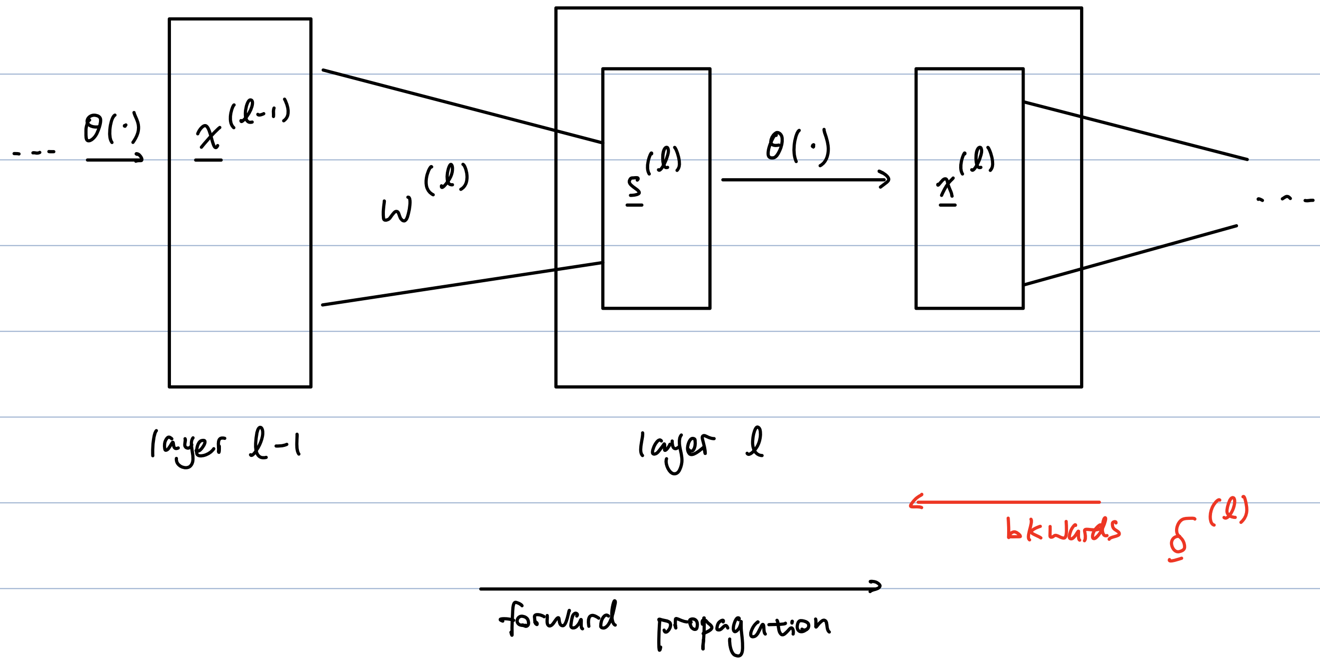
Lec 16 Backward propagation

1. Recall:



↳ $\Omega = \{ w^{(1)}, w^{(2)}, \cdots w^{(L)} \}$

↳ Loss function

$e_n(\Omega) \overset{e.g}{=} g(\underline{x}^{(L)}, y_n)$

↳ Goal:

compute $\dfrac{\partial e_n(\Omega)}{\partial w_{i,j}^{(\ell)}} \quad \forall\ i, j, \ell$

But difficult $\Longrightarrow$ Sol: backward propagation

2. Backward propagation Algo: (1960's)

↳ (1986, Hinton), similar to dynamic programing

we'll use $e(\Omega)$ instead of $e_n(\Omega)$ here since same analysis for any n.

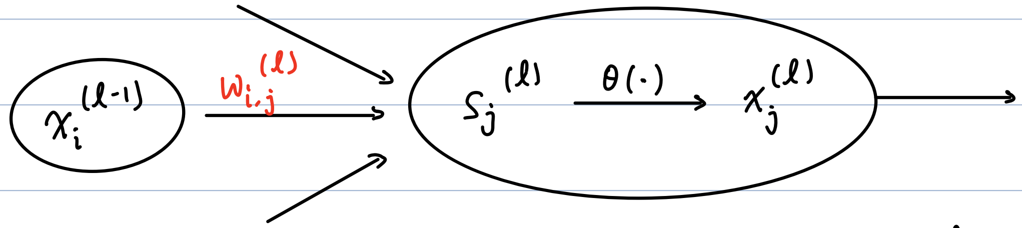
↳. Note:

① Memoryless: store all info up to layer ℓ.

$e(\Omega) = e(\underline{s}^{(\ell)}, w^{(\ell+1)}, \cdots w^{(L)})$

$= e((s_1^{(\ell)}, s_j^{(\ell)}, \cdots s_d^{(\ell)}), w^{(\ell+1)}, \cdots w^{(L)})$

②



Node j in layer ℓ

- Only $s_j^{(\ell)}$ depends on $w_{i,j}^{(\ell)}$

$$\Rightarrow \frac{\partial e(\Omega)}{\partial w_{i,j}^{(\ell)}} = \frac{\partial e(\Omega)}{\partial s_j^{(\ell)}} \cdot \frac{\partial s_j^{(\ell)}}{\partial w_{i,j}^{(\ell)}}$$

$$\because s_j^{(\ell)} = w_{0,j}^{(\ell)} + \sum_{k=1}^{d^{(\ell-1)}} w_{k,j}^{(\ell)} \, x_k^{(\ell-1)}$$

$$\therefore \frac{\partial s_j^{(\ell)}}{\partial w_{i,j}^{(\ell)}} = x_i^{(\ell-1)}$$

Let $\delta_j^{(\ell)} \triangleq \dfrac{\partial e(\Omega)}{\partial s_j^{(\ell)}}$ (sensitivity of $e(\Omega)$ to layer ℓ input)

$$\therefore \boxed{\frac{\partial e(\Omega)}{\partial s_j^{(\ell)}} = x_i^{(\ell-1)} \cdot \delta_j^{(\ell)}}$$

Define $\dfrac{\partial e(\Omega)}{\partial w^{(\ell)}} \triangleq \left[\dfrac{\partial e(\Omega)}{\partial w_{i,j}^{(\ell)}} \right]$ $\begin{matrix} 0 \le i \le d^{(i-1)} \\ 1 \le j \le d^{(\ell)} \end{matrix}$

$$\because \underline{x}^{(\ell-1)} = \begin{bmatrix} x_0^{(\ell-1)} \\ x_1^{(\ell-1)} \\ \vdots \\ x_{d^{(\ell-1)}}^{(\ell-1)} \end{bmatrix}, \quad \underline{\delta}^{(\ell)} = \begin{bmatrix} \delta_0^{(\ell)} \\ \vdots \\ \delta_{d^{(\ell-1)}}^{(\ell)} \end{bmatrix}$$

$$\therefore \boxed{\frac{\partial e(\Omega)}{\partial w^{(\ell)}} = \underset{\text{col}}{\underset{\uparrow}{\underline{x}^{(d-1)}}} \cdot \underset{\text{row}}{\underset{\uparrow}{\underline{\delta}^{(\ell)}}}{}^{T}}$$

2. Compute $\delta^{(\ell)}$ "backwards":

① consider $\ell = L$

Let $d^{(L)} = 1$ for simpility.

$$s^{(L)} \longrightarrow \theta \longrightarrow x^{(L)} \qquad g(x^{(L)}, y) = e(\Omega)$$

same activation func θ for rest for simpility

$$\therefore \delta^{(L)} = \frac{\partial e(\Omega)}{\partial s^{(L)}} = \frac{\partial g(x^{(L)}, y)}{\partial s^{(L)}}$$

$$= \frac{\partial g(x^{(L)}, y)}{\partial x^{(L)}} \cdot \frac{\partial x^{(L)}}{\partial s^{(L)}}$$

