A4-Q2: Gradient of Matrix Product

Consider $L(Y) \in \mathbb{R}$, for $Y \in \mathbb{R}^{D\times N}$

Suppose you have $\nabla_Y L \in \mathbb{R}^{D\times N}$.

Let $Y = H \cdot W$ for $H \in \mathbb{R}^{D\times M}$ & $W \in \mathbb{R}^{M\times N}$

Show that

$$\underset{D\times M}{\nabla_H L} = \underset{D\times N}{\nabla_Y L} \cdot \underset{N\times M}{W^T} \qquad (1)$$

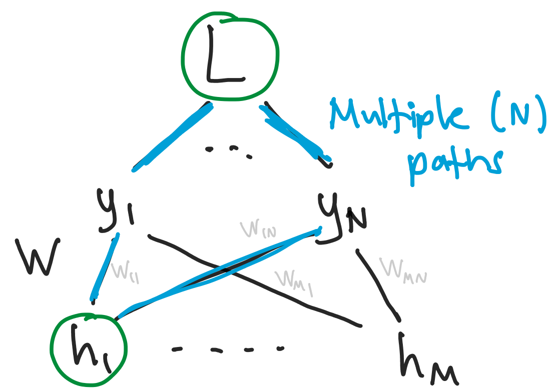
and

$$\underset{M\times N}{\nabla_W L} = \underset{M\times D}{H^T} \cdot \underset{D\times N}{\nabla_Y L} \qquad (2)$$

Pf (of (1))

Consider $\dfrac{\partial L}{\partial H_{dm}}$

$$= \sum_{n=1}^{N} \frac{\partial L}{\partial Y_{dn}} \frac{\partial Y_{dn}}{\partial H_{dm}}$$



$$= \underbrace{\left[\frac{\partial L}{\partial Y_{d1}} \cdots \frac{\partial L}{\partial Y_{dN}}\right]}_{} \begin{bmatrix} \frac{\partial Y_{d1}}{\partial H_{dm}} \\ \vdots \\ \frac{\partial Y_{dN}}{\partial H_{dm}} \end{bmatrix}$$

Thinking through just 1 node of 1 sample.

$$y_n = \sum_{m=1}^{M} h_m W_{mn}$$

$$\therefore \frac{\partial y_n}{\partial h_m} = W_{mn}$$

This is where the W^T comes from?

$$\therefore \frac{\partial L}{\partial H_{dm}} = \nabla_{Y_{d\cdot}} L \cdot \begin{bmatrix} W_{m1} \\ \vdots \\ W_{mN} \end{bmatrix}$$

this is the transpose of the m^{th} row of W

Put all of these, $m=1, \ldots, M$, in one row vector...

$$\therefore \left[\underset{m=1}{\frac{\partial L}{\partial H_{d1}}} \cdots \underset{m=M}{\frac{\partial L}{\partial H_{dM}}}\right] = \nabla_{Y_{d\cdot}} L \cdot \left[W_{1\cdot} \middle| \cdots \middle| W_{M\cdot} \right] = \nabla_{Y_{d\cdot}} L \cdot W^T$$

M^{th} col. of W^T.

Stacking for $d = 1, \ldots, D$

$$\begin{array}{c} d=1 \rightarrow \\ \\ \\ d=D \rightarrow \end{array} \underbrace{\begin{bmatrix} \frac{\partial L}{\partial H_{11}} & \cdots & \frac{\partial L}{\partial H_{1M}} \\ \vdots & & \vdots \\ \frac{\partial L}{\partial H_{D1}} & \cdots & \frac{\partial L}{\partial H_{DM}} \end{bmatrix}}_{\nabla_H L} = \underbrace{\begin{bmatrix} \nabla_{Y_{1\cdot}} L \\ \hline \vdots \\ \hline \nabla_{Y_{D\cdot}} L \end{bmatrix}}_{\nabla_Y L} \cdot W^T$$

This is how the D samples factor into the formula?

$$\boxed{\therefore \nabla_H L = \nabla_Y L \cdot W^T}$$

∎

Pf (of ②)

Need to show that $\nabla_W L = H^T \cdot \nabla_Y L$

Consider $\frac{\partial L}{\partial W_{ij}}$. W_{ij} is the weight from H_{di} to Y_{dj}

$$Y_{dj} = \sum_{i=1}^{M} H_{di} W_{ij} \qquad \therefore \frac{\partial Y_{dj}}{\partial W_{ij}} = H_{di} \quad \forall j = 1, \dots, N$$

