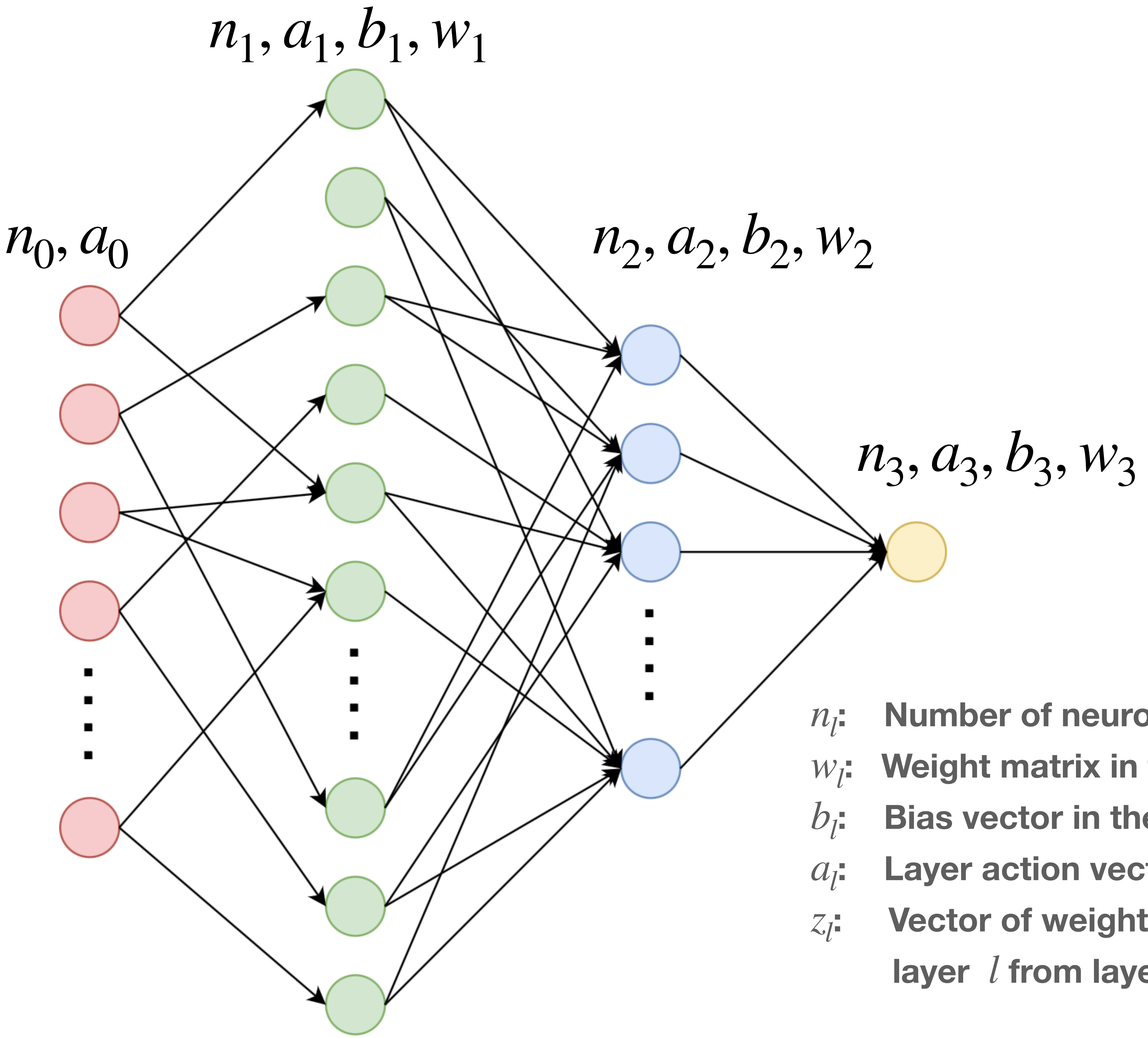
Gradient Descent and Backpropagation: Artificial Neural Networks

