26.9.20

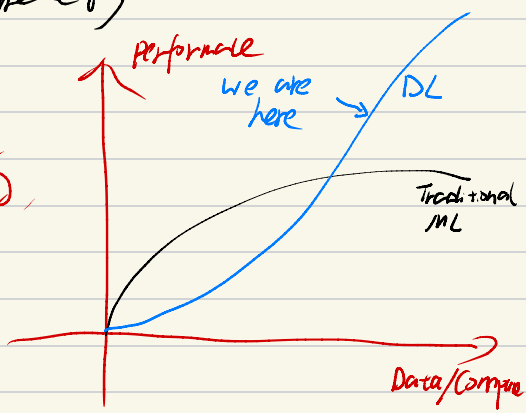
DL: $\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i, \theta))$

$f(X, \theta) \in DNN$

X: features

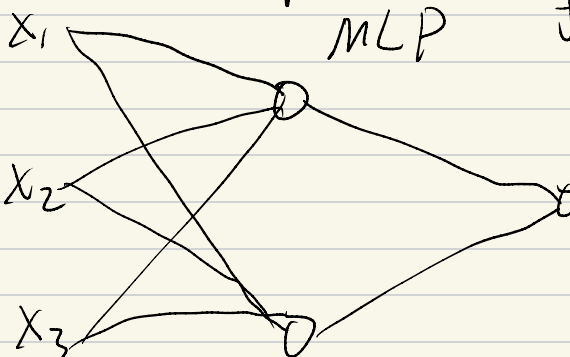
θ → parameters

(0) Why it matters.



(1) What is $f(X, \theta)$

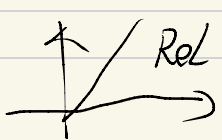
MLP



$f(X, \theta) = \sigma_2(W_2 \sigma_1(W_1 X + b_1) + b_2)$

$= P_\theta[Y=1 \mid X]$

$\sigma(WX+b)$ activation function



(2) What is the loss $\mathcal{L}(\cdot)$

Regression: $\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$

Classification: $\mathcal{L}(\hat{y}, y) = -y \log \hat{y} - (1-y) \log(1-\hat{y})$

$$\hat{y} = f(x, \theta)$$

(3) How do we find $\hat{\theta} = \arg\min_{\theta} \mathcal{L}(y, \hat{y})$

Gradient Deslent: $\hat{\theta}_{t+1} = \hat{\theta}_t - \nabla_\theta \mathcal{L}(\hat{\theta}_t)$

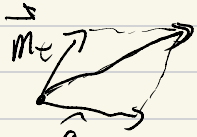
(4) How do we find $\nabla_\theta \mathcal{L}(\hat{\theta}_t)$

BP $y = f(z)$, $z = g(x)$

$$\frac{dy}{dx} = \frac{dy}{dz}\Big|_{z=g(x)} \cdot \frac{dz}{dx}\Big|_{x}$$

(ii) SGD $\{1, \dots B\} \subseteq \{1, \dots n\}$ random

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \mathcal{L}(y_i, f(x_i, \hat{\theta}_t))$$



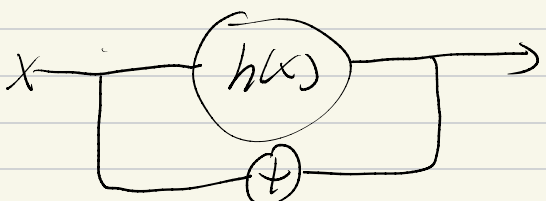
$$\vec{m}_{t+1} = \mu \vec{m}_t + (1-\mu)\vec{g}_t$$

$$\vec{\theta}_{t+1} = \vec{\theta}_t - \alpha \cdot \vec{m}_t$$

α: learning rate

Adam = Momentum + Adaptive learning Rate

(iii) Optimization



ResNet / Skip Connection

(1) Regularize

Optimization landscape

(b) Gradient Vanishing / Explosion

$$g_1 \cdot g_2 \cdots g_k \; ; \; g_i = \nabla f_i \qquad f_i = x + h_i(x)$$

$$\nabla f_i = I + \nabla h_i$$

(b) Initialization of He-Initialization

(c) Batch Normalization

---

2 arms / Ads ad 1 ad 2

θ_1 θ_2

Prior $\theta_1 \sim Beta(\alpha_1, \beta_1)$; $\theta_2 \sim Beta(\alpha_2, \beta_2)$

Pdf $x^{\alpha_1 - 1}(1-x)^{\beta_1 - 1}$ $x \in [0,1]$

$\alpha_1 = \beta_1 = \alpha_2 = \beta_2 \Rightarrow \theta_1, \theta_2 \sim Uniform[0,1]$

TS: Period $\alpha_1(t)$, $\alpha_2(t)$

$\beta_1(t)$, $\beta_2(t)$

Sample $\theta_1(t) \sim Beta(\alpha_1(t), \beta_1(t))$

$\theta_2(t) \sim Beta(\alpha_2(t), \beta_2(t))$

$$a_t^* = \begin{cases} 1 & \text{if } \theta_1(t) \geq \theta_2(t) \\ 2 & \text{if o.w.} \end{cases}$$

$$\alpha_{a_t^*}(t+1) = \alpha_{a_t^*}(t) + R_t$$

$$\beta_{a_t^*}(t+1) = \beta_{a_t^*}(t) + 1 - R_t$$