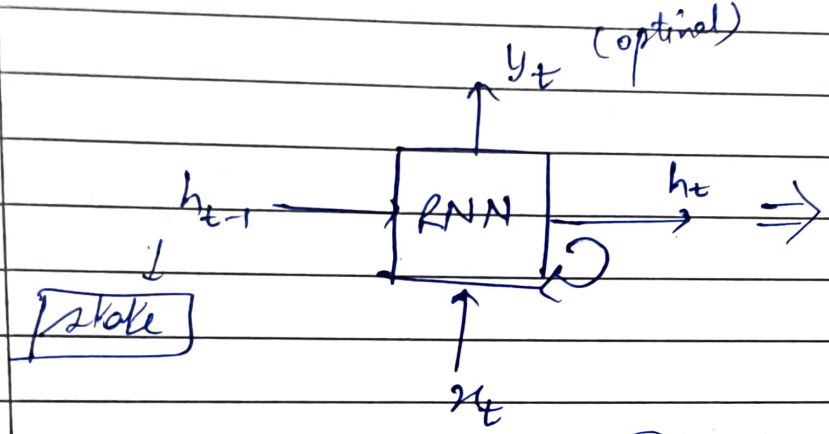
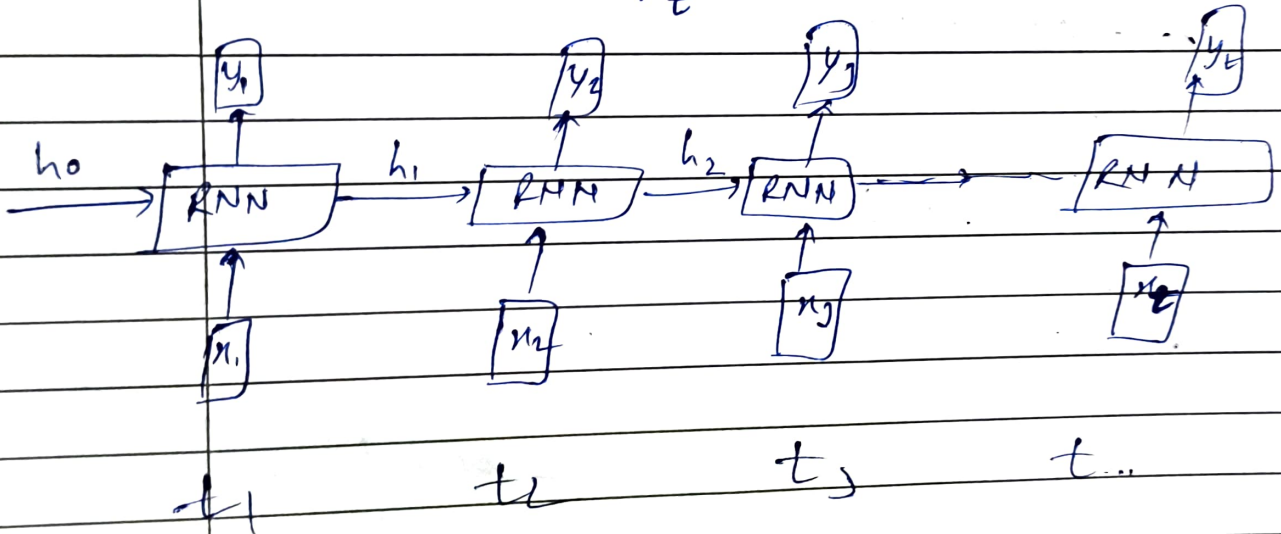
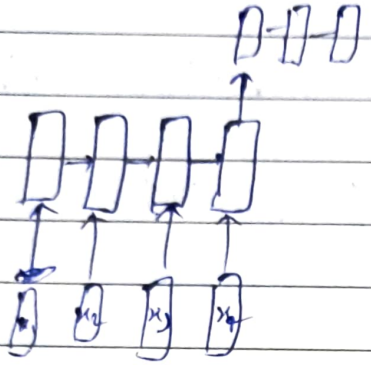
16/9/25


y_t (optimal)
h_{t-1}
RNN
h_t
state
x_t

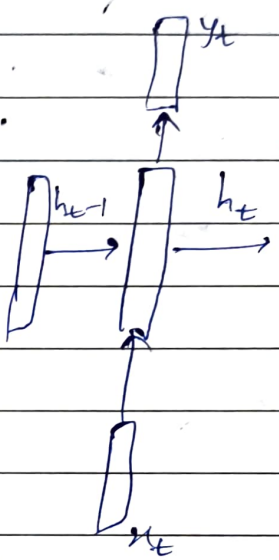

y_1
y_2
y_3
y_t
h_0
RNN
h_1
RNN
h_2
RNN
RNN
x_1
x_2
x_3
x_t
t_1
t_2
t_3
t_n

Machine translation example (english → french)



$$h_t = f_W(h_{t-1}, x_t)$$

$$h_t = \tanh\left((W)_{hh} h_{t-1} + W_{xh} x_t\right)$$



$$\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{4\times 1} = \tanh\left(\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{W_{hh}\ 4\times 4} \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{h_{t-1}\ 4\times 1} + \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{W_{xh}\ 4\times 3} \begin{bmatrix} \\ \\ \\ \end{bmatrix}_{3\times 1} \right)$$

$$y_t = W_{hy} h_t$$

$$\begin{bmatrix} \\ \\ \end{bmatrix}_{2\times 1} = \begin{bmatrix} \\ \\ \end{bmatrix}_{2\times 4} \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix}_{4\times 1}$$

① h_t

② y_t

cursor.ai

17/9/25

Q: $T = 2$ (two time steps)

$X^{(1)}$: $x_1 \in \mathbb{R}^2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $x_2 \in \mathbb{R}^2 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

input dim $= 2 \Rightarrow x_t \in \mathbb{R}^2$

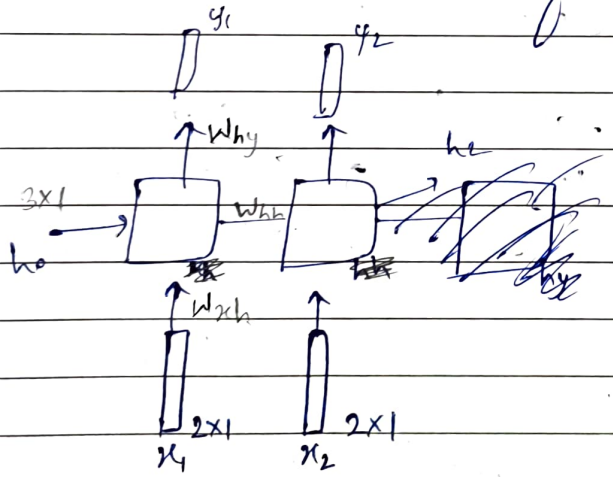
hidden state dim $= 3 \Rightarrow h_t \in \mathbb{R}^3$

$h_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1}$, $W_{xh} = \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times2}$, $W_{hh} = \begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.7 \end{bmatrix}_{3\times3}$

$y_t \in \mathbb{R}^2$ $\quad W_{hy} = \begin{bmatrix} 1 & -1 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}_{2\times3}$

what will be dimensions of all these?

