Gridient and Jacobian

$$y = f(x) \quad \text{where} \quad x = \begin{bmatrix} u \\ v \\ w \end{bmatrix}$$

① Gradient

$$\nabla f = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix}$$

The direction of ∇f is the direction of the steepest ascent.

$$\begin{aligned} f(x + \Delta x) &= f(x) + \frac{\partial f}{\partial u} \Delta u + \frac{\partial f}{\partial v} \Delta v + \frac{\partial f}{\partial w} \Delta w \\ &= f(x) + (\nabla f)^t \Delta x \\ &= f(x) + \|\nabla f\| \, \|\Delta x\| \cos\theta \end{aligned}$$

$\cos\theta = 1$ when ∇f and Δx have the same direction.

② Jacobian

$$Y = \begin{bmatrix} f(x) \\ g(x) \\ h(x) \end{bmatrix} \triangleq \bar{f}(x)$$

$$J = \begin{bmatrix} \left(\frac{\partial f}{\partial x}\right)^t \\ \left(\frac{\partial g}{\partial x}\right)^t \\ \left(\frac{\partial h}{\partial x}\right)^t \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} & \frac{\partial f}{\partial w} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} & \frac{\partial g}{\partial w} \\ \frac{\partial h}{\partial u} & \frac{\partial h}{\partial v} & \frac{\partial h}{\partial w} \end{bmatrix}$$

Back propagation algorithm using standard matrix notations

2017. 5. 23
金昌洙.

$$\begin{tikzcd} y^0 \arrow[r] & \boxed{f^1} \arrow[r] & y^1 \arrow[r] & \cdots \arrow[r] & y^{r-1} \arrow[r] & \boxed{f^r} \arrow[r] & y^r \arrow[r] & \cdots \arrow[r] & y^{L-1} \arrow[r] & \boxed{f^L} \arrow[r] & y^L \\ & \theta^1 \arrow[u] & & & & \theta^r \arrow[u] & & & & \theta^L \arrow[u] & \end{tikzcd}$$

$$y^r = f^r(y^{r-1}, \theta^r) \qquad y^r = \begin{bmatrix} y_1^r \\ \vdots \\ y_{k_r}^r \end{bmatrix} \quad k_r = \text{\# of neurons in layer } r$$

$$\theta^r = \begin{bmatrix} \theta_1^r \\ \vdots \\ \theta_{j_r}^r \end{bmatrix} \quad j_r = \text{\# of parameters in layer } r$$

Cost function $\mathcal{E} = g(y^L, \underbrace{y}_{\text{ground truth}})$ for example $\mathcal{E} = \| y^L - y \|$

Jacobian matrix at layer r

$$J^r \triangleq [\underbrace{J_y^r}_{\text{input}} \ \underbrace{J_\theta^r}_{\text{parameter}}] \triangleq \begin{bmatrix} \frac{\partial f_1^r}{\partial y_1^{r-1}} & \cdots & \frac{\partial f_1^r}{\partial y_{k_{r-1}}^{r-1}} & \frac{\partial f_1^r}{\partial \theta_1^r} & \cdots & \frac{\partial f_1^r}{\partial \theta_{j_r}^r} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial f_{k_r}^r}{\partial y_1^{r-1}} & \cdots & \frac{\partial f_{k_r}^r}{\partial y_{k_{r-1}}^{r-1}} & \frac{\partial f_{k_r}^r}{\partial \theta_1^r} & \cdots & \frac{\partial f_{k_r}^r}{\partial \theta_{j_r}^r} \end{bmatrix}$$

" evaluated at y^{r-1}, θ^r

Back propagation algorithm

(A) Layer L

$$\frac{\partial \mathcal{E}}{\partial \theta^L} = \begin{bmatrix} \frac{\partial \mathcal{E}}{\partial \theta_1^L} \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial \theta_{j_L}^L} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1^L}{\partial \theta_1^L} \frac{\partial \mathcal{E}}{\partial y_1^L} + \cdots + \frac{\partial f_{k_L}^L}{\partial \theta_1^L} \frac{\partial \mathcal{E}}{\partial y_{k_L}^L} \\ \\ \frac{\partial f_1^L}{\partial \theta_{j_L}^L} \frac{\partial \mathcal{E}}{\partial y_1^L} + \cdots + \frac{\partial f_{k_L}^L}{\partial \theta_{j_L}^L} \frac{\partial \mathcal{E}}{\partial y_{k_L}^L} \end{bmatrix}$$

$$= (J_\theta^L)^T \frac{\partial \mathcal{E}}{\partial y^L}$$

Therefore $\quad \dfrac{\partial \mathcal{E}}{\partial \theta^L} = (J_\theta^L)^T \dfrac{\partial \mathcal{E}}{\partial y^L} \quad \cdots ①$

