DSA5202 Lec6

Recap:

Lec5

① FCNN $\begin{cases} \text{shallow} \\ \text{deep} \to \begin{cases} f(x) = V^T f_T(x) \\ f_{t+1}(x) = \sigma(W_t f_t(x) + b_t) \end{cases} \end{cases}$

$t = 0, 1, \dots, T-1$

② Empirical Risk Minimization (ERM) $\Rightarrow$ $\boxed{\text{Training set}}$

$$\boxed{\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \ \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), y_i)}$$

optimize $\to \begin{cases} \text{GD} \\ \text{SGD} \\ \text{Momentum GD} \end{cases}$

③ BP Algorithm to calculate gradient $\begin{cases} \text{forward} \\ \text{backward} \end{cases}$

(Batch update in practice)

forward: $x \to x_1 \to \cdots \to x_T \to x_{T+1}$

backward:

$$\begin{array}{ccccc}
x_{T+1} \to p_{T+1} := \nabla_{x_{T+1}} \ell & \longrightarrow & \nabla_{W_T} \ell & \leftarrow & x_T \\
\downarrow & x_T = g_{T-1}(x_{T-1}, W_{T-1}) & & & \\
p_T & \longrightarrow & \nabla_{W_{T-1}} \ell & \leftarrow & x_{T-1} \\
\downarrow & & & & \\
\vdots & & & & \\
\downarrow & & & & \\
p_1 & \longrightarrow & \nabla_{W_0} \ell & \leftarrow & x_0 = x
\end{array}$$

①

$$p_t := \nabla_{x_t} \ell(x_{T+1}, y)$$
$$= \nabla_{x_t} g_t(x_t, W_t) \cdot p_{t+1}$$

②

$$\nabla_{W_t} \ell(x_{T+1}, y)$$
$$= \nabla_{W_t} g_t(x_t, W_t) \cdot p_{t+1}$$

$\boxed{\text{depth}}$

→ **Issue**: when T is large, the gradient is potential to $\begin{cases} \text{vanish} \\ \text{explode} \end{cases}$

$\boxed{\text{especially for RNN!}}$

↳ comes from the nature of BP

$$P_t = \underbrace{\prod_{i=t+1}^{T+1} \gamma_i}_{\text{tend to } \begin{cases} \text{explode} \\ \text{vanish} \end{cases}} P_{T+1}$$

④ CNN ⟶ convolution operation

Today's lecture :

1. Gradient Vanishing / Gradient Exploding

$\boxed{\text{E.g. (intuition)}}$ (toy example)

↓

FCNN : $x_{t+1} = \sigma(w_t \cdot x_t)$ $\xrightarrow{}$ $\boxed{\text{omit bias}}$ $t = 0, 1, \dots, T$

$w_t, x_t \in \mathbb{R}$

consider ⟹ $P_t = \nabla_{x_t} g_t(x_t, w_t) \cdot P_{t+1}$

$$= w_t \cdot \sigma'(w_t \cdot x_t) \cdot P_{t+1} \qquad t = 0, 1, \dots, T$$

initialization ← w_t; $\sigma'(w_t \cdot x_t)$ → activation function

$$P_{T+1} = \nabla_{x_{T+1}} \ell(x_{T+1}, y)$$

Insight : $\boxed{\text{gradient vanishing / exploding}}$ ⟺ $\begin{cases} \text{activation function} \\ \text{weight initialization} \end{cases}$

↳ also depends on the $\boxed{\text{choice of architecture}}$

In this example : we use $\boxed{\text{FCNN}}$ (do not utilize architecture like skip-connection)

↓

$\boxed{\text{most naive one}}$

2. choice of $\boxed{\text{activation function}}$

$\downarrow$ $\boxed{\text{1d FCNN}}$ (omit bias)

Recap last example : $P_t = w_t \cdot \boxed{\sigma'(w_t \cdot x_t)} \cdot P_{t+1}$

$\downarrow$

$\boxed{\text{ideally}}$, we want $|\sigma'| \approx 1$

less likely for GV

Relu : $\sigma(z) = \max\{0, z\} \Rightarrow \sigma'(z) = \begin{cases} 1, & z > 0 \\ 0, & z < 0 \end{cases}$

Sigmoid : $\sigma(z) = \dfrac{1}{1+\exp(-z)} \Rightarrow \sigma'(z) = \dfrac{\exp(-z)}{(1+\exp(-z))^2}$

potential GV

$\Downarrow$

$0 \le \sigma'(z) \le \dfrac{1}{1+\exp(-z)} < 1$

may lead to $\boxed{\text{Gradient Vanish}}$

3. choice of $\boxed{\text{initialization}}$

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

$\Downarrow$

$\gamma_t := \dfrac{P_t}{P_{t+1}} = w_t \cdot \sigma'(w_t \cdot x_t)$ $\Rightarrow$ $\boxed{\text{idea}}$: we want to control this rate

→ Random Initialization : $w_t \sim \text{Gaussian}(0, \gamma^2)$

Here, we choose (assume) $\sigma \rightarrow$ ReLu

$\Rightarrow$ Question : How to choose γ^2 (variance) ?