Ans



$h_t = \tanh\left(\underset{3\times1}{W_{hh}}\ \underset{3\times3}{h_{t-1}} + \underset{3\times2}{W_{xh}}\ \underset{2\times1}{x_t}\right)$

$y_t = W_{hy}\ h_t$

$2\times1 \quad 2\times3 \quad 3\times1$

Time step $= 1$, $t = 1$

$h_t = \tanh\left(W_{hh}\ h_{t-1} + W_{xh}\ x_t\right)$

$h_1 = \tanh\left(W_{hh}\ h_0 + W_{xh}\ x_1\right)$

$= \tanh([\,])$

$= [\,]_{3\times1}$, $y_t = W_{hy}\ h_t$

$y_1 = W_{hy}\ h_1$

Timestep = 2 , t = 2

$$h_t = \tanh\left(W_{hh} h_{t-1} + W_{xh} x_t\right)$$

At t = 1

$$h_1 = \tanh\left(\begin{bmatrix} 0.1 & 0.4 & 0 \\ -0.2 & 0.3 & 0.2 \\ 0.05 & -0.1 & 0.2 \end{bmatrix}_{3\times3} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1} + \begin{bmatrix} 0.5 & -0.3 \\ 0.8 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}_{3\times2} \begin{bmatrix} 1 \\ 2 \end{bmatrix}_{2\times1}\right)$$

$$h_1 = \tanh\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1} + \begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}_{3\times1}\right)$$

$$h_1 = \tanh\left(\begin{bmatrix} -0.1 \\ 1.2 \\ 0.9 \end{bmatrix}\right) = \begin{bmatrix} 0.0995 \\ 0.834 \\ 0.7167 \end{bmatrix} \approx \begin{bmatrix} 0.099 \\ 0.83 \\ 0.716 \end{bmatrix}$$

$$y_1 = W_{hy} h_1$$

$$y_1 = \begin{bmatrix} 1 & -1 & 0.5 \\ 0.5 & 0.5 & -0.5 \end{bmatrix}_{2\times3} \begin{bmatrix} 0.0995 \\ 0.834 \\ 0.7167 \end{bmatrix}_{3\times1}$$

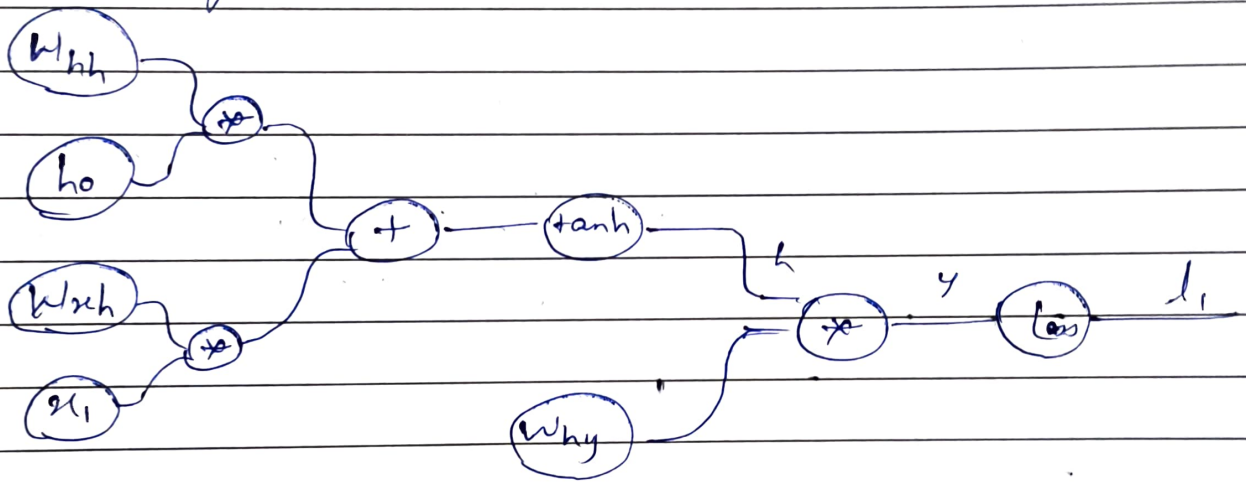
$$y_1 = \begin{bmatrix} -0.37615 \\ 0.1084 \end{bmatrix} \approx$$

0.09975
+ 0.417
− 0.35835
0.1084

Word 2 Vec

Ejp. → Simplified character level Language Model RNN

computational graph



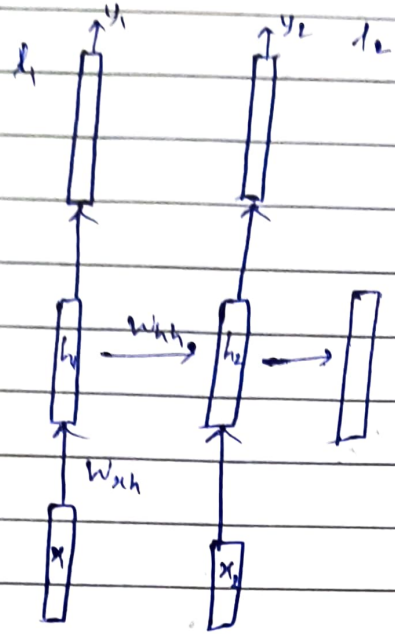
$$\frac{\partial l_1}{\partial W_{hh}} + \frac{\partial l_2}{\partial W_{hh}} + \cdots \rightarrow \frac{\partial l}{\partial W_{hh}}$$

