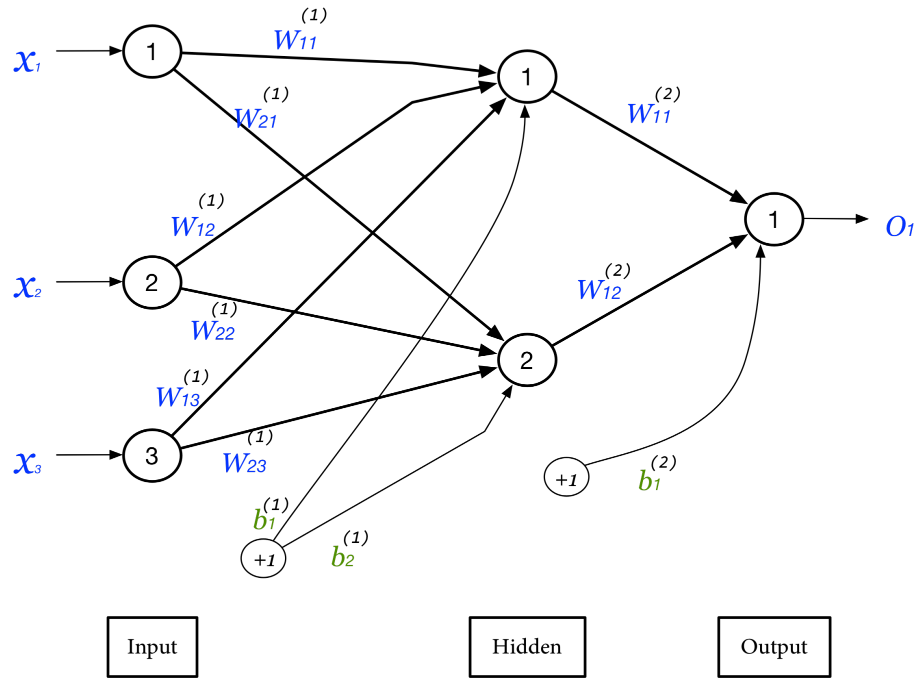
# TD1 - Back Propagation Algorithm

## 1 The algorithm

A Back Propagation network learns by example. Back Propagation is the algorithm used to train the network by changing the network's weights in order to minimize a predefined loss function.

## 2 Working example

In this section we show all the calculations for a full sized network with 3 inputs, 2 hidden layer neurons and 1 output neuron as shown in fig. 3.2.



Assume that the Sigmoid activation function is applied to hidden layer and output layer. Assume that the network has the following initial weights ($\eta$ is the learning rate)

$$\eta = \begin{pmatrix} 0.9 \end{pmatrix}$$

$$w1 = \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} = \begin{pmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \\ -0.5 & 0.2 \end{pmatrix}$$

$$w2 = \begin{pmatrix} w_{11} \\ w_{12} \end{pmatrix} = \begin{pmatrix} -0.3 \\ -0.2 \end{pmatrix}$$

$$b1 = \begin{pmatrix} b_1 & b_2 \end{pmatrix} = \begin{pmatrix} -0.4 & 0.2 \end{pmatrix}$$

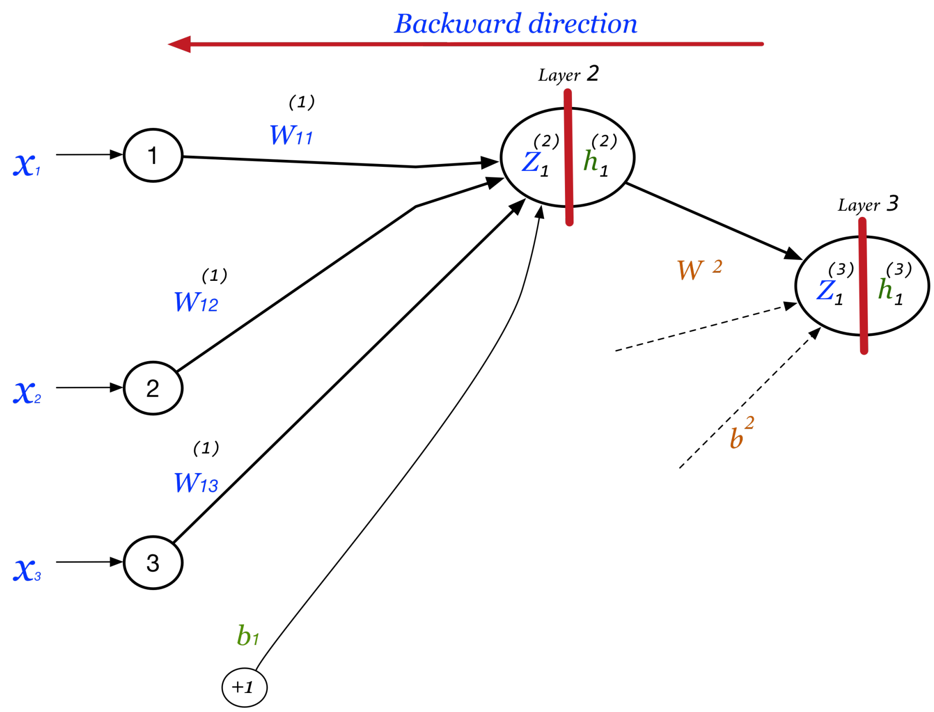
$$b2 = \begin{pmatrix} b_2 \end{pmatrix} = \begin{pmatrix} 0.1 \end{pmatrix}$$

