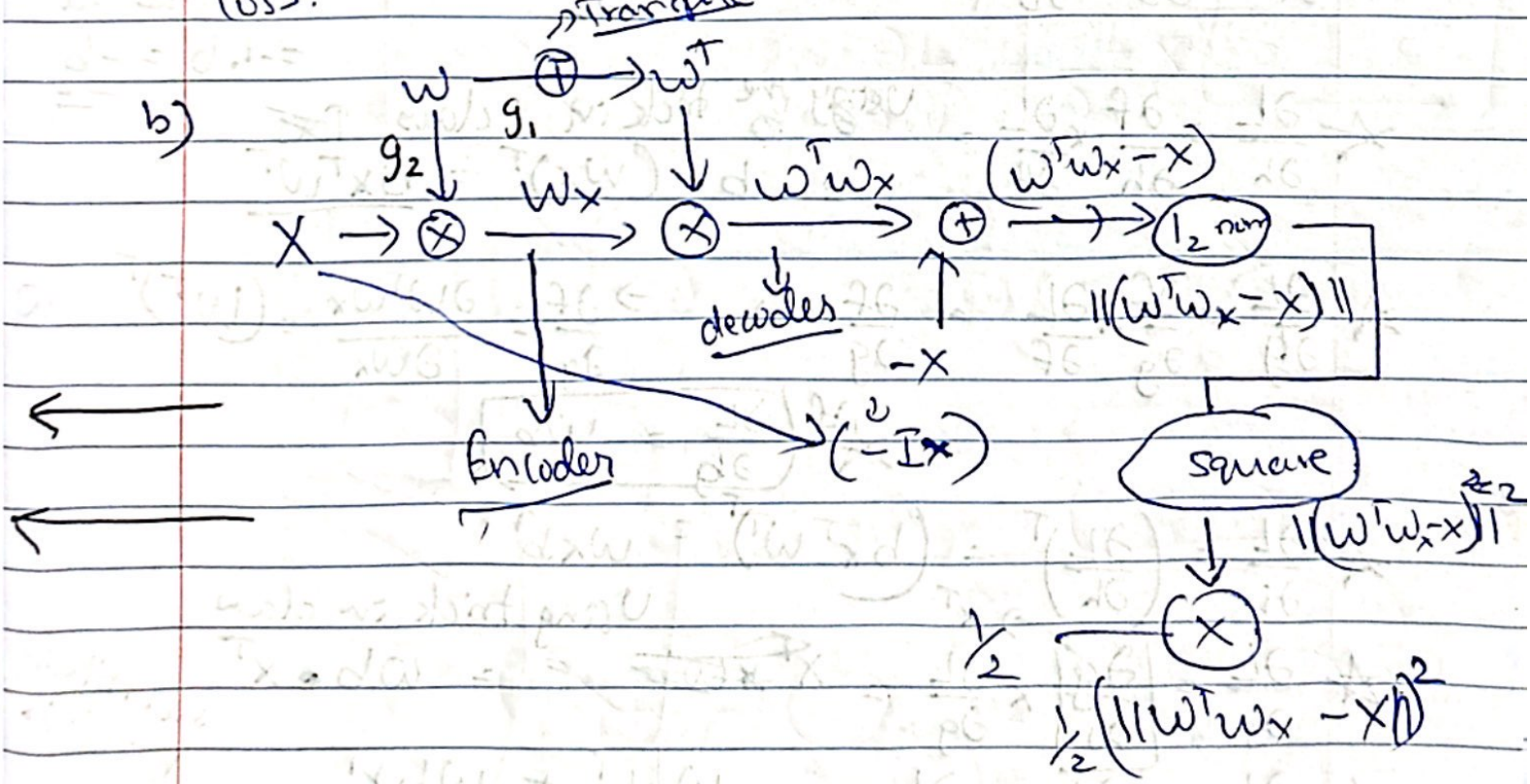
ECE 147

Homework 3 KRISH PATEL

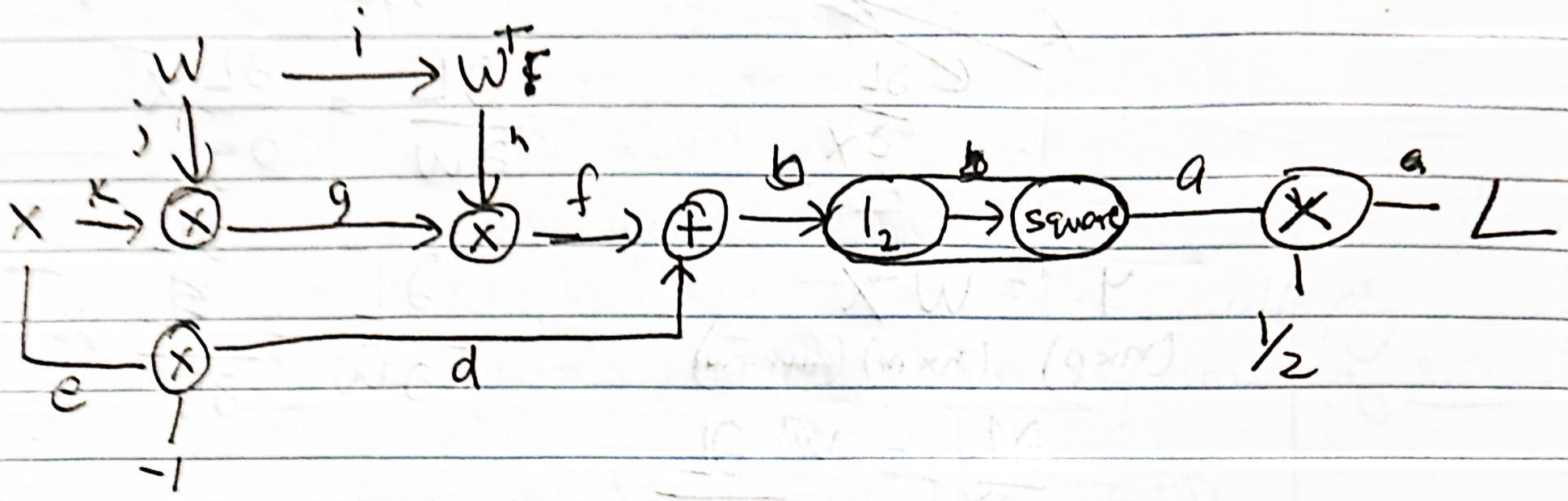
(Q1) $W \in \mathbb{R}^{m \times n}$

a) When W is a square matrix, (assumption) W^TW would minimize the value of the transformation (If W^T is orthogonal, W^TW would be equal to I, thus minimizing $(W^TWx - x)$. Now, moving to this example, W reduces the dimensionality of x, W^T is used for the reconstruction (similar to PCA). By minimizing the difference of the x' (reduced x dimensionity, which is the reconstructed to x's original dimension) and x, we are minimizing information loss.

b)



count



c) W has 2 paths, one through Wx and one through $W \to ① \to W^T \to W^TWx$.
These converge at W^TWx
Thus, we can just add the ~~two~~ →

$$\frac{\partial L}{\partial W} = \frac{\partial g_1}{\partial W} \cdot \frac{\partial L}{\partial g_1} + \frac{\partial g_2}{\partial W} \cdot \frac{\partial L}{\partial g_2}$$ where g_1 and g_2 are defined in the figure behind.

$a = \frac{1}{2} b^T b$ $\quad 2b \cdot \frac{1}{2} = b$

d) $\frac{\partial L}{\partial a} = \frac{1}{2}$ $\quad \frac{\partial L}{\partial b} = \frac{\partial a}{\partial b} \times \frac{\partial L}{\partial a} = 2 \cdot \frac{2b \times 1}{2} = b$

$\frac{\partial L}{\partial f} = \frac{\partial f}{\partial b} \times \frac{\partial L}{\partial b} = 1 \times b = b \therefore \frac{\partial L}{\partial b} = b$ $\quad \frac{\partial L}{\partial e} = \frac{\partial d}{\partial e} \times \frac{\partial L}{\partial d} = -1 \cdot b = -b$

$\frac{\partial L}{\partial h} = \frac{\partial f}{\partial h} \times \frac{\partial L}{\partial f} =$ Using the b trick in class,
$= b \times (Wx)^T = bx^TW^T$

