# Derivation of Backpropagation

## Introduction (Basic derivation)

Let $\hat{y} = w_1 x_1 + w_2 x_2 + w_3 x_3 + b$

and $Loss = (y - \hat{y})^2$ ← squared error loss

we can rewrite the above in "matrix" format.

$$\underset{1\times1}{[y]} = \underset{1\times3}{[x_1 \;\; x_2 \;\; x_3]} \underset{3\times1}{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}} + \underset{1\times1}{[b]}$$

What we want is to learn the best W and b, and for backprop we need derivative of learnable parameter w.r.t. Loss.

notice gradients have same shape as the original tensors!

so, $$\frac{dL}{dW} = \underset{3\times1}{\begin{bmatrix} dL/dw_1 \\ dL/dw_2 \\ dL/dw_3 \end{bmatrix}} \qquad \frac{dL}{db} = \underset{1\times1}{[dL/db]}$$

Lets do $dL/dw_1$ first!

chain rule

$$\frac{dL}{dw_1} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_1}$$

if $L = (y-\hat{y})^2$, $\frac{dL}{d\hat{y}} = -2(y-\hat{y})$

if $\hat{y} = w_1x_1 + w_2x_2 + w_3x_3 + b$, $\frac{d\hat{y}}{dw_1} = x_1$

$$\therefore \frac{dL}{dw_1} = -2(y-\hat{y})x_1$$

More generally, $\frac{dL}{dw} = \begin{bmatrix} dL/dw_1 \\ dL/dw_2 \\ dL/dw_3 \end{bmatrix} = \underset{3\times1}{\begin{bmatrix} -2(y-\hat{y})x_1 \\ -2(y-\hat{y})x_2 \\ -2(y-\hat{y})x_3 \end{bmatrix}}$

$$= \underset{3\times1}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} \underset{1\times1}{\begin{bmatrix} -2(y-\hat{y}) \end{bmatrix}} = \vec{x}^T \frac{dL}{d\hat{y}}$$

Remember! $\vec{x}$ was a row vector, now its a column vector!

Now we need our gradient for the intercept term!

$$\frac{dL}{db} = \frac{dL}{d\hat{y}} \frac{d\hat{y}}{db}$$

$\frac{dL}{d\hat{y}} = -2(y-\hat{y})$ (same as before)

$\frac{d\hat{y}}{db} = 1$ $\therefore \frac{dL}{db} = \frac{dL}{d\hat{y}} \cdot 1$

$$\boxed{\text{Final Results:} \quad \frac{dL}{dw} = \vec{X}^T \frac{dL}{d\hat{y}} \qquad \frac{dL}{db} = \frac{dL}{d\hat{y}}}$$

## Adding a Batch Dimension

A thing we omitted in the past derivation was the batch dimension! In neural networks we typically pass in N samples and then propagate back the average loss!

So lets rewrite the problem to have a batch size of 2.

Let:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \oplus \underset{1\times1}{[b]}$$

Element-wise sum! same $b_{1\times1}$ is added to each sample in the batch independently (broadcasting)

and our loss is now Mean Squared Error

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \text{where } N=2 \text{ in our case.}$$

just like before, we need to compute $\frac{dL}{d\hat{y}}$, but we will compute this for all $\hat{y}_i$ as we have multiple samples in our batch.

move constant into sum.

$$\frac{dL}{d\hat{y}}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2\right)$$

Derivative of a sum is the sum of the derivatives.

$$=\frac{dL}{d\hat{y}}\sum_{i=1}^{N}\frac{1}{N}(y_i-\hat{y}_i)^2$$

$$=\sum_{i=1}^{N}\frac{dL}{d\hat{y}_i}\frac{1}{N}(y_i-\hat{y}_i)^2$$

$$\frac{dL}{d\hat{y}}=\sum_{i=1}^{N}\underbrace{-\frac{2}{N}(y_i-\hat{y}_i)}_{\text{easy to compute}}$$

We haven't dealt w/ this sum yet, but we will soon!

derivative same as before!

So now, just as before,

$$\frac{dL}{dW} = \begin{bmatrix} dL/dw_1 \\ dL/dw_2 \\ dL/dw_3 \end{bmatrix}$$

Lets do $dL/dw_1$ first.

$$\frac{dL}{dw_1} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dw_1} = \sum_{i=1}^{2} -\frac{2}{2}(y_i - \hat{y}_i)\frac{d\hat{y}_i}{dw_1}$$

how do we differentiate $\hat{y}_i$ wrt $w_1$?

$$\hat{y}_1 = w_1 x_{11} + w_2 x_{12} + w_3 x_{13} + b$$
$$\hat{y}_2 = w_1 x_{21} + w_2 x_{22} + w_3 x_{23} + b$$

then, $\frac{d\hat{y}_1}{dw_1} = x_{11}$ and $\frac{d\hat{y}_2}{dw_1} = x_{21}$

so then

$$\frac{dL}{dw_1} = \sum_{i=1}^{2} -\frac{2}{2}(y_i - \hat{y}_i)\frac{d\hat{y}_i}{dw_1}$$
$$= -(y_1 - \hat{y}_1)x_{11} - (y_2 - \hat{y}_2)x_{21}$$

Similarly:

$$\frac{dL}{dW_2} = -(y_1-\hat{y}_1)X_{12} - (y_2-\hat{y}_2)X_{22}$$

$$\frac{dL}{dW_3} = -(y_1-\hat{y}_1)X_{13} - (y_2-\hat{y}_2)X_{23}$$

All together!

