

1) Ghazal Kariani

Zhaoxin Li

2.1)

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{w}^T \mathbf{x} + b \geq 0 \\ 0 & \text{if } \mathbf{w}^T \mathbf{x} + b < 0 \end{cases}$$

Find $\mathbf{w} \nparallel b$ Try $\mathbf{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, b = -2$

Pair	x_1	x_2	$f_{\text{AND}}(\mathbf{x})$
1	0	0	0
2	0	1	0
3	1	0	0
4	1	1	1

$$\text{Pair 1: } \mathbf{w}^T \mathbf{x} + b = [1 \ 1] \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 2 = 0 - 2 = 0 \Rightarrow f(x) = 0 \checkmark$$

$$\text{Pair 2: } [1 \ 1] \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 2 = 1 - 2 = -1 \Rightarrow f(x) = 0 \checkmark$$

$$\text{Pair 3: } [1 \ 1] \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 2 = 1 - 2 = -1 \Rightarrow f(x) = 0 \checkmark$$

$$\text{Pair 4: } [1 \ 1] \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 2 = 2 - 2 = 0 \Rightarrow f(x) = 1 \checkmark$$

By inspection, $\mathbf{w}_{\text{AND}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \nparallel b_{\text{AND}} = -2$ work

Now, for OR:

Try $\mathbf{w}_{\text{OR}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, b_{\text{OR}} = 0.5$

Pair	x_1	x_2	$f_{\text{OR}}(\mathbf{x})$
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	1

$$\text{Pair 1: } \mathbf{w}^T \mathbf{x} + b = [1 \ 1] \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.5 = -0.5 \Rightarrow f(x) = 0 \checkmark$$

$$\text{Pair 2: } [1 \ 1] \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 0.5 = 1 - 0.5 = 0.5 \Rightarrow f(x) = 1 \checkmark$$

2.1) cont'd

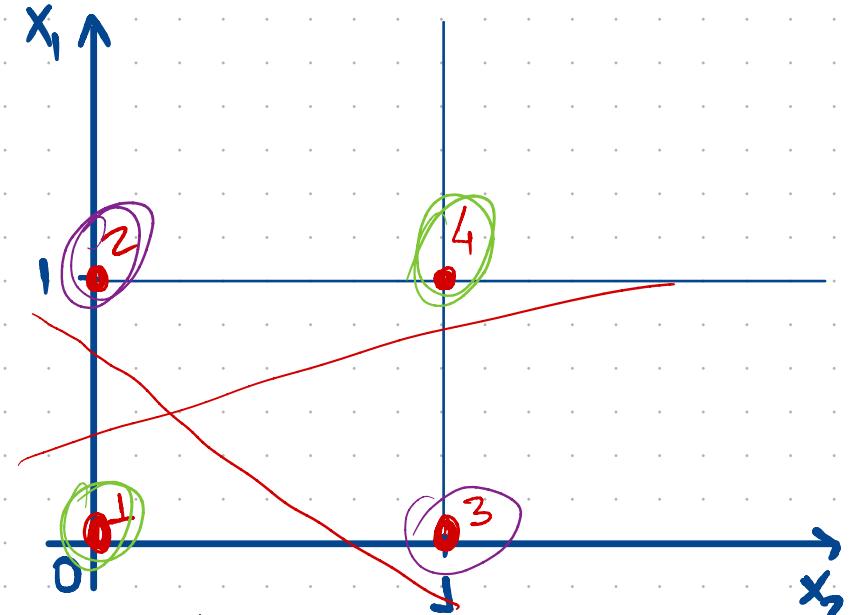
Pair 3: $[1 \ 1] \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0.5 = 1 - 0.5 = 0.5 \Rightarrow f(x) = 1 \checkmark$

Pair 4: $[1 \ 1] \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.5 = 2 - 0.5 = 1.5 \Rightarrow f(x) = 1 \checkmark$

Similarly, by inspection, $\boxed{W_{02} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad b_{02} = -0.5}$ work

2.2)