$$\frac{\partial l_1}{\partial W_{xh}}$$

$$W'_{hh} = W'_{hh} - \alpha \frac{\partial l}{\partial W_{hh}}$$

22/9/25

- Sum the whole loss and compute the gradient



$$L = l_1 + l_2 + l_3 \dots l_T$$

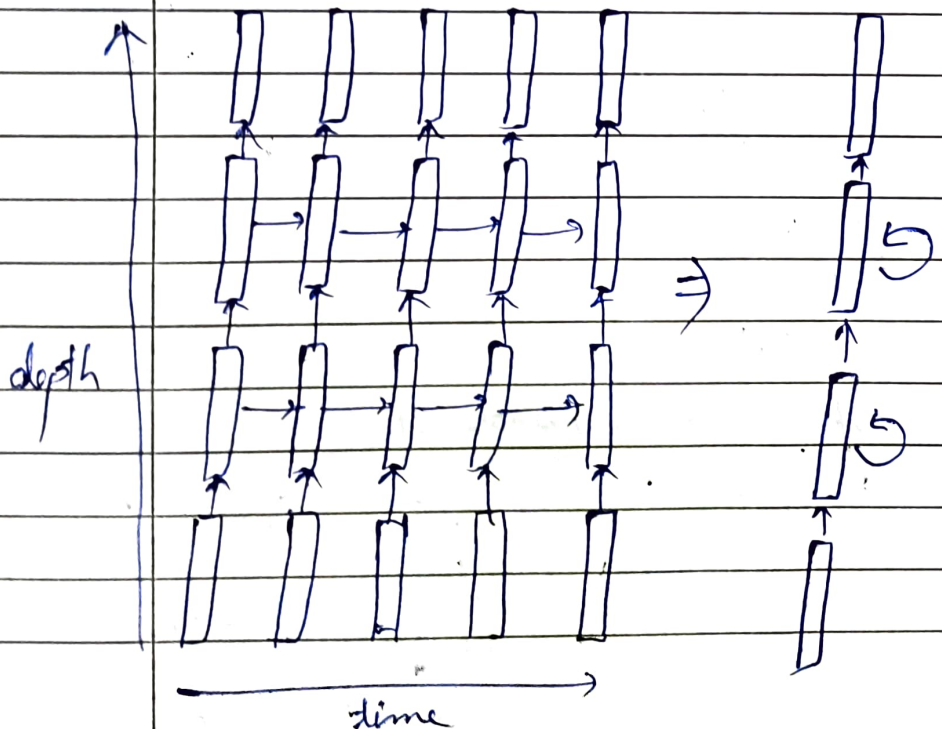
$$\frac{\partial L}{\partial W_{hh}} \qquad W_{hh} \propto \frac{\partial L}{\partial W_{hh}}$$

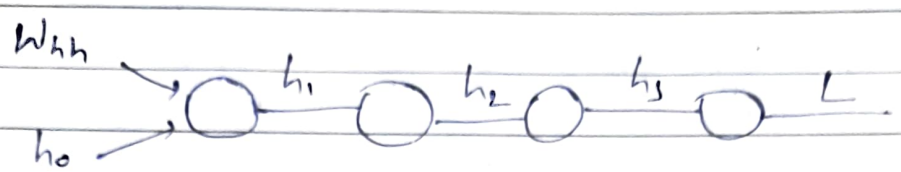
$$\frac{\partial L}{\partial W_{xh}}$$

$$\frac{\partial L}{\partial W_{hy}}$$

BPTT → Backpropagation Through time

Multilayer RNN





$$\frac{\partial L}{\partial W_{hh}} = \frac{\partial L}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W_{hh}}$$

$$= \frac{\partial L}{\partial h_3} \left(\frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \right) \cdot \frac{\partial h_1}{\partial W_{hh}}$$

$$= \frac{\partial L}{\partial h_3} \left(\prod_{t=2}^{T=3} \frac{\partial h_t}{\partial h_{t-1}} \right) \frac{\partial h_1}{\partial W_{hh}}$$ ← generalized form

for T time steps

$$\frac{\partial L_T}{\partial W_{hh}} = \frac{\partial L_T}{\partial h_t} \cdot \left(\prod_{t}^{T} \frac{\partial h_t}{\partial h_{t-1}} \right) \frac{\partial h_1}{\partial W_{hh}}$$

$$h_t = \tanh\left(W_{hh} h_{t-1} + W_{xh} x_t \right)$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh'\left(W_{hh} h_{t-1} + W_{xh} x_t \right) W_{hh}$$

$$\frac{\partial L}{\partial W_{hh}} = \frac{\partial L_T}{\partial h_t} \left(\prod_{t=2}^{T} \tanh'\left(W_{hh} h_{t-1} + W_{xh} x_t \right) W_{hh}^{T-1} \right) \frac{\partial h_1}{\partial W_{hh}}$$

If no non-linearity

$$h_t = W_{hh} h_{t-1} + W_{xh} x_t$$

$$\frac{\partial h_t}{\partial h_{t-1}} = W_{hh}$$

*high vanishing gradient problem

to overcome this, LSTM introduced

24/3/25 LSTM

Q.

$x_t = [0.5, -0.1]$

$C_{t-1} = [0.2, -0.2]$

$h_{t-1} = [0.0, 0.1]$

$W_{xi} = \begin{bmatrix} 0.5 & -0.3 \\ 0.4 & 0.1 \end{bmatrix}$

$W_{xo} = \begin{bmatrix} 0.3 & 0.25 \\ -0.2 & 0.2 \end{bmatrix}$

$W_{ho} = \begin{bmatrix} 0.15 & 0.05 \\ 0.1 & -0.2 \end{bmatrix}$

$W_{hi} = \begin{bmatrix} 0.1 & 0.2 \\ -0.2 & 0.05 \end{bmatrix}$

$W_{xf} = \begin{bmatrix} -0.4 & 0.2 \\ 0.3 & 0.3 \end{bmatrix}$

$W_{hf} = \begin{bmatrix} 0.05 & -0.1 \\ 0.2 & 0.1 \end{bmatrix}$

$W_{xg} = \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}$

$W_{hg} = \begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}$

$h_t = ?$

$c_t = ?$

ans.

$$h_t = o_t \odot \tanh(C_t)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t$$

$$f_t = \sigma(W_{hf} h_{t-1} + W_{xf} x_t)$$

$$f_t = \sigma\left(\begin{bmatrix}0.05 & -0.1\\ 0.2 & 0.1\end{bmatrix}\begin{bmatrix}0\\0.1\end{bmatrix} + \begin{bmatrix}-0.4 & 0.2\\ 0.3 & 0.7\end{bmatrix}\begin{bmatrix}0.5\\-0.1\end{bmatrix}\right)$$