Perumadura De Silva

Necessary Formulas

- Reciprocal Rule:
  - $\frac{d}{dx}\left(\frac{1}{f(x)}\right) = -\frac{f'(x)}{f(x)^2}$
- Exponent rule:
  - $\frac{d(e^x)}{dx} = e^x$
- Function composition
  - $\frac{d[f(g(x))]}{dx} = \frac{df(g(x))}{dg}\frac{dg(x)}{dx}$
  - Hence Chain rule:
    - $\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}$



n_l: Number of neurones in layer l

w_l: Weight matrix in the layer l

b_l: Bias vector in the layer l

a_l: Layer action vector in layer l

z_l: Vector of weighted inputs to the layer l from layer $l-1$

Gradient Descent & Back prop

n :- Numbers of neurons

W :- layer weight matrix

b :- layer bias vector

a :- layer activation vector

z :- vector of weighted inputs from the ~~previolay~~ previous layer before activation

g :- layer activation function

↳ $\sigma(x) = \dfrac{1}{1+e^{-x}}$; sigmoid

C :- cost function ; ~~se~~ ; $\frac{1}{2}\sum(y-\hat{y})^2$ square error

$\hat{y} = h(x)$

* Backprop goal : Find the partial derivatives ~~w.r.t~~ of weighs & biases w.r.t prediction cost

* with partial derivatives

↳ $w^l_{k+1} = w^l_k - \alpha \dfrac{\partial C}{\partial w^l_k}$

↳ $b^l_{k+1} = b^l_k - \alpha \dfrac{\partial C}{\partial b^l_k}$

* Forward pass

$$\boxed{a^{(0)} = x}$$

$$z^1 = w^1 a^0 + b^1 \qquad \frac{\partial C}{\partial w^1}, \frac{\partial C}{\partial b^1}$$

$$\frac{\partial C}{\partial w^1} = \frac{\partial C}{\partial a^3}\frac{\partial a^3}{\partial z^3}\frac{\partial z^3}{\partial a^2}\frac{\partial a^2}{\partial z^2}\frac{\partial z^2}{\partial a^1}\frac{\partial a^1}{\partial z^1}\frac{\partial z^1}{\partial w^1}$$

$$\boxed{a^1 = \sigma(z^1)}$$

$$z^2 = w^2 a^1 + b^2 \qquad \frac{\partial C}{\partial w^2}, \frac{\partial C}{\partial b^2}$$

$$\boxed{a^2 = \sigma(z^2)}$$

$$z^3 = w^3 a^2 + b^3 \qquad \frac{\partial C}{\partial w^3}, \frac{\partial C}{\partial b^3}$$

$$\boxed{a^3 = \sigma(z^3) = \hat{y}}$$

$$C = \frac{1}{2}(y - \hat{y})^2$$; ~~Mea~~ Square error per sample

$$a(z(w^2, b^2, a(z(w^1, b^1, x))))$$

→ Because of the function composition property, the chain rule can be used to find the weigh, bias partial derivatives

* Partial derivatives required for back prop of gradients

1 $\dfrac{\partial C}{\partial z^L}$ $\qquad z^L \longrightarrow \underset{\sigma(z^L)}{O} \longrightarrow \overset{y \downarrow}{C}$

2 $\dfrac{\partial C}{\partial z^l}$ $\qquad z^l \longrightarrow \underset{\sigma(z^l)}{O} \cdots\cdots \underset{\sigma(z^L)}{O} \longrightarrow C$