Pair	x_1	x_2	$f_{\text{XOR}}(x)$
1	0	0	0
2	0	1	1
3	1	0	1
4	1	1	0



by inspection in the

2D plane, there is

no possible line to separate the pairs corresponding to the same f_{XOR} value \Rightarrow the XOR

function isn't linearly separable

Formally, we try to find rules for a line using $w \nparallel b$ (Assume it is separable — proof by contradiction):

$$\text{Pair 1} < \text{Pair 2} \quad (f_{\text{XOR}}(\text{Pair 1}) = 0; f_{\text{XOR}}(\text{Pair 2}) = 1)$$

$$w^T [0] + b < w^T [1] + b \Rightarrow b < w_1 \cdot 0 + w_2 \cdot 1 + b$$

$$\underline{\underline{w_2 > 0}} \quad \textcircled{I} \quad 1^{\text{st}} \text{ rule}$$

$$\text{Pair 4} < \text{Pair 3} \quad (f_{\text{XOR}}(\text{Pair 4}) = 0; f_{\text{XOR}}(\text{Pair 3}) = 1)$$

$$w^T [1] + b < w^T [0] + b \Rightarrow w_1 + w_2 < w_1$$

$$\underline{\underline{w_2 < 0}} \quad \textcircled{II} \quad \text{Rules } \textcircled{I} \text{ and } \textcircled{II} \text{ lead to a contradiction} \blacksquare$$

3.3) Objective: Design an LSTM that maps x to y in the XOR operation.

Usual notation (x for input, y for output), and taking advantage of the hint:

$$(x \wedge \bar{y}) \vee (\bar{x} \wedge y) = \text{XOR operation} \rightarrow \text{Notation + because it's a recursive scheme}$$

$$y_t = (x_t \wedge \bar{y}_{t-1}) \vee (\bar{x}_t \wedge y_{t-1}) \quad \text{Remind gate-to-algebra}$$

$$= x_t * \bar{y}_{t-1} + \bar{x}_t * y_{t-1} \quad \textcircled{I}$$

Mapping and matching with eqn(5) from the PS:

$$\textcircled{I} \models (5) : C_t = x_t * y_{t-1} + \bar{x}_t * y_{t-1} = f_t * C_{t-1} + i_t * \bar{C}_t$$

Yielding: $C_t = y_t = \text{XOR output}$

$$f_t = \bar{x}_t, \quad i_t = x_t \quad \text{II}$$

As per DNN's nature, there is also the hidden state h_t , which in this case maps to C_t (and y_t consequently). $\rightarrow C_t = y_t = h_t$

Now we can match \textcircled{II} and eqn (2)

$$3.3 \text{ cont'd) eqn (2): } f_t = \sigma(w_f \cdot [h_{t-1} \ x_t] + b_f)$$

mapping & matching with \textcircled{I} and \textcircled{II} :

$$f_t = \bar{x}_t = \sigma(w_f \cdot [h_{t-1} \ x_t] + b_f)$$

$$\begin{aligned} x_t = 0 &\Rightarrow \bar{x}_t = 1 = f_t \\ x_t = 1 &\Rightarrow \bar{x}_t = 0 = f_t \end{aligned} \quad \left. \right\} \textcircled{III}$$

Inspecting:

$$\begin{bmatrix} w_{f1} \\ w_{f2} \end{bmatrix} \left[\begin{bmatrix} h_{t-1} & x_t \end{bmatrix} \right] + b_f = w_{f1} h_{t-1} + w_{f2} x_t + b_f = f_t$$

↑ ↑ ↑
 w_{f1} w_{f2} b_f

didn't even have to touch the σ

Setting $w_{f1} = 0 \Rightarrow w_{f2}$ has to be opposite to b_f
So the rule \textcircled{III} holds

$$\therefore w_f = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, b_f = 1$$

Similar matching, but with eqn. (3)

$$i_t = x_t = \sigma(w_i \cdot [h_{t-1} \ x_t] + b_i)$$

$$x_t = i_t = 0 \quad \& \quad x_t = i_t = 1$$

Inspecting: $\begin{bmatrix} w_{i1} \\ w_{i2} \end{bmatrix} \left[\begin{bmatrix} h_{t-1} & x_t \end{bmatrix} \right] + b_i = w_{i1} h_{t-1} + w_{i2} x_t + b_i$

3.3 cont'd)

$$w_{i1}h_{t-1} + w_{i2}x_t + b_i = x_t$$

0 1 0

Also by inspection:

$$w_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, b_i = 0$$

Again no need to mess with Γ

Continuing the same process for eqn (4):

$$\bar{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c)$$

$$c_t = y_t = h_t \Rightarrow \bar{c}_t = \bar{y}_t = \bar{h}_t \quad (\text{Again try to ab everything inside the activation func})$$

$$w_{c1}h_{t-1} + w_{c2}x_t + b_c = \bar{h}_t \rightarrow \tanh = 0 \Leftrightarrow x = 0$$

