Example:

$$y_1 = f(x) = x W_1 , \quad x \in R^{1\times D}, \quad W_1 \in R^{D\times F}$$

$$\underset{\in R^{1\times G}}{y_2} = g(x) = f(x) W_2 , \quad f(x) \in R^{1\times F}, \quad W_2 \in R^{F\times G}$$

$$L = loss = \sum_i y_{2i} \in R$$

How to compute $\frac{\partial L}{\partial W_2} \in R^{F\times G}$, if we already have $\frac{\partial L}{\partial y_2}$

$$\frac{\partial L}{\partial W_2} = \underbrace{f(x)^T}_{R^{F\times 1}} \underbrace{\frac{\partial L}{\partial y_2}}_{\in R^{1\times G}}$$

k-term

[o o o o o]

Sum over rows

chain rule: $L(y_1, y_2, y_3, \dots)$

$$y_{2k} = \sum_r f(x)_r W_{2rk}$$

Why? Because:

$$\frac{\partial L}{\partial W_{2ij}} = \sum_k \frac{\partial L}{\partial y_{2k}} \cdot \boxed{\frac{\partial y_{2k}}{\partial W_{2ij}}}$$

$$\rightarrow \frac{\partial y_{2k}}{\partial W_{2ij}} = \begin{cases} f(x)_i & j = k \\ 0 & j \neq k \end{cases}$$

Zero because W_{2ij} not in this column

$$= \sum_k \frac{\partial L}{\partial y_{2k}} \cdot \underbrace{(j=k)}_{\text{only 1 when } j=k} \cdot f(x)_i$$

$$= \frac{\partial L}{\partial y_{2j}} f(x)_i$$

Putting together :

$$\frac{\partial L}{\partial W_{2ij}} = \frac{\partial L}{\partial y_{2j}} f(x)_i$$

to matrix

$$\Rightarrow \frac{\partial L}{\partial W_2} = \underbrace{f(X)^T}_{\begin{bmatrix} | \end{bmatrix} \in R^{F\times 1}} \boxed{\frac{\partial L}{\partial y_2}} \quad \text{from upstream}$$

$$\begin{bmatrix} \text{---} \end{bmatrix} \in R^{1\times G}$$

Extension : What if $f(X) \in R^{B\times F}$ is a batch? (each row an example)

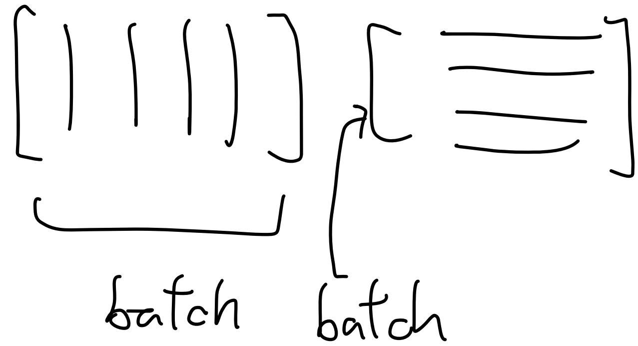
$$\frac{\partial L}{\partial W_2} = f(\underbrace{X_i}_{\text{row } i})^T \frac{\partial L}{\underbrace{\partial Y_{2i}}_{\text{row } i}} \quad \text{for example } i$$

We need to take the mean of $\frac{\partial L}{\partial i}$ gradient

which is :

$$\frac{\partial L}{\partial W_2} = \frac{1}{B} \sum_i \frac{\partial L}{\partial x_i} = \frac{1}{B} \sum f(x_i)^T \frac{\partial L}{\partial Y_{2i}}$$

$$= \frac{1}{B} f(X)^T \frac{\partial L}{\partial Y_2} \in \mathbb{R}^{B \times G}$$



