Jonathan Nusantara jan 26s

Problem 1. (15 points). Consider one layer of a ReLU network. The feature vector is d dimensional $\vec{x}$. The linear transformation is a $m \times d$ dimensional matrix W. The output of the ReLU network is a m dimensional vector y given by $\max\{\mathbf{0}, W\vec{x}\}$. This is a component-wise max function.

- Suppose $\vec{x}$ is fixed, and all its entries are non-zero.
- Suppose the entries in W are all independent, and distributed accoding to a Gaussian distribution with mean 0, and standard deviation 1 (a $N(0,1)$ distribution).

1. Show that the expected number of non-zero entries in the output is $m/2$.
2. Suppose $\|\vec{x}\|_2^2 = \sigma^2$, what is the distribution of each entry in Wx (the output before applying ReLU function)?
3. What is the mean of each entry in y (after ReLU function)?

(1)

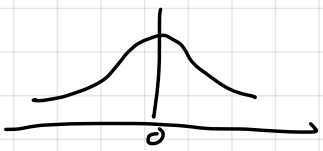
$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \quad W = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{md} \end{bmatrix} \quad W\bar{x} = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{md} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$d \times 1$ $\quad$ $m \times d$

$$= \begin{bmatrix} x_1 w_{11} + \cdots + x_d w_{1d} \\ \vdots \\ x_1 w_{m1} + \cdots + x_d w_{md} \end{bmatrix}$$

$m \times 1$

We know that W has distribution, where mean = 0



let $W\bar{x} = q \Rightarrow \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix}$

So, $\mathbb{E}(w) = 0$

What is $\mathbb{E}(W\bar{x})$? We know that $\bar{x}$ is fixed or constant.

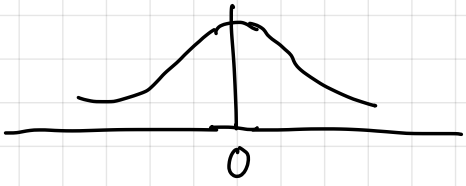
$$\begin{aligned} \mathbb{E}(q_1) &= \mathbb{E}(x_1 w_{11} + \cdots + x_1 w_{1d}) \\ &= \mathbb{E}(x_1 w_{11}) + \cdots + \mathbb{E}(x_d w_{1d}) \quad \text{property of } \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) \\ &= x_1 \mathbb{E}(w_{11}) + \cdots + x_d \mathbb{E}(w_{1d}) \quad \mathbb{E}(cX) = c\,\mathbb{E}(X) \\ &= x_1(0) + \cdots + x_d(0) \\ &= 0 \end{aligned}$$

$\rightarrow$ This is true for all $\mathbb{E}(q_i)$ $\quad i = 1, \ldots, m$

So, $\mathbb{E}(q) = \mathbb{E}(W\bar{x}) = 0$

Since W has a Gaussian dist.

$W\bar{x}$ where $\bar{x}$ is constant also has Gaussian dist



, mean / expected value $= 0$

From plot above, we can see that half > 0

half < 0.

$Y = \max\{0, W\bar{x}\}$. If $W\bar{x} \geq 0$, $Y > 0$

If $W\bar{x} < 0$, $Y = 0$

there's half chance that $W\bar{x} > 0$ and $Y > 0$

So, the expected value of $Y > 0$, is $\frac{1}{2} \times m$ (# of entries)

$= \boxed{\frac{m}{2}}$

② $\sigma^2 = \|\bar{x}\|_2^2 = x_1^2 + x_2^2 + \ldots + x_d^2$ (mean, variance)

We know $W \sim N(0,1) = N(\mu, \sigma^2)$

Based on properties:

$\mathbb{E}(cX) = c\,\mathbb{E}(X)$

so, $\mathbb{E}(Wx) = x\,\mathbb{E}(W)$, since x is fixed / constant

$V(cX) = c^2 V(X)$

So, $V(Wx) = x^2 V(W)$

$Wx \sim N(\mu, \sigma^2)$

Find the distribution of each entry of $W\bar{x} \rightarrow q_1, q_2 \ldots, q_m$

$q_1 = x_1 W_{11} + \ldots + x_d W_{1d}$

$W_{1i}\bar{x}_i = x_1 W_{11} + \ldots + x_d W_{1d}$

$W_{1i}\bar{x}_i \sim N(\mu, \sigma^2)$ ← find the distribution

From Q1, we have this proof:

$$\begin{aligned}
\mathbb{E}(q_1) &= \mathbb{E}(x_1 W_{11} + \ldots + x_1 W_{1d}) \\
&= \mathbb{E}(x_1 W_{11}) + \ldots + \mathbb{E}(x_d W_{1d}) \\
&= x_1 \mathbb{E}(W_{11}) + \ldots + x_d \mathbb{E}(W_{1d}) \\
&= x_1 (0) + \ldots + x_d (0) \\
&= \underline{\underline{0}} \qquad \rightarrow \text{this is the case for all } \mathbb{E}(q_i),\ i = 1, \ldots, m
\end{aligned}$$

