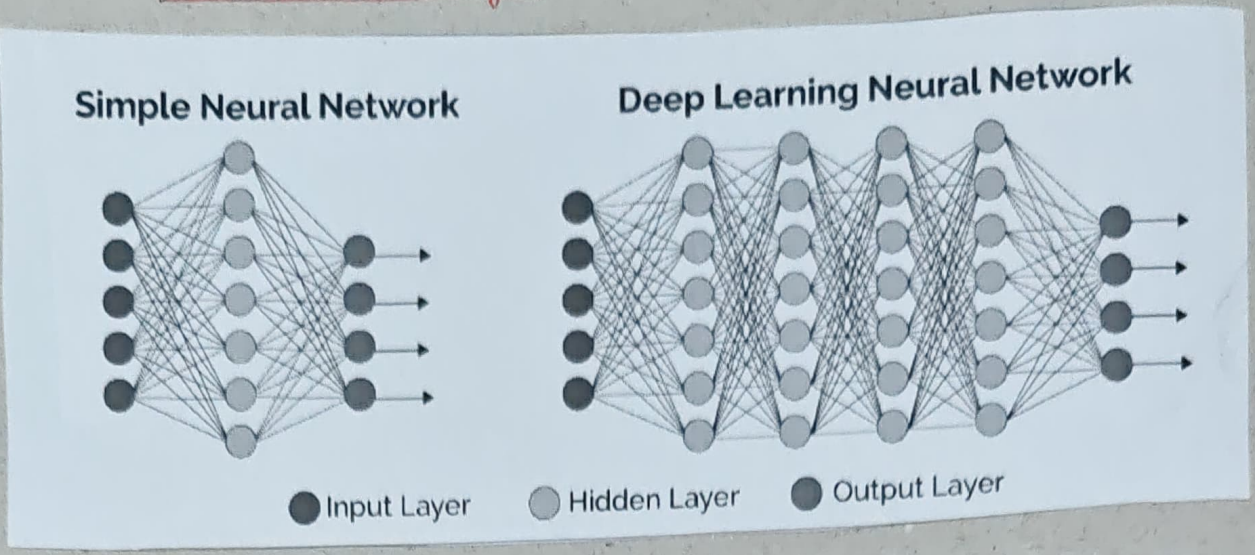
11. Deep Learning

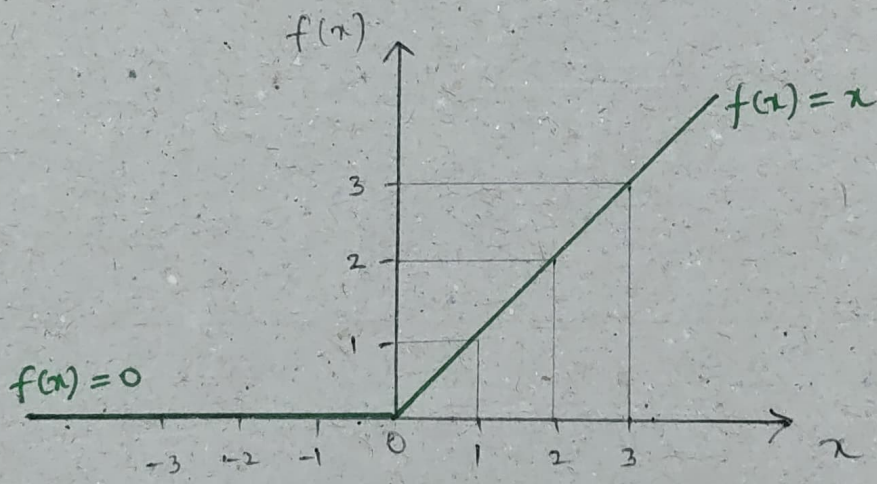


- Deep neural nets comprise of multiple layers of neurons, termed as Hidden layers.
- Different layers can employ different activation functions.