Using the following first training example x = $[1\ ;\ 0\ ;\ 1]^T$ whose class label $y = 1$

1. Perform a forward pass on the network.
2. Perform a reverse pass (training) once using $(x_1; y_1)$ .
3. Perform a further forward pass
4. Comment on the result.

# 3 Solution

## 3.1 Forward pass




**Input Layer 1**

**Matrix based operation**

$\Rightarrow$ **General formula** : $Z^{l+1} = W^l h^l + b^l$

$\Rightarrow$ First layer : $Z^2 = W^1 h^1 + b^1$ with $h^1 = X$ for first hidden layer

$\Rightarrow$ Second layer : $Z^3 = W^2 h^2 + b^2$ with $h^2 = f(Z^2)$

$$Z^2 = X^T W^1 + b^1 = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} + \begin{pmatrix} b_1 & b_2 \end{pmatrix}$$

$$Z^2 = X^T W^1 + b^1 = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.2 & -0.3 \\ 0.4 & 0.1 \\ -0.5 & 0.2 \end{pmatrix} + \begin{pmatrix} -0.4 & 0.2 \end{pmatrix} = \begin{pmatrix} -0.7 & 0.1 \end{pmatrix}$$

**Output Layer 1**

Apply activation function $\sigma(.)$ foreach node in the hidden layer. For matrix based operation we can apply the activation function to each element of the matrix.

$\Rightarrow \sigma(Z^2) = \sigma(X^T W^1 + b^1)$

$$h^2 = \sigma(Z^2) = \sigma(\begin{pmatrix} -0.7 & 0.1 \end{pmatrix}) = \begin{pmatrix} 0.332 & 0.525 \end{pmatrix}$$

**Input layer2** (this is the output-layer in this example)

$$Z^3 = h^2 W^2 + b^2 = \begin{pmatrix} h_1^2 & h_2^2 \end{pmatrix} \begin{pmatrix} w_{11} \\ w_{12} \end{pmatrix} + \begin{pmatrix} b_1 \end{pmatrix}$$

$$Z^3 = \begin{pmatrix} 0.332 & 0.525 \end{pmatrix} \begin{pmatrix} -0.3 \\ -0.2 \end{pmatrix} + \begin{pmatrix} 0.1 \end{pmatrix} = -0.105$$

$$h^3 = \sigma(Z^3) = \sigma(\begin{pmatrix} -0.105 \end{pmatrix}) = \begin{pmatrix} \mathbf{0.474} \end{pmatrix}$$

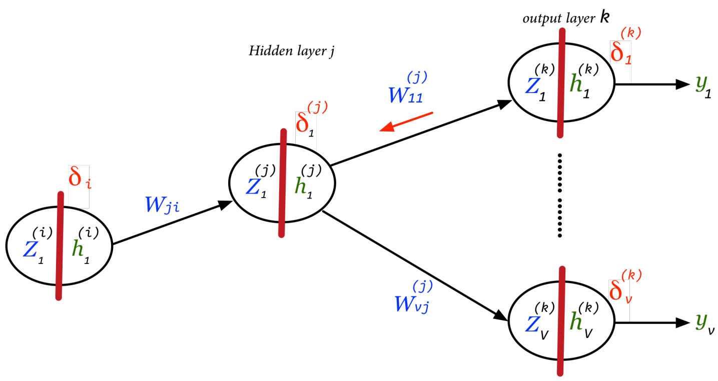
This is to compare to the target ($\mathbf{y = 1}$)

output error : $E = \frac{1}{2}(0.474 - 1)^2 = 0.138338$

## 3.2 Backward pass

### Backward for output Layer

The output neuron $m$ of the output layer $k$ is $y_m = h_m^{(k)}$




$$\frac{\partial E}{\partial w_{11}^{(j)}} = \frac{\partial E}{\partial h_1^{(k)}} \ \frac{\partial h_1^{(k)}}{\partial Z_1^{(k)}} \ \frac{\partial Z_1^{(k)}}{\partial w_{11}^{(j)}}$$