Let's find similar proof for Variance

$$\begin{aligned}
V(q_1) &= V(x_1 W_{11} + \ldots + x_d W_{1d}) \\
&= V(x_1 W_{11}) + \ldots + V(x_d W_{1d}) \qquad \text{since entries of } W \text{ are all independent.} \\
&= x_1^2 V(W_{11}) + \ldots + x_d^2 V(W_{1d}) \qquad V(cX) = c^2 V(X) \text{ and the } x\text{'s in our application is constant.} \\
&= x_1^2 (1) + \ldots + x_d^2 (1) \\
&= x_1^2 + x_2^2 + \ldots + x_d^2 \\
&= \underline{\underline{\sigma^2}}
\end{aligned}$$

↑ this is the case for $V(q_1), V(q_2) \ldots, V(q_m)$

So, distribution for each entry of $Wx \rightarrow \boxed{N(0, \sigma^2)}$

③

$$\hat{y} = \begin{bmatrix} \max(0, x_1 w_{11} + \dots + x_d w_{1d}) \\ \vdots \\ \max(0, x_1 w_{m1} + \dots + x_d w_{md}) \end{bmatrix} = \begin{bmatrix} \max(0, q_1) \\ \vdots \\ \max(0, q_m) \end{bmatrix}$$

Find μ of each entry in $\hat{y}$:

We have a Gaussian dist for y, since it is a linear comb of gaussian and linear comb of gaussian is a gaussian by definition.

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\mathbb{E}(y_i) = \mathbb{E}(\max(0, q_i)) \qquad \sigma^2 = \sigma^2, \ \mu = 0$$

$$= \int_{-\infty}^{\infty} \max(0, q_i)\, p(q_i)\, dq_i$$

$$= \int_{-\infty}^{\infty} \max(0, q_i) \frac{e^{-(q_i-0)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} dq_i$$

$$= \int_{0}^{\infty} q_i \frac{e^{-q_i^2/2\sigma^2}}{\sigma\sqrt{2\pi}} dq_i$$

← bound change to $(0, \infty)$ because when $q_i < 0$, value $= 0$

$$= \left| -\frac{\sigma e^{-q_i^2/2\sigma^2}}{\sqrt{2\pi}} \right|_0^{\infty}$$

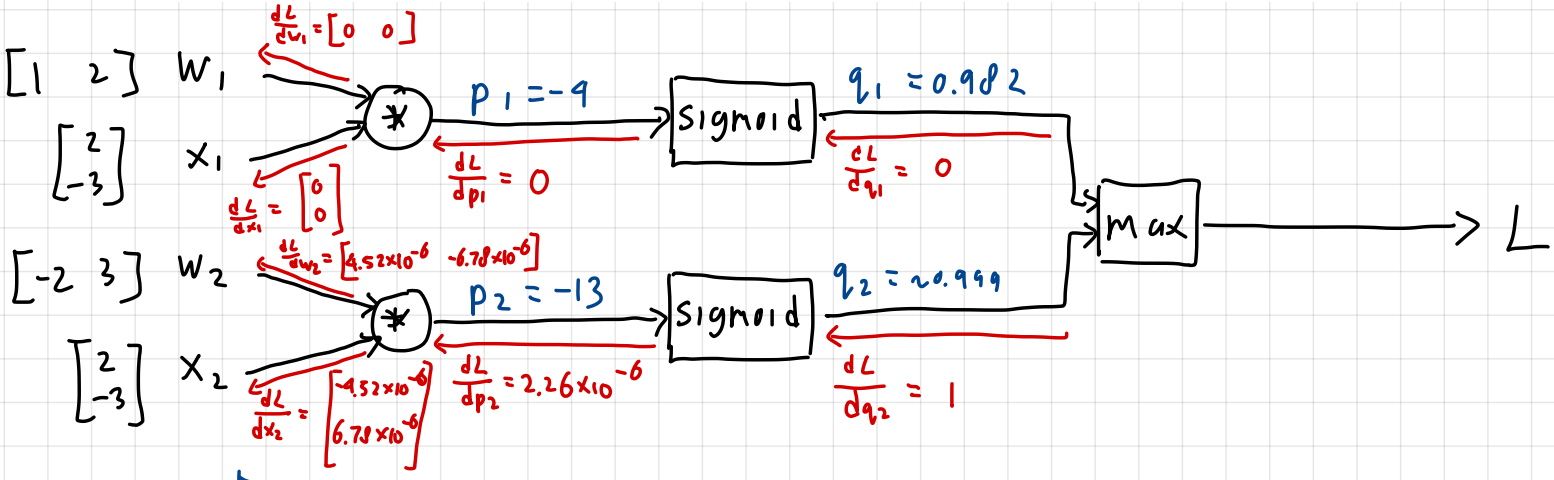
$$= 0 - \left(-\frac{\sigma(1)}{\sqrt{2\pi}}\right)$$

