FYS-STK 4155, OCT 6, 2022

Gradient methods

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)}) \times g(\beta^{(n)})$$

$$\approx \beta^{(n)} - \gamma^{(n)} g(\beta^{(n)})$$

$\gamma^{(n)}$: Learning rate

- efficient gradient evaluation $\Rightarrow$ Stochastic GD + momentum GD
- Learning rate tuning
  - linear $\gamma^{(n)}$ in terms of iterations n
  - exponential
  - other

- Adagrad
- RMS prop
- Adam

Steepest descent

$$\gamma^{(n)} = \frac{g^{T(n)} g^{(n)}}{g^{T(n)} H^{(n)} g^{(n)}}$$

Adagrad $\gamma^{(n)} \sim \frac{1}{g^{(n)} g^{T(n)}}$

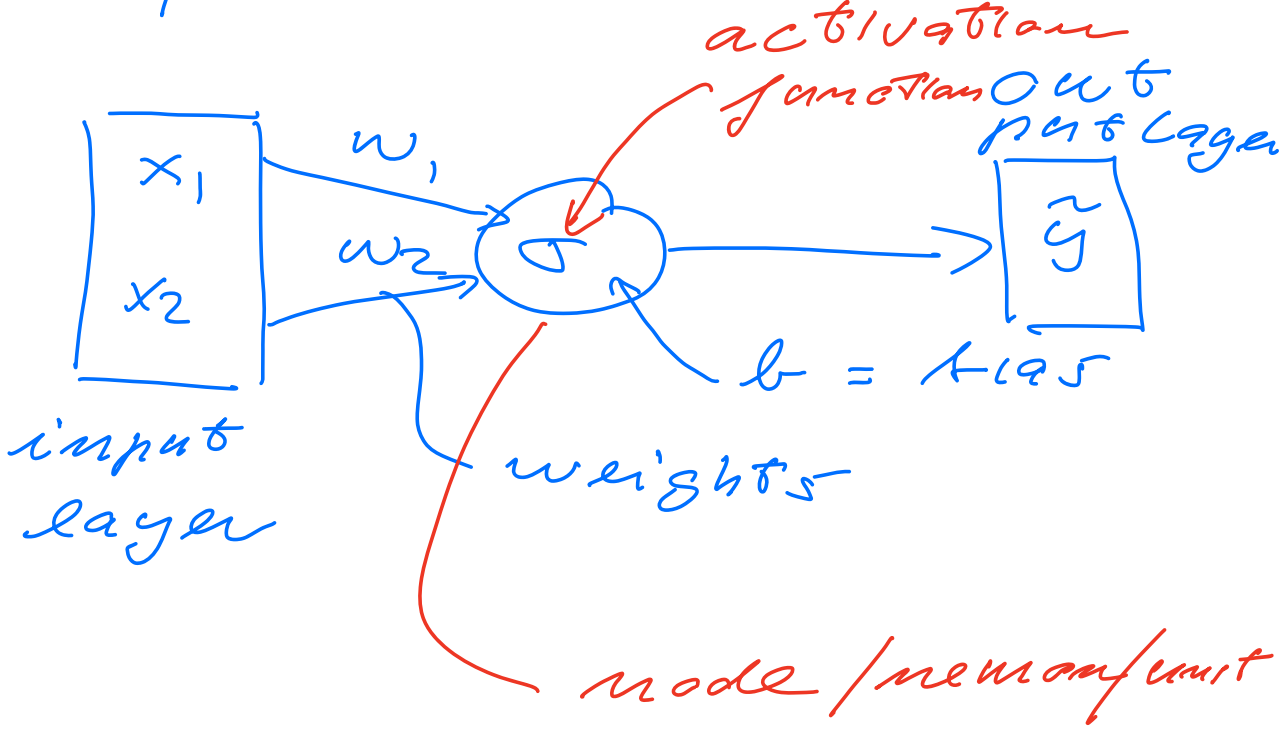
Neural Networks

Consider a simple case

$y \to t$ (target)