0 -1

$$\tanh = 1 \Leftrightarrow x > 0$$

Potentially degenerate
so won't zero out b_c

Try $b_c = 1$

$$\bar{h}_t = \tanh(-h_{t-1} + 1) \rightarrow \begin{cases} h_{t-1} = 0 \Rightarrow h_t = 1 \\ h_{t-1} = 1 \Rightarrow h_t = 0 \end{cases} \checkmark$$

By inspection:

$$w_c = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, b_c = 1$$

3.3 cont'd) Last to match and inspect is eqn. (6) & (7)

$$h_t = O_t * \tanh(C_t), \quad O_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = y_t = C_t$$

$\tanh(C_t)$ will dominate

$$\tanh(C_t=1) = 1 = h_t \quad \Rightarrow \text{leave } O_t = 1$$

$$\tanh(C_t=0) = 0 = h_t$$

$$\sigma(W_o[h_{t-1}, x_t] + b_o) = 1$$

$$W_{o1}h_{t-1} + W_{o2}x_t + b_o = 1$$

Isolate since we want O_t fixed at 1

$$\therefore W_o = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad b = 1$$

Copying all other learned weights to the same pg for simplicity:

$$W_f = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad b_f = 1$$

$$W_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad b_i = 0$$

$$W_c = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad b_c = 1$$

DESIGN COMPLETE!

3.4) Went to prove that B_t (highest scoring beam)
Scores worse than or equal to the best _{$\leq t$}

Let us define $f(\cdot)$ as the score

Thus far we have $f(\text{best}_{\leq t})$ as the highest
By the claim, $f(\text{best}_{\leq t}) \geq$ any other $f(\cdot)$

Assume $B_t(y')$, which happens after $\text{best}_{\leq t}$

i.e., t

$$f(\text{best}_{\leq t}) \geq f(B_t)$$

Assume another beam at another time later
than t (t^+)

By the claim: $f(B_{t^+}) \leq f(B_t)$

Scores are upper bounded by nature

Hence $f(B_{t^+}) \leq f(\text{best}_{\leq t})$ ■

3.5) Eqn(14): $h_t = W^T h_{t-1}$ is recurring
 Going backwards, we can see the first hidden state established relationship: $h_1 = W^T h_0$

Carrying forward: $h_t = W \cdot W^T \dots h_0$
 $h_t = (W^T)^t h_0$

We can observe that anything that is contained in W will propagate in a potentially huge scale downstream because it is to the power t , i.e., to the power of the timesteps

Intuitively, we know that this means that big numbers can lead to overflow (exploding gradients) and small can lead to vanishing.

We can see how this relates to the eigenvalues by performing eigendecomposition / diagonalization (common, not SVD) of W :

$$W = V \Lambda V^{-1} \quad (\Lambda \text{ is a diagonal matrix containing } W\text{'s eigenvalues in the diagonal})$$

$$\begin{aligned} h_t &= (V \Lambda V^{-1})^T h_0 \\ &= ((V^{-1})^T \Lambda^T V^T)^t h_0 = (V^{-T} \Lambda^T V^T)^t \\ &\text{expanding} \quad \rightarrow \text{continued} \quad \rightarrow \end{aligned}$$

3.5 cont'd) expanding:

$$h_t = [(V^{-T} \underbrace{\Lambda^T V^T}_{\Lambda^t}) \cdot (V^{-T} \underbrace{\Lambda^T V^T}_{\Lambda^t}) \cdots (V^{-T} \underbrace{\Lambda^T V^T}_{\Lambda^t})]$$

$$V^T V^{-T} = I$$

$$h_t = V^{-T} \underbrace{\Lambda^t}_t V \longrightarrow \text{eigenvectors cancel out eventually, but we are still left with } \Lambda^t$$

Confirming the intuition: All the "essential information" is "stored" and passed via a matrix's eigenvalues, which are the numbers that are to the power t .

More formally, there are 3 cases:

(Let us refer to the eigenvalues that constitute Λ as σ_i)

1. σ_i are too small.

$$t \rightarrow \infty \Rightarrow \sigma_i \rightarrow 0 \quad (\text{vanishing gradients})$$

2. σ_i are too big (not necessarily positive).

$$t \rightarrow \infty \Rightarrow \sigma_i \rightarrow \infty \quad (\text{exploding gradients})$$

3. Not 1 and Not 2, which is where we want to stay.