$$\mu = \frac{\sigma}{\sqrt{2\pi}}$$

Problem 2. (10 points). Consider the setting as in the previous problem, with $m = 2$, and $d = 2$. Let

$$W = \begin{bmatrix} 1 & 2 \\ -2 & 3 \end{bmatrix}, \vec{x} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}. \qquad \sigma = \frac{1}{1+e^{-x}}$$

Consider the function $L = \max\left\{\sigma(W_{(1)}\vec{x}), \sigma(W_{(2)}\vec{x})\right\}$, where σ is the Sigmoid function and $W_{(i)}$ denotes the ith row of W. Please draw the computational graph for this function, and compute the gradients (which will be Jacobians at some nodes!).



Forward :

$$p_1 = [1 \quad 2] \cdot \begin{bmatrix} 2 \\ -3 \end{bmatrix} = 2 + (-6) = \boxed{-4}$$

$$p_2 = [-2 \quad 3] \cdot \begin{bmatrix} 2 \\ -3 \end{bmatrix} = -4 + (-9) = \boxed{-13}$$

$$q_1 = \frac{1}{1+e^{-4}} = \boxed{0.98201379}$$

$$q_2 = \frac{1}{1+e^{-13}} = \boxed{0.99999773}$$

$$L = \max(q_1, q_2) = q_2 = \boxed{0.99999773}$$

Backward :

$$\frac{dL}{dq_1} = \boxed{0}, \quad \text{because } q_2 > q_1$$

$$\frac{dL}{dq_2} = \boxed{1}, \quad \text{because } q_2 > q_1$$

$$\frac{dL}{dp_1} = \frac{dL}{dq_1} \cdot \frac{dq_1}{dp_1} = 0 \cdot \left(\sigma(q_1)(1-\sigma(q_1))\right) = 0 \cdot \left(\frac{1}{1+e^{-4}} \cdot \left(1 - \frac{1}{1+e^{-4}}\right)\right)$$

$$= 0 \cdot 0.01766 = \boxed{0}$$

$$\frac{dL}{dp_2} = \frac{dL}{dq_2} \cdot \frac{dq_2}{dp_2} = 1 \cdot \left(\sigma(q_2)(1-\sigma(q_2))\right) = 1 \cdot \left(\frac{1}{1+e^{-13}} \cdot \left(1 - \frac{1}{1+e^{-13}}\right)\right) = \boxed{2.26 \times 10^{-6}}$$

$$\frac{dL}{dW_1} = \frac{dL}{dp_1} \cdot \frac{dp_1}{dW_1} = \frac{dL}{dp_1} \cdot \begin{bmatrix} \frac{dp_1}{dW_{1(1)}} & \frac{dp_1}{dW_{1(2)}} \end{bmatrix} = 0 \cdot \begin{bmatrix} X_{1(1)} & X_{1(2)} \end{bmatrix}$$

$$= 0 \cdot \begin{bmatrix} 2 & -3 \end{bmatrix} = \boxed{\begin{bmatrix} 0 & 0 \end{bmatrix}}$$

same dim to $W_{(1)}$

$$\frac{dL}{dx_1} = \frac{dL}{dp_1} \cdot \frac{dp_1}{dx_1} = \frac{dL}{dp_1} \cdot \begin{bmatrix} \frac{dp_1}{dx_{1(1)}} \\ \frac{dp_1}{dx_{1(2)}} \end{bmatrix} = 0 \cdot \begin{bmatrix} W_{1(1)} \\ W_{1(2)} \end{bmatrix} = 0 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \boxed{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}$$

