Gradient Descent

① Network parameters $\theta = \{w_1, \cdots b_1, b_2 \cdots\}$

$\nabla L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial w_1} \\ \vdots \\ \frac{\partial L(\theta)}{\partial b} \end{bmatrix}$

compute $\nabla L(\theta^0)$ $\quad \theta^1 = \theta^0 - \eta \nabla L(\theta^0)$

compute $\nabla L(\theta^1)$ $\quad \theta^2 = \theta^1 - \eta \nabla L(\theta^1)$

……

Starting parameters $\theta^0 \longrightarrow \theta^1 \longrightarrow \theta^2 \longrightarrow \cdots$

But we have millions of parameters — Then we should use back propagation

② recall that Chain Rule

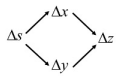
Chain Rule

Case 1 $y = g(x) \quad z = h(y)$

$\Delta x \to \Delta y \to \Delta z$ $\quad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$

Case 2

$x = g(s) \quad y = h(s) \quad z = k(x, y)$

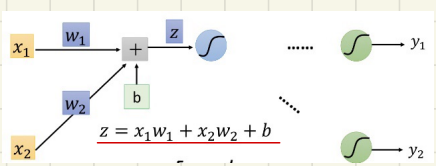


$\frac{dz}{ds} = \frac{\partial z}{\partial x}\frac{dx}{ds} + \frac{\partial z}{\partial y}\frac{dy}{ds}$

对于 $x^n \longrightarrow \boxed{NN\ \theta} \longrightarrow y^n \longleftrightarrow \hat{y}^n$

C^n ↳ y 与 $\hat{y}$ 之间的 Cross Entropy

$L(\theta) = \sum_{n=1}^{N} C^n(\theta) \longrightarrow \frac{\partial L(\theta)}{\partial w} = \sum_{n=1}^{N} \boxed{\frac{\partial C^n(\theta)}{\partial w}}$ 为最终计算



Forward pass: compute $\frac{\partial z}{\partial w}$ for all parameters

Backwark pass: compute $\frac{\partial C}{\partial z}$ for all activation function inputs z.

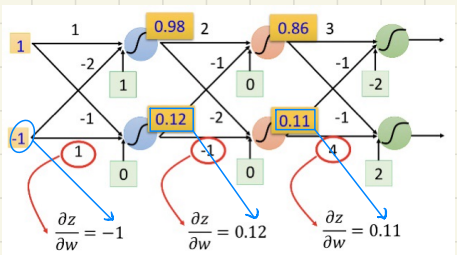
$\frac{\partial C}{\partial w} = \frac{\partial z}{\partial w}\frac{\partial C}{\partial z}$

① compute $\frac{\partial z}{\partial w}$ for all parameters forward pass

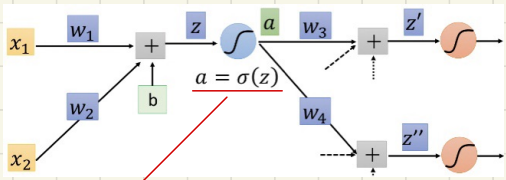
$\frac{\partial z}{\partial w_1} = x_1$

$\frac{\partial z}{\partial w_2} = x_2$

规律: The value of the input connected by the weight

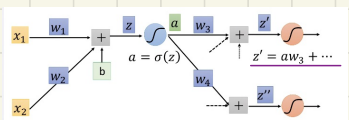


② compute $\frac{\partial C}{\partial z}$ for all parameters Backward pass



$$\frac{\partial C}{\partial z} = \frac{\partial a}{\partial z} \cdot \frac{\partial C}{\partial a}$$

$\sigma'(z)$



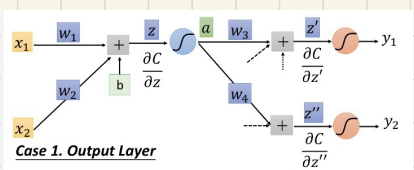
(假设 a 有两个 σ，所以 a 有两项)

$$\frac{\partial C}{\partial a} = \frac{\partial z'}{\partial a}\frac{\partial a}{\partial z'} + \frac{\partial z''}{\partial a}\frac{\partial a}{\partial z''}$$

$$= w_3 \frac{\partial a}{\partial z'} + w_4 \frac{\partial a}{\partial z''}$$

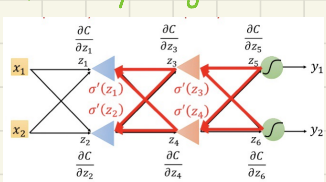
$$\frac{\partial C}{\partial z} = \sigma'(z)\left[w_3 \frac{\partial C}{\partial z'} + w_4 \frac{\partial a}{\partial z''}\right]$$

Case1: Output layer



$$\frac{\partial C}{\partial z'} = \frac{\partial y_1}{\partial z'}\frac{\partial C}{\partial y_1} \qquad \frac{\partial C}{\partial z''} = \frac{\partial y_2}{\partial z''}\frac{\partial C}{\partial y_2}$$

Case 2: Not output Layer



compute $\frac{\partial C}{\partial z}$ recursively

Util we reach output layer...

反向的 Recrual work

Summary: