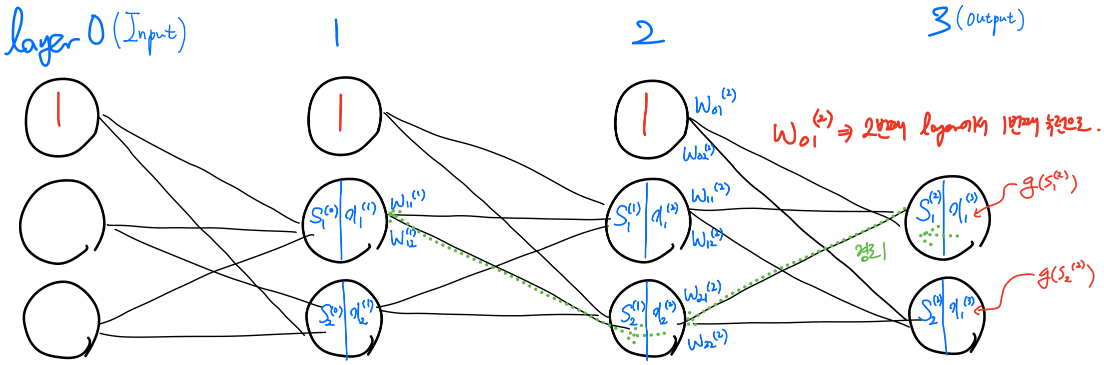


$$x_1^{(3)} = g(S_1^{(2)}),\quad S_1^{(2)} = W_{01}^{(2)} + W_{11}^{(2)} x_1^{(2)} + W_{21}^{(2)} x_1^{(2)}$$

$$x_2^{(3)} = g(S_2^{(2)}),\quad S_2^{(2)} = W_{02}^{(2)} + W_{12}^{(2)} x_1^{(2)} + W_{22}^{(2)} x_2^{(2)}$$

$\times\ W_{02}^{(2)}$

Error function

$$f = (y_1 - \underbrace{x_1^{(3)}}_{\text{Output1}})^2 + (y_2 - \underbrace{x_2^{(3)}}_{\text{Output2}})^2$$

$$\delta_1^{(2)} \triangleq \frac{\partial f}{\partial S_1^{(2)}}$$

$$\frac{\partial f}{\partial W_{21}^{(2)}} = \boxed{\frac{\partial f}{\partial x_1^{(3)}} \cdot \frac{\partial x_1^{(3)}}{\partial S_1^{(2)}}} \cdot \underbrace{\frac{\partial S_1^{(2)}}{\partial W_{21}^{(2)}}}_{x_2^{(2)}}$$

$$\delta_2^{(2)} \triangleq \frac{\partial f}{\partial S_2^{(2)}}$$

$$\frac{\partial f}{\partial W_{22}^{(2)}} = \boxed{\frac{\partial f}{\partial x_2^{(3)}} \cdot \frac{\partial x_2^{(3)}}{\partial S_2^{(2)}}} \cdot \underbrace{\frac{\partial S_2^{(2)}}{\partial W_{22}^{(2)}}}_{x_2^{(2)}}$$

W_{12}가 2가지 output에 영향을 주므로

경로1 / 경로 2

$$\frac{\partial f}{\partial W_{12}^{(1)}} = \overbrace{\underbrace{\boxed{\frac{\partial f}{\partial x_1^{(3)}} \cdot \frac{\partial x_1^{(3)}}{\partial S_1^{(2)}}}}_{\delta_1^{(2)}} \cdot \underbrace{\boxed{\frac{\partial S_1^{(2)}}{\partial x_2^{(2)}}}}_{W_{21}^{(2)}} \cdot \frac{\partial x_2^{(2)}}{\partial S_2^{(1)}} \cdot \underbrace{\boxed{\frac{\partial S_2^{(1)}}{\partial W_{12}^{(1)}}}}_{x_1^{(1)}}}^{\text{경로1}} + \overbrace{\underbrace{\boxed{\frac{\partial f}{\partial x_2^{(3)}} \cdot \frac{\partial x_2^{(3)}}{\partial S_2^{(2)}}}}_{\delta_2^{(2)}} \cdot \underbrace{\boxed{\frac{\partial S_2^{(2)}}{\partial x_2^{(1)}}}}_{W_{22}^{(2)}} \cdot \frac{\partial x_2^{(2)}}{\partial S_2^{(1)}} \cdot \underbrace{\boxed{\frac{\partial S_2^{(1)}}{\partial W_{12}^{(1)}}}}_{x_1^{(1)}}}^{\text{경로 2}}$$

$$= \underbrace{\begin{bmatrix} \delta_1^{(2)} & \delta_2^{(2)} \end{bmatrix}}_{\delta^{(2)}} \underbrace{\begin{bmatrix} W_{21}^{(2)} \\ W_{22}^{(2)} \end{bmatrix}}_{W_2^{(2)}} \frac{\partial x_2^{(2)}}{\partial S_2^{(1)}} x_1^{(1)} = \underbrace{\delta^{(2)T} \cdot W_2^{(2)} \cdot \frac{\partial x_2^{(2)}}{\partial S_2^{(1)}}}_{\text{다 더하면.}} \cdot x_1^{(1)}$$

$$= \frac{\partial f}{\partial S_2^{(1)}} = \delta_2^{(1)}$$

이 값들은 forward 계산이 되어야 구할 수 있음.

$$\frac{\partial f}{\partial W_{ij}^{(l)}} = \delta_j^{(l)} x_i^{(l)}$$

$$\delta_j^{(l)} = (\delta^{(l+1)})^T W_j^{(l+1)} \cdot \frac{\partial x_j^{(l+1)}}{\partial S_j^{(l)}} \quad (l \le L-2)$$

$$W_{ij}^{(l)} \leftarrow W_{ij}^{(l)} - \alpha \frac{\partial T}{\partial W_{ij}^{(l)}}$$

- Initial Weight

초기값 ⇒ 0이면 안됨.

- Stochastic Gradient Descent.

오차함수 f_i는 $f_1, f_2, \cdots f_i$로 여러 레이어를 Input 값을 때마다 얻은 것이다.

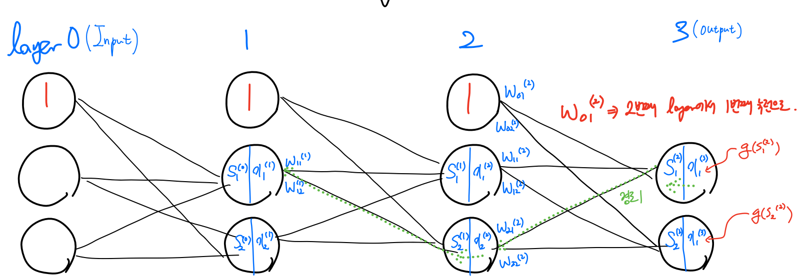
때문에 $W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \sum_n \frac{\partial f_n}{\partial W_{ij}}$로 계산한다. 하지만 이는 계산량이 너무 多

그래서 Sum 하지 않고 1개로 가중치를 update.

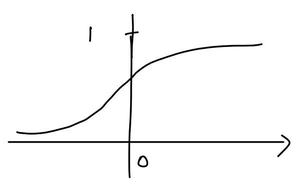
- Minibatch SGD

n개씩 조금씩 사용해서 update.

⊙ Gradient Vanishing.



Sigmoid



0 미서의 미분값은 0.25이다.

가장 큰 기울기.

$$\frac{\partial f}{\partial W_{ij}^{(l)}} = \delta_j^{(l)} a_i^{(l)}$$

$$\delta_j^{(l)} = (\delta^{(l+1)})^T W_j^{(l+1)} \cdot \frac{\partial a_J^{(l+1)}}{\partial S_J^{(l)}} \quad (l \le L-2)$$

통과하기 전값은 sigmoid 통과 값으로 미분.

∴ 최대값은 기껏해야 0.25

$$W_{ij}^{(l)} \leftarrow W_{ij}^{(l)} - \alpha \frac{\partial f}{\partial W_{ij}^{(l)}}$$

...

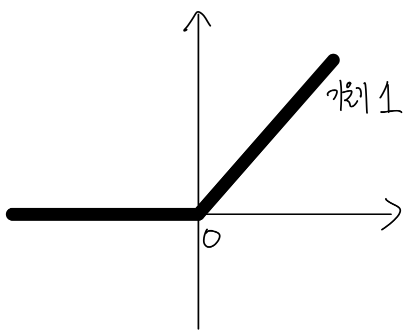
· DNN에서 첫번째 layer의 weight를 갱신할 때 $\delta_1^{(1)}$를 구하려면 뒤쪽의 layer에서 계산된 δ들이 계속 곱해지는 것이다.

그러면 기울기 $\frac{\partial x_J^{(l+1)}}{\partial s_J^{(l)}}$가 계속 곱해지게 되는데 deep한 Net에서는 최대 기울기 0.25가 곱해진다고 해도 δ_1에 이르러서는 gradient가 거의 0이 되기 때문에

Gradient Vanishing 문제가 발생하는 것이다.

⊙ ReLU

· 그래서 기울기가 더 큰 Activation으로 개발된 함수이다.



※ 비선형요소를 빼지 않고 기울기가 1인 linear func를 쓰면

linear Regression이 된다.

ReLU는 linear Regression 이다.

$$x \xrightarrow{W_1} h_1 \xrightarrow{W_2} h_2 \xrightarrow{W_3} h_3 \xrightarrow{W_4} y \qquad Cost = f = \frac{1}{2}(y-t)^2$$

$$h = g(x \cdot w_1) \qquad y = g(h \cdot w_4)$$

$$W_4 = W_4 - \lambda \frac{\partial Cost(W_4)}{\partial W_4} \qquad W_4 = W_4 - h_3^T \lambda \delta_4$$

$$\frac{\partial Cost}{\partial W_4} = \underbrace{\frac{\partial Cost}{\partial y} \cdot \frac{\partial y}{\partial (h_3 \cdot W_4)}}_{\delta_4} \cdot \frac{\partial (h_3 \cdot W_4)}{\partial W_4} = (y-t)\, g(h \cdot w_4)(1-g(h \cdot w_4)) \cdot h_3 = h_3^T \underbrace{(y-t)y(1-y)}_{\delta_4}$$

$$W_3 = W_3 - \lambda \frac{\partial Cost}{\partial W_3} \qquad W_3 = W_3 - h_2^T \lambda \delta_3$$

$$\frac{\partial Cost}{\partial W_3} = \underbrace{\frac{\partial Cost}{\partial y} \cdot \frac{\partial y}{\partial (h_3 \cdot W_4)} \cdot \frac{\partial (h_3 \cdot W_4)}{\partial h_3} \cdot \frac{\partial h_3}{\partial (h_2 \cdot W_3)}}_{\delta_3} \cdot \frac{\partial (h_2 \cdot W_3)}{\partial W_3} = h_2^T \overbrace{\underbrace{(y-t)y(1-y)}_{\delta_4} W_4^T h_3(1-h_3)}^{\delta_3}$$

$$W_2 = W_2 - \lambda \frac{\partial Cost}{\partial W_2} \qquad W_2 = W_2 - h_1^T \lambda \delta_2$$

$$\frac{\partial Cost}{\partial W_2} = \underbrace{\frac{\partial Cost}{\partial y} \cdot \frac{\partial y}{\partial (h_3 \cdot W_4)} \cdot \frac{\partial (h_3 \cdot W_4)}{\partial h_3} \cdot \frac{\partial h_3}{\partial (h_2 \cdot W_3)} \cdot \frac{\partial (h_2 \cdot W_3)}{\partial h_2} \cdot \frac{\partial h_2}{\partial (h_1 \cdot W_2)}}_{\delta_2} \cdot \frac{\partial (h_1 \cdot W_2)}{\partial W_2} = h_1^T \delta_3 \cdot W_3^T h_2(1-h_2)$$

- 일반화 식

$$W_t = W_t - h_{t-1}^T \lambda \delta_{t+1}$$

$$\delta_t = \delta_{t+1} \cdot W_{t+1}^T h_t(1-h_t)$$

↳ sigmoid 미분.

- Stochastic Gradient Descent.

  Initial weight. $\Rightarrow$ Gaussian Dist.