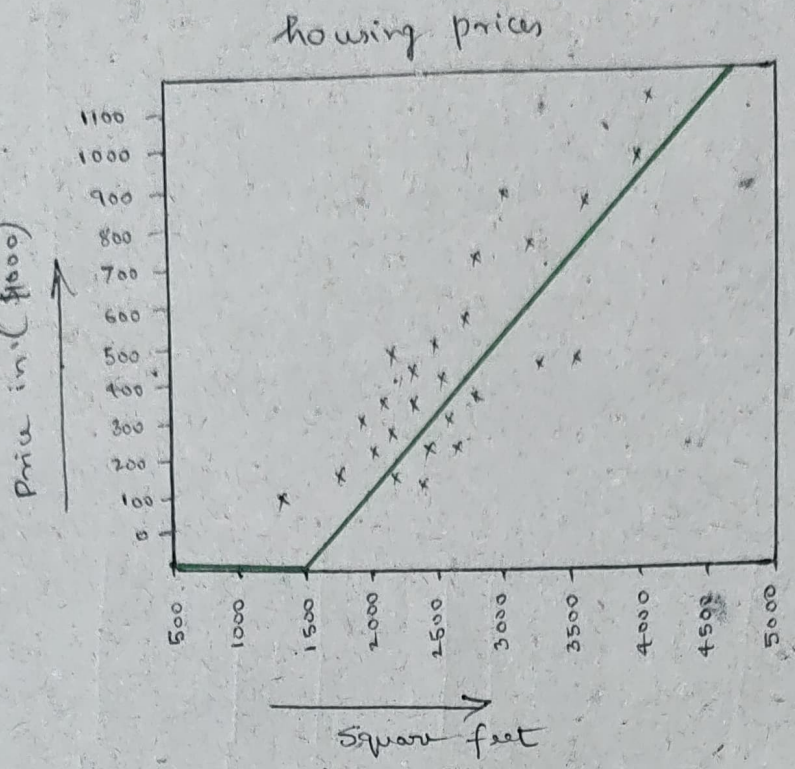
ReLU activation

- Rectified Linear Unit (ReLU) activation used by the inner layers.

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



⊙ Why ReLU ?



ReLU best captures non-linearities (i.e) kinks as shown in data.

⊙ Backpropagation

We derive the relevant expressions for backpropagation in a Deep NN.

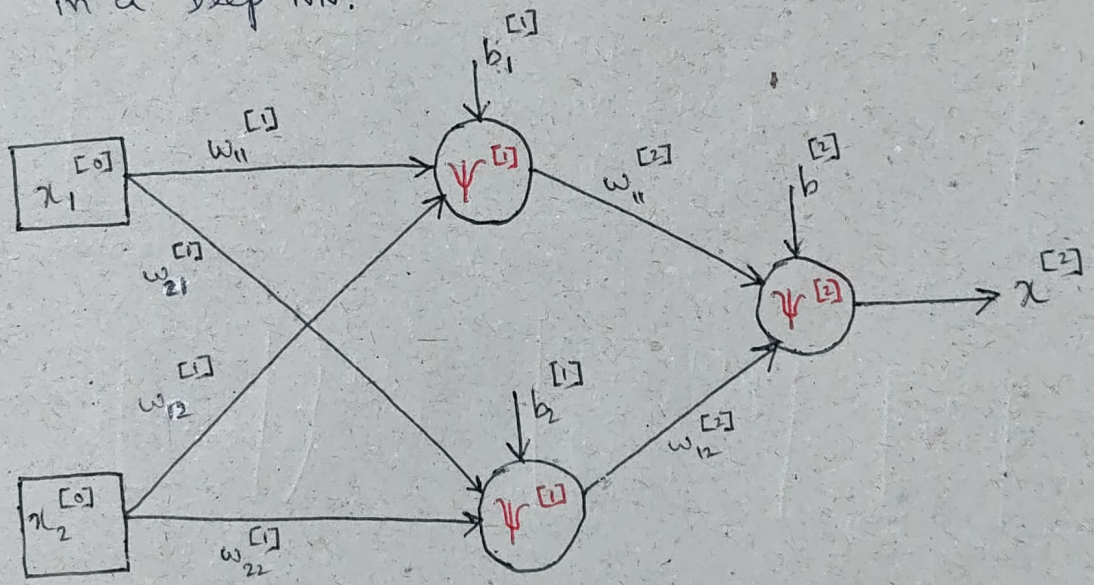


Fig: 2 Layer NN - Backpropagation.

