Suman Nepal.
6.036 pset 4.

1. Neural Networks.

$[x_1, x_2] = [3, 14]$

$z_1 = 3+0-1 = 2$; $f(z_1) = 2$.

$z_2 = 14-1 = 13$; $f(z_2) = 13$.

$z_3 = -3-1 = -4$; $f(z_3) = 0$

$z_4 = -14-1 = -15$: $f(z_4) = 0$

$u_1 = 2+13 = 15$; $f(u_1) = 15$

$u_2 = -2-13-2 = -17$: $f(u_2) = 0$

$o_1 = \frac{e^{15}}{e^{15}+1}$; $o_2 = \frac{1}{e^{15}+1}$

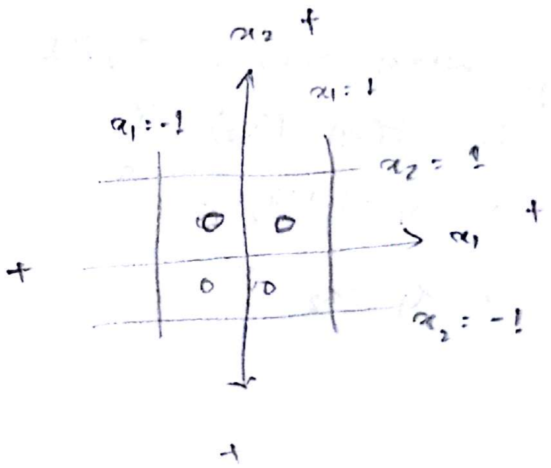
b. Draw decisions boundaries.

$z_1 = 0$ $\Rightarrow \boxed{x_1 - 1 = 0}$

$z_2 = 0$ $\boxed{x_2 - 1 = 0}$

$z_3 = 0$ $\boxed{-x_1 - 1 = 0}$

$z_4 = 0$ $\boxed{-x_2 - 1 = 0}$



c.

$f(z_1) + f(z_2) + f(z_3) + f(z_4) = 0$

Since, $f(x)$ is all positive, the above expression becomes zero when all of them are zero.

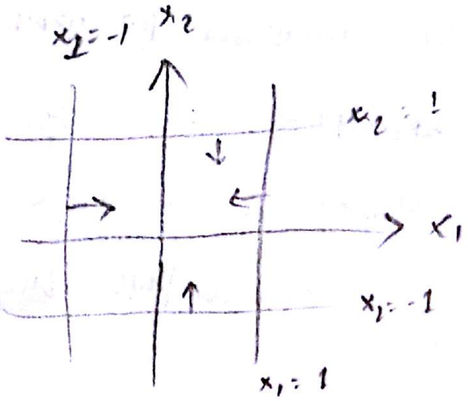
$f(z_1) = 0 \quad \forall \; x_1 - 1 \le 0 \Rightarrow x_1 \le 1$

$f(z_2) = 0 \quad \forall \; x_2 - 1 \le 0 \Rightarrow x_2 \le 1$

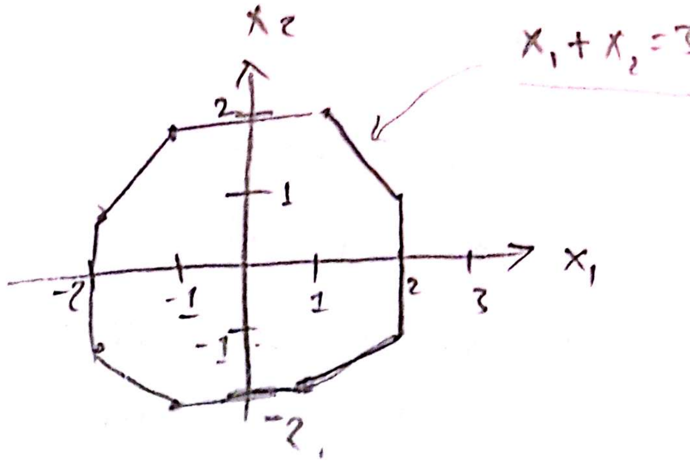
$f(z_3) = 0 \quad \forall \; -x_1 - 1 \le 0 \Rightarrow x_1 \ge -1$

$f(z_4) = 0 \quad \forall \; -x_2 - 1 \le 0 \Rightarrow x_2 \ge -1$

So, the region bounded is



$f(z_1) + f(z_2) + f(z_3) + f(z_4) = 1$



for $z_1 > 1, z_2 > 1$; $f(z_3) = 0$ $f(z_4) = 0$

$f(z_1) + f(z_2) = 1$

$\Rightarrow x_1 - 1 + x_2 - 1 = 1$

$\Rightarrow x_1 + x_2 = 3.$

Similarly, we can obtain eq^{n}s for different segments.

