Lecture 3

UAT (Universal Approximation Theorem)

Cybenko 1989

$\forall f \in C([0,1]^d) \qquad f: \mathbb{R}^d \to \mathbb{R}$

σ – activation f.

$\sigma: \mathbb{R} \to \mathbb{R}$

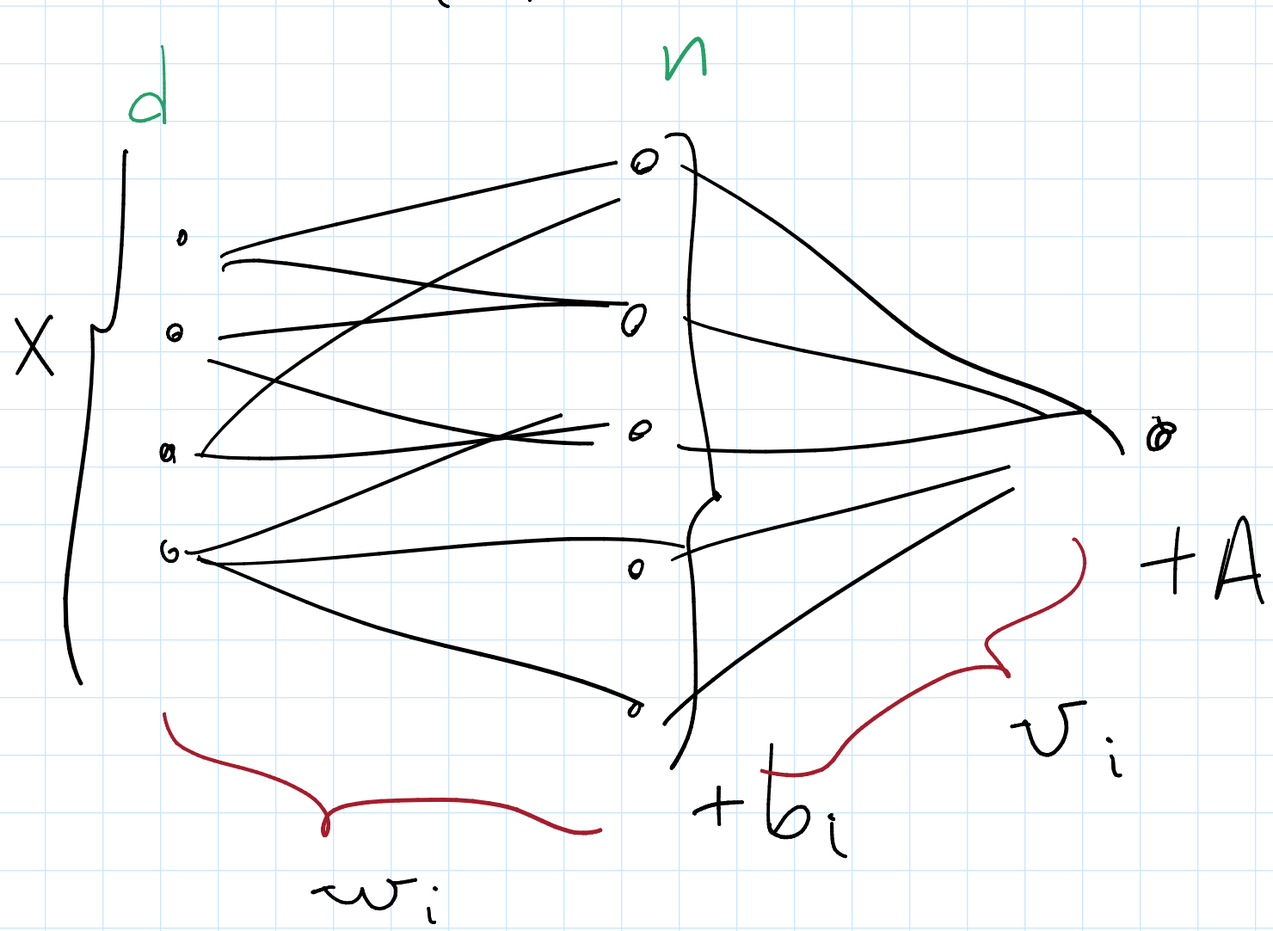
$$\begin{cases} \lim\limits_{x \to -\infty} \sigma(x) = 0 \\ \lim\limits_{x \to +\infty} \sigma(x) = 1 \end{cases} \Rightarrow \text{non-linear}$$

