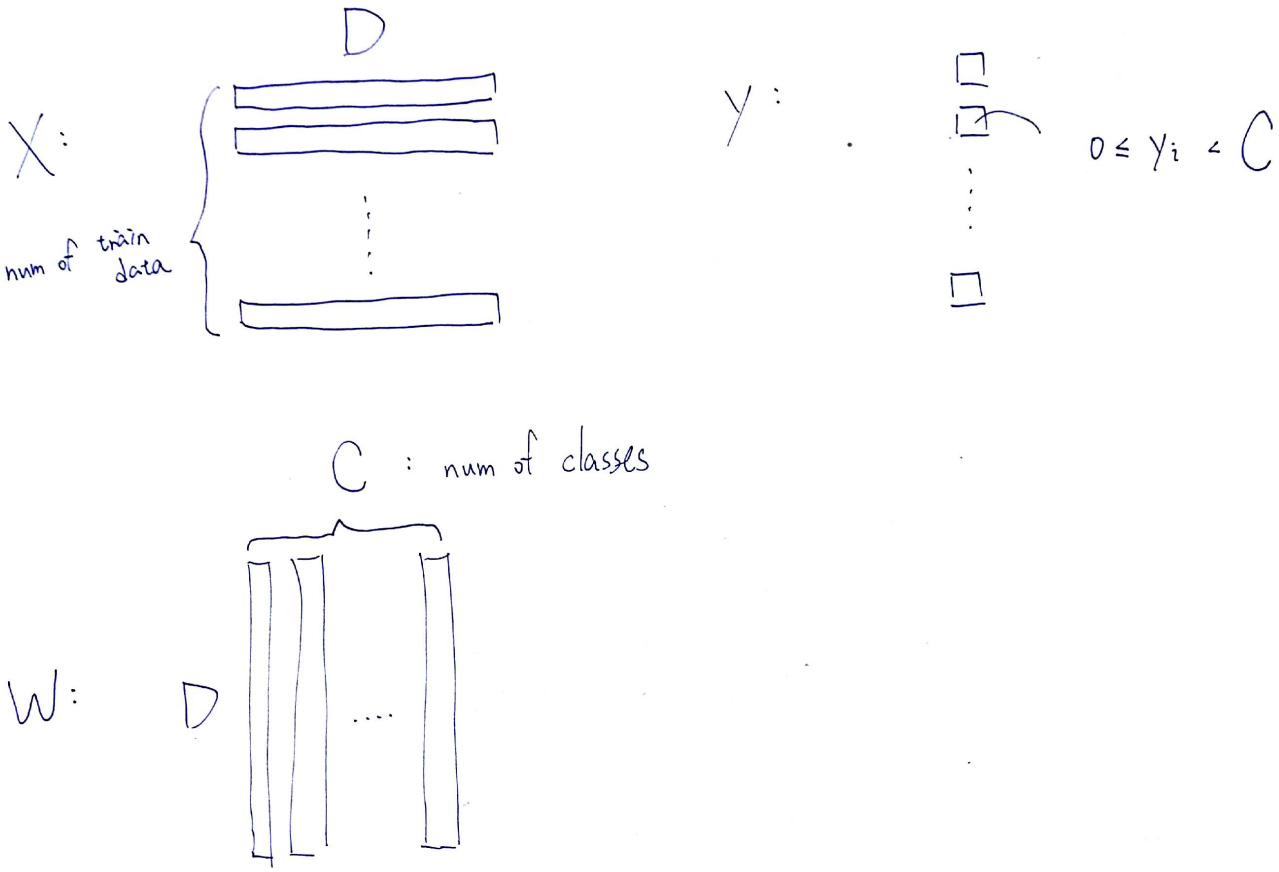


For one data point i, its data loss is computed as follow:

$$L_i = \sum_{\substack{j=0 \\ j \neq y_i}}^{C-1} \max\left(0,\ S_j - S_{y_i} + 1\right) \quad \text{where} \quad S_j = X_i W_j$$

score for class j

S_j : 1 □ = X_i : 1 ▭ (D) * W_j : ▯ (D, 1)

For its scores for all classes,

$S[i]$: 1 ▭ (C) = X_i : 1 ▭ (D) * W : ▯ (D, C)

For all data points,

S : ▯ (num of train data, C) = X : ▯ (num of train data, D) * W : ▯ (D, C)

To compute gradient

$$\nabla_{S_j} L_i = \mathbb{1}\left(s_j - s_{y_i} + 1 > 0\right) \quad \text{where } \mathbb{1}(\text{true}) = 1,\ \mathbb{1}(\text{false}) = 0$$