Note :

(i) What does $w_{48}^{[5]}$ represent ?

Weight from 8^{th} neuron in Layer 4 to 4^{th} neuron in Layer 5

(ii) What does $b_2^{[3]}$ represent ?

Bias of 2^{nd} neuron of layer 3.

2 layer NN model

⊙ Input to Layer 1 is

$$\underbrace{\begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix}}_{\bar{z}^{[1]}} = \underbrace{\begin{bmatrix} w_{11}^{[1]} & w_{12}^{[1]} \\ w_{21}^{[1]} & w_{22}^{[1]} \end{bmatrix}}_{W^{[1]}} \underbrace{\begin{bmatrix} x_1^{[0]} \\ x_2^{[0]} \end{bmatrix}}_{\bar{x}^{[0]}} + \underbrace{\begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \end{bmatrix}}_{\bar{b}^{[1]}}$$

$\therefore \bar{Z}^{[1]}$

Note :

⊙ Size of $W^{[1]}$ is $m \times n$.

$m \rightarrow$ No. of Neurons in current layer

$n \rightarrow$ No. of Neurons in previous layer

⊙ Size of Bias Vector

$m \rightarrow$ No. of Neurons in current layer

⊙ Output of layer 1 is

Activation function of Layer 1

$$\underbrace{\begin{bmatrix} x_1^{[1]} \\ x_2^{[1]} \end{bmatrix}}_{\bar{x}^{[1]}} = \begin{bmatrix} \psi^{[1]}(z_1^{[1]}) \\ \psi^{[1]}(z_2^{[1]}) \end{bmatrix} = \psi^{[1]} \left(\underbrace{\begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix}}_{\bar{z}^{[1]}} \right)$$

O/p $\Rightarrow$ input to next node

⊙ In Layer 2 we have

$$z_1^{[2]} = \underbrace{\begin{bmatrix} w_{11}^{[2]} & w_{12}^{[2]} \end{bmatrix}}_{W^{[2]}} \underbrace{\begin{bmatrix} x_1^{[1]} \\ x_2^{[1]} \end{bmatrix}}_{\bar{x}^{[1]}} + b^{[2]}$$

$$x^{[2]} = \psi^{[2]}\left(z_1^{[2]}\right)$$

Note:
Size of $W^{[2]} = 1 \times 2$
(i) 1 neuron in current layer
2 neuron in previous layer.

Note:
What is the size of $W^{[3]}$ for the Deep Learning NN shown in image?

7×7

⊙ The Loss function is given as

$$L = \frac{1}{2}\left(y - x^{[2]}\right)^2 = \frac{1}{2}\left(y - \psi^{[2]}\left(z_1^{[2]}\right)\right)^2$$

⊙ Gradient for layer 2

The gradient w.r.t $W^{[2]}$ is given as

$$\frac{\partial}{\partial W^{[2]}} L = \underbrace{-\left(y - x^{[2]}\right) \times \left(\psi^{[2]}\right)'\left(z_1^{[2]}\right)}_{\phi^{[2]}} \times \left(\bar{x}^{[1]}\right)^T$$

$$= \phi^{[2]} \times \left(\bar{x}^{[1]}\right)^T$$

$$\frac{\partial}{\partial b^{[2]}} L = \phi^{[2]}$$

⊙ Gradient for layer 1

The gradient w.r.t $W^{[1]}$ can be evaluated as follows.

$$\frac{\partial}{\partial W^{[1]}} L = \left((W^{[2]})^T \phi^{[2]} \odot (\psi^{[1]})' (\bar{z}^{[1]}) \right) \times (\bar{x}^{[0]})^T$$

$$\frac{\partial}{\partial b^{[1]}} L = (W^{[2]})^T \phi^{[2]} \odot (\psi^{[1]})' (\bar{z}^{[1]})$$

(Detailed derivation in Appendix).