To clean up the notation a bit:

$$y = xW \;,\; x \in R^{1\times F} \;,\; W \in R^{F\times D} \;,\; y \in R^{1\times D}$$

$$L(y) = \cdots\cdots \quad \text{Scalar}$$

Assume: $\frac{\partial L}{\partial y}$ is computed already

$$\frac{\partial L}{\partial W} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \frac{\partial L}{\partial w_{ij}} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

by chain rule: L is affected by $y_1, y_2\, y_3, \cdots y_D$

$$\frac{\partial L}{\partial W_{ij}} = \sum_K \frac{\partial L}{\partial y_k} \cdot \frac{\partial y_K}{\partial W_{ij}}$$

$$\left(\begin{tikzcd} & y_1 \arrow[dl] & \\ L & y_2 \arrow[l] & W_{ij} \arrow[ul] \arrow[l] \arrow[dl] \\ & y_3 \arrow[ul] & \end{tikzcd}\right)$$

$$= \sum_k \frac{\partial L}{\partial y_K} \cdot (k=j) \cdot x_r \qquad y_K = \sum_r x_r W_{rK}$$

$$= \frac{\partial L}{\partial y_K} \cdot x_r \qquad \Rightarrow \frac{\partial y_K}{\partial W_{ij}} = \begin{cases} x_r & k=j \\ 0 & k\neq j \end{cases}$$

Putting it to vector form

$$\frac{\partial L}{\partial W} = x^T \frac{\partial L}{\partial y}$$

If we introduce batch dimension:

$$Y = XW, \quad X \in R^{B \times F}, \quad W \in R^{F \times D},$$

$$L(Y) = \text{scalar} \qquad Y \in R^{B \times D}$$

Assume $\frac{\partial L}{\partial Y} \in R^{B \times D}$ is already computed

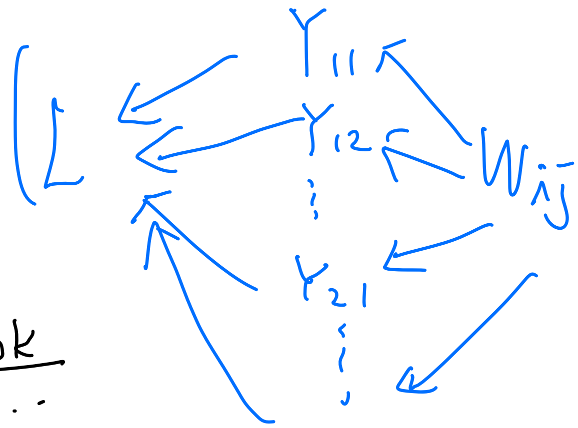
Compute:

$$\frac{\partial L}{\partial W} = \begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \frac{\partial L}{\partial W_{ij}} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix}$$

known

chain rule

$$\frac{\partial L}{\partial W_{ij}} = \sum_{b=1}^{B} \sum_{k=1}^{D} \boxed{\frac{\partial L}{\partial Y_{bk}}} \frac{\partial Y_{bk}}{\partial W_{ij}}$$



First compute $\frac{\partial Y_{bk}}{\partial W_{ij}}$, to compute this, expand Y_{bk}

$$Y_{bk} = X_{b,:} @ W_{:,k}$$

$$= \sum_{r=1}^{F} X_{br} W_{rk}$$

$$\frac{\partial Y_{bk}}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} \sum_{r=1}^{F} X_{br} W_{rk}$$

$$= \begin{cases} X_{bi} \dfrac{\partial W_{ij}}{\partial W_{ij}} & k = j \\ 0 & k \neq j \end{cases}$$

Plug this back:

$$\frac{\partial L}{\partial W_{ij}} = \sum_{b=1}^{B} \sum_{k=1}^{D} \frac{\partial L}{\partial Y_{bk}} \cdot (k=j) \cdot X_{bi}$$

when $k \neq j$, terms are zero

$$= \sum_{b=1}^{B} \frac{\partial L}{\partial Y_{bj}} \cdot X_{bi}$$

left dim

right

B is the contracting dim

Putting this to vector: $\dfrac{\partial L}{\partial W} = X^T \dfrac{\partial L}{\partial Y}$

$$X^T \in R^{F \times B}, \quad \frac{\partial L}{\partial Y} \in R^{B \times D}$$