$2\times2 \quad 2\times1 \quad 2\times2 \quad 2\times1$

$$= \sigma\left(\begin{bmatrix}-0.01\\0.01\end{bmatrix} + \begin{bmatrix}-0.22\\0.12\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}-0.23\\0.13\end{bmatrix}\right) = \begin{bmatrix}0.555\\0.467\end{bmatrix}$$

$$= \sigma\left(\begin{bmatrix}-0.01\\0.01\end{bmatrix} + \begin{bmatrix}-0.22\\0.12\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}-0.23\\0.13\end{bmatrix}\right) = \begin{bmatrix}0.44\\0.532\end{bmatrix}$$

$$i_t = \sigma\left(\begin{bmatrix}0.1 & 0.2\\ -0.2 & 0.05\end{bmatrix}\begin{bmatrix}0.0\\0.1\end{bmatrix} + \begin{bmatrix}0.5 & -0.3\\ 0.4 & 0.1\end{bmatrix}\begin{bmatrix}0.5\\-0.1\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}0.02\\0.005\end{bmatrix} + \begin{bmatrix}0.28\\0.19\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}0.30\\0.195\end{bmatrix}\right) = \begin{bmatrix}0.57\\0.548\end{bmatrix}$$

$$O_t = \sigma\left(\begin{bmatrix}0.15 & 0.05\\ 0.1 & -0.2\end{bmatrix}\begin{bmatrix}0.0\\0.1\end{bmatrix} + \begin{bmatrix}0.3 & 0.25\\ -0.2 & 0.2\end{bmatrix}\begin{bmatrix}0.5\\-0.1\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}0.005\\-0.02\end{bmatrix} + \begin{bmatrix}0.125\\-0.12\end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix}0.13\\-0.14\end{bmatrix}\right) = \begin{bmatrix}0.532\\0.465\end{bmatrix}$$

$$g_t = \tanh\left(\begin{bmatrix} 0.2 & 0.1 \\ -0.1 & 0.05 \end{bmatrix}\begin{bmatrix} 0.0 \\ 0.1 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 \\ 0.2 & -0.3 \end{bmatrix}\begin{bmatrix} 0.5 \\ -0.1 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} 0.01 \\ 0.005 \end{bmatrix} + \begin{bmatrix} -0.29 \\ 0.13 \end{bmatrix}\right)$$

$$= \tanh\left(\begin{bmatrix} -0.28 \\ 0.135 \end{bmatrix}\right) = \begin{bmatrix} -0.27 \\ 0.134 \end{bmatrix}$$

$$C_t = \begin{bmatrix} 0.44 \\ 0.532 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.57 \\ 0.508 \end{bmatrix}\begin{bmatrix} -0.27 \\ 0.134 \end{bmatrix}$$

$$= \begin{bmatrix} 0.088 \\ -0.1064 \end{bmatrix} + \begin{bmatrix} -0.1539 \\ 0.0734 \end{bmatrix} = \begin{bmatrix} -0.0659 \\ -0.033 \end{bmatrix}$$

Vanishing gradient

$h_t = o_t \odot \tanh(c_t)$

$c_t = f_t \odot c_{t-1} + i_t \odot g_t$

will not go to 0 as no σ or $\tanh$

$\therefore h_t$ will not go to 0

__/__/____

RNN

$$\frac{\partial h_t}{\partial h_{t-1}} = \tanh'\left(W_{hh} h_{t-1} + W_{xh} x_{t-1}\right)$$

< 1

LSTM

$$\frac{\partial h_t}{\partial h_{t-1}} = o_t \tanh'(c_t) \boxed{\frac{\partial c_t}{\partial h_{t-1}}} + \frac{\partial o_t}{\partial h_{t-1}} \tanh(c_t)$$

$$\frac{\partial c_t}{\partial h_{t-1}} = f_t \frac{\partial c_{t-1}}{\partial h_{t-1}} + c_{t-1} \frac{\partial f_t}{\partial h_{t-1}} + g_t \frac{\partial i_t}{\partial h_{t-1}} + i_t \frac{\partial g_t}{\partial h_{t-1}}$$

(expand)

25/9/25

$$\frac{\partial L}{\partial h_{t-1}} = \frac{\partial L}{\partial h_t} \frac{\partial h_t}{\partial h_{t-1}}$$

$$\frac{\partial h_t}{\partial h_{t-1}} = o_t \left(1 - \tanh^2(c_t)\right) \frac{\partial c_t}{\partial h_{t-1}} + \frac{\partial o_t}{\partial h_{t-1}} \tanh(c_t)$$

$$\frac{\partial c_t}{\partial h_{t-1}} = \frac{\partial}{\partial h_{t-1}} \left[f_t \odot c_{t-1} + i_t \odot g_t\right]$$

$$= f_t \frac{\partial c_{t-1}}{\partial h_{t-1}} + c_{t-1} \frac{\partial f_t}{\partial h_{t-1}} + g_t \frac{\partial i_t}{\partial h_{t-1}} + i_t \frac{\partial g_t}{\partial h_{t-1}}$$

