Lec 7 → Gradient Vanishing ⟷ Model Architecture (stability)

Re-cap:

① $\begin{cases} \nabla_{W_t} \ell(f_\theta(x_{T+1}), y) = \nabla_{W_t} x_{t+1} \nabla_{x_{t+1}} \ell = \nabla_{W_t} g_t(x_t; W_t) \cdot p_{t+1} \\ p_t := \nabla_{x_t} \ell = \nabla_{x_t} x_{t+1} \cdot \nabla_{x_{t+1}} \ell = \nabla_{x_t} g_t(x_t; W_t) \cdot p_{t+1} \end{cases}$

⇓

for small t, $\boxed{p_t \to 0}$ since $p_t = \prod_{i=t}^{T} \gamma_i \; p_{T+1}$, $\prod_{i=t}^{T} \gamma_i \to 0$

② For 1-dim FCNN without bias, we have

$$p_t = \frac{d}{dx_t} \cdot [\sigma(W_t x_t)] \cdot p_{t+1}$$

$$= \underbrace{W_t}_{\text{weight init.}} \cdot \underbrace{\sigma'(w_t \cdot x_t)}_{\text{choice of activation function}} \cdot p_{t+1}$$

☆ $\boxed{GV}$

→ for activation function, we hope: $\sigma'(x) \approx 1$ ⟷ ReLu

we don't want $\sigma'(x) \ll 1$ ⟷ sigmoid

→ for initialization of weight, we are interested in

$$\gamma_t := \frac{p_t}{p_{t+1}} = w_t \cdot \sigma'(w_t \cdot x_t) \qquad x_0$$

↓

Suppose $W_t \sim N(0, \gamma^2)$ here, w_t & x_t are R.V.

Conclusion: a) $\mathbb{E}_{w_t, x_t}[\gamma_t] = 0$

→ $\boxed{\text{suggest } \gamma^2 = 2}$

b) $\mathbb{E}_{w_t, x_t}[\gamma_t^2] = \frac{\gamma^2}{2}$

→ a more general result: FCNN → depth T, → width d_t in t-th layer

⇓

☆ stablize forward prop. & backward prop.

$$w_t^{ij} \sim N(0, \frac{2}{d_t}) \qquad \boxed{W_t \in \mathbb{R}^{d_{t+1} \times d_t}}$$

→ $\boxed{\text{idea}}$: stablize $\mathbb{E}_{w_t, x_t}[(x_{t+1}^i)^2]$

Today's lecture :

1. All previous discussion is based on trivial FCNN / CNN

$$\Rightarrow \begin{cases} x_{t+1} = \sigma(w_t \cdot x_t + b_t) & t = 0, 1, \dots, T-1 \\ x_{T+1} = v^T x_T \end{cases}$$

Question: can we modify the ARCHITECTURE to solve Gradient Vanishing Problem ?

2. Intuition : for 1-dim FCNN without bias example :

$$P_t = \boxed{w_t \, \sigma'(w_t \cdot x_t)} \, P_{t+1}$$

→ may goes to 0

idea : can we achieve $P_t = \boxed{(1 + *)} \cdot P_{t+1}$?

→ goes to 1

ResNet Architecture :

$$\begin{cases} x_{t+1} = x_t + \boxed{F_t(x_t)} & t = 0, 1, \dots, T-1 \\ x_{T+1} = v^T x_T \end{cases}$$

→ t-th residual block

→ can be complicated

→ $\boxed{x_{t+1} - x_t} = F_t(x_t)$

↓

"residual"

compute gradient for 1-dim case :

$$\longrightarrow \frac{dP}{dx_t} = \frac{dP}{dx_{t+1}} \cdot \frac{dx_{t+1}}{dx_t} = \frac{dP}{dx_{t+1}} \cdot \left(1 + \frac{dF_t}{dx_t}\right)$$

$$= \frac{dP}{dx_{T+1}} \prod_{k=t}^{T} \left(1 + \frac{dF_k}{dx_k}\right)$$

not ≈ 0

Remark:

1. $\boxed{\text{"1"}}$ plays crucial role in preserving the magnitude of gradient !!! $\longrightarrow$ $\boxed{\text{why it will help solve GD}}$

2. However, it may cause Gradient Exploding

   $\to$ **Solution**:

   ① In practice, we can use Normalization Layer (BN....)