$\frac{\partial L}{\partial g} = \frac{\partial f}{\partial g} \times \frac{\partial L}{\partial f} = \frac{\partial f}{\partial g} \times b \quad \to \frac{\partial f}{\partial g} = \frac{\partial W^TWx}{\partial Wx} = (W^T)^T = W$

$\frac{\partial L}{\partial b} = Wb$

$\frac{\partial L}{\partial i} = \left(\frac{\partial L}{\partial h}\right)^T = (bx^TW^T)^T = Wxb^T \to x^T$

Using trick in class

$\frac{\partial L}{\partial j} = \frac{\partial g}{\partial j} \times \frac{\partial L}{\partial g} = Wb \cdot x^T$

$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial j} + \frac{\partial L}{\partial i} = Wxb^T + Wbx^T$ substituting $b = W^TWx - x$

$= \boxed{Wx(W^TWx - x)^T + W(W^TWx - x)x^T}$

(02) I am a ECE 147 student

(03) NNDL to the Rescue!!!!

a) Input layer → D neurons.
Hidden layer → H neurons
Output layer → C neurons = 7

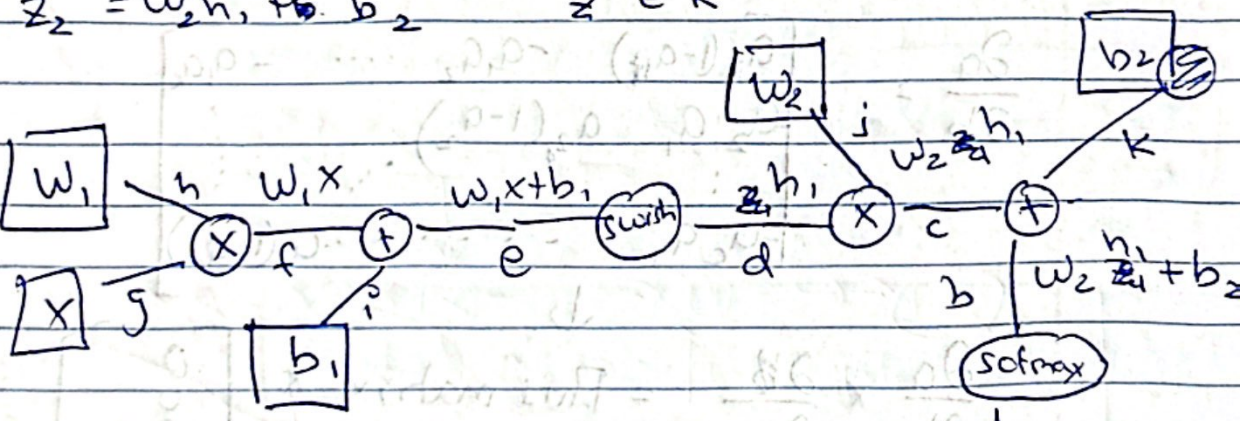
$$\text{swish}(k) = \frac{k}{1+e^{-k}} = k\,\sigma(k)$$

↓ sigmoid

$z_1 = W_1 x + b_1, \quad z \in R^H$

$h_1 = \text{swish}(z_1)$

$z_2 = W_2 h_1 + b_2 \quad z \in R^C$



$\text{Loss} = -\log P_k$

b) $\nabla_{W_2} L = ?, \quad \nabla_{b_2} L$

where $b = W_2 h_1 + b_2$

$\frac{\partial L}{\partial a} = ?$

$\frac{\partial L}{\partial b} = \frac{\partial a}{\partial b} \times \frac{\partial L}{\partial a}$:= ON NEXT PAGE

$= \frac{\partial(-\log P_k)}{\partial a}$

$\frac{\partial L}{\partial a} = [0 \; 0 \; \cdots \; -1/P_k \; 0 \; 0 \; 0]$ ← where $-1/P_k$ is for the correct class.

$$\frac{\partial a}{\partial b} = \begin{bmatrix} \quad \end{bmatrix} \in \mathbb{R}^{C \times C}$$

↓

Jacobian matrix

$$a_{ij} = \frac{e^{b_{ij}}}{\sum_{k=1}^{C} e^{b_{ik}}}$$

when $j = l$ $\quad \frac{\partial a_{ij}}{\partial b_{il}} = a_{ij}(1 - a_{ij})$

