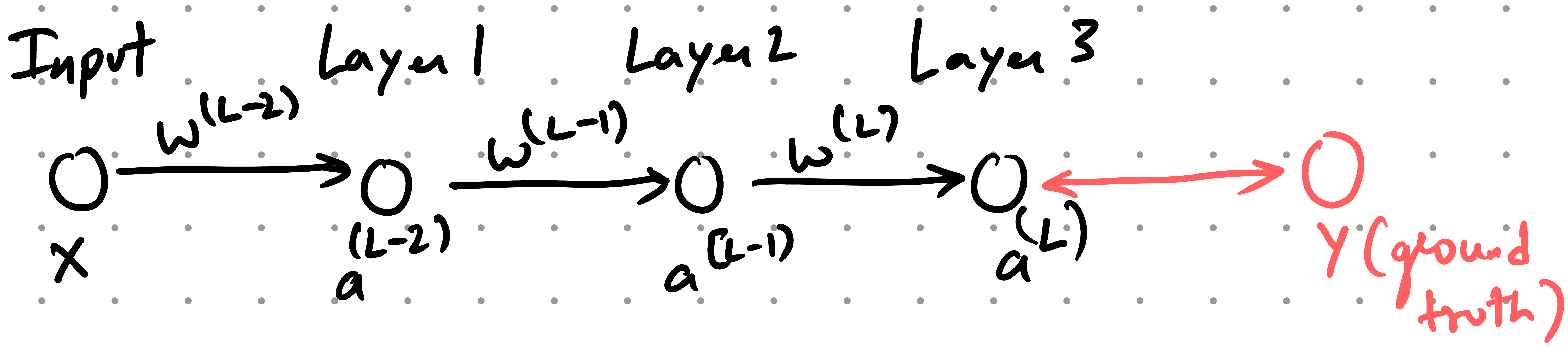
Backpropagation :

→ Assume a neural network with 3 layers and each layer has a single neuron.



→ Now, lets consider just the last layer :

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

$$C = (a^{(L)} - y)^2$$

(Least square cost)

→ Now, what are the parameters we can tweak to reduce the cost?

weight ($w^{(L)}$), bias ($b^{(L)}$) and previous layer's output ($a^{(L-1)}$)

→ So, now, the change in cost w.r.t the above parameters can be represented as : $\frac{\partial C}{\partial w^{(L)}}$, $\frac{\partial C}{\partial b^{(L)}}$, $\frac{\partial C}{\partial a^{(L-1)}}$

→ Now, using the chain rule,

$$\frac{\partial C}{\partial w^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial w^{(L)}}$$

$$\frac{\partial C}{\partial b^{(L)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial b^{(L)}}$$

$$\frac{\partial C}{\partial a^{(L-1)}} = \frac{\partial C}{\partial a^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial z^{(L)}}{\partial a^{(L-1)}}$$

Equations:

$$C = (a^{(L)} - y)^2$$

$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$

→ By differentiation:

$$\frac{\partial C}{\partial a^{(L)}} = 2(a^{(L)} - y)$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = a^{(L-1)}$$

$$\frac{\partial z^{(L)}}{\partial b^{(L)}} = 1$$

$$\frac{\partial z^{(L)}}{\partial a^{(L-1)}} = w^{(L)}$$

→ Therefore,

$$\frac{\partial C}{\partial w^{(L)}} = 2(a^{(L)} - y) \cdot \sigma'(z^{(L)}) \cdot a^{(L-1)}$$

$$\frac{\partial C}{\partial b^{(L)}} = 2(a^{(L)} - y) \cdot \sigma'(z^{(L)})$$

$$\frac{\partial C}{\partial a^{(L-1)}} = 2(a^{(L)} - y) \cdot \sigma'(z^{(L)}) \cdot w^{(L)}$$

→ Similarly, you can get the change in cost (∂C) w.r.t other parameters

$$\frac{\partial C}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdot \frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \sigma'(z^{(L-1)}) \cdot a^{(L-2)}$$

$$\frac{\partial C}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdot \frac{\partial z^{(L-1)}}{\partial b^{(L-1)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \sigma'(z^{(L-1)})$$

$$\frac{\partial C}{\partial a^{(L-2)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \frac{\partial a^{(L-1)}}{\partial z^{(L-1)}} \cdot \frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} = \frac{\partial C}{\partial a^{(L-1)}} \cdot \sigma'(z^{(L-1)}) \cdot w^{(L-1)}$$

$$\frac{\partial C}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L-2)}} \cdot \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \cdot \frac{\partial z^{(L-2)}}{\partial w^{(L-2)}} = \frac{\partial C}{\partial a^{(L-2)}} \cdot \sigma'(z^{(L-2)}) \cdot a^{(L-2)}$$

$$\frac{\partial C}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L-2)}} \cdot \frac{\partial a^{(L-2)}}{\partial z^{(L-2)}} \cdot \frac{\partial z^{(L-2)}}{\partial b^{(L-2)}} = \frac{\partial C}{\partial a^{(L-2)}} \cdot \sigma'(z^{(L-2)})$$

Note: We do not worry about $\frac{\partial C}{\partial a^{(L-3)}}$ which is the input x as it is not a parameter that can be changed

→ For multiple samples of x, just take the average of the parameter changes across all the samples of x.