Fix error level $\forall \varepsilon > 0$:

$\exists n \in \mathbb{N}$, $\{w_i\}_{i=1}^n \quad w_i \in \mathbb{R}^d \quad \{b_i\}_{i=1}^n \quad b_i \in \mathbb{R}$

$\{v_i\}_{i=1}^n \quad v_i \in \mathbb{R} \quad A \in \mathbb{R}$

$$g(x) = \sum_{i=1}^{n} v_i \sigma(\langle w_i, x \rangle + b_i) + A$$



$$\sup_{x \in [0,1]^d} |f(x) - g(x)| \leq \varepsilon$$

Optimization

$$Q^t(w) = \frac{1}{B} \sum_{i=1}^{B} L\left(f_w(x_{t_i}), y_{t_i}\right)$$

t – opt. step

$t_i \sim U(\{1, \dots, \ell\})$

$\{(x_i, y_i)\}_{i=1}^{\ell}$

$$\begin{cases} g_t = \nabla_w Q^t(w_{t-1}) \\ m_t = \mu m_{t-1} + g_t \\ w_t = w_{t-1} - \eta_t \cancel{g_t}\ m_t \end{cases}$$

SGD + momentum

η_t — learning rate

$m_0 = 0$, $0 < \mu \leq 1$ $\mu = 0,9$

$\mu m_{t-1} + g_t (\mu - 1)$

$$Q^t(w) \to Q_\lambda^t(w) = Q^t(w) + \frac{\lambda}{2}\|w\|_2^2$$

$$\nabla_w \|w\|_2^2 = 2w$$

$$\begin{cases} g_t = \nabla_w Q^t(w_{t-1}) + \lambda w_{t-1} \\ m_t = \mu m_{t-1} + g_t \\ w_t = w_{t-1} - \eta_t m_t = \end{cases}$$

$$= -w_{t-1} - \eta_t \left(\mu m_{t-1} + \nabla_w Q^t(w_{t-1}) + \lambda w_{t-1} \right) =$$

$$= w_{t-1}\underbrace{(1 - \eta_t \lambda)}_{<1} - \eta_t \underbrace{\left(\mu m_{t-1} + \nabla_w Q^t(w_{t-1}) \right)}_{m_t^{old}}$$

Weight decay $\quad \lambda \sim 10^{-4}, 10^{-5}$

Adam

1. Adaptive learning rate — AdaGrad, RMSProp
2. Momentum

$$\begin{cases} g_t = \nabla_w Q^t(w_{t-1}) + \lambda w_{t-1} \\ m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t \odot g_t \\ \hat{m}_t = \dfrac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \dfrac{v_t}{1-\beta_2^t} \\ w_t = w_{t-1} - \eta_t \dfrac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \end{cases}$$

$\odot$ — element-wise prod.

$m_0 = 0$

$v_0 = 1$

$\varepsilon, \beta_1, \beta_2$ — hyper parameters $\quad \varepsilon \approx 10^{-8}$

