

## Softmax

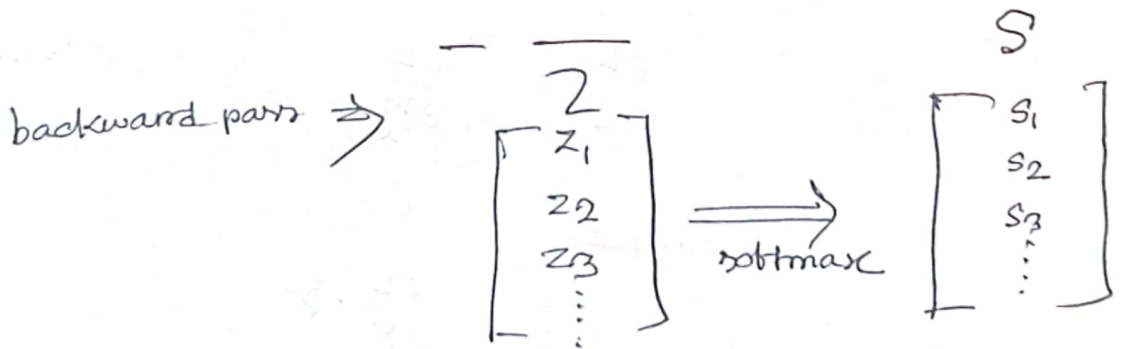
given outputs from prev. layer =  $\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \end{bmatrix}$

for every  $z_i$ ,

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{z=1}^n e^{z_z}}$$

$$= \frac{e^{z_i - \max(z)}}{\sum_{z=1}^n e^{z_z - \max(z)}}$$

[to stop overflow]



$$\frac{\partial s_i}{\partial z_j} = \begin{cases} s_i(1-s_i) & \text{when } i=j \\ -s_i s_j & \text{when } i \neq j \end{cases}$$

$$J_n = \text{diag}(S) - SS^T$$

# Batch normalization

forward

$$\hat{X} = \frac{X - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$Y = \gamma \cdot \hat{X} + \beta$$

backward

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial Y} \times \frac{\partial Y}{\partial \gamma}$$

$$= \frac{\partial L}{\partial Y} \left( \sum_{\text{batch}} \hat{X}_i \right) \times \frac{\partial Y}{\partial \gamma_i}$$

$$= \sum \frac{\partial L}{\partial Y_i} \hat{X}_i$$

sum because gamma has effect on all the samples in the batch.

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^m \frac{\partial L}{\partial Y_i} \hat{X}_i$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^m \frac{\partial L}{\partial Y_i} \quad [m = \text{batch size}]$$

for a sample,

$$\frac{\partial L}{\partial \hat{X}} = \frac{\partial L}{\partial Y} \cdot \gamma$$

$$\begin{aligned}\frac{\partial L}{\partial \sigma^2} &= \frac{\partial L}{\partial \gamma} \times \frac{\partial \gamma}{\partial \hat{X}} \times \frac{\partial \hat{X}}{\partial \sigma^2} \\ &= \frac{\partial L}{\partial \gamma} \times \frac{\partial \gamma}{\partial \hat{X}} \times (x - \mu) \times -\frac{1}{2} \times \left( \sqrt{\sigma^2 + \epsilon} \right)^{-3/2}\end{aligned}$$

for many samples  $\Rightarrow$

$$\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^m \frac{\partial L}{\partial \gamma_i} \times \frac{\partial \gamma_i}{\partial \hat{X}_i} \times (x_i - \mu) \times -\frac{1}{2} \times \left( \sqrt{\sigma^2 + \epsilon} \right)^{-3/2}$$

same way (chain rule and partial derivative),

we can find,  $\frac{\partial L}{\partial X}$  and  $\frac{\partial L}{\partial \mu}$ .

sum over  $x_i$ ,  $\hat{X}$ , or  $\gamma$  is important for  $\gamma$ ,  $\sigma^2$ ,  $\beta$ ,  $\mu$  because they are same over all samples.