$\hat{R} \quad \mathbb{R}$

$L(\cdot)$ $\quad \mathbb{R}^m \to \mathbb{R}$

$$\frac{\partial \hat R}{\partial \vec y_3} = \begin{pmatrix} \frac{\partial \hat R}{\partial y_3[1]} & \frac{\partial \hat R}{\partial y_3[2]} & \cdots & \frac{\partial \hat R}{\partial y_3[m]} \end{pmatrix} \quad (1\times m)$$

$(m\times 1)$ $\quad \nabla_{\vec y_3} \hat R$ ← engineering notation

Actually $\frac{\partial R\ (\text{scalar})}{\partial \vec u\ (\text{col-vec})}$ is a row-vec (Numerator Layout)

sometimes we write col-vec for engineering requirement

$\vec y_3 \quad \mathbb{R}^m$

$h(\cdot)$

$$\frac{\partial \vec y_3}{\partial \vec z_3} = \begin{pmatrix} h'(z_3[1]) & 0 & \cdots & 0 \\ 0 & h'(z_3[2]) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h'(z_3[m]) \end{pmatrix} \quad (m\times m)$$

diagonal

$$\underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec z_3}} = \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec y_3}} \underset{(m\times m)}{\frac{\partial \vec y_3}{\partial \vec z_3}} = \begin{pmatrix} h'(z_3[1]) \cdot \partial \hat R/\partial y_3[1] \\ h'(z_3[2]) \cdot \partial \hat R/\partial y_3[2] \\ \cdots \\ h'(z_3[m]) \cdot \partial \hat R/\partial y_3[m] \end{pmatrix}^{T}$$

post-hoc marker because at first we wrote $\frac{\partial R}{\partial \vec y}$ as col-vec

or $h'(\vec z_3) \odot \nabla_{\vec y_3} \hat R$

elementwize-product

$\vec z_3 \quad \mathbb{R}^m$

W_3 $(m\times n)$

b_3 $(m\times 1)$

$$\frac{\partial \vec z_3}{\partial \vec y_2} = \begin{pmatrix} \partial z_3[1]/\partial y_2[1] & \partial z_3[1]/\partial y_2[2] & \cdots & \partial z_3[1]/\partial y_2[n] \\ \partial z_3[2]/\partial y_2[1] & \partial z_3[2]/\partial y_2[2] & \cdots & \partial z_3[2]/\partial y_2[n] \\ \vdots & \vdots & \ddots & \vdots \\ \partial z_3[m]/\partial y_2[1] & \partial z_3[m]/\partial y_2[2] & \cdots & \partial z_3[m]/\partial y_2[n] \end{pmatrix} = \begin{pmatrix} W_3[1,1] & W_3[1,2] & \cdots & W_3[1,n] \\ W_3[2,1] & W_3[2,2] & \cdots & W_3[2,n] \\ \vdots & \vdots & \ddots & \vdots \\ W_3[m,1] & W_3[m,2] & \cdots & W_3[m,n] \end{pmatrix} = W_3$$

$$\underset{(1\times n)}{\frac{\partial \hat R}{\partial \vec y_2}} = \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec z_3}} \cdot \underset{(m\times n)}{W_3}$$

$$\frac{\partial \hat R}{\partial b_3} = 1 \cdot \frac{\partial \hat R}{\partial \vec z_3}$$

$\frac{\partial \vec z_3}{\partial W_3}$ = shit! it's a 3D tensor!

$$\begin{pmatrix} z_3[1] \\ z_3[2] \\ \vdots \\ z_3[m] \end{pmatrix} = \begin{pmatrix} W_3[1,1] & W_3[1,2] & \cdots & W_3[1,n] \\ W_3[2,1] & W_3[2,2] & \cdots & W_3[2,n] \\ \vdots & \vdots & \ddots & \vdots \\ W_3[m,1] & W_3[m,2] & \cdots & W_3[m,n] \end{pmatrix} \begin{pmatrix} y_2[1] \\ y_2[2] \\ \cdots \\ y_2[n] \end{pmatrix}$$