Sum over batch is handled within the matrix multiplication

$$\frac{dL}{dW} = \underset{3\times 1}{\begin{bmatrix} dL/dW_1 \\ dL/dW_2 \\ dL/dW_3 \end{bmatrix}} = \underset{3\times 1}{\begin{bmatrix} -(y_1-\hat{y}_1)X_{11} - (y_2-\hat{y}_2)X_{21} \\ -(y_1-\hat{y}_1)X_{12} - (y_2-\hat{y}_2)X_{22} \\ -(y_1-\hat{y}_1)X_{13} - (y_2-\hat{y}_2)X_{23} \end{bmatrix}}$$

$$= \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \end{bmatrix} \begin{bmatrix} -(y_1-\hat{y}_1) \\ -(y_2-\hat{y}_2) \end{bmatrix}$$

$$= X^T \begin{bmatrix} dL/d\hat{y}_1 \\ dL/d\hat{y}_2 \end{bmatrix} = X^T \frac{dL}{d\hat{y}}$$

$$\frac{dL}{db} = \underbrace{\frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{db}}_{} = \sum_{i=1}^{2} -(y_i - \hat{y}_i) \cancel{\frac{d\hat{y}_i}{db}} \rightarrow 1$$

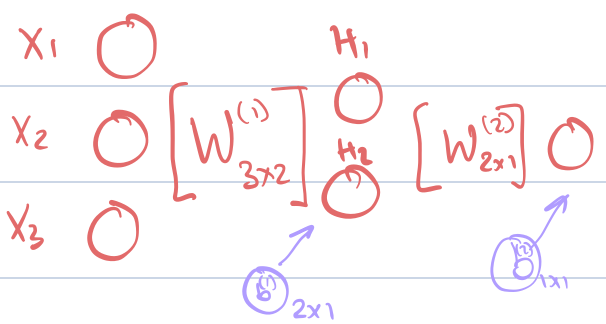
always 1

$$= \sum_{i=1}^{2} \frac{dL}{d\hat{y}_i}$$

$$\therefore \boxed{\frac{dL}{dw} = X^T \frac{dL}{dy} \qquad \frac{dL}{db} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}_i}}$$

## Stacking Layers!

Until now we have done a very simple projection with a Weight $W$ and bias $b$. But deep learning stacks multiple layers, what does that look like?




$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \end{bmatrix} \qquad H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$$

$$W^{(1)} = \left[ W_{3\times 2} \right] \qquad W^{(2)} = \left[ W_{2\times 1} \right]$$

$$b^{(1)} = \left[ b_{2\times 1} \right] \qquad b^{(2)} = \left[ b_{1\times 1} \right.$$

$$\hat{y} = [XW^{(1)} + b^{(1)}]W^{(2)} + b^{(2)}$$

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Lets split this into 2 composed functions:

$$\hat{y} = HW^{(2)} + b^{(2)} \qquad H = XW^{(1)} + b^{(1)}$$

So lets find the gradients of $W^{(2)}$ and $b^{(2)}$

based on what we did previously:

$$\frac{dL}{dW^{(2)}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dW^{(2)}} = H^T \frac{dL}{d\hat{y}}$$

$$\frac{dL}{db^{(2)}} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}_i}$$

Now that we have gradients for $W^{(2)}$ and $b^{(2)}$

we must backpropagate to $W^{(1)}$ and $b^{(1)}$

$$\frac{dL}{dW^{(1)}} = \frac{dL}{dH} \cdot \frac{dH}{dW^{(1)}} = X^T \underbrace{\frac{dL}{dH}}$$ Using previous formula

$$\frac{dL}{dH} = \underbrace{\frac{dL}{d\hat{y}}}_{\text{we know this!}} \cdot \underbrace{\frac{d\hat{y}}{dH}}_{\text{we don't know this}}$$

we have not computed this intermediate, but we need it for backprop to a new layer!