3 $\dfrac{\partial z^l}{\partial w^l}$ $\qquad w^l a^{l-1} \longrightarrow z^l$

4 $\dfrac{\partial z^l}{\partial b^l}$ $\qquad b^l \longrightarrow z^l$

1. $$\frac{\partial C}{\partial z^L} = \frac{\partial C}{\partial a^L}\frac{\partial a^L}{\partial z^L} = \frac{\partial C}{\partial a^L} \odot \sigma'(z^L)$$

derivative of activation

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

$\hookrightarrow$ $$\frac{\partial a^L}{\partial z^L}; \quad a^L = \sigma(z^L) = \frac{1}{1+e^{-z^L}}$$

$$\frac{\partial a^L}{\partial z^L} = \frac{\partial \sigma(z^L)}{\partial z^L} = \frac{\partial}{\partial z^L}\left(\frac{1}{1+e^{-z^L}}\right)$$

* using Reciprocal rule,

$\hookrightarrow$ $$\frac{d}{dx}\left(\frac{1}{f(x)}\right) = -\frac{\frac{df(x)}{dx}}{f(x)^2}$$ | $$\frac{\partial a^L}{\partial z^L} = -\frac{\frac{\partial(1+e^{-z^L})}{\partial z^L}}{(1+e^{-z^L})^2}$$

$$= -\frac{1}{(1+e^{-z^L})^2}\frac{\partial(1+e^{-z^L})}{\partial z^L}$$

* Exponent rule, $\left|\frac{de^x}{dx} = e^x\right.$

$$\frac{\partial a^L}{\partial z^L} = -\frac{1}{(1+e^{-z^L})^2} e^{-z^L}\left(\frac{\partial -z^L}{\partial z^L}\right)$$

$$= \frac{e^{-z^L}}{(1+e^{-z^L})^2}$$

$$\hookrightarrow \frac{\partial a^L}{\partial z^L} = \frac{e^{-z^L} + 1 - 1}{(e^{-z^L} + 1)^2}$$

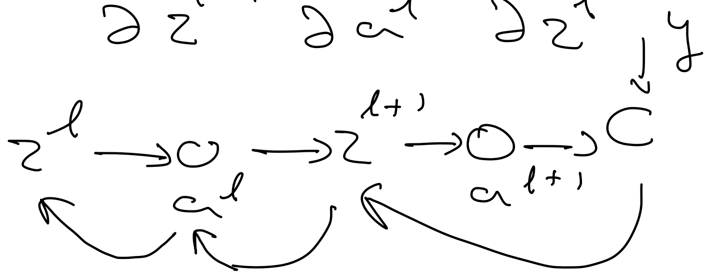
$$= \frac{1}{(1 + e^{-z^L})} \left[\frac{e^{-z^L} + 1}{e^{-z^L} + 1} - \frac{1}{e^{-z^L} + 1} \right]$$

$$= \sigma(z^L)\left[1 - \sigma(z^L)\right]$$

So $$\frac{\partial C}{\partial z^L} = -(y - \hat{y})\,\sigma(z^L)\left[1 - \sigma(z^L)\right]$$

$$= -(y - a^L)\,a^L\left[1 - a^L\right]$$

2. $$\frac{\partial C}{\partial z^l} = \frac{\partial C}{\partial z^{l+1}} \frac{\partial z^{l+1}}{\partial a^l} \frac{\partial a^l}{\partial z^l}$$



$$z^{l+1} = w^{l+1} a^l + b^{l+1}$$

$$\frac{\partial z^{l+1}}{\partial a^l} = w^{l+1}$$

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

$$\hookrightarrow \text{so } \frac{\partial C}{\partial z^l} = \left[(w^{l+1})^T \frac{\partial C}{\partial z^{l+1}} \right] \odot \sigma'(z^l)$$

$$3/4.\ \frac{\partial C}{\partial z^l} = \frac{\partial C}{\partial z^{l+1}} \frac{\partial z^{l+1}}{\partial a^l} \frac{\partial a^l}{\partial z^l} \left(\frac{\partial z^l}{\partial w^l} \Big/ \frac{\partial z^l}{\partial b^l} \right)$$

$$\hookrightarrow \frac{\partial z^l}{\partial w^l} = w^l a^{l-1} + b^l$$

$$= a^{l-1}$$

$$\hookrightarrow \frac{\partial z^l}{\partial b^l} = 1$$