Problem:

What is the size of

$$\underbrace{(W^{[2]})^T}_{2 \times 1} \underbrace{\phi^{[2]}}_{1 \times 1} \odot \underbrace{(\psi^{[1]})' (\bar{z}^{[1]})}_{2 \times 1}$$

Solution:

Size : 2×1

$\odot \rightarrow$ Element-by-element Multiplication.

Appendix (2 layer NN model)

⊙ Input to layer 1 is

$$\underbrace{\begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix}}_{\bar{z}^{[1]}} = \underbrace{\begin{bmatrix} w_{11}^{[1]} & w_{12}^{[1]} \\ w_{21}^{[1]} & w_{22}^{[1]} \end{bmatrix}}_{W^{[1]}} \underbrace{\begin{bmatrix} x_1^{[0]} \\ x_2^{[0]} \end{bmatrix}}_{\bar{x}^{[0]}} + \underbrace{\begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \end{bmatrix}}_{\bar{b}^{[1]}}$$

⊙ Output of layer 1 is

$$\underbrace{\begin{bmatrix} x_1^{[1]} \\ x_2^{[1]} \end{bmatrix}}_{\bar{x}^{[1]}} = \psi^{[1]} \left(\underbrace{\begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix}}_{\bar{z}^{[1]}} \right)$$

$$= \psi^{[1]} \left(\begin{bmatrix} w_{11}^{[1]} x_1^{[0]} + w_{12}^{[1]} x_2^{[0]} + b_1^{[1]} \\ w_{21}^{[1]} x_1^{[0]} + w_{22}^{[1]} x_2^{[0]} + b_2^{[1]} \end{bmatrix} \right)$$

⊙ In layer 2, we have

$$z_1^{[2]} = \underbrace{\begin{bmatrix} w_{11}^{[2]} & w_{12}^{[2]} \end{bmatrix}}_{W^{[2]}} \underbrace{\begin{bmatrix} x_1^{[1]} \\ x_2^{[1]} \end{bmatrix}}_{\bar{x}^{[1]}} + b^{[2]}$$

$$x_1^{[2]} = \psi^{[2]} \left(z_1^{[2]} \right)$$

⊙ The <u>Loss function</u> is given as

$$L = \frac{1}{2} \left(y - x^{[2]} \right)^2 = \frac{1}{2} \left(y - \psi^{[2]} \left(z_1^{[2]} \right) \right)^2$$

$$= \frac{1}{2} \left(y - \psi^{[2]} \left(w_{11}^{[2]} x_1^{[1]} + w_{12}^{[2]} x_2^{[1]} + b^{[2]} \right) \right)^2$$

⊙ Gradient for Layer 2

* The gradient w.r.t $W^{[2]}$ is given as

$$L = \frac{1}{2}\left(y - \psi^{[2]}\left(w_{11}^{[2]} x_1^{[1]} + w_{12}^{[2]} x_2^{[1]} + b^{[2]}\right)\right)^2$$

$$\frac{\partial}{\partial W^{[2]}} = -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]}) \begin{bmatrix} x_1^{[1]} & x_2^{[1]} \end{bmatrix}$$

$$= -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]}) \times (\bar{x}^{[1]})^T$$

$$= \phi^{[2]} \times (\bar{x}^{[1]})^T$$

* The gradient w.r.t $b^{[2]}$ is given as

$$L = \frac{1}{2}\left(y - \psi^{[2]}\left(w_{11}^{[2]} x_1^{[1]} + w_{12}^{[2]} x_2^{[1]} + b^{[2]}\right)\right)^2$$

$$\frac{\partial}{\partial b^{[2]}} = -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]})$$

$$= \phi^{[2]}$$

⊙ Gradient for Layer 1

*. The gradient w.r.t $W^{[1]}$ can be evaluated as follows.

