$$X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \end{pmatrix}_{2\times 2} \qquad W = \begin{pmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{2,2} & w_{2,3} \end{pmatrix}_{2\times 3}$$

$$Y = XW = \begin{pmatrix} x_{1,1}w_{1,1} + x_{1,2}w_{2,1} & x_{1,1}w_{1,2} + x_{1,2}w_{2,2} & x_{1,1}w_{1,3} + x_{1,2}w_{2,3} \\ x_{2,1}w_{1,1} + x_{2,2}w_{2,1} & x_{2,1}w_{1,2} + x_{2,2}w_{2,2} & x_{2,1}w_{1,3} + x_{2,2}w_{2,3} \end{pmatrix}$$

$$\sigma(Y) = \begin{pmatrix} \sigma(y_{11}), & \sigma(y_{12}), & \sigma(y_{13}) \\ \sigma(y_{21}), & \sigma(y_{22}), & \sigma(y_{23}) \end{pmatrix}$$

$$L(\sigma(y)) \in R$$

Note: "the gradient of sigmoid" means the derivative of the sigmoid function wrt its input, Y

Thus, we are looking to calculate $\frac{\partial L}{\partial Y} = \frac{\partial L}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial Y}$

$$\frac{\partial L}{\partial \sigma} = \begin{bmatrix} \frac{\partial L}{\partial \sigma_{11}} & \frac{\partial L}{\partial \sigma_{12}} & \frac{\partial L}{\partial_{13}} \\ \frac{\partial L}{\partial \sigma_{21}} & \frac{\partial L}{\partial_{22}} & \frac{\partial L}{\partial_{23}} \end{bmatrix}_{2\times 3}$$ Because $L(\sigma) \in R$

we want to calculate $\frac{\partial \sigma}{\partial Y}$, and we will do it by breaking it up to steps, for each entry of $\sigma(Y)$

This is due to the facts that we're dealing with scalars

$$\frac{\partial L}{\partial y_{11}} = \sum_{i}^{n} \sum_{j}^{m} \frac{\partial L}{\partial \sigma_{ij}} \cdot \frac{\partial \sigma_{ij}}{\partial y_{11}},$$ which could visualise

as a dot product (Elementwise multiplication and summation)

$$\begin{pmatrix} \frac{\partial L}{\partial \sigma_{11}} & \frac{\partial L}{\partial \sigma_{12}} & \frac{\partial L}{\partial \sigma_{13}} \\ \frac{\partial L}{\partial \sigma_{21}} & \frac{\partial L}{\partial \sigma_{22}} & \frac{\partial L}{\partial \sigma_{23}} \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial \sigma_{11}}{\partial y_{11}} & \frac{\partial \sigma_{12}}{\partial y_{11}} & \frac{\partial \sigma_{13}}{\partial y_{11}} \\ \frac{\partial \sigma_{21}}{\partial y_{11}} & \frac{\partial \sigma_{22}}{\partial y_{11}} & \frac{\partial \sigma_{23}}{\partial y_{11}} \end{pmatrix}$$

and since $\frac{\partial \sigma}{\partial y} = \sigma(y)(1 - \sigma(y))$ For Scalars!!!

$$\begin{pmatrix} \frac{\partial L}{\partial \sigma_{11}} & \frac{\partial L}{\partial \sigma_{12}} & \frac{\partial L}{\partial \sigma_{13}} \\ \frac{\partial L}{\partial \sigma_{21}} & \frac{\partial L}{\partial \sigma_{22}} & \frac{\partial L}{\partial \sigma_{23}} \end{pmatrix} \cdot \begin{pmatrix} \sigma_{11}(1-\sigma_{11}) & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\longrightarrow \frac{\partial L}{\partial y_{11}} = \frac{\partial L}{\partial \sigma_{11}} \cdot \sigma_{11}(1 - \sigma_{11}),$$

and in general

$$\frac{\partial L}{\partial y} = \begin{pmatrix} \frac{\partial L}{\partial \sigma_{11}} \cdot \sigma_{11}(1-\sigma_{11}), & \frac{\partial L}{\partial \sigma_{12}} \cdot \sigma_{12}(1-\sigma_{12}), & \frac{\partial L}{\partial \sigma_{13}} \cdot (1-\sigma_{13}) \\ \frac{\partial L}{\partial \sigma_{21}} \cdot (1-\sigma_{21}), & \frac{\partial L}{\partial \sigma_{22}} \cdot \sigma_{22}(1-\sigma_{22}), & \frac{\partial L}{\partial \sigma_{23}} \sigma_{23}(1-\sigma_{23}) \end{pmatrix}$$

And How could we compute it in code?

$$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial \sigma} \odot \sigma(1-\sigma)$$

where $\odot$ is elementwise multiplication.

Proof and reasoning

For simplicity let us define $L(X) = \sum_{i}^{k} 2x_i$

$$L(\sigma) = 2\sigma_{11} + 2\sigma_{12} + 2\sigma_{13} + 2\sigma_{21} + 2\sigma_{22} + 2\sigma_{23}$$

$$\frac{\partial L(\sigma)}{\partial y_{11}} = \frac{\partial 2\sigma(y_{11})}{\partial y_{11}} + \frac{\partial 2\sigma(y_{12})}{\partial y_{11}} + \frac{\partial 2\sigma(y_{13})}{\partial y_{11}} + \frac{\partial 2\sigma(y_{21})}{\partial y_{11}} + \frac{\partial 2\sigma(y_{22})}{\partial y_{11}} + \frac{\partial 2\sigma(y_{23})}{\partial y_{11}}$$

$$\frac{\partial L(\sigma)}{\partial y_{11}} = \frac{\partial L}{\partial \sigma(y_{11})} \cdot \frac{\partial \sigma(y_{11})}{\partial y_{11}} + 0 + \dots + 0 = 2(\sigma_{11}(1-\sigma_{11}))$$