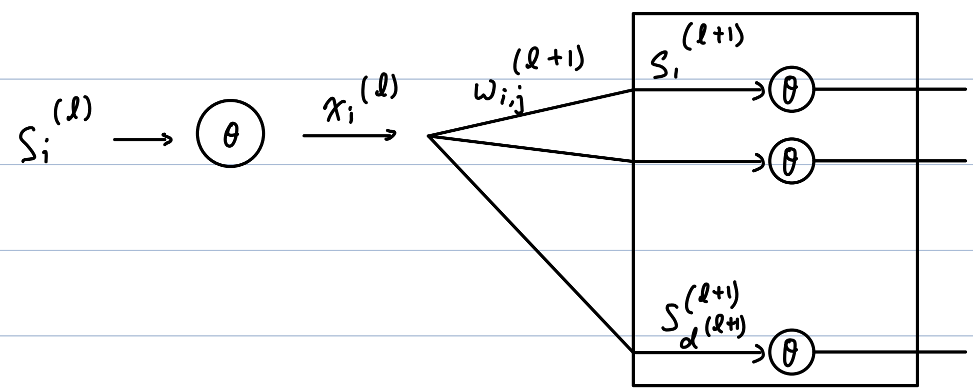
$$= \frac{\partial g(x^{(L)}, y)}{\partial x^{(L)}} \cdot \theta(s^{(L)})$$

e.g. Regression:

$$g(x^{(L)}, y) = (x^{(L)} - y)^2$$

$$\delta^{(L)} = 2(x^{(L)} - y)\, \theta'(s^{(L)})$$

② Intermediate node: $\delta_i^{(\ell)}$



$$\delta_i^{(\ell)} = \frac{\partial e(\Omega)}{\partial s_i^{(\ell)}} = \frac{\partial e(\Omega)}{\partial x_i^{(\ell)}} \cdot \frac{\partial x_i^{(\ell)}}{\partial s_i^{(\ell)}} = \frac{\partial e(\Omega)}{\partial x_i^{(\ell)}} \cdot \theta'(s_i^{(\ell)})$$

// Recall the chain rule:

if $z = f(x, y)$, and $\begin{cases} x = g(s, t) \\ y = h(s, t) \end{cases}$

then $\dfrac{\partial z}{\partial s} = \dfrac{\partial z}{\partial x} \cdot \dfrac{\partial x}{\partial s} + \dfrac{\partial z}{\partial y} \cdot \dfrac{\partial y}{\partial s}$

$\dfrac{\partial z}{\partial t} = \cdots$

$$\therefore \frac{\partial e(\Omega)}{\partial x_i^{(\ell)}} = \sum_{j=1}^{d^{(\ell+1)}} \frac{\partial e(\Omega)}{\partial s_j^{(\ell+1)}} \cdot \frac{\partial s_j^{(\ell+1)}}{\partial x_i^{(\ell)}}$$

$$= \sum_{j=1}^{d^{(\ell+1)}} \delta_j^{(d+1)} \cdot w_{i,j}^{(\ell+1)}$$

$$\therefore \boxed{\delta_i^{(\ell)} = \left[\sum_{j=1}^{d^{(\ell+1)}} \delta_j^{(\ell+1)} \, w_{i,j}^{(\ell+1)} \right] \cdot \theta'\left(s_i^{(\ell)}\right)}$$

Pack every i together:

$$\therefore \underline{\delta}^{(\ell)} = \begin{bmatrix} \delta_1^{(\ell)} \\ \vdots \\ \delta_{d^{(\ell)}}^{(\ell)} \end{bmatrix}, \quad \theta'\left(\underline{s}^{(\ell)}\right) = \begin{bmatrix} \theta'\left(s_1^{(\ell)}\right) \\ \vdots \\ \theta'\left(s_{d^{(\ell)}}^{(\ell)}\right) \end{bmatrix}.$$

$$\hat{W}^{(\ell+1)} \triangleq \begin{bmatrix} w_{1,1}^{(\ell+1)} & \cdots & w_{1,d^{(\ell+1)}}^{(\ell+1)} \\ \vdots & & \vdots \\ w_{d^{(\ell)},1}^{(\ell+1)} & \cdots & w_{d^{(\ell)},d^{(\ell+1)}}^{(\ell+1)} \end{bmatrix}$$

hat since remove bias

$$\therefore \boxed{\underline{\delta}^{(\ell)} = \left[\hat{W}^{(\ell+1)} \, \underline{\delta}^{(\ell+1)} \right] \otimes \theta'\left(\underline{s}^{(\ell)}\right)}$$

pointwise multiplication.