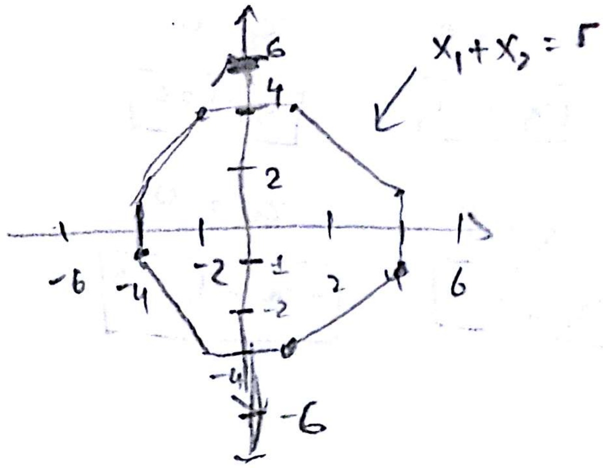
$f(z_1) + f(z_2) + f(z_3) + f(z_4) = 3$

Similarly, when $z_1 > 1$, $z_2 > 1 \Rightarrow f(z_3) = 0$, $f(z_4) = 0$

ie $f(z_1) + f(z_2) = 3$

$\Rightarrow x_1 - 1 + x_2 - 1 = 3$

$\Rightarrow x_1 + x_2 = 5$



d. <u>Modified softmax</u>: with β.

This increases the probability range for the higher values outputted. ie if $\beta = 2$, the exponential normalization of output covers higher range than $\beta = 1$.

It concentrates higher prob to higher activation.

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_{i=1}^{k} \frac{\partial C}{\partial z_i^{l+1}} \cdot \frac{\partial z_i^{l+1}}{\partial z_j^l}$$

$$= \sum_k \delta_k^{l+1} \frac{\partial}{\partial z_i^l} \left(a_k^l \cdot w_{kj}^{l+1} + b_j^{l+1} \right)$$

$$= \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot \frac{\partial}{\partial z_i^l} f(z_i^l)$$

$$= \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot f'(z_i^l)$$

(ii). $$\frac{\partial C}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial w_{jk}^l} = \delta_j^l \cdot \frac{\partial}{\partial w_{jk}^l} \left(w_{jk}^l a^{l-1} + b^l \right)$$

$$= \delta_j^l \cdot \cancel{w_{jk}^l}\, a_k^{l-1}$$

$$\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial b_j^l} = \delta_j^l \cdot 1 = \delta_j^l$$

(iii) Calculate feed forward signals.

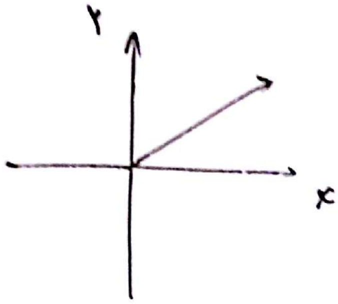
Calculate output Error & cost / error 'C'.

Backpropagate the error by weights in previous layer and gradients of associated activation function.

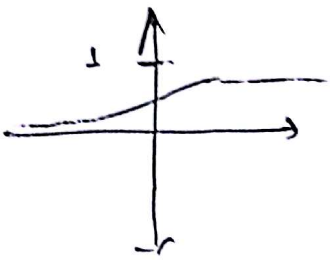
Calculating the gradients $\frac{\partial C}{\partial w}$ on backpropagated ~~ep~~ signals.

$w = w - \eta \frac{\partial C}{\partial w}$ for every weight matrix.

(ii).

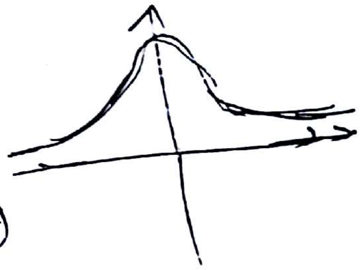


Relu function.



sigmoid function.

$$\frac{\partial}{\partial x}\frac{1}{(\bar{e}^{x}+1)} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}}{(1+e^{-x})^2} - \frac{1}{(1+e^{-x})^2}$$



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

(b)

(iii). It is a smooth function almost linear at its mean.
It's nonlinearity and have finite limits.

(c) $$\frac{\partial C}{\partial w_1} = \frac{\partial}{\partial w_1}\tfrac{1}{2}(y-t)^2 = (y-t)\cdot\frac{\partial y}{\partial z_2}\cdot\frac{\partial z_2}{\partial a_1}\cdot\frac{\partial a_1}{\partial z_1}\cdot\frac{\partial z_1}{\partial w_1}$$

$$= (y-t)\cdot\sigma(z_2)(1-\sigma(z_2))\cdot w_2\cdot w_1$$

$$\frac{\partial C}{\partial w_2} = \frac{\partial}{\partial w_2}(y-t)^2/2 = (y-t)\cdot\frac{\partial y}{\partial z_2}\cdot\frac{\partial z_2}{\partial w_2} = (y-t)\cdot\sigma(z_2)(1-\sigma(z_2))\cdot a_1$$

Part. 2

Input.	h_0	h_1	h_2	h_3	h_4	h_5
1	0	0	0.76	-0.76	0.76	-0.381
2.	0.76	-0.76	0.381	0.76	-0.76	X

(b) The visible state carries information about the previous states during the training since we have this recurrent loop.

(c). $W_{cc} = -50$

d. $t = 1$

$x = 3$ $\qquad\qquad\qquad\qquad$ $y =$

$w_1 = 0.01$

$w_2 = -5$

$b = -1.$

$$\frac{\partial C}{\partial w_1} = (y-t)\,\sigma(z_1)\,[1-\sigma(z_1)]\,.\,w_2 w_1$$

$$= \cancel{1-0.6528}\ 0.6588$$

$$w_1 \leftarrow w_1 - 0.6588\eta$$

$$\cancel{w_1 \leftarrow w_1 - 0.6588\eta}$$

Similarly,

$$w_2 \leftarrow w_2 + 0.0013\,\eta$$