   $\Downarrow$

   try to normalize the features after each layer

   ② modify $\dfrac{d\mathcal{F}_k}{dx_k} \longrightarrow \alpha_k \cdot \dfrac{d\mathcal{F}_k}{dx_k}$

   $\Rightarrow$ scaling k-th residual blocks with factor α_k

3. Scaling k-th residual block to avoid Gradient Exploding

$\longrightarrow$ $\boxed{\text{Example}}$: simple ResNet with width d (fixed)

$$\begin{cases} x_{t+1} = x_t + \boxed{W_t \cdot \sigma(x_t)} & t = 0, 1, \dots, T-1 \\ x_{T+1} = T_{fc} \cdot x_T \\ x_0 = W_{in} \cdot x \end{cases}$$

$\boxed{W_t \in \mathbb{R}^{d \times d}}$ → $\mathcal{F}_t(x_t)$

$\boxed{W_t \in \mathbb{R}^{d\times d}}$, we assume $W_t^{ij} \sim N(0, \frac{\gamma_t^2}{d})$

$\longrightarrow$ Our interest is: $\mathbb{E}_{W_t, x_t}[(x_{t+1}^i)^2]$

Calculation: $(x_{t+1}^i)^2 = \left(x_t^i + \sum_{j=1}^{d} W_t^{ij} \cdot \sigma(x_t^j)\right)^2$

Therefore, $\mathbb{E}_{W_t, x_t}[(x_{t+1}^i)^2]$ $\qquad$ $x_t = f(W_{t-1}, \dots, W_0, W_{in}; x)$

$$= \mathbb{E}_{\cancel{W_t},\, x_t}\left[(x_t^i)^2 \right] + \mathbb{E}_{W_t,\, x_t}\left[\left(\sum_{j=1}^{d} w_t^{ij}\, \sigma(x_t^j) \right)^2 \right]$$

$$+\ 2 \cdot \mathbb{E}_{W_t,\, x_t}\left[x_t^i \cdot \sum_{j=1}^{d} w_t^{ij}\, \sigma(x_t^j) \right]$$

$\boxed{\text{second term}}$

$$\mathbb{E}_{W_t,\, x_t}\left[\left(\sum_{j=1}^{d} w_t^{ij}\, \sigma(x_t^j) \right)^2 \right]$$

$$= \mathbb{E}_{x_t}\left\{ \mathbb{E}_{W_t}\left[\sum_{1 \le j, j' \le d} w_t^{ij} w_t^{ij'}\, \sigma(x_t^j) \cdot \sigma(x_t^{j'}) \,\middle|\, x_t \right] \right\}$$

$$= \mathbb{E}_{x_t}\left\{ \mathbb{E}_{W_t}\left[\sum_{j=1}^{d} (w_t^{ij})^2 \cdot \sigma(x_t^j)^2 \,\middle|\, x_t \right] \right\}$$

$$= \mathbb{E}_{x_t}\left\{ \sum_{j=1}^{d} \mathbb{E}_{W_t}\left[(w_t^{ij})^2 \cdot \sigma(x_t^j)^2 \,\middle|\, x_t \right] \right\}$$

$\boxed{\text{independence}}$

$$x_t = f(W_{t-1}, \dots, W_0) \perp\!\!\!\perp W_t$$

$$= \sum_{j=1}^{d} \mathbb{E}_{x_t, W_t}\left[(w_t^{ij})^2 \cdot \sigma(x_t^j)^2 \right]$$

Note:

$\{x_t^j\}_{j=1}^{d} \longrightarrow$ same distribution

(not independent)