$f(x_1, x_2) \rightarrow f(x) \rightarrow \sigma$

$\beta \rightarrow \Theta$



$$\tilde{y} = \sigma(\underbrace{x_1, x_2}_{\text{inputs}}, \underbrace{w_1, w_2, b}_{\Theta})$$

$$X$$

$$X = [x_1, x_2]^T \qquad w = [w_1, w_2]^T$$

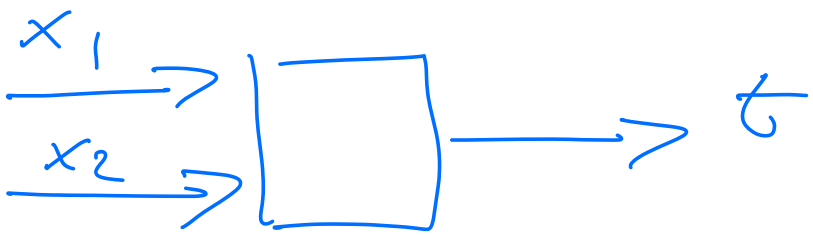
$$\tilde{y} = X^T w + b$$

parameters

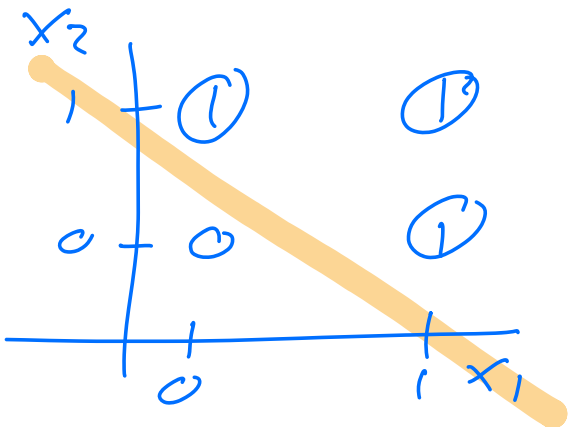
Θ

OR, XOR & AND gate

OR gate



x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	1



$$\tilde{y}_i = x_1(i)\, w_1 + x_2(i)\, w_2 + b$$

$$X = \{ \underbrace{[0,0]^T}_{x_1\ x_2}, [0,1]^T, [1,0]^T, [1,1]^T \}$$

$t=0$: $\begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b$

$t=1$; $\begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b$

$t=1$; $\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b$

$t=1$; $\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + b$

$b \quad x_1 \quad x_2$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

$$\Theta = \begin{bmatrix} b, & w_1, & w_2 \end{bmatrix}$$

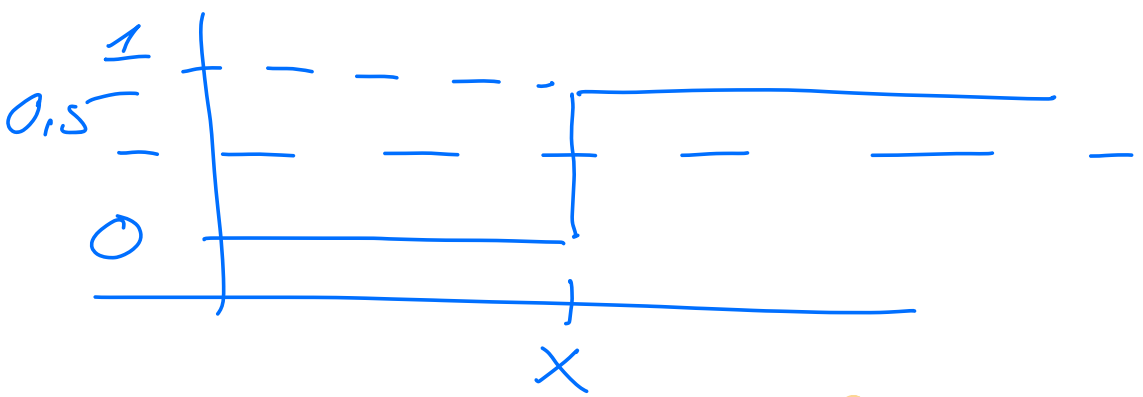
$$(X^T X)^{-1} = \begin{bmatrix} 3/4 & -1/2 & -1/2 \\ -1/2 & 1 & 0 \\ -1/2 & 0 & 1 \end{bmatrix}$$