$\beta_1 = 0,9 \qquad \beta_2 = 0,999$

$$m_t = (1-\beta_1) g_t + \beta_1 m_{t-1} = (1-\beta_1) g_t + \beta_1 (1-\beta_1) g_{t-1}$$

$$+ \beta_1^2 m_{t-2} = \sum_{i=0}^{t-1} (1-\beta_1) \beta_1^i g_{t-i}$$

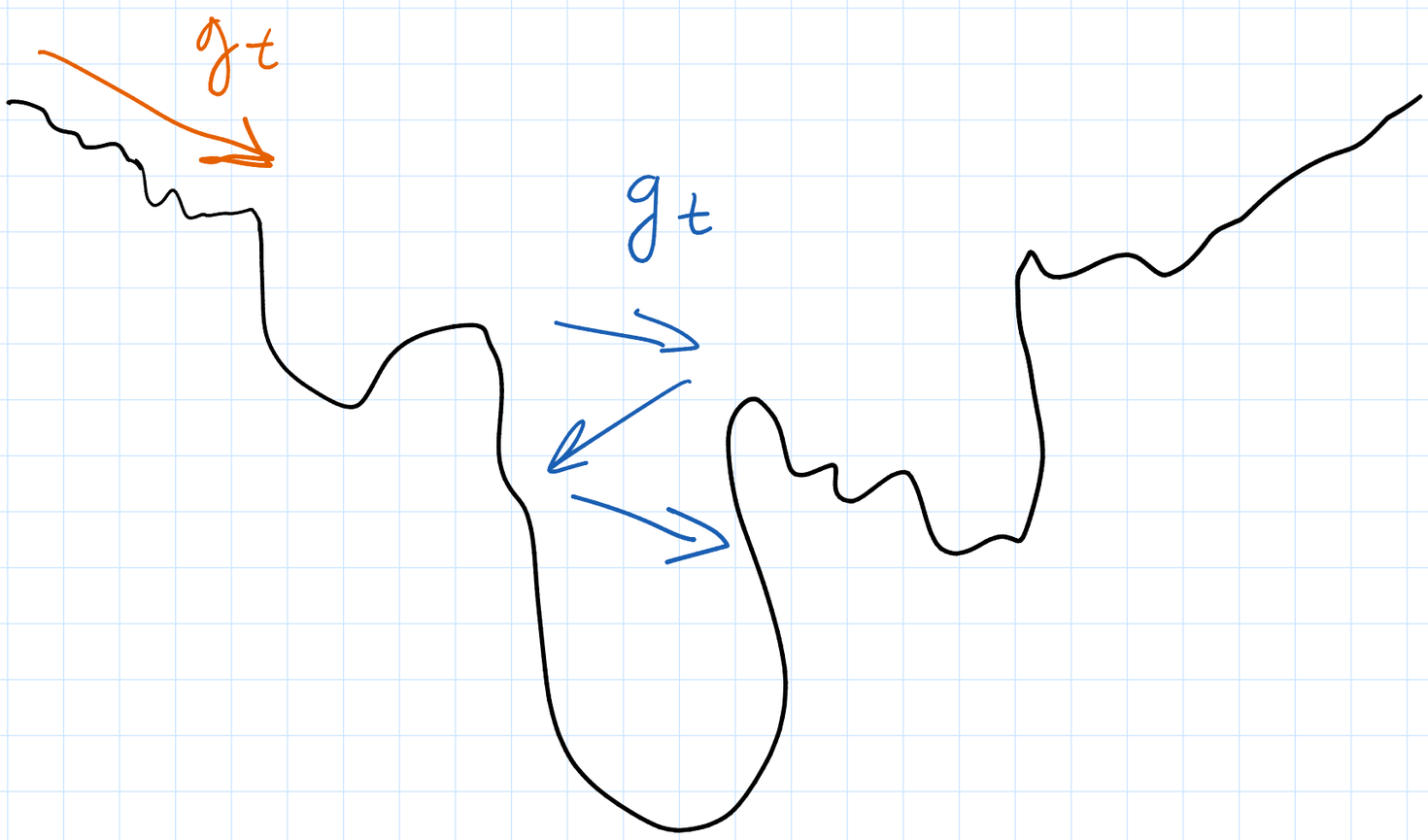
$$\hat{m}_t = \frac{1}{1-\beta_1^t} m_t = \frac{1-\beta_1}{1-\beta_1^t} \sum_{i=0}^{t-1} \beta_1^i g_{t-i} =$$

