Backpropagation by hand

Layer 2: $\frac{\partial \mathcal{L}}{\partial w_j^{(2)}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w_j^{(2)}} = (\hat{y} - y) \cdot \frac{\partial \hat{y}}{\partial w_j^{(2)}}$

expanding the first summation:

$$\hat{y} = w_1^{(2)} \underbrace{\sum_i x_i w_{i1}^{(1)}}_{\frac{\partial \hat{y}}{\partial w_1^{(2)}}} + w_2^{(2)} \underbrace{\sum_i x_i w_{i2}^{(1)}}_{\frac{\partial \hat{y}}{\partial w_2^{(2)}}}$$

So, $\frac{\partial \mathcal{L}}{\partial w_j^{(2)}} = (\hat{y} - y) \sum_i x_i w_{ij}^{(1)}$

Matrix form: $\hat{y} = X W^{(1)} W^{(2)}$, $\frac{\partial \hat{y}}{\partial W^{(2)}} = X W^{(1)}$

$\therefore \frac{\partial \mathcal{L}}{\partial W^{(2)}} = (\hat{y} - y) X W^{(1)}$

Layer 1: $\frac{\partial \mathcal{L}}{\partial w^{(1)}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h} \frac{\partial h}{\partial w^{(1)}}$, where $h = X W^{(1)}$

input to hidden layer

$= (\hat{y} - y) W^{(2)^T} X$

I missed this part! $\frac{\partial (X W^{(1)} W^{(2)})}{\partial (X W^{(1)})} = W^{(2)^T}$

with bias terms:

$$\hat{y} = C + 1\, b^{(2)} \qquad \frac{\partial \hat{y}}{\partial b^{(2)}} = 1$$

$$\hat{y} = (xw^{(1)} + b^{(1)})\,w^{(2)} + b^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial b^{(2)}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial b^{(2)}} = \frac{\partial \mathcal{L}}{\partial \hat{y}} = \hat{y} - y$$

with activations: $\hat{y} = f_2\big(\underbrace{f_1(\overbrace{xw^{(1)} + b^{(1)}}^{z_1})\,w^{(2)} + b^{(2)}}_{z_2}\big)$

$$\frac{\partial \mathcal{L}}{\partial w^{(2)}} = \underset{1}{\frac{\partial \mathcal{L}}{\partial \hat{y}}}\ \underset{2}{\frac{\partial \hat{y}}{\partial f_2}}\ \underset{3}{\frac{\partial f_2}{\partial z_2}}\ \underset{4}{\frac{\partial z_2}{\partial w^{(2)}}}$$

since $\hat{y} = f_2$

$$= \underset{1}{(\hat{y} - y)}\ \underset{2}{(1)}\ \underset{3}{\frac{\partial f_2}{\partial z_2}}\ \underset{4}{f_1(xw^{(1)} + b^{(1)})}$$

assume both sigmoid

depends on function, e.g. $f_2(z) = \sigma(z) = f_1(z)$

$$\frac{\partial f_2}{\partial z} = \sigma(z)(1 - \sigma(z))$$

Layer 1:

store this

$$\frac{\partial \mathcal{L}}{\partial w^{(1)}} = \left(\frac{\partial \mathcal{L}}{\partial f_2} \frac{\partial f_2}{\partial z_2}\right) \frac{\partial z_2}{\partial f_1} \frac{\partial f_1}{\partial z_1} \frac{\partial z_1}{\partial w^{(1)}}$$

$$= (\hat{y} - y) \frac{\partial f_2}{\partial z_2}\, w^{(2)^T} \frac{\partial f_1}{\partial z_1}\, X$$

$$= (\hat{y} - y)\, \sigma(z_2)(1 - \sigma(z_2))\, w^{(2)^T} \sigma(z_1)(1 - \sigma(z_1))\, X$$