↳ same dim to x_1

$$\frac{dL}{dW_2} = \frac{dL}{dp_2} \cdot \frac{dp_2}{dW_2} = \frac{dL}{dp_2} \cdot \begin{bmatrix} \frac{dp_2}{dW_{2(1)}} & \frac{dp_2}{dW_{2(2)}} \end{bmatrix} = 2.26 \times 10^{-6} \cdot \begin{bmatrix} X_{2(1)} & X_{2(2)} \end{bmatrix}$$

$$= 2.26 \times 10^{-6} \cdot \begin{bmatrix} 2 & -3 \end{bmatrix}$$

same dim to $W_{(2)}$ ← 1×2 ← $= \boxed{\begin{bmatrix} 4.52 \times 10^{-6} & -6.78 \times 10^{-6} \end{bmatrix}}$

$$\frac{dL}{dx_2} = \frac{dL}{dp_2} \cdot \frac{dp_2}{dx_2} = \frac{dL}{dp_2} \cdot \begin{bmatrix} \frac{dp_2}{dx_{2(1)}} \\ \frac{dp_2}{dx_{2(2)}} \end{bmatrix} = 2.26 \times 10^{-6} \cdot \begin{bmatrix} W_{2(1)} \\ W_{2(2)} \end{bmatrix} = 2.26 \times 10^{-6} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} -4.52 \times 10^{-6} \\ 6.78 \times 10^{-6} \end{bmatrix}}$$

same dim to x_2 ← 2×1 ←

Problem 3. (10 points). Given inputs $z_1, z_2 \in \mathbb{R}$, the softmax function is the following:

$$\hat{y} = \frac{e^{z_1}}{e^{z_1} + e^{z_2}}.$$

Let $y \in \{0, 1\}$, then define the cross-entropy loss between y and $\hat{y}$ be

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}).$$

Prove that:

$$\frac{\partial L(y, \hat{y})}{\partial z_1} = \hat{y} - y, \frac{\partial L(y, \hat{y})}{\partial z_2} = y - \hat{y}.$$

Proof $\frac{\partial L(y,\hat{y})}{\partial z_1} = \hat{y} - y$

$L(y,\hat{y}) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$

$\frac{\partial L(y,\hat{y})}{\partial z_1} = \frac{\partial}{\partial z_1}(-y \log(\hat{y})) - \frac{\partial}{\partial z_1}(1-y)\log(1-\hat{y})$

$= -y \frac{\partial}{\partial z_1}\left(\log\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right) - (1-y)\frac{\partial}{\partial z_1}\left(\log\left(1 - \frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right)$

$= -y \frac{\partial}{\partial z_1}\left(\log\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right) - (1-y)\frac{\partial}{\partial z_1}\left(\log\left(\frac{e^{z_2}}{e^{z_1}+e^{z_2}}\right)\right)$

$= -y \frac{\partial}{\partial z_1}\left(\log(e^{z_1}) - \log(e^{z_1}+e^{z_2})\right) - (1-y)\frac{\partial}{\partial z_1}\left(\log(e^{z_2}) - \log(e^{z_1}+e^{z_2})\right)$

$= -y\left(z_1 - \log(e^{z_1}+e^{z_2})\right) - (1-y)\left(z_2 - \log(e^{z_1}+e^{z_2})\right)$

$= -y\left(\frac{\partial}{\partial z_1} z_1 - \frac{\partial}{\partial z_1}\log(e^{z_1}+e^{z_2})\right) - (1-y)\left(\frac{\partial}{\partial z_1} z_2 - \frac{\partial}{\partial z_1}\log(e^{z_1}+e^{z_2})\right)$

$= -y(1 - \hat{y}) - (1-y)(0 - \hat{y})$

$= -y + y\hat{y} + \hat{y} - y\hat{y}$

$= \hat{y} - y$ (proven)

$\frac{\partial}{\partial z_1}\log(e^{z_1}+e^{z_2}) = \frac{1}{e^{z_1}+e^{z_2}}(e^{z_1}) = \frac{e^{z_1}}{e^{z_1}+e^{z_2}} = \hat{y}$

Proof $\frac{\partial L(y,\hat{y})}{\partial z_2} = y - \hat{y}$

$$L(y,\hat{y}) = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\frac{\partial L(y,\hat{y})}{\partial z_2} = \frac{\partial}{\partial z_2}(-y\log(\hat{y})) - \frac{\partial}{\partial z_2}(1-y)\log(1-\hat{y})$$

$$= \frac{\partial}{\partial z_2}(-y\log(\hat{y})) - \frac{\partial}{\partial z_2}(1-y)\log(1-\hat{y})$$

$$= -y\frac{\partial}{\partial z_2}\left(\log\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right) - (1-y)\frac{\partial}{\partial z_2}\left(\log\left(1-\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right)$$

$$= -y\frac{\partial}{\partial z_2}\left(\log\left(\frac{e^{z_1}}{e^{z_1}+e^{z_2}}\right)\right) - (1-y)\frac{\partial}{\partial z_2}\left(\log\left(\frac{e^{z_2}}{e^{z_1}+e^{z_2}}\right)\right)$$