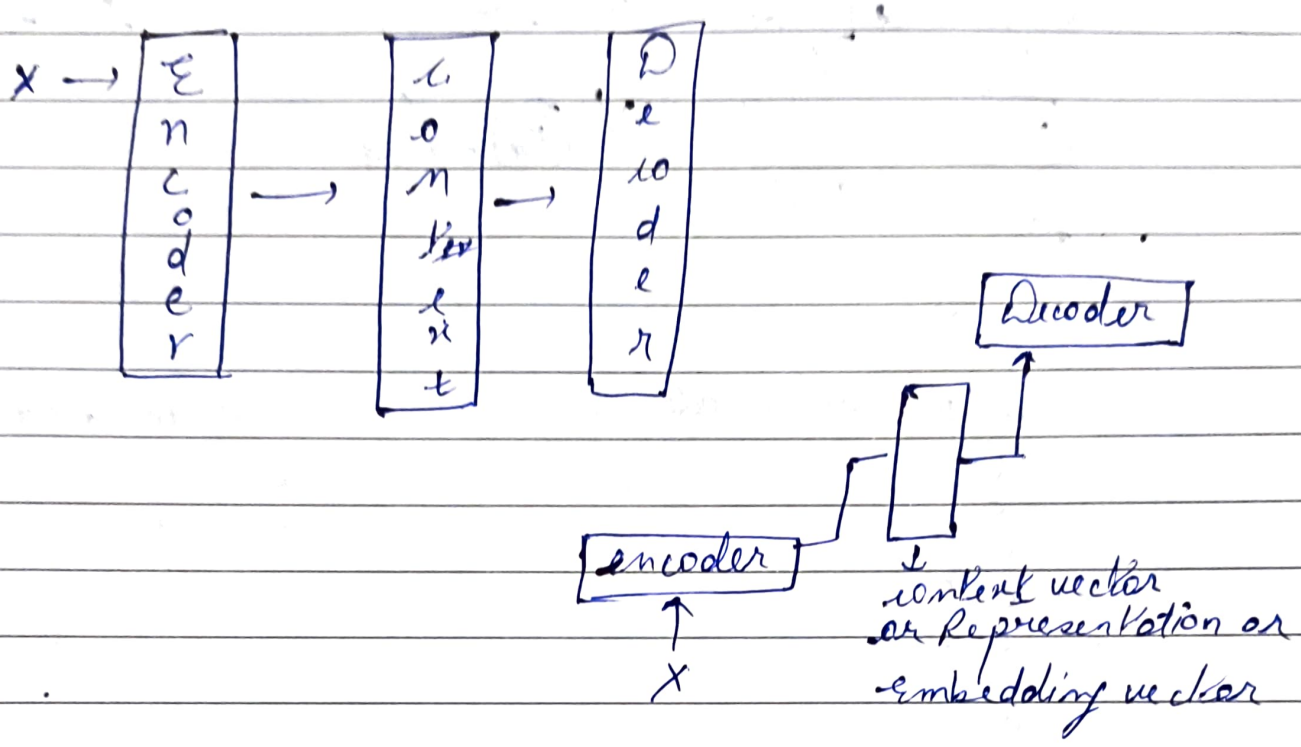
put in $\frac{\partial h_t}{\partial h_{t-1}}$

we can see there are many additive terms

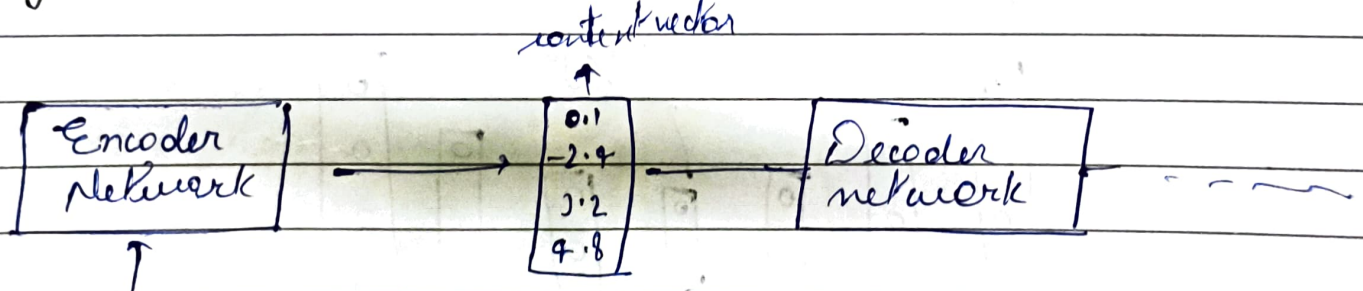
$\therefore$ vanishing gradient is avoided # theoretically can go to 0

New

Encoder Decoder Networks



Eg# Machine Translation Task



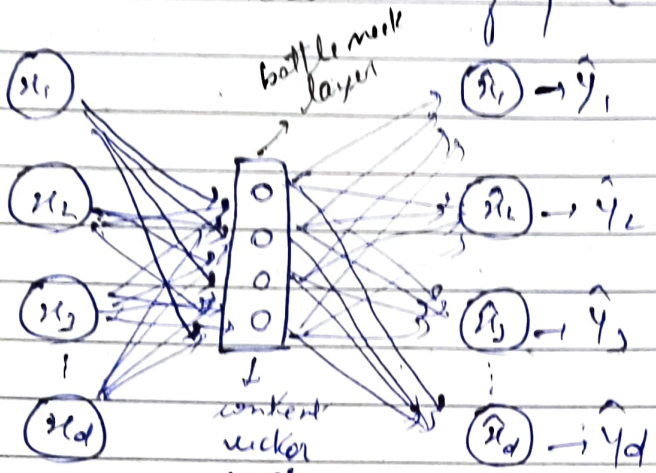
→ sequence to sequence learning

· Dataset
· ML to simulate data

~~Datasets~~

Linear / non-linear
· PCA
· tSNE

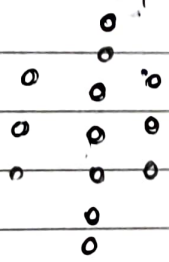
1) Auto encoders → to better representation of input (main purpose)



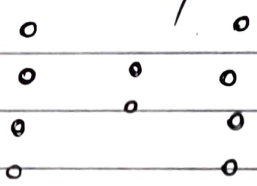
→ Learn to reconstruct input.