$$= \frac{\sum_{i=0}^{t-1} \beta_1^i g_{t-i}}{\sum_{j=0}^{t-1} \beta_1^j} \Leftarrow (1-\beta_1)(1 + \beta_1 + \dots + \beta_1^{t-1})$$

$$\alpha_i = \frac{\beta_1^i}{\sum_{j=0}^{t-1} \beta_1^j} \qquad \sum_i \alpha_i = 1$$

$$\Leftarrow \sum_{i=0}^{t-1} \alpha_i g_{t-i} = \mathbb{E} g$$

$$\hat{v}_t = \mathbb{E}[g \odot g] = Var(g) + \mathbb{E}g \odot \mathbb{E}g$$



1. Far from optimum

$$Var(g) \ll |g|^2$$

$$w_t = w_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} \approx w_{t-1} - \eta_t$$

$$\frac{\mathbb{E}g}{\sqrt{Var(g) + |\mathbb{E}g \odot \mathbb{E}g + \varepsilon}} \approx 1$$

≈ 0

2. Close to optimum

$$Var(g) \gg 0$$

$$\frac{\mathbb{E}g}{\sqrt{Var\, g + |\mathbb{E}g \odot \mathbb{E}g + \varepsilon}} < 1$$

SGD + momentum — CNN

Adam — RNN, Transformer

$$w_t = w_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}} = w_{t-1} - \eta_t \frac{m_t}{(1-\beta_1^t)\sqrt{\dots}} =$$

$$= w_{t-1} - \eta_t \frac{\beta_1 m_{t-1} + \nabla_w Q^t(w_{t-1}) + \lambda w_{t-1}}{(1-\beta_1^t)\sqrt{\dots}} =$$

$$= w_{t-1}\left(1 - \frac{\eta_t \lambda}{(1-\beta_1^t)\sqrt{\hat{v}_t + \varepsilon}}\right)$$

AdamW

$$
\begin{cases}
g_t = \nabla_w Q^t(w_{t-1}) + \cancel{\lambda w_{t-1}} \\
m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t \odot g_t \\
\hat{m}_t = \dfrac{m_t}{1-\beta_1^t} \qquad \hat{v}_t = \dfrac{v_t}{1-\beta_2^t} \\
w_t = w_{t-1}(1-\lambda\eta_t) - \eta_t \dfrac{\hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}}
\end{cases}
$$

η_t — ? How to choose learning rate ?

- $\eta_t = \eta_0$ — const scheduler Adam $\eta_t = 10^{-3}$

- Step LR

$$
\eta_t = \begin{cases} \eta_0 & 0 < t \le t_0 \\ \eta_1 & t_0 < t \le t_1 \\ \vdots & \end{cases}
$$