So,
$$\frac{\partial L}{\partial W_{ij}} = \sum_{d=1}^{D} \frac{\partial L}{\partial Y_{dj}} \frac{\partial Y_{dj}}{\partial W_{ij}}$$

Consider a term for a single sample (d):

$$\begin{bmatrix} \frac{\partial L}{\partial W_{11}} & \cdots & \frac{\partial L}{\partial W_{1N}} \\ \vdots & & \\ \frac{\partial L}{\partial W_{M1}} & \cdots & \frac{\partial L}{\partial W_{MN}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial Y_{d1}} \frac{\partial Y_{d1}}{\partial W_{11}} & \cdots & \frac{\partial L}{\partial Y_{dN}} \frac{\partial Y_{dN}}{\partial W_{1N}} \\ \frac{\partial L}{\partial Y_{d1}} \frac{\partial Y_{d1}}{\partial W_{21}} & & \vdots \\ \vdots & & \\ \frac{\partial L}{\partial Y_{d1}} \frac{\partial Y_{d1}}{\partial W_{M1}} & & \frac{\partial L}{\partial Y_{dN}} \frac{\partial Y_{dN}}{\partial W_{MN}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial Y_{d1}} H_{d1} & \cdots & \frac{\partial L}{\partial Y_{dN}} H_{d1} \\ \frac{\partial L}{\partial Y_{d1}} H_{d2} & & \vdots \\ \vdots & & \\ \frac{\partial L}{\partial Y_{d1}} H_{dM} & \cdots & \frac{\partial L}{\partial Y_{dN}} H_{dM} \end{bmatrix}$$

It's ok if they just took this from the lectures.

$$= \begin{bmatrix} H_{d1} \\ H_{d2} \\ \vdots \\ H_{dM} \end{bmatrix} \begin{bmatrix} \frac{\partial L}{\partial Y_{d1}} & \frac{\partial L}{\partial Y_{d2}} & \cdots & \frac{\partial L}{\partial Y_{dN}} \end{bmatrix} = h^T \cdot \nabla_y L$$

$$= (H_{d\cdot})^T \cdot \nabla_{Y_{d\cdot}} L$$

They could point out this outer-product for a single sample?

Adding over d ...

(rank-1)

$$\nabla_W L = \sum_{d=1}^{D} (H_{d\cdot})^T \cdot \nabla_{Y_{d\cdot}} L$$

Then combine all the rank-1 products together?

$$= \left[\begin{array}{c|c|c} (H_{1\cdot})^T & \cdots & (H_{D\cdot})^T \end{array}\right] \begin{bmatrix} \nabla_{Y_{1\cdot}} L \\ \hline \vdots \\ \hline \nabla_{Y_{D\cdot}} L \end{bmatrix}$$

$$\boxed{\therefore\ \nabla_W L = H^T \cdot \nabla_Y L}$$

$\blacksquare$

Alternatively, consider the function $\bar{L}(Y^T) = L(Y)$.
It's just a version of L that operates on the transpose of Y. Everything is transposed for $\bar{L}$.

Thus, $\nabla_{Y^T} \bar{L} = (\nabla_Y L)^T$, and $\nabla_{W^T} \bar{L} = (\nabla_W L)^T$.

And $$\nabla_{W^T} \bar{L}(Y^T) = \nabla_{W^T} \bar{L}(W^T H^T)$$

Let $A = W^T$, $B = H^T$.

According to the proof of (1), we know

$$\nabla_A \bar{L}(AB) = \nabla_{Y^T} \bar{L} \cdot B^T$$
$$= (\nabla_Y L)^T \cdot H$$

$$\therefore\ \nabla_{W^T} \bar{L} = (\nabla_Y L)^T \cdot H$$

But $\nabla_W L = (\nabla_{W^T} \bar{L})$

$$= \left[(\nabla_Y L)^T H \right]^T = H^T \cdot \nabla_Y L \quad \text{as required.} \quad \blacksquare$$

A note about proof design.

The solution must prove (1) <u>and</u> (2).

One of them must be proven outright, but the proof of the second one can use the first (already proven) result.