Answer: we want to analyze $r_t := \frac{P_t}{P_{t+1}} = w_t \cdot \sigma'(w_t \cdot x_t)$

$$\sigma'(z) = \begin{cases} 1, & z > 0 \\ 0, & z < 0 \end{cases}$$

x_t is also random

Conclusion: $\begin{cases} \mathbb{E}_{w_t, x_t}[r_t] = 0 \\ \mathbb{E}_{w_t, x_t}[r_t^2] = \frac{\gamma^2}{2} \end{cases}$ $\Longrightarrow$ suggest us to choose $\boxed{\gamma^2 = 2}$

Derivation: b) $\mathbb{E}_{w_t, x_t}[w_t^2 \cdot \sigma'(w_t \cdot x_t)^2]$

$= \mathbb{E}_{w_t, x_t}[w_t^2 \cdot \mathbb{1}\{w_t \cdot x_t > 0\}]$

if $w_t \sim p(\cdot)$ symmetric.

then: $\mathbb{E}_{w_t}[f(w_t)]$
$= \mathbb{E}_{w_t}[f(-w_t)]$

$= \mathbb{E}_{w_t, x_t}[(-w_t)^2 \mathbb{1}\{(-w_t) \cdot x_t > 0\}]$

$\Rightarrow 2\,\mathbb{E}_{w_t, x_t}[r_t^2] = \boxed{\mathbb{E}_{w_t, x_t}[w_t^2]} = \gamma^2$

$\mathbb{E}_{w_t}[w_t^2]$

$\Rightarrow \mathbb{E}_{w_t, x_t}[r_t^2] = \frac{\gamma^2}{2}$

$\longrightarrow$ Generally, the initialization scheme is:

In DNNs with width d, this becomes $\gamma_d^2 = 2/d$. This is known as Kaiming (or He) initialization scheme. More generally, with different widths d_t, we have

$$W_t^{ij} \sim \mathcal{N}(0, \frac{2}{d_t}).$$

$\rightarrow$ This also solves a problem of vanishing/exploding during forward propagation!!

not only for the gradient back-propagation

for general width d_t (previously, $d_t = 1$)

Consider a FCNN:

$$x_{t+1} = \sigma(W_t \cdot x_t), \quad W_t \in \mathbb{R}^{d_{t+1} \times d_t}$$

Goal →

Here, $W_t^{ij} \sim N(0, \gamma_t^2)$ → we want to determine γ_t^2

x_t is also random!

$$\longrightarrow x_t \in \mathbb{R}^{d_t} \longrightarrow x_t = \begin{pmatrix} x_t^1 \\ \vdots \\ x_t^{d_t} \end{pmatrix} \in \mathbb{R}^{d_t}$$

$$\Rightarrow \boxed{x_{t+1}^i = \sigma\left(\sum_{j=1}^{d_t} W_t^{ij} x_t^j\right)} \rightarrow \text{scaler form}$$

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

since $W_t^{i1}, \ldots, W_t^{id_t} \sim N(0, \gamma_t^2)$

$$\Rightarrow \sum_{j=1}^{d_t} x_t^j \cdot W_t^{ij} \,\Big|\, x_t \sim N\left(0, \sum_{j=1}^{d_t} (x_t^j)^2 \cdot \gamma_t^2\right)$$

$$\Rightarrow \mathbb{E}_{W_t^{i\cdot}, x_t}\left[(x_{t+1}^i)^2\right] = \mathbb{E}_{x_t}\left[\mathbb{E}_{W_t^{i\cdot}}\left[\left\{\sigma\left(\sum_{j=1}^{d_t} W_t^{ij} x_t^j\right)\right\}^2 \Big| x_t\right]\right]$$

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

$$= \boxed{\frac{\gamma_t^2}{2} \cdot \mathbb{E}_{x_t}\left[\|x_t\|_2^2\right]}$$

→ independent with respect to i

$$= \frac{\gamma_t^2}{2} \cdot d_t \, \mathbb{E}_{x_t^i}\left[(x_t^i)^2\right]$$

choice of node in next layer

Lemma: if $z \sim N(0, \alpha^2)$

then $\mathbb{E}[\sigma^2(z)] = \frac{\alpha^2}{2}$

$\Downarrow$

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

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

$$= \frac{\alpha^2}{2}$$

$\Downarrow$

we want $\boxed{\frac{\gamma_t^2 dt}{2} = 1}$ $\rightsquigarrow$ stablize

$$\Updownarrow$$

$$\gamma_t^2 = \frac{2}{dt}$$

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

$$\Downarrow$$

$$r_t := \frac{P_t}{P_{t+1}} = w_t \cdot \sigma'(w_t \cdot x_t)$$

①

$$\Rightarrow \mathbb{E}_{w_t, x_t}[r_t] = \mathbb{E}_{x_t}\left\{\mathbb{E}_{w_t}[r_t \mid x_t]\right\}$$

$$= \mathbb{E}_{x_t}\left\{\mathbb{E}_{w_t}[w_t \cdot \mathbb{1}\{w_t x_t > 0\} \mid x_t]\right\}$$

$$= \mathbb{E}_{x_t}[f(x_t)]$$

$$f(x_t) = \begin{cases} \mathbb{E}_{w_t}[w_t \mathbb{1}\{w_t > 0\}] & \text{if } \underline{x_t > 0} \\ \mathbb{E}_{w_t}[w_t \mathbb{1}\{w_t < 0\}] & \text{if } \underline{x_t < 0} \end{cases}$$

$$\left(\text{Assume : } \begin{bmatrix} P(x_t > 0) = \frac{1}{2} \\ P(x_t < 0) = \frac{1}{2} \end{bmatrix}\right)$$

$$= \frac{1}{2}\left(\mathbb{E}_{w_t}[w_t \mathbb{1}\{w_t > 0\}] + \mathbb{E}_{w_t}[w_t \mathbb{1}\{w_t < 0\}]\right)$$

$$= \frac{1}{2}\mathbb{E}_{w_t}[w_t] = \underline{0}$$