$$X \in R^{N \times D_{in}}$$
$$Y \in R^{N \times D_{out}}$$
$$W_1 \in R^{D_{in} \times H}$$
$$W_2 \in R^{H \times D_{out}}$$
$$H = XW_1$$
$$H_{relu} = max(H, 0)$$
$$Y_{pred} = H_{relu}W_2$$
$$loss = ||Y_{pred} - Y||_F^2$$

1.求 $\frac{\partial loss}{\partial Y_{pred}}$

因为 $d(loss) = d(||Y_{pred} - Y||_F^2) = tr(2(Y_{pred} - Y)^\top dY_{pred})$

所以 $\frac{\partial loss}{\partial Y_{pred}} = 2(Y_{pred} - Y)$

2.求 $\frac{\partial loss}{\partial W_2}$

$$\begin{aligned}
\text{因为} d(loss) &= d(||Y_{pred} - Y||_F^2) \\
&= d(||H_{relu}W_2 - Y||_F^2) \\
&= tr(2(H_{relu}W_2 - Y)^\top d(H_{relu}W_2 - Y)) \\
&= tr(2(H_{relu}W_2 - Y)^\top dH_{relu}W_2) \\
&= tr(2(H_{relu}W_2 - Y)^\top H_{relu}dW_2)
\end{aligned}$$

$$\begin{aligned}
\text{所以} \frac{\partial loss}{\partial H_{relu}} &= 2H_{relu}^\top(H_{relu}W_2 - Y) \\
&= H_{relu}^\top \frac{\partial loss}{\partial Y_{pred}}
\end{aligned}$$

3.求 $\frac{\partial loss}{\partial H_{relu}}$

$$\begin{aligned}
\text{因为} d(loss) &= d(||Y_{pred} - Y||_F^2) \\
&= d(||H_{relu}W_2 - Y||_F^2) \\
&= tr(2(H_{relu}W_2 - Y)^\top d(H_{relu}W_2 - Y)) \\
&= tr(2(H_{relu}W_2 - Y)^\top dH_{relu}W_2) \\
&= tr(2W_2(H_{relu}W_2 - Y)^\top dH_{relu})
\end{aligned}$$

$$\begin{aligned}
\text{所以} \frac{\partial loss}{\partial H_{relu}} &= 2(H_{relu}W_2 - Y)W_2^\top \\
&= \frac{\partial loss}{\partial Y_{pred}}W_2^\top
\end{aligned}$$

4.求 $\frac{\partial loss}{\partial H}$

因为 $d(loss) = tr(2W_2(H_{relu}W_2 - Y)^\top dH_{relu})$

且 $H_{relu} = max(H, 0)$

若 $H > 0$，$H_{relu} = H$，$d(loss) = tr(2W_2(H_{relu}W_2 - Y)^\top dH)$

$$\frac{\partial loss}{\partial H} = 2(H_{relu}W_2 - Y)W_2^\top = \frac{\partial loss}{\partial H_{relu}}$$

若 $H <= 0$，$H_{relu} = 0, d(loss) = 0$

$$\frac{\partial loss}{\partial H} = 0$$

$$5.\text{求}\frac{\partial loss}{\partial W_1}$$

$$\begin{aligned}\text{因为}d(loss) &= tr(2W_2(H_{relu}W_2-Y)^\top dH)\\ &= tr(2W_2(H_{relu}W_2-Y)^\top dXW_1)\\ &= tr(2W_2(H_{relu}W_2-Y)^\top XdW_1)\end{aligned}$$

$$\text{所以}\frac{\partial loss}{\partial W_1} = X^\top 2(H_{relu}W_2-Y)W_2^\top = X^\top\frac{\partial loss}{\partial H}$$