$$= -y\frac{\partial}{\partial z_2}\left(\log(e^{z_1}) - \log(e^{z_1}+e^{z_2})\right) - (1-y)\frac{\partial}{\partial z_2}\left(\log(e^{z_2}) - \log(e^{z_1}+e^{z_2})\right)$$

$$= -y\left(\frac{\partial}{\partial z_2}z_1 - \frac{\partial}{\partial z_2}\log(e^{z_1}+e^{z_2})\right) - (1-y)\left(\frac{\partial}{\partial z_2}z_2 - \frac{\partial}{\partial z_2}\log(e^{z_1}+e^{z_2})\right)$$

$$= -y(0 - (1-\hat{y})) - (1-y)(1-(1-\hat{y}))$$

$$= -y(-1+\hat{y}) - (1-y)(\hat{y})$$

$$= y - y\hat{y} - \hat{y} + y\hat{y}$$

$$= y - \hat{y} \quad (\text{proven})$$

$$\frac{\partial}{\partial z_2}\log(e^{z_1}+e^{z_2}) = \frac{1}{e^{z_1}+e^{z_2}}(e^{z_2}) = \frac{e^{z_2}}{e^{z_1}+e^{z_2}} = 1-\hat{y}$$

Problem 4. (15 points). Consider datapoints in Figure ??: $(-2, 0), (2, 0)$ are crosses, and $(0, 2), (0, -2)$ are circles. Let the crosses be labeled $+1$, and the circles be labeled -1. In this problem the goal

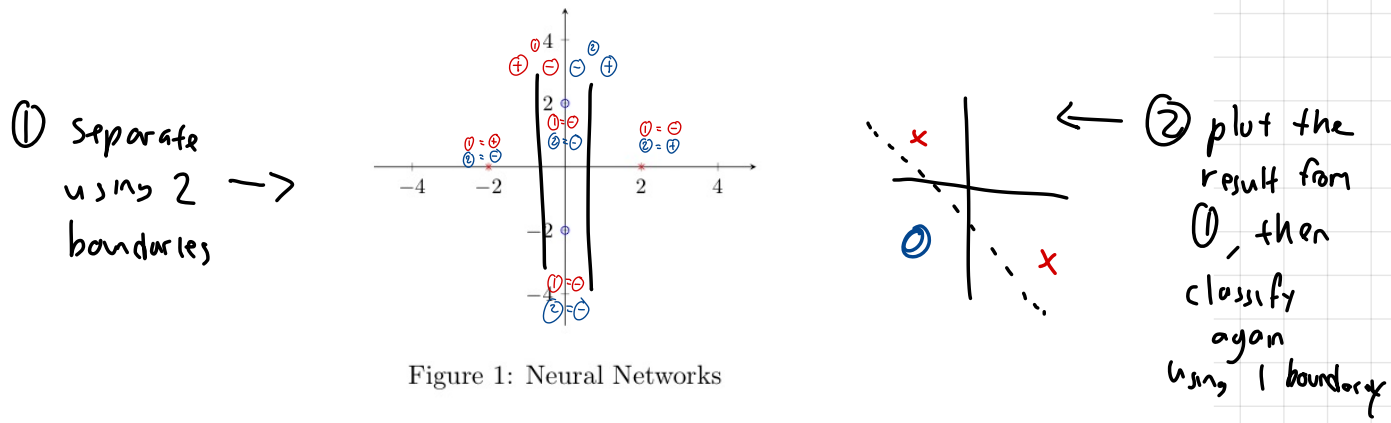
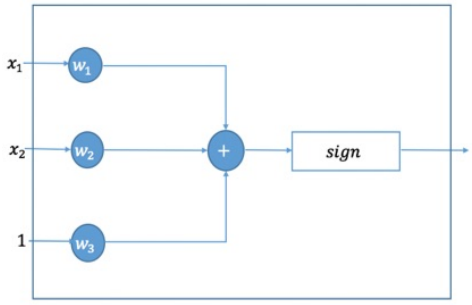


Figure 1: Neural Networks

is to design a neural network with no error on this dataset.

To make things simple, consider the following generalization. We first append a $+1$ to each input and form a new dataset as follows: $(-2, 0, 1), (2, 0, 1)$ are labeled $+1$, and $(0, 2, 1), (0, -2, 1)$ are labeled -1. Note that the last feature is redundant.

We consider the following basic units for our neural networks: Linear transformation followed by hard thresholding. Each unit has three parameters w_1, w_2, w_3. The output of the unit is the sign of the inner product of the parameters with the input.



1. Design a neural network with these units that make no error on the datapoints above. (Hint: You can take two units in the first layer, and one in the output layer, a total of three units).