$y = f(x)$
f = identity function

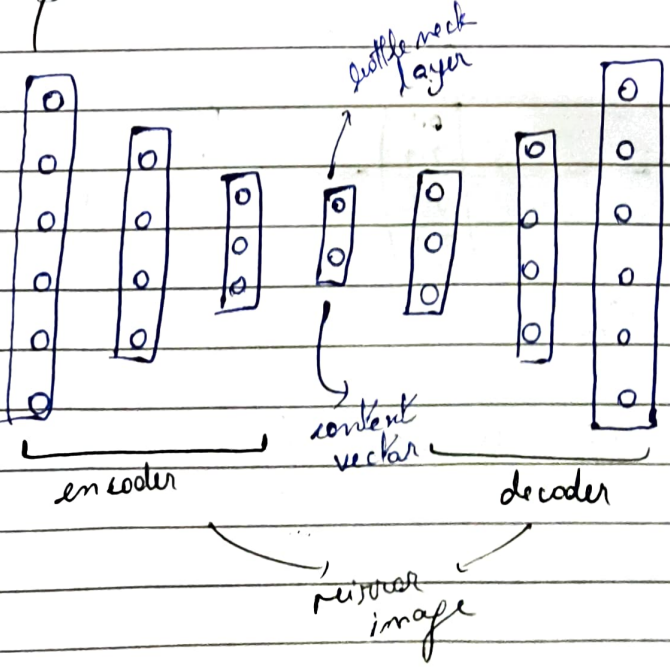
over complete autoencoder



under complete autoencoder

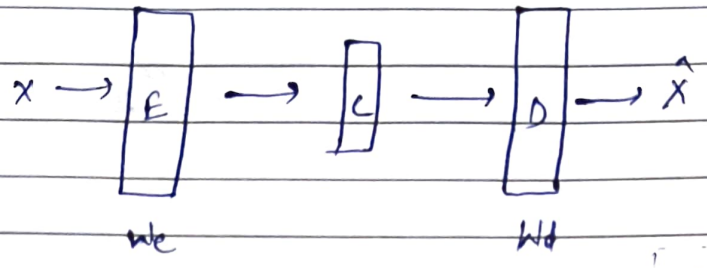


2) Deep auto encoders



$$L = \frac{1}{2}\|x - \hat{x}\|_2$$

30/9/25



Q. $x \in \mathbb{R}^2$, $x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

Autoencoder = Loss → MSE
If = PCA = Linear (encoder)
= Linear (Decoder)

$h \in R^1$ (context vector)

We and Wd = ?

We: 1×2 $= [0.5 \quad -1.0]_{1\times 2}$

Wd: 2×1 $= \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix}_{2\times 1}$

One forward pass and Loss = ?

$$h = \underset{2\times 1}{\begin{bmatrix} 2 \\ 0 \end{bmatrix}} \underset{1\times 2}{\begin{bmatrix} 0.5 & -1.0 \end{bmatrix}} = \underset{2\times 2}{\begin{bmatrix} 1 & -2 \\ 0 & 0 \end{bmatrix}}$$

$\hat{x} =$

$$h = \begin{bmatrix} 0.5 & -1.0 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \end{bmatrix} = [1]$$

$$\hat{x} = h\,Wd = [1] \times \begin{bmatrix} 1.0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

$$\text{Loss} = \frac{1}{2}\left((1-2)^2 + (0.5-0)^2\right)$$

$$= \frac{1}{2}(1 + 0.25) = \frac{1.25}{2}$$

$$= 0.625$$

7.4
2.7
__/__/____

Eg:- If $h_1 = [1, 0, 1]$, $h_2 = [0, 1, 1]$, $h_3 = [1, 1, 0]$
$S_{t-1} = [1, 0, 1]$, score function = dot product.
$C_t = ?$

Ans:

$$C_t = \sum_{j=1}^{T} \alpha_{t,j} h_j$$

$$C_t = \alpha_{t,1} h_1 + \alpha_{t,2} h_2 + \alpha_{t,3} h_3$$

$$\alpha_{t,1} = \frac{\exp(\text{score}(S_{t-1}, h_1))}{\exp(\text{score}(S_{t-1}, h_1)) + \exp(\text{score}(S_{t-1}, h_2)) + \exp(\text{score}(S_{t-1}, h_3))}$$

$$= \frac{7.4}{12.8} = 0.58$$

$$\alpha_{t,2} = \frac{2.7}{12.8} = 0.2$$

$$\alpha_{t,3} = \frac{2.7}{12.8} = 0.2$$

$$C_t = 0.58 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + 0.2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + 0.2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.58 \\ 0 \\ 0.58 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.2 \\ 0.2 \end{bmatrix} + \begin{bmatrix} 0.2 \\ 0.2 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.78 \\ 0.4 \\ 0.78 \end{bmatrix}$$

9/10/25

Thinking

Thinking

① p_1 = Thinking — q_1 = Machines
k_1 = Thinking — k_1 = Thinking
v_1 = Thinking — v = Machines

② p_1' = Thinking — q_1' = Machines
k_2 = Machines — k_2 = Machines
v = Thinking — v = Machines

i/p = Playing Outside ⇒ z_1, z_2

Playing:

$q_1 = [0.212 \quad 0.04 \quad 0.63 \quad 0.36]^T$

$k_1 = [0.31 \quad 0.84 \quad 0.963 \quad 0.57]^T$

$v_1 = [0.36 \quad 0.63 \quad 0.1 \quad 0.78]^T$

Outside:

$q_2' = [0.1 \quad 0.14 \quad 0.86 \quad 0.77]^T$

$k_2 = [0.45 \quad 0.94 \quad 0.73 \quad 0.58]^T$

$v_2 = [0.31 \quad 0.36 \quad 0.19 \quad 0.72]^T$

$Z_1 = \begin{bmatrix} 0.1334 & 0.588 & 0.496 & 0.5594 \\ [0.1136 & 0.6046 & 0.1431 & 0.5930] \end{bmatrix}$

Score $q \cdot k_1$

$= 0.06572 + 0.0336 + 0.60669 + 0.2052$

$= 0.911121$

$q_1 * k_2$

$= 0.0954 + 0.03768 + 0.4593 + 0.2088$

$= 0.8017$

Divided by $\sqrt{d_k} = \frac{\sqrt{4}}{2}$

$= \frac{0.911121}{2}$

$= 0.455605$

≈ 0.45

$\frac{0.8017}{2}$

$= 0.40085$

≈ 0.4

Softmax $\frac{1.57}{1.57+1.43} = \frac{1.57}{3.06}$

$\frac{1.43}{1.57+1.43} = \frac{1.43}{3.06}$

$= 0.513$

0.487

Softmax x Value $= 0.513 \cdot \begin{bmatrix} 0.36 \\ 0.83 \\ 0.1 \\ 0.38 \end{bmatrix} + 0.487 \begin{bmatrix} 0.31 \\ 0.36 \\ 0.19 \\ 0.72 \end{bmatrix}$

$= \begin{bmatrix} 0.18468 \\ 0.42579 \\ 0.0513 \\ 0.19494 \end{bmatrix} + \begin{bmatrix} 0.15097 \\ 0.17532 \\ 0.09253 \\ 0.35067 \end{bmatrix}$

$= \begin{bmatrix} 0.33565 \\ 0.60111 \\ 0.14383 \\ 0.54558 \end{bmatrix}$

Flux Balance

__/__/__

Diffusion model

$$x_0 \in \mathbb{R}^d \sim q(x_0) \quad \text{ground truth dist.}$$

Problem: sample from p_{x_0} (unknown)