$$\psi^{[2]}\left(w_{11}^{[2]} x_1^{[1]} + w_{12}^{[2]} x_2^{[1]} + b^{[2]}\right)$$

$$= \psi^{[2]}\left(w_{11}^{[2]} \psi^{[1]}\left(w_{11}^{[1]} x_1^{[0]} + w_{12}^{[1]} x_2^{[0]} + b_1^{[1]}\right)\right.$$

$$\left. + w_{12}^{[2]} \psi^{[1]}\left(w_{21}^{[1]} x_1^{[0]} + w_{22}^{[1]} x_2^{[0]} + b_2^{[1]}\right) + b^{[2]}\right)$$

$$\frac{\partial}{\partial W^{[1]}} = -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]}) \times \begin{bmatrix} w_{11}^{[2]} (\psi^{[1]})'(z_1^{[1]}) \\ w_{12}^{[2]} (\psi^{[1]})'(z_2^{[1]}) \end{bmatrix}$$

$$\times \begin{bmatrix} x_1^{[0]} & x_2^{[0]} \end{bmatrix}$$

$$= -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]}) \times \left((W^{[2]})^T \odot (\psi^{[1]})'(\bar{z}^{[1]}) \right) \times (\bar{x}^{[0]})^T$$

* Observe that, this can be written as

$$\frac{\partial}{\partial W^{[1]}} = -(y - x^{[2]}) \times (\psi^{[2]})'(z_1^{[2]}) \times \left((W^{[2]})^T \odot (\psi^{[1]})'(\bar{z}^{[1]}) \right) \times (\bar{x}^{[0]})^T$$

$$= \left((W^{[2]})^T \phi^{[2]} \odot (\psi^{[1]})'(\bar{z}^{[1]}) \right) \times (\bar{x}^{[0]})^T$$

* Furthermore,

$$\frac{\partial}{\partial \bar{b}^{[1]}} = \left((W^{[2]})^T \phi^{[2]} \odot (\psi^{[1]})'(\bar{z}^{[1]}) \right)$$

$$\underline{\qquad\qquad} \times \underline{\qquad\qquad}$$

Multi-layer Neural nets

Consider a general K layer Neural network. The NN is mathematically modeled as

$$\bar{x}^{[1]} = \psi^{[1]}\left(W^{[1]}\bar{x}^{[0]} + \bar{b}^{[1]}\right)$$

$$\bar{x}^{[2]} = \psi^{[2]}\left(W^{[2]}\bar{x}^{[1]} + \bar{b}^{[2]}\right)$$

$$\vdots$$

$$\bar{x}^{[K]} = \psi^{[K]}\left(W^{[K]}\bar{x}^{[K-1]} + \bar{b}^{[K]}\right)$$

Back propagation rules

⊙ The Loss function for this NN is defined as

$$L = \frac{1}{2}\left(y - x^{[K]}\right)^2$$

($x^{[K]}$: Actual o/p of layer K; y: desired o/p)

⊙ Start with $k = K$

$$\phi^{[k]} = -\left(y - x^{[k]}\right) \times \left(\psi^{[k]}\right)'\left(z_1^{[k]}\right)$$

$$\frac{\partial}{\partial W^{[k]}} L = \phi^{[k]} \times \left(\bar{x}^{[k-1]}\right)^T$$

$$\frac{\partial}{\partial b^{[k]}} L = \phi^{[k]}$$

⊙ Now decrement $k = k - 1$

⊙ Perform

$$\phi^{[k]} = \left(W^{[k+1]}\right)^T \cdot \phi^{[k+1]} \odot \left(\psi^{[k]}\right)'\left(\bar{z}^{[k]}\right)$$

$$\frac{\partial}{\partial W^{[k]}} L = \phi^{[k]} \times \left(\bar{x}^{[k-1]}\right)^T$$

$$\frac{\partial}{\partial \bar{b}^{[k]}} L = \phi^{[k]}$$

⊙ Repeat till $k = 1$.

Problem: What is the size of $\phi^{[k]}$?

$$\frac{\partial}{\partial W^{[k]}} L = \phi^{[k]} \times \left(\bar{x}^{[k-1]}\right)^T$$

→ $\left(\bar{x}^{[k-1]}\right)^T$ is a row vector. Hence $\phi^{[k]}$ is a column vector.

→ Size equals number of rows in $W^{[k]}$.

(i) No. of neurons in layer k.