$\frac{\partial z_3[1]}{\partial W_3[1,1]} = y_2[1]$

$\frac{\partial z_3[1]}{\partial W_3[1,2]} = y_2[2]$

$\cdots$

$\frac{\partial z_3[1]}{\partial W_3[1,n]} = y_2[n]$

but we have $\boxed{\frac{\partial z_3[m]}{\partial W_3[m,n]} = y_2[n], \quad \frac{\partial z_3[m]}{\partial W_3[p,n]}\ (p \neq m) = 0}$

$$\frac{\partial \hat R}{\partial W_3} = \begin{pmatrix} \partial \hat R/\partial W_3[1,1] & \partial \hat R/\partial W_3[1,2] & \cdots & \partial \hat R/\partial W_3[1,n] \\ \partial \hat R/\partial W_3[2,1] & \partial \hat R/\partial W_3[2,2] & \cdots & \partial \hat R/\partial W_3[2,n] \\ \vdots & \vdots & \ddots & \vdots \\ \partial \hat R/\partial W_3[m,1] & \partial \hat R/\partial W_3[m,2] & \cdots & \partial \hat R/\partial W_3[m,n] \end{pmatrix}^{T}$$

$$= \underset{(n\times 1)}{\vec y_2} \cdot \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec z_3}}$$

(this is a $n\times m$ matrix, but $\nabla_W \hat R$ is a $m\times n$, for substraction compatibility)

post-hoc marker

$$\frac{\partial \hat R}{\partial W_3[m,n]} = \frac{\partial \hat R}{\partial z_3[m]} \frac{\partial z_3[m]}{\partial W_3[m,n]} = \frac{\partial \hat R}{\partial z_3[m]} \cdot y_2[n]$$

$\vec y_2 \quad \mathbb{R}^n$

$g(\cdot)$

$$\underset{(n\times n)}{\frac{\partial \vec y_2}{\partial \vec z_2}} = \operatorname{diag}(g'(\vec z_2))$$

$$\frac{\partial \hat R}{\partial \vec z_2} = \frac{\partial \hat R}{\partial \vec y_2} \cdot \frac{\partial \vec y_2}{\partial \vec z_2}$$

$\vec z_2 \quad \mathbb{R}^n$

W_2 $(n\times p)$

b_2 $(n\times 1)$

$$\frac{\partial \vec z_2}{\partial \vec y_1} = W_2$$

$$\underset{(1\times p)}{\frac{\partial \hat R}{\partial \vec y_1}} = \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec y_3}} \cdot \underset{(m\times m)}{\operatorname{diag}(h'(\vec z_3))} \cdot \underset{(m\times n)}{W_3} \cdot \underset{(n\times n)}{\operatorname{diag}(g'(\vec z_2))} \cdot \underset{(n\times p)}{W_2}$$

$(m\times m) \to \partial \vec y_3/\partial \vec z_3$ $\quad (m\times n) \to \partial \vec z_3/\partial \vec y_2$ $\quad (n\times n) \to \partial \vec y_2/\partial \vec z_2$ $\quad (n\times p) \to \partial \vec z_2/\partial \vec y_1$

$$\underset{(p\times n)}{\frac{\partial \hat R}{\partial W_2}} = \vec y_1 \cdot \frac{\partial \hat R}{\partial \vec z_2}$$

$$= \underset{(p\times 1)}{\vec y_1} \cdot \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec z_3}} \cdot \underset{(m\times n)}{W_3} \cdot \underset{(n\times n)}{\operatorname{diag}(g'(\vec z))}$$

$\vec y_1 \quad \mathbb{R}^p$

$f(\cdot)$ $\quad \operatorname{diag}(f'(\vec z_1))$

$\vec z_1 \quad \mathbb{R}^p \xleftarrow{W_1\ (p\times q) \quad b_1\ (p\times 1)} x \quad \mathbb{R}^q$

$$\frac{\partial \vec z_1}{\partial x} = W_1$$

layer $l-1$: p

layer l: n

layer $l+1$: m

$$(1\times n) \leftarrow \frac{\partial \hat R}{\partial \vec y_l} = \underset{(1\times m)}{\frac{\partial \hat R}{\partial \vec z_{(l+1)}}} \cdot \underset{(m\times n)}{W_{(l+1)}}$$

$$(1\times n) \leftarrow \frac{\partial \hat R}{\partial \vec z_l} = \underset{(1\times n)}{\frac{\partial \hat R}{\partial \vec y_l}} \cdot \underset{(n\times n)}{\operatorname{diag}(f'(\vec z_l))}$$

$$(p\times n) \leftarrow \frac{\partial \hat R}{\partial W_l} = \underset{(p\times 1)}{\vec y_{l-1}} \cdot \underset{(1\times n)}{\frac{\partial \hat R}{\partial \vec z_l}}$$

$$\underset{(1\times n)}{\frac{\partial \hat R}{\partial b_l}} = \frac{\partial \hat R}{\partial \vec z_l}$$