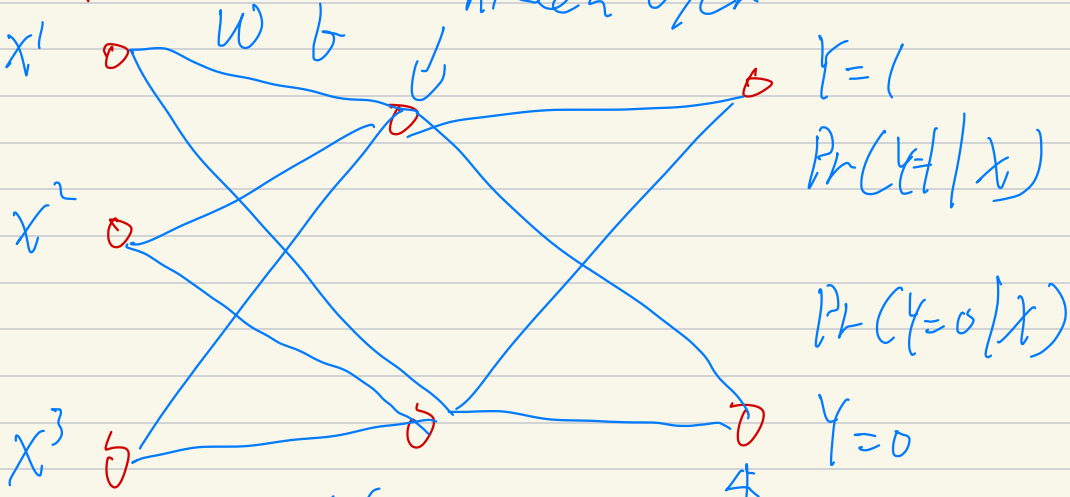
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i;\theta)) \triangleq \mathcal{L}(\theta)$$

$f(x, \theta)$ is a DNN,

features

parameters



$\sigma(Wx+b)$

Input Layer

Activation

Output Layer

ReLU

$$f(x;\theta) = \sigma_2(\sigma_1(W_1 x + b_1) W_2 + b_2)$$

$\vdots$

sigmoid / tanh

① What is the loss $L(y, \hat{y})$

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

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

② How do we find $\hat{\theta} \in \arg\min_{\theta} \mathcal{L}(\theta)$

Gradient Descent

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \cdot \nabla_\theta \mathcal{L}(\theta)$$

③ How do we estimate $\nabla_\theta \mathcal{L}(\theta)$

(i): Chain Rule / BP

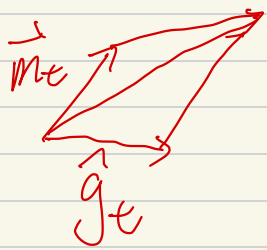
$y = f(z) \quad z = g(x)$

$$\frac{dy}{dx} = \left.\frac{df}{dz}\right|_{y=f(z)} \cdot \left.\frac{dz}{dx}\right|_{x,}$$

(ii) SGD $\{1, 2, \cdots B\} \subseteq \{1, \cdots n\}$

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

Momentum:



$$\vec{m}_{t+1} = \mu \cdot \vec{m}_t + \vec{g}_t$$

Adam = Momentum + RMSP

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

(i) Initialization $\{(W, b) \sim N(0, \frac{1/2}{n_G})$ → Dist. of $\hat{\theta}_0$

(ii) B/Normalization $\frac{x_i - \bar{x}_i}{SD(x_i)}$

(iii) Skip Connection

$g_1 g_2 \cdots g_k$

$g_1 \; g_2 \; g_3 \cdots \; g_k$: Gradient Vanishing

$$\underbrace{-x + f_i(x)}_{} + \underbrace{x}_{I}$$