Gradient of w.r.t the Softmax

$S = QK^T \qquad P = \text{softmax}(S) \qquad O = PV$

Imagine we have a row vector $S_i = S[i,:] \in R^N$

$P_i = \text{softmax}(S_i) \in R^N$

$$\text{softmax}(P_{ij}) = \frac{e^{S_{ij}}}{\sum_{j=1}^{N} e^{S_{ij}}}$$

$$\frac{\partial \phi}{\partial S_i} = \frac{\partial \phi}{\partial P_i} \cdot \frac{\partial P_i}{\partial S_i}$$

each element in this matrix will be the derivative of an element of P_i w.r.t an element of S_i

$$\frac{\partial P_{ij}}{\partial S_{ik}} = \frac{\partial \left[\frac{e^{S_{ij}}}{\sum_{j=1}^{N} e^{S_{ij}}}\right]}{\partial S_{ik}}$$

We know that the derivative of the ratio of two functions is as follows:

$$\left[\frac{f(x)}{g(x)}\right]' = \frac{f'(x)\,g(x) - g'(x)\,f(x)}{[g(x)]^2}$$

$$S_i = [S_{i1} \quad S_{i2} \quad S_{i3}] \xrightarrow{\text{softmax}} P_i = [P_{i1} \quad P_{i2} \quad P_{i3}]$$

So we will have two cases: $j = k$, $j \neq k$

$$\frac{\partial \left[\frac{e^{S_{ij}}}{\sum_{j=1}^{N} e^{S_{ij}}}\right]}{\partial S_{ik}}$$

$$j = k \Rightarrow \frac{e^{S_{ij}}\left(\sum_{j=1}^{N} e^{S_{ij}}\right) - e^{S_{ik}} e^{S_{ij}}}{\left(\sum_{j=1}^{N} e^{S_{ij}}\right)^2} = \frac{e^{S_{ij}}\left(\sum_{j=1}^{N} e^{S_{ij}} - e^{S_{ik}}\right)}{\left(\sum_{j=1}^{N} e^{S_{ij}}\right)^2} =$$

$$\frac{e^{S_{ij}}}{\left(\sum_{j=1}^{N} e^{S_{ij}}\right)} \cdot \frac{\left(\sum_{j=1}^{N} e^{S_{ij}} - e^{S_{ik}}\right)}{\left(\sum_{j=1}^{N} e^{S_{ij}}\right)} = P_{ij} \cdot (1 - P_{ik})$$

$$j \neq k \Rightarrow \frac{0 - e^{S_{ik}} e^{S_{ij}}}{\left(\sum_{j=1}^{N} e^{S_{ij}}\right)^2} = -P_{ik} P_{ij}$$

To summarize:

$$\frac{\partial P_{ij}}{\partial S_{ik}} \begin{cases} j = k \implies P_{ij}(1 - P_{ik}) \\ j \neq k \implies -P_{ik} P_{ij} \end{cases}$$

$P_{11} - P_{11} \cdot P_{11}$ $\qquad$ $P_{12} - P_{12} P_{12}$ $\qquad$ Symmetric

$$\frac{\partial P_{ij}}{\partial S_{ik}} = \begin{bmatrix} P_{11}(1-P_{11}) & -P_{11}P_{12} & -P_{11}P_{13} & \cdots & -P_{11}P_{1N} \\ -P_{12}P_{11} & P_{12}(1-P_{12}) & -P_{12}P_{13} & \cdots & -P_{12}P_{1N} \\ \vdots & & & & \\ -P_{1N}P_{11} & -P_{1N}P_{12} & \cdots & & P_{1N}(1-P_{1N}) \end{bmatrix} =$$

OUTER PRODUCT

$$\mathrm{diag}(P_i) - P_i \otimes P_i =$$

$$\mathrm{diag}(P_i) - P_i P_i^T$$

Only if you consider P_i a column vector

Using the fact that the Jacobian of $y = \mathrm{softmax}(x)$ is $\mathrm{diag}(y) - yy^T$