$j \neq l$ $\quad \frac{\partial a_{ij}}{\partial b_{il}} = -a_{ij} \cdot a_{il}$

$$\frac{\partial a}{\partial b} = \begin{bmatrix} a_1(1-a_1) & -a_1 a_2 & \cdots & -a_1 a_C \\ -a_2 a_1 & a_2(1-a_2) & \ddots & \vdots \\ \vdots & & & \\ -a_C a_1 & \cdots & \cdots & a_C(1-a_C) \end{bmatrix}$$

↓

$$\frac{\partial a}{\partial b} \times \frac{\partial L}{\partial a} = \text{This matrix} \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1/p_k \\ 0 \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} -a_1 a_k \\ -a_2 a_k \\ \vdots \\ (a_k)(1-a_k) \\ \vdots \\ -a_C a_k \end{bmatrix} \times \frac{-1}{p_k}$$

$$= \begin{bmatrix} -\sigma(x_1)\sigma(x_k) \\ -\sigma(x_2)\sigma(x_k) \\ \vdots \\ \sigma(x_k)(1-\sigma(x_k)) \\ \vdots \\ \sigma(x_c)(\sigma(x_k)) \end{bmatrix} \begin{pmatrix} -1 \\ \overline{\sigma_k} \end{pmatrix} = \begin{bmatrix} \sigma(x_1) \\ \sigma(x_2) \\ \vdots \\ 1(1-\sigma(x_k)) \\ \vdots \\ \sigma(x_c) \end{bmatrix}$$

$$= \left[\text{softmax}(b) - \begin{bmatrix} 0 \\ 0 \\ \vdots \\ k \\ \vdots \end{bmatrix} \right]$$

$$\frac{\partial L}{\partial b} = \text{softmax}(b) - e_k$$

Substituting $b = z_2$

THUS, !!!

$$\frac{\partial L}{\partial b_2} = \frac{\partial b}{\partial b_2} \times \frac{\partial L}{\partial b} = 1 \times (\text{softmax}(z_2) - e_k) \times$$

($\frac{\partial b}{\partial b_2} \downarrow 1$; $z_2 \downarrow$ forward pass value.)

$$\frac{\partial L}{\partial W_2} = \begin{bmatrix} \frac{\partial b}{\partial j} \\ \downarrow \\ h_1 \end{bmatrix}^T \times \frac{\partial b}{\partial c} \times \frac{\partial L}{\partial b} = k(\text{softmax}(z_2) - e_k) h_1^T$$

($\frac{\partial b}{\partial c} \downarrow 1$)

Using the weird shortcut rule in Hans.

c) Now that we have $\frac{\partial L}{\partial d} = W_2^T(\text{softmax}(b) - e_k)$

we can back prop.

Back prop through swish. ∂

test

$$\text{swish}(k) = \frac{k}{1+e^{-k}}$$

derivative: $\frac{\partial(\text{swish}(k))}{\partial k}$

$$= \frac{1\cdot(1+e^{-k}) - k(-1)\cdot(1+e^{-k})}{(1+e^{-k})^2}$$

$$= \frac{1(1+e^{-k}) + k(e^{-k})}{(1+e^{-k})^2}$$

$$\Rightarrow = \left(\frac{1}{\sigma(k)} + k\left(\frac{1}{\sigma(k)} - 1\right)\right)\sigma(k)^2$$

$$= \sigma(k) + k(\sigma(k) - \sigma(k)^2)$$

$$= \sigma(k) + k\sigma(k)(1-\sigma(k))$$

Substituting $z_1 = w_1 x + b$ and $h_1 = \text{swish}(w_1 x + b)$ and $k = z_1$

$$= (h_1 + \sigma(z_1)\odot(1-h_1))$$

$$\sigma(z_1) + h_1 - h_1 * \sigma(z_1)$$
$$= (h_1 + \sigma(z_1)(1-h_1))$$

Now, $\frac{\partial L}{\partial h} = \frac{\partial f}{\partial h} \times \frac{\partial L}{\partial f} =$

$$= \left((h_1 + \sigma(z_1)\odot(1-h_1))\odot\left(W_2^T(\text{softmax}(z_2) - e_k)\right)\right)^T$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial f} = \left((h_1 + \sigma(z_1)\odot(1-h_1))\odot\left(W_2^T(\text{softmax}(z_2) - e_k)\right)\right)$$