2. Show that if you design a neural network with ONLY one such unit, then the points cannot be all classified correctly.

①

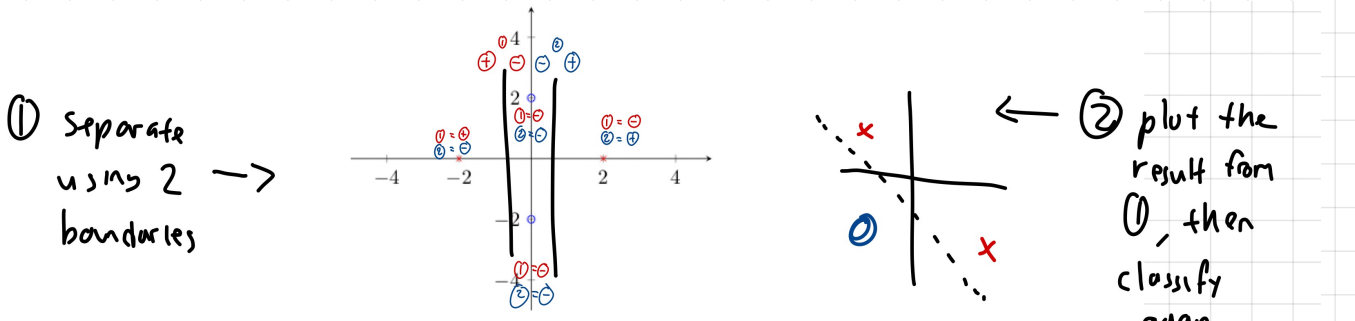
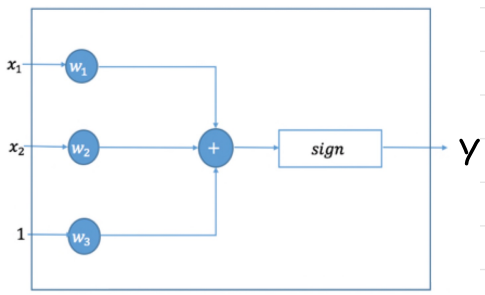
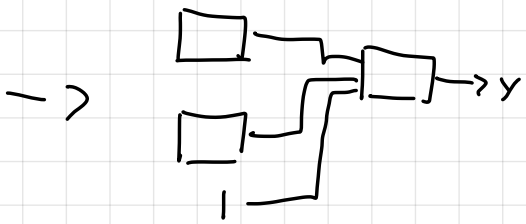


Figure 1: Neural Networks



$Y = \text{Sign}(x_1 w_1 + x_2 w_2 + w_3)$



$\bar{x} \in \mathbb{R}^3$, $\{x_1, x_2, x_3 = 1\}$

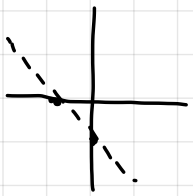
Plane 1: $x_1 \geq -1 \cdot x_3$, where $x_1 > -1 \Rightarrow \ominus$ and $x_1 < -1 \Rightarrow \oplus$

So, $W = \{1, 0, -1\}$ (w_1, w_2, w_3)

Plane 2: $x_1 \geq 1 \cdot x_3 = 1$, where $x_1 > 1 \Rightarrow \oplus$ and $x_1 < -1 \Rightarrow \ominus$

So, $W = \{-1, 0, -1\}$

Plane 3:



$x_1 = -1 \Rightarrow x_2 = 0$

$x_1 = 0 \Rightarrow x_2 = -1$

$$\frac{x_1+1}{0+1} = \frac{x_2-0}{-1-0}$$

$$x_1 + 1 = \frac{x_2}{-1}$$

$$-1 - x_1 = x_2$$

$W = \{1, 1, 1\}$

Test our

$(0, 2, 1)$: $y = -1$

$q_1 = \text{sign}(0 \times 1 + 2 \times 0 + 1 \times -1) = -1$

$q_2 = \text{sign}(0 \times -1 + 2 \times 0 + 1 \times -1) = -1$

$y = \text{sign}(-1 \times 1 + \overset{-1}{-1} \times 1 + 1 \times 1) = -1$ ✓

$(-2, 0, 1)$: $y = +1$

$q_1 = \text{sign}(-2 \times 1 + \overset{-3}{0} \times 0 + 1 \times -1) = -1$

$q_2 = \text{sign}(-2 \times -1 + \overset{1}{0} \times 0 + 1 \times -1) = 1$

$y = \text{sign}(-1 \times 1 + \overset{1}{1} \times 1 + 1 \times 1) = 1$ ✓

$(0, -2, 1)$: $y = -1$

$q_1 = \text{sign}(0 \times 1 + -2 \times 0 + 1 \times -1) = -1$

$q_2 = \text{sign}(0 \times -1 + -2 \times 0 + 1 \times -1) = -1$

$y = \text{sign}(-1 \times 1 + \overset{-1}{-1} \times 1 + 1 \times 1) = -1$ ✓

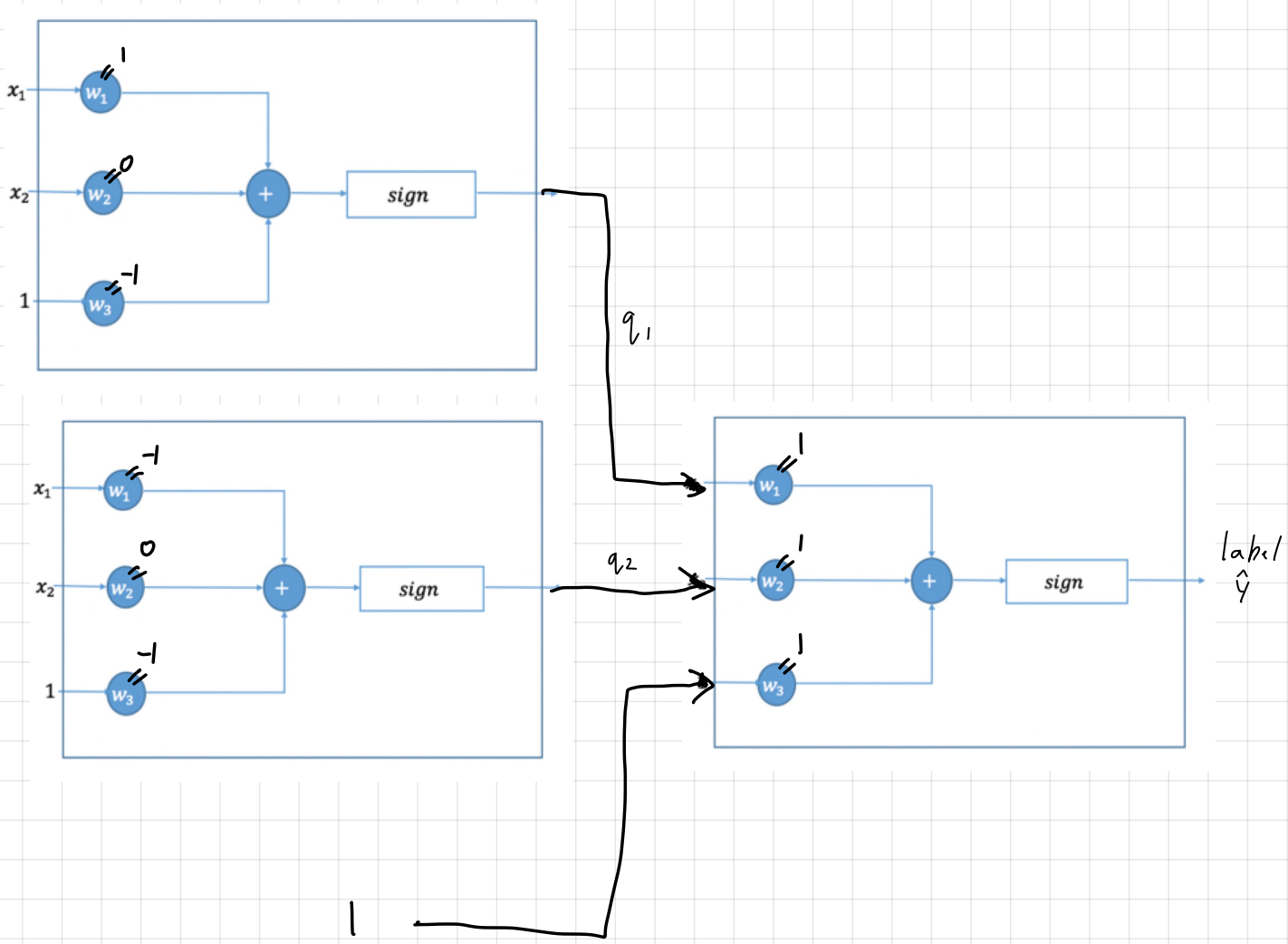
$(2, 0, 1)$: $y = +1$

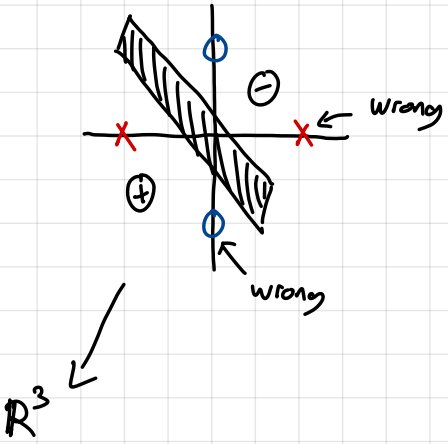
$q_1 = \text{sign}(2 \times 1 + \overset{1}{0} \times 0 + 1 \times -1) = 1$

$q_2 = \text{sign}(2 \times -1 + \overset{-3}{0} \times 0 + 1 \times -1) = -1$

$y = \text{sign}(1 \times 1 + \overset{1}{-1} \times 1 + 1 \times 1) = 1$ ✓

NO errors!