Suppose p_θ is the "model" distribution

$$D_{KL}(p_{x_0} \| p_\theta) = \int_{x_0} p_{x_0} \log \frac{p_{x_0}}{p_\theta} dx_0$$

$$\theta^* = \arg\min_\theta D_{KL}\left[p_{x_0} \| p_\theta\right]$$

$$= \arg\min_\theta \left[\int p_{x_0} \log p_{x_0} - \int p_{x_0} \log p_\theta\right]$$

$$= \arg\max_\theta \int_{x_0} p_{x_0} (\log p_\theta) dx_0$$

$$= \arg\max_\theta \mathop{\mathbb{E}}_{p_{x_0}} \log p_\theta(x_0)$$

$x \sim p_x$ $\int_x f(x) p_x(x) dx$

$\mathop{\mathbb{E}}_{p_x}(f(x))$

$$\mathop{\mathbb{E}}_{p_x} f(x) \approx \frac{1}{N} \sum_{i=1}^{N} f(x_i) \qquad x_i \sim p_x$$

LLN $N \to \infty$

$$\approx \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i)$$

$x_i \sim p_{x_0}$

MLE

Diffusion Models

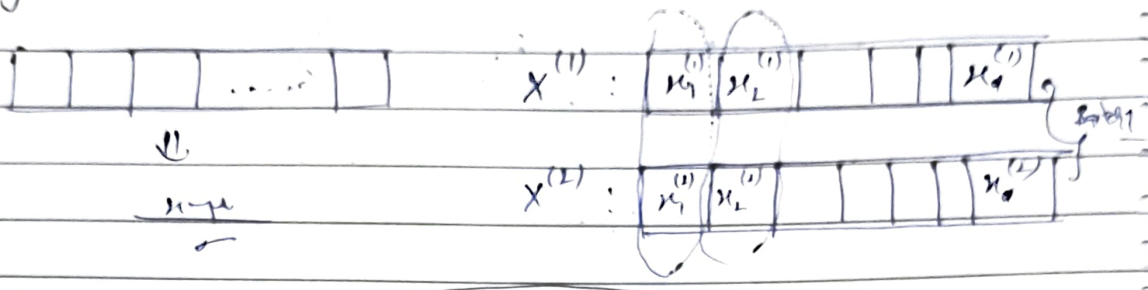
p_θ $\quad x_0 \xrightarrow{\text{Enc}} z \xrightarrow{\text{Dec}} x_0$ $\quad z \sim \mathcal{N}(0, I)$

↳ latent variable

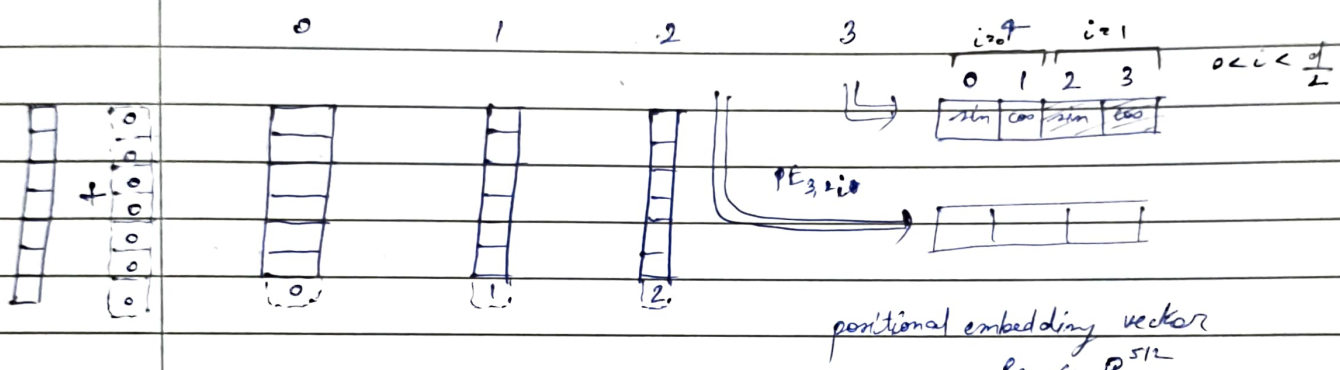
Latent model

13/10/25

Layer Normalization ⇒ Batch Normalization



Key components of Transformers architecture



positional embedding vector
$PE \in \mathbb{R}^{512}$

Positional Encoding

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{emb}}}}\right)$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{emb}}}}\right)$$

eg:

I am a robot

$PE : d_{24} \Rightarrow \mathbb{R}^4$

I	→	0	→	P_{00}	P_{01}	P_{02}	P_{03}
am	→	1	→	P_{10}	P_{11}	P_{12}	P_{13}
a	→	2					
robot	→	3	→	P_{30}	P_{31}	P_{32}	P_{33}

$$P_{00} = \sin\left(\frac{0}{10000^{\frac{2\times0}{4}}}\right)$$

$$= \sin(0) = 0$$

$$P_{01} = \cos\left(\frac{0}{10000^{\frac{2\times0}{4}}}\right) = \cos\left(\frac{0}{1}\right)$$

$$= \cos(0) = 1$$

$$P_{02} = \sin\left(\frac{0}{10000^{\frac{2\times1}{4}}}\right) = \sin(0) = 0$$

$$P_{03} = \cos(0) = 1$$

$$P_{10} = \sin\left(\frac{1}{10000^{\frac{2\times0}{4}}}\right) = \sin(1) = 0.84$$

$$P_{11} = \cos\left(\frac{1}{10000^{\frac{2\times0}{4}}}\right) = \cos(1) = 0.54$$

$$P_{12} = \sin\left(\frac{1}{10000^{\frac{2\times1}{4}}}\right) = \sin\left(\frac{1}{100}\right) = \sin\left(\frac{1}{10}\right) = 0.0998$$

$$P_{13} = \cos\left(\frac{1}{10000^{\frac{2\times1}{4}}}\right) = \cos\left(\frac{1}{10}\right) = 0.995$$

similarly for rest

~~###~~ P_{500} & dot product = P_{501} [ ]

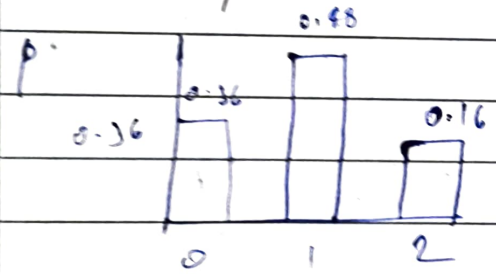
P_{10000} & dot product = P_{10001} [ ]

input : [ ] [i/p] + [PE]