$\eta_0 > \eta_1 > \dots$

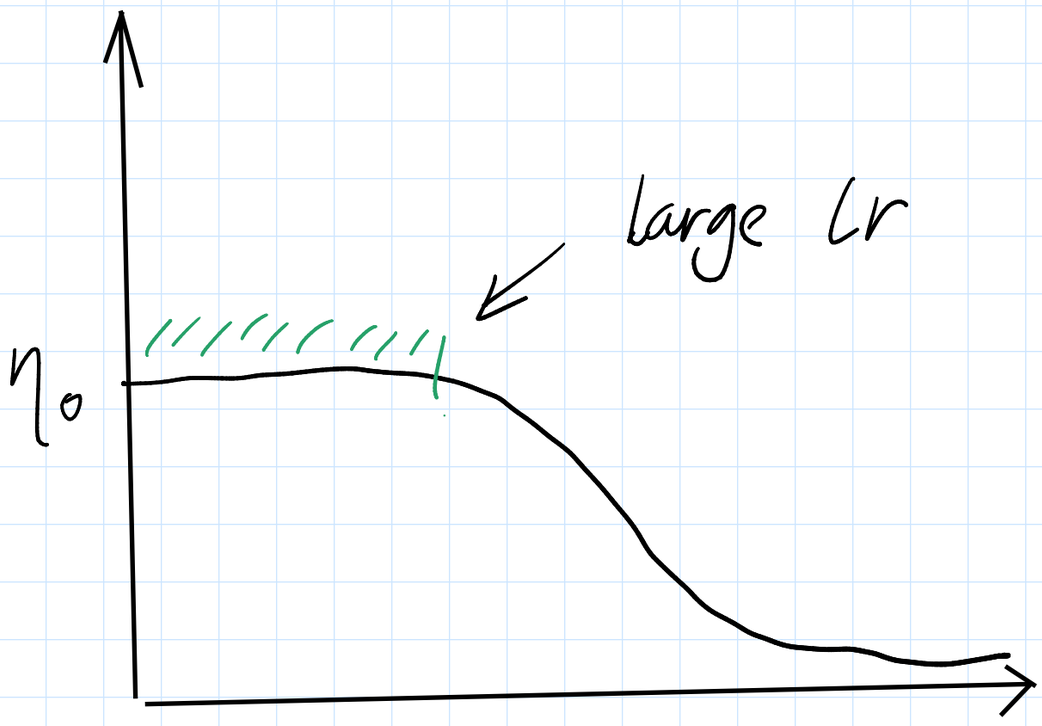
- Exponential LR

$$\eta_t = \gamma \cdot \eta_{t-1} \qquad \gamma < 1$$

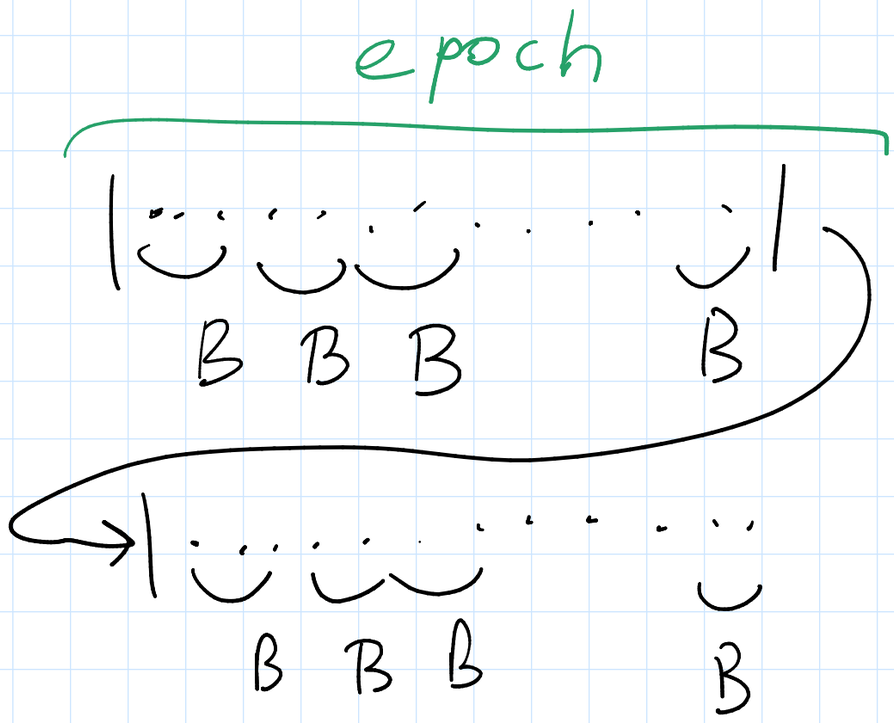
- Cosine LR

$$\eta_t = \eta_0 \frac{1}{2} \cos\left(\frac{\pi t}{T}\right)$$

T — number of steps



- Warmup LR



- Reduce LR on platean

adaptive, on validation set