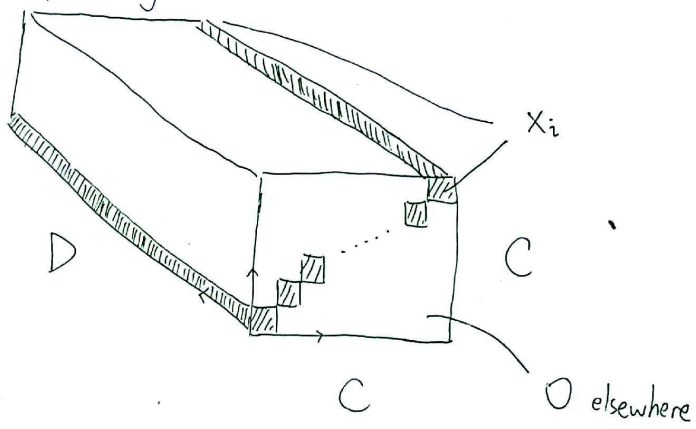
$$\nabla_{S_{y_i}} L_i = -\sum_{j \neq y_i} \mathbb{1}\left(s_j - s_{y_i} + 1 > 0\right)$$

$$\nabla_{S[i]} L_i = \quad 1 \quad \boxed{} \; C$$

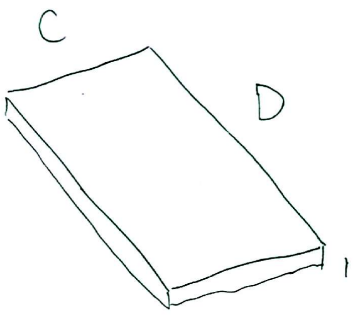
$\nabla_{w_j} S_j = x_i$

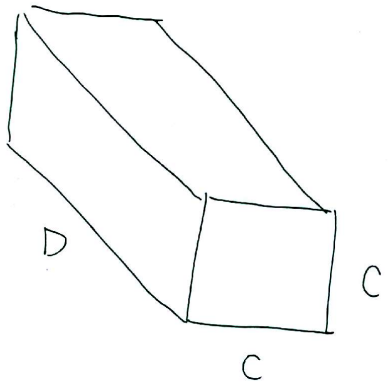
$\nabla_{w_j} S_k = 0 \quad k \neq j$

$\nabla_w S[i] =$

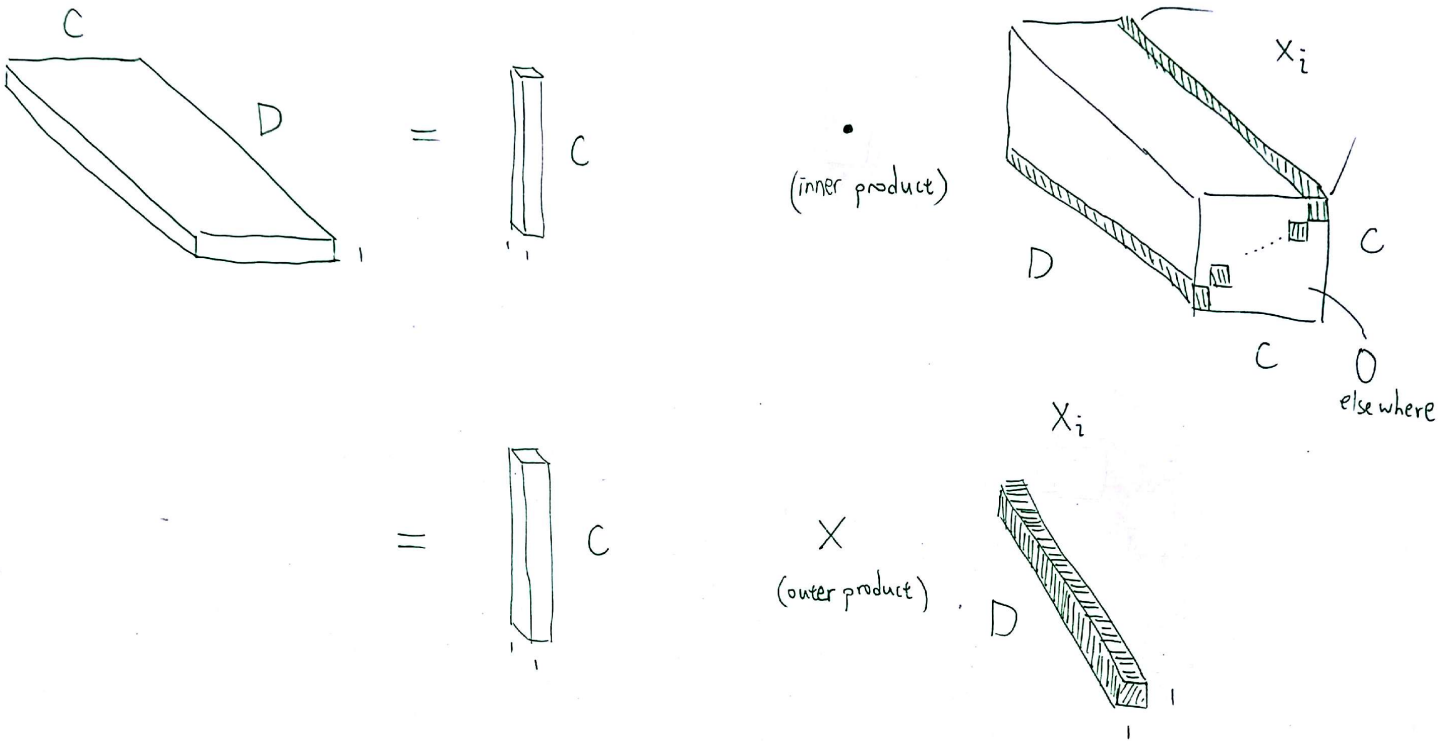


$$\nabla_w L_i = \nabla_{S[i]} L_i * \nabla_w S[i]$$





$\nabla_w L_i$ $\nabla_{s[i]} L_i$ $\nabla_w s[i]$



... which simplifies computation a bit.