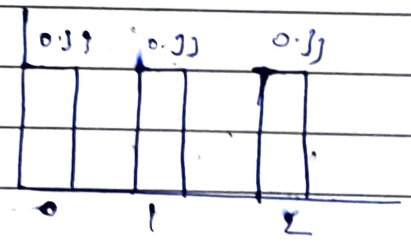
(in the paper)

output = [ ]

KL example



Binomial distance



Uniform dist

Distribution x	0	1	2
$P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
$Q(x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

$$D_{KL}(P \| Q) = \sum_{x \in X} P(x=x) \ln \frac{P(x=x)}{Q(x=x)}$$

$$= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right)$$

$$\approx 0.0852936$$

$$D_{KL}(Q \| P) = \frac{1}{3} \ln\left(\frac{1/3}{9/25}\right) + \frac{1}{3} \ln\left(\frac{1/3}{12/25}\right) + \frac{1}{3} \ln\left(\frac{1/3}{4/25}\right)$$

$$= 0.097455$$

Variational Divergence minimization

opt fun^{n}

$f(u) = u \log u - (u+1) \log(u+1)$

Generative Adversial Network (GAN)

zero-sum game

opt. this

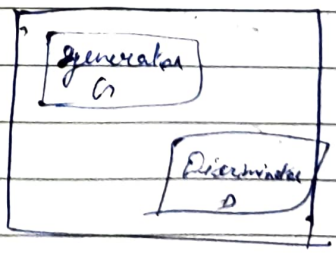
$X = \{x_1, x_2, \ldots x_n\} \sim$ iid

$\sim p_x$ real or True

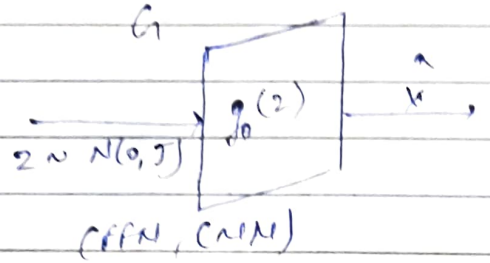
$z \sim N(0, I)$ → $g_\theta(z)$ → $\hat{x} \approx p_\theta$ fake

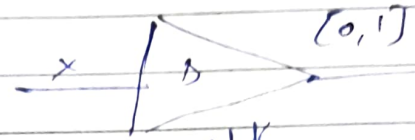
g_θ = generator

$x \sim p_x$, $\hat{x} \sim p_\theta$, $p_x \approx p_\theta$



$x \rightarrow D_w(x) \rightarrow [1, 0]$





$$J_{GAN}(\theta, w) = \mathop{\mathbb{E}}_{x \sim P_x} \log D_w(x) + \mathop{\mathbb{E}}_{\hat{x} \sim P_\theta} \log\left(1 - D_w(\hat{x})\right)$$

$$\hat{x} = g_\theta(z)$$

input : $\{x_1, x_2, \ldots x_n\} \sim iid \sim P_x$

$B_1 : x_1, x_2 \ldots x_{B_1} \sim P_x$

$B_2 : \hat{x}_1, \hat{x}_2 \ldots \hat{x}_{B_2} \sim P_\theta$

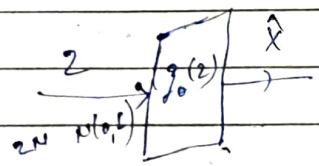
$$\min_{G} \max_{D} \left[\frac{1}{B_1} \sum_{i=1}^{B_1} \log D_w(x_i) + \frac{1}{B_2} \sum_{j=1}^{B_2} \log\left(1 - D_w(\underbrace{g_\theta(z_j)}_{\hat{x}_j})\right) \right]$$

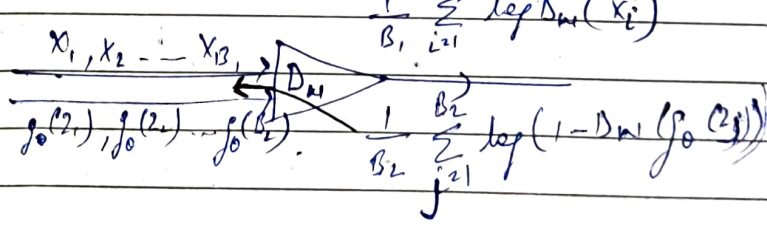
$$J_{GAN}(\theta, w) = \underset{\text{argmax}}{} \left[\frac{1}{B_1} \sum_{i=1}^{B_1} \log D_w(x_i) + \frac{1}{B_2} \sum_{j=1}^{B_2} \log\left(1 - D_w(g_\theta(z_j))\right) \right]$$

~~Gen~~

To Discriminator

Keep θ as constant





$$\nabla_w J \quad w^{t+1} = w^t + \alpha_t \nabla_w J$$

$$\theta^* = \underset{\theta}{\text{argmin}}\ J_{GAN}(\theta, w)$$

$$\approx \underset{\theta}{\text{argmin}} \left[\frac{1}{B_1} \sum_{i=1}^{B_1} \log D_w(x_i) + \frac{1}{B_2} \sum_{j=1}^{B_2} \log\left(1 - D_w(g_\theta(z_j))\right) \right]$$

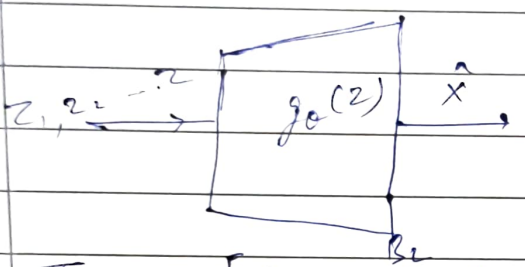
$\hat{x}_j$

__/__/____

$$\theta^* = \arg\min_{\theta} \left[\frac{1}{B_2} \sum_{j=1}^{B_2} \log\left(1 - D_W(g_\theta(z_j))\right) \right]$$

$$\theta : \theta - \alpha_2 \nabla_\theta J(\theta, W)$$

To train generator



$$J_{GAN} = \left[\frac{1}{B_2} \sum_{j=1}^{B_2} \log\left(1 - D_W\left(g_\theta(z_i)\right)\right) \right]$$

$$\theta^{t+1} = \theta^t - \alpha_2 \nabla_\theta J$$