$$(\beta) = \Theta = (X^T X)^{-1} X^T y$$

$$y = [0 \; 1 \; 1 \; 1]^T$$

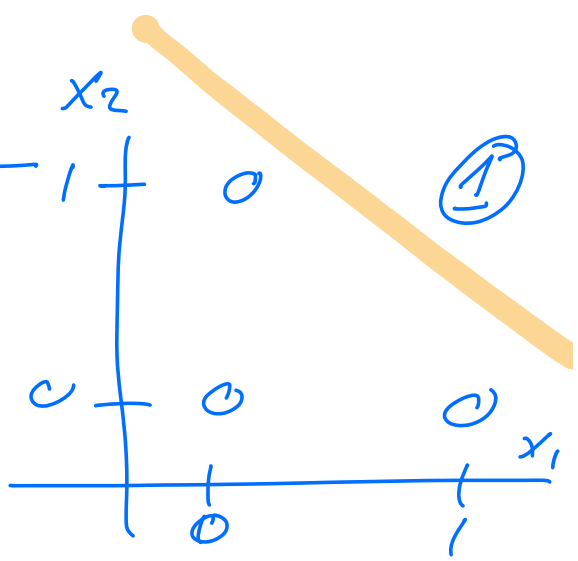
$$\Theta = \left[\frac{1}{4}, \frac{1}{2}, \frac{1}{2}\right]^T$$

$$\tilde{y} = X\Theta = [1/4, 3/4, 3/4, 5/4]$$



AND Gate

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1



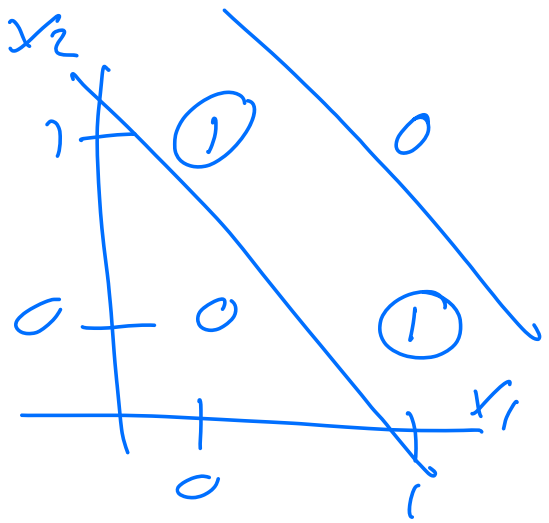
$$\theta = \begin{bmatrix} -1/4 & 1/2 & 1/2 \end{bmatrix}^T$$