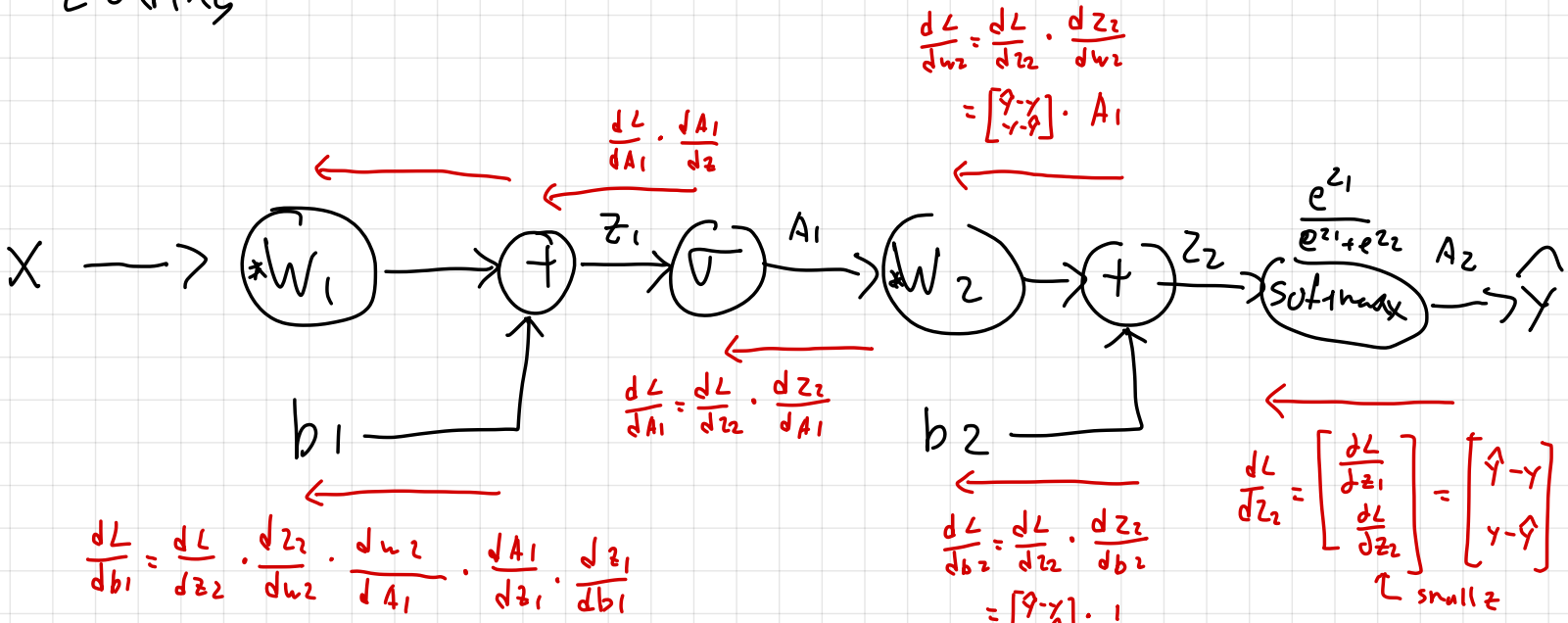


② 

→ there is no single 3D hyperplane that can corretly classify the 4 ponts, as the points cannot be classified with a single boundary. As shown in Q_1 explanation, we need at least 2 boundaries, which can be achieved using 2 boundaries.

Since using a single unit means a single boundary, so the points cannot be correctly classified. Thus, we need more than one unit.

Coding:



$$\frac{dw_2}{dz_1} = \frac{dw_2}{dA_1} \cdot \frac{dA_1}{dz_1}$$

$$=$$

$A_1 = \sigma(z_1) \quad = (20 \times n)$

$A_1 \quad (20 \times n)$

$$\frac{dL}{dw_1} = \frac{dL}{dz_2} \cdot \frac{dz_2}{A_1} \cdot \frac{dA_1}{dz_1}$$

$20 \times n \quad \cdot \quad 20 \times n$