$\Downarrow$

$\boxed{\text{same expectation}}$

$$= \sum_{j=1}^{d} \mathbb{E}_{x_t^j}\left[\sigma(x_t^j)^2 \right] \cdot \mathbb{E}_{w_t^{ij}}\left[(w_t^{ij})^2 \right]$$

$$= \sum_{j=1}^{d} \cdot \mathbb{E}_{x_t^1}\left[\sigma(x_t^1)^2 \right] \cdot \frac{\gamma_t^2}{d}$$

$$= \gamma^2 \cdot \mathbb{E}_{x_t^1}\left[\sigma(x_t^1)^2 \right]$$

$\boxed{\text{Third term}}$

$$\mathbb{E}_{W_t,\, x_t}\left[x_t^i \sum_{j=1}^{d} w_t^{ij}\, \sigma(x_t^j) \right]$$

$$= \sum_{j=1}^{d} \underbrace{\mathbb{E}_{w_t^{ij}}\left[w_t^{ij} \right]}_{=0} \cdot \mathbb{E}_{x_t}\left[x_t^i \cdot \sigma(x_t^j) \right]$$

(independence similarly)

$$= 0$$

Therefore: $\mathbb{E}_{\substack{W_t, x_t \\ \{W_k\}_{k<t}}} [(x^i_{t+1})^2]$

$$= \mathbb{E}_{x^i_t} [(x^i_t)^2] + \gamma_t^2 \, \mathbb{E}_{x^i_t} [\sigma (x^i_t)^2]$$

Note: given x, $x_0 = W_{in} x \Rightarrow x^i_0$ is symmetric w.r.t pdf

since $x^i_0 = \sum_{j=1}^{d} x^j \, w^{ij}_{in} \rightarrow \boxed{\text{Gaussian}(0, \frac{\gamma^2}{d})}$

Now, $x_0 = W_{in} x \longrightarrow \boxed{\text{symmetric}}$

$x_1 = x_0 + W_0 \, \sigma(x_0) \longrightarrow$ symmetric $\boxed{???}$ $\downarrow$!!! $\boxed{\text{shown later}}$

$$= \mathbb{E}_{x^i_t} [(x^i_t)^2] + \frac{\gamma_t^2}{2} \cdot \mathbb{E}_{x^i_t} [(x^i_t)^2]$$

the conclusion if x_t symmetry

$$= \left(1 + \frac{\gamma_t^2}{2}\right) \mathbb{E}_{x^i_t} [(x^i_t)^2]$$

$$= \cdots = \boxed{\left(1 + \frac{\gamma_t^2}{2}\right)^t} \mathbb{E}_{x^i_1} [(x^i_1)^2]$$

$\downarrow$

exploding behavior !!!!

$$\boxed{W^{ij}_t \sim \mathcal{N}(0, \frac{\gamma_t^2}{d})}$$

$\longrightarrow$ Solution: intuition: initialize the weights whose variance is decreasing with depth

① We can chose $\gamma_t^2 = \frac{1}{t} \longrightarrow$ decrease with depth

then $\mathbb{E}_{W_t, x_t} [(x^i_{t+1})^2] = \left(1 + \frac{1}{2t}\right)^t \mathbb{E} [(x^i_1)^2]$

$\longrightarrow e^{\frac{1}{2}}$ as $t \to \infty$

$\downarrow$

$\boxed{\text{no longer exploding}}$

② Moreover, we can modify the architecture to solve:

$$\begin{cases} x_{t+1} = x_t + \lambda_t \cdot W_t \cdot \sigma(x_t) \\ W_t^{ij} \sim \mathcal{N}\left(0, \frac{\gamma^2}{d}\right) \end{cases}$$

then $\mathbb{E}_{x_t, w_t}[(x_{t+1}^i)^2] = \left(1 + \frac{\gamma^2 \cdot \lambda_t^2}{2}\right)^T \cdot \mathbb{E}[(x_1^i)^2]$

we can choose $\lambda_t^2 = \frac{1}{T}$ to avoid exploding

$$\Rightarrow \boxed{x_{t+1} = x_t + \frac{1}{\sqrt{T}} \cdot W_t \cdot \sigma(x_t)}$$

