1 Backprop

The main neural network equations are as follows:

$$\begin{aligned} z^l &= w^l a^{l-1} + b^l \\ a^l &= \sigma(z^l) \\ C &= \frac{1}{2}\|a^l - y\|^2 \end{aligned} \tag{1}$$

The equations for backprop are as follows. The idea is to see the change in cost function wrt weights and biases. Below equations use the notations of 3Blue1Brown series. The changes are given by:

$$\begin{aligned} \frac{\partial C}{\partial w^l} &= \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l}\frac{\partial z^l}{\partial w^l} \\ \frac{\partial C}{\partial b^l} &= \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l}\frac{\partial z^l}{\partial b^l} \end{aligned} \tag{2}$$

Now, $\frac{\partial C}{\partial a^l} = a^l - y$ (assuming quadratic cost function). And $\frac{\partial a^l}{\partial z^l} = \sigma'(z^l)$. Also, $\frac{\partial z^l}{\partial w^l} = a^{l-1}$, while $\frac{\partial z^l}{\partial b^l} = 1$.

Rewriting the equations acc. to the notations in the NN&DL Book Chapter 2.

Consider

$$\begin{aligned} \delta^l &= \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l} \\ &= \frac{\partial C}{\partial a^l} \odot \sigma'(z^l) \end{aligned} \tag{3}$$

The equation for the error δ^l wrt error of the next layer is

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \tag{4}$$

This back-propagates the gradient values through the network. We use the equations in (2) to calculate changes in cost function. The change in cost function wrt bias is given by

$$\begin{aligned} \frac{\partial C}{\partial b^l} &= \frac{\partial C}{\partial a^l}\frac{\partial a^l}{\partial z^l} \\ &= \delta^l (from 3) \end{aligned} \tag{5}$$

Since $\frac{\partial z^l}{\partial b^l} = 1$. Similarly calculating change in cost function wrt weights:

$$\frac{\partial C}{\partial w^l} = a^{l-1}\delta^l \tag{6}$$

Since $\frac{\partial z^l}{\partial w^l} = a^{l-1}$.

(Proof for eqn 4 to be done later.)

Backprop algo:

- Input set of training exs
- For each training ex:
  - Feedfwd: calculate z, a
  - Error: calculate δ using (3)
  - Backprop: For each of the steps backwards, calculate δ using (4)
- Gradient descent: for each of the steps backwards, update:

$$w^l = w^l - \frac{\eta}{m}\delta^l(a^{l-1})^T$$
$$b^l = b^l - \frac{\eta}{m}\delta^l$$