$$b \quad w_1 \quad w_2$$

$$\tilde{y} = X\theta = \begin{bmatrix} -1/4 & 1/4 & 1/4 & 3/4 \end{bmatrix}^T$$

XOR

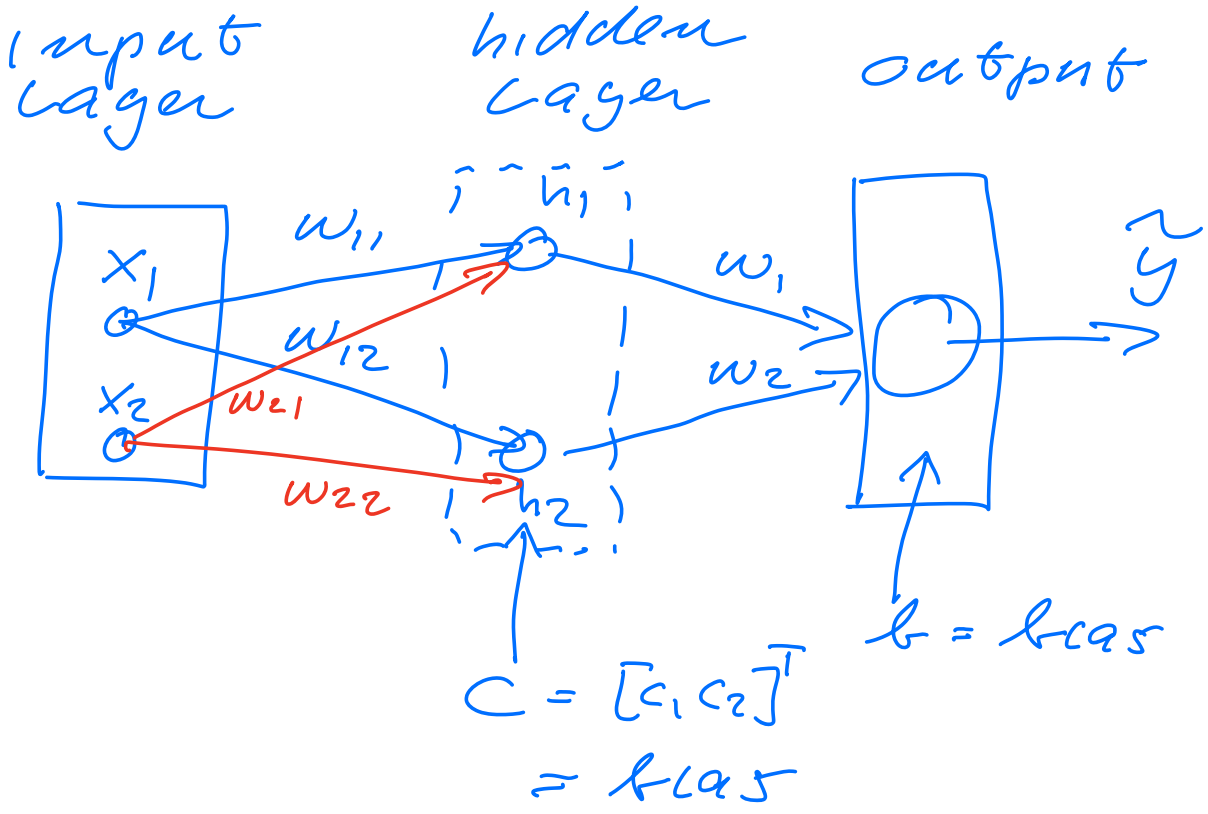
x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0



$$\theta = \left[\frac{1}{2}, 0, 0\right]^T$$

$$\tilde{y} = X\theta = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}^T$$

First NN encounter.

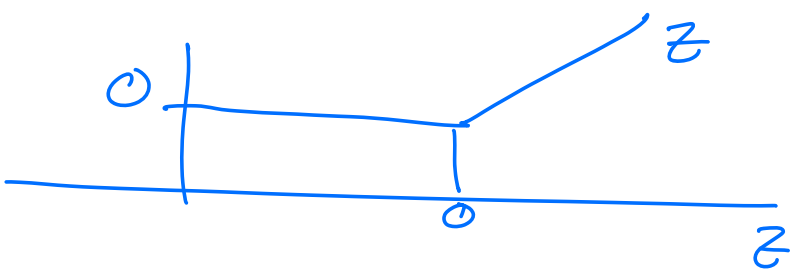


- Model : number of hidden layers
  # of nodes

- activation functions : outputs from h_i and outputs from output layer.

- Examples are
  - step function
  - sigmoid $\sigma = \frac{e^t}{1+e^t}$
  - tanh
  - RELU
  - ELU ...

$\sigma(z) = \max(0, z)$



- From input to hidden

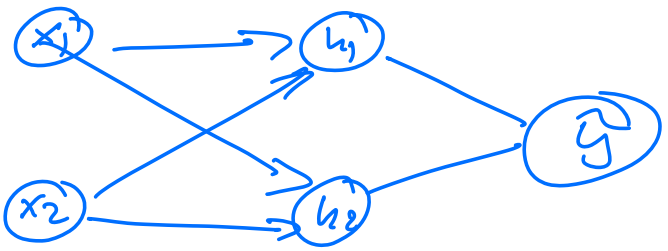
$$h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + c$$

$$\sigma^{(1)}(h) = \sigma^{(1)}(x; W, c)$$

$$\hat{y} = \sigma^{(2)}(h; w, b)$$

$$= \sigma^{(2)}\left(\sigma^{(1)}(x; W_c); w_b\right)$$

Simple model



$$W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$c = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \qquad b = 0$$

$$w = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$XW = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$

XOR - gate

$X \in \mathbb{R}^{2 \times 4}$

$$X = \begin{matrix} x_1 & x_2 \end{matrix}$$
$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Add bias vector

$$h = \begin{bmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

ReLU function application

$$= \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{bmatrix}$$

multiply with w

$$\tilde{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} ?$$

XOR gate