→ modification on architecture

Show the symmetry of x_t

① $x_0 = W_{in}\, x \Rightarrow x_0^i = \sum_{j=1}^{d} x^j \cdot W_{in}^{ij}$ → symmetric (Gaussian)

$$\boxed{\begin{aligned} &\mathbb{P}(aX + bY \le c) \\ &= \mathbb{E}_{X,Y}[\mathbb{1}\{aX + bY \le c\}] \\ &= \mathbb{E}_{X,Y}[\mathbb{1}\{-aX - bY \le c\}] \\ &= \mathbb{P}(aX + bY \ge -c) \end{aligned}}$$

$\Rightarrow x_0^i$ is symmetric RV $i = 1, 2, 3, \dots, d$

naturally symmetric

$(\cdot)^i \,|\, x_t = \sum_{j=1}^{d} \sigma(x_t^j) W_t^{ij} \,\Big|\, x_t \sim \mathcal{N}(\cdot,\cdot)$

② $x_{t+1} = \boxed{x_t} + \boxed{W_t \cdot \sigma(x_t)}$ → symmetric

$\mathbb{P}\left(\sum_{j=1}^{d} \sigma(x_t^j) W_t^{ij} \le x\right)$

$$= \mathbb{E}\left[\mathbb{1}\left\{ \sum_j \sigma(x_t^j)\, W_t^{ij} \le x \right\} \right]$$

$$= \mathbb{E}_{x_t}\left[\mathbb{E}_{W_t}\left[\mathbb{1}\left\{ \sum_j \sigma(x_t^j) W_t^{ij} \le x \right\} \,\middle|\, x_t \right] \right]$$

$$= \mathbb{E}_{x_t}\left[\mathbb{E}_{W_t}\left[\mathbb{1}\left\{ \sum \sigma(x_t^j) W_t^{ij} \ge -x \right\} \,\middle|\, x_t \right] \right]$$

since $\sum \sigma(x_t^j) W_t^{ij} \mid x_t$ is symmetric

$$= \mathbb{P}\left(\sum_{j=1}^{d} \sigma(x_t^j)\, W_t^{ij} \ge -x \right)$$

Recap: $x_{t+1} = \underbrace{x_t}_{\text{symmetric}} + \underbrace{W_t \cdot \sigma(x_t)}_{\text{symmetric}} \Rightarrow x_{t+1} \longrightarrow$ symmetric

☆☆☆

$\longrightarrow$ **Remark**:

Up to now, we have explored recipes for Gradient Vanishing

① Activation function $\longrightarrow \sigma'(z) \approx 1 \Longleftrightarrow \boxed{\text{ReLu}}$

② Architecture $\longrightarrow$ skip-connection (Residual Block)

③ Initialization:

☆ for the first time of calculating forward propagation $\to$ the first update

a) FCNN: $W_t^{ij} \sim N\left(0, \frac{2}{d_t}\right)$

$\to$ maintain $\mathbb{E}\left[(x_{t+1}^i)^2 \right] = \mathbb{E}\left[(x_1^i)^2 \right]$ in forward prop.

b) ResNet: $W_t^{ij} \sim N\left(0, \frac{1}{d_t} \cdot \boxed{\frac{1}{t}}\right)$ $\to$ decay for later layer

$\longrightarrow$ maintain $\mathbb{E}\left[(x_{t+1}^i)^2 \right] = C \cdot \mathbb{E}\left[(x_1^i)^2 \right] \to$ not explode as $t \to \infty$

Lemma: if $z \to$ symmetric R.V.

then $\mathbb{E}_z[\sigma^2(z)]$

$$= \mathbb{E}_z[z^2 \cdot \mathbb{1}\{z \geq 0\}]$$

$$= \frac{1}{2}\mathbb{E}_z[z^2]$$