(B) Layer $r = L-1, L-2, \dots, 1$

$$\begin{tikzcd} y^{r-1} \arrow[r] & \boxed{f^r} \arrow[r] & y^r \\ & \theta^r \arrow[u] & \end{tikzcd}$$

$$\frac{\partial \mathcal{E}}{\partial \theta^r} = \begin{bmatrix} \frac{\partial \mathcal{E}}{\partial \theta_1^r} \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial \theta_{j_r}^r} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1^r}{\partial \theta_1^r}\frac{\partial \mathcal{E}}{\partial y_1^r} + \cdots + \frac{\partial y_{k_r}^r}{\partial \theta_1^r}\frac{\partial \mathcal{E}}{\partial y_{k_r}^r} \\ \vdots \\ \frac{\partial y_1^r}{\partial \theta_{j_r}^r}\frac{\partial \mathcal{E}}{\partial y_1^r} + \cdots + \frac{\partial y_{k_r}^r}{\partial \theta_{j_r}^r}\frac{\partial \mathcal{E}}{\partial y_{k_r}^r} \end{bmatrix}$$

$$= \left(J_\theta^r\right)^T \frac{\partial \mathcal{E}}{\partial y^r}$$

Also note that

$$\frac{\partial \mathcal{E}}{\partial y^r} = \begin{bmatrix} \frac{\partial \mathcal{E}}{\partial y_1^r} \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial y_{k_r}^r} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1^{r+1}}{\partial y_1^r}\frac{\partial \mathcal{E}}{\partial y_1^{r+1}} + \cdots + \frac{\partial y_{k_{r+1}}^{r+1}}{\partial y_1^r}\frac{\partial \mathcal{E}}{\partial y_{k_{r+1}}^{r+1}} \\ \vdots \\ \frac{\partial y_1^{r+1}}{\partial y_{k_r}^r}\frac{\partial \mathcal{E}}{\partial y_1^{r+1}} + \cdots + \frac{\partial y_{k_{r+1}}^{r+1}}{\partial y_{k_r}^r}\frac{\partial \mathcal{E}}{\partial y_{k_{r+1}}^{r+1}} \end{bmatrix}$$

$$= \left(J_y^{r+1}\right)^T \frac{\partial \mathcal{E}}{\partial y^{r+1}}$$

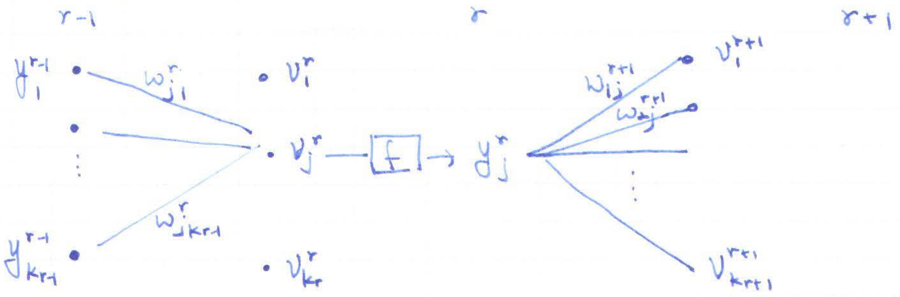
To summarize, we have

$$\frac{\partial \mathcal{E}}{\partial \theta^r} = \left(J_\theta^r\right)^T \frac{\partial \mathcal{E}}{\partial y^r} \quad \cdots (2)$$

(① is a special case of ②.)

$$\frac{\partial \mathcal{E}}{\partial y^r} = \left(J_y^{r+1}\right)^T \frac{\partial \mathcal{E}}{\partial y^{r+1}} \quad \cdots (3)$$

Example) Textbook



$$\frac{\partial \mathcal{E}}{\partial w_j^r} = \frac{\partial v_j^r}{\partial w_j^r}\frac{\partial y_j^r}{\partial v_j^r}\frac{\partial \mathcal{E}}{\partial y_j^r}$$ (this is a part of ②)

$$= y^{r-1}\frac{\partial \mathcal{E}}{\partial v_j^r}$$

$$= y^{r-1}\delta_j^r$$ (in textbook)

Also

$$\frac{\partial \mathcal{E}}{\partial y_j^r} = w_{1j}^{r+1} f'(v_1^{r+1}) \frac{\partial \mathcal{E}}{\partial y_1^{r+1}} \cdots$$

$$= \sum_l w_{lj}^{r+1} f'(v_l^{r+1}) \frac{\partial \mathcal{E}}{\partial y_l^{r+1}}$$ (this is a part of ③')

In textbook, this is given by

$$\delta_j^r = \frac{\partial \mathcal{E}}{\partial v_j^r} = \frac{\partial y_j^r}{\partial v_j^r}\frac{\partial \mathcal{E}}{\partial y_j^r}$$

$$= f'(v_j^r)\sum_l w_{lj}^{r+1} f'(v_l^{r+1})\frac{\partial \mathcal{E}}{\partial y_l^{r+1}}$$

$$= f'(v_j^r)\underbrace{\boxed{\sum_l w_{lj}^{r+1}\delta_l^{r+1}}}_{= e_j^r}$$