Remember, the $H$ in our case is $\begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix}$ b/c we have a batch size of 2

$B \times 2$

$$\frac{d\hat{y}}{dH} = \begin{bmatrix} dL/dh_{11} & dL/dh_{12} \\ dL/dh_{21} & dL/dh_{22} \end{bmatrix}$$

Let's do $\dfrac{dL}{dh_{11}} = \dfrac{dL}{d\hat{y}_1} \cdot \dfrac{d\hat{y}_1}{dh_{11}}$

if $\hat{y}_1 = W_1^{(2)} h_{11} + W_2^{(2)} h_{12} + b^{(2)}$

Then $\dfrac{d\hat{y}_1}{dh_{11}} = W_1^{(2)}$

Similarly, $\dfrac{d\hat{y}_1}{dh_{12}} = W_2^{(2)}$ $\quad \dfrac{d\hat{y}_2}{dh_{21}} = W_1^{(1)}$ and $\dfrac{d\hat{y}_2}{dh_{22}} = W_2^{(2)}$

Then $\dfrac{d\hat{y}}{dH} = \begin{bmatrix} d\hat{y}_1/dh_{11} & d\hat{y}_1/dh_{12} \\ d\hat{y}_2/dh_{21} & d\hat{y}_2/dh_{22} \end{bmatrix} = \begin{bmatrix} W_1^{(2)} & W_2^{(2)} \\ W_1^{(2)} & W_2^{(2} \end{bmatrix}$

$$\frac{dL}{dH} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dH} = \begin{bmatrix} \frac{dL}{d\hat{y}_1} \cdot \frac{d\hat{y}_1}{dh_{11}} & \frac{dL}{d\hat{y}_1} \cdot \frac{d\hat{y}_1}{dh_{12}} \\ \frac{dL}{d\hat{y}_2} \cdot \frac{d\hat{y}_2}{dh_{21}} & \frac{dL}{d\hat{y}_2} \cdot \frac{d\hat{y}_2}{dh_{22}} \end{bmatrix}$$

$$\overset{2\times 2}{= \begin{bmatrix} \frac{dL}{d\hat{y}_1} \cdot W_1^{(2)} & \frac{dL}{d\hat{y}_1} W_2^{(2)} \\ \frac{dL}{d\hat{y}_2} W_1^{(2)} & \frac{dL}{d\hat{y}_2} W_2^{(2)} \end{bmatrix}} = \underset{2\times 1}{\begin{bmatrix} dL/d\hat{y}_1 \\ dL/d\hat{y}_2 \end{bmatrix}} \underset{1\times 2}{\begin{bmatrix} W_1^{(2)} & W_2^{(2)} \end{bmatrix}}$$

$$= \frac{dL}{d\hat{y}} \cdot W^T$$

$$\therefore \frac{dL}{dw^{(1)}} = \underbrace{\frac{dL}{d\hat{y}}}_{\text{Known}} \cdot \frac{d\hat{y}}{dH} \cdot \underbrace{\frac{dH}{dw^{(1)}}}_{\text{we know the formula}}$$

$$\therefore \frac{dL}{dw^{(1)}} = X^T \frac{dL}{d\hat{y}} W^T$$

Similarly $\frac{dL}{db^{(1)}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{dH} \cdot \cancel{\frac{dH}{db^{(1)}}}$ always 1

$$\frac{dL}{db^{(1)}} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}_i} \cdot \frac{d\hat{y}_i}{dH}$$

So all the formulas we need are.

if $\hat{y} = XW + b$ and $L$ is some loss.

$$\boxed{\frac{dL}{dW} = X^T \frac{dL}{d\hat{y}} \qquad \frac{dL}{db} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}_i}}$$

and for backprop on multiple stacked layers,

we need $\boxed{\frac{dL}{dX} = \frac{dL}{d\hat{y}} W^T}$

# Adding an Activation Function

All we did so far is stack 2 linear layers, but what if we also add an activation function?

$$\hat{y} = \sigma\left(\left[XW^{(1)} + b^{(1)}\right]\right)W^{(2)} + b^{(2)}$$

So our composition of functions now is:

$$H = XW^{(1)} + b^{(1)}$$
$$S = \sigma(H)$$
$$\hat{y} = SW^{(2)} + b^{(2)}$$

First, find gradients on $W^{(2)}$ and $b^{(2)}$

$$\frac{dL}{dW^{(2)}} = S^T \frac{dL}{d\hat{y}} \qquad \frac{dL}{db^{(2)}} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}_i}$$

$$\frac{dL}{dS} = \frac{dL}{d\hat{y}} \cdot W^{(2)T}$$

now we need to backpropagate but through the activation function.

new derivative for sigmoid

$$\frac{dL}{dW^{(1)}} = \frac{dL}{dS} \cdot \frac{dS}{dH} \cdot \frac{dH}{dW^{(1)}}$$

$$= \frac{dL}{d\hat{y}} W^{(2)T} \underbrace{\left( \sigma(H)(1-\sigma(H)) \right)}_{(\sigma(x))' = \sigma(x)(1-\sigma(x))} \frac{dH}{dW^{(1)}}$$

$$= X^T \frac{dL}{d\hat{y}} W^{(2)T} \left( \sigma(H)(1-\sigma(H)) \right)$$

$$\frac{dL}{db^{(2)}} = \sum_{i=1}^{N} \frac{dL}{d\hat{y}} W^{(2)T} \left( \sigma(H)(1-\sigma(H)) \right)$$

## Summary!

At every linear layer compute

$\frac{dL}{dW}$, $\frac{dL}{db}$ and use them to update weights and intercept.

also compute $\frac{dL}{dX}$ as it is needed for chain rule for more previous layers.