$$\frac{\partial Z_1^{(k)}}{\partial w_{11}^{(j)}} = h_1^{(j)}$$

$$\frac{\partial E}{\partial h_1^{(k)}} \ \frac{\partial h_1^{(k)}}{\partial Z_1^{(k)}} \ = (h_1^{(k)} - t_1) \ \ \sigma'(Z_1^{(j)})$$

$$\frac{\partial h_1^{(k)}}{\partial Z_1^{(k)}} = \sigma'(Z_1^{(j)})$$

$$\frac{\partial E}{\partial h_1^{(k)}} = (h_1^{(k)} - t_1) \text{ Since } E = \frac{1}{2}\sum_k (h_1^{(k)} - t_k)^2 \rightarrow \frac{\partial E}{\partial h_1^{(k)}} = h_1^{(k)} - t_k$$

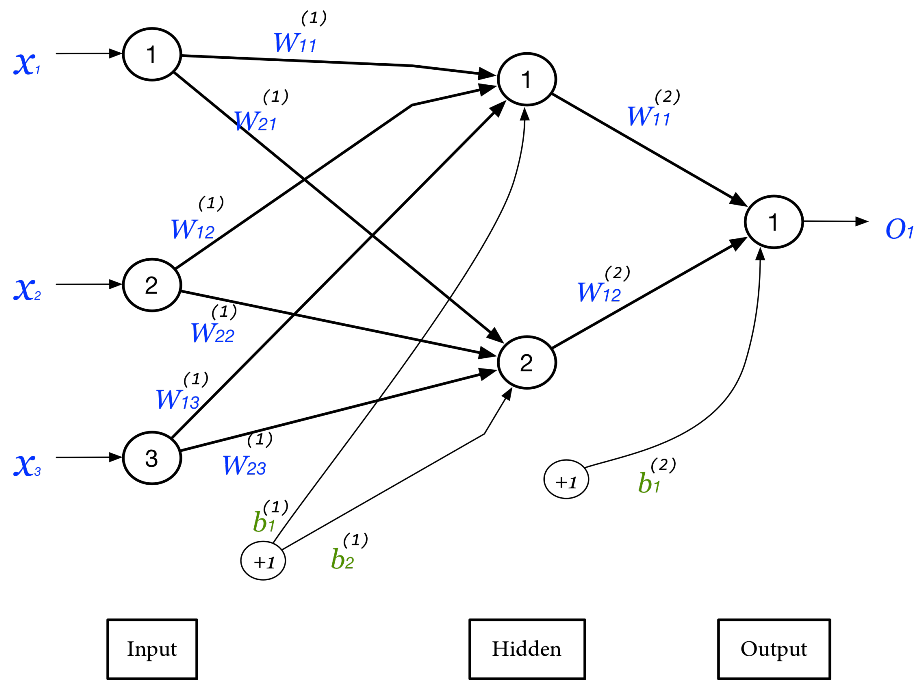
Let $\delta_1^{(k)} = \frac{\partial E}{\partial h_1^{(k)}} \ \frac{\partial h_1^{(k)}}{\partial Z_1^{(k)}} = (h_1^{(k)} - t_1) \ \ \sigma'(Z_1^{(k)})$

$$\frac{\partial E}{\partial w_{11}^{(j)}} = \delta_1^{(k)} h_1^{(j)}$$

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = h_j^{(l)} \delta_i^{(l+1)} \quad with \quad \delta_i^{(l)} = \sigma'(Z_i^{(l)})(h_i^{(l)} - t_i)$$

$$\frac{\partial E}{\partial b_i^{(l)}} = \delta_i^{(l+1)}$$

For our network :




Output Layer :

$$\frac{\partial E}{\partial w_{ij}^{(l)}} = h_j^{(l)} \delta_i^{(l+1)} \quad with \quad \delta_i^{(l)} = \sigma^{'}(Z_i^{(l)})(h_i^{(l)} - t_i)$$

$k = 3$ in our example

$\frac{\partial MSE}{\partial w_{11}^{(2)}} = h_1^{(2)} \delta_1^{(3)}\ with\ \ \delta_1^{(3)} = \sigma^{'}(Z_1^{(3)})(h_1^{(3)} - t_1)$

$h_1^{(3)} = y_1 = 0.474\ \ ;\ \ t_1 = 1;\ \ Z_1^{(3)} = -0.105\ \ h_1^{(2)} = 0.332$

$\delta_1^{(3)} = \sigma(-0.105)(1 - \sigma(-0.105))(0.474 - 1)$

$$\delta_1^{(3)} = (0.474)(1 - 0.474)(-0.526) = -0.1311$$

$$\frac{\partial MSE}{\partial w_{11}^{(2)}} = (-0.1311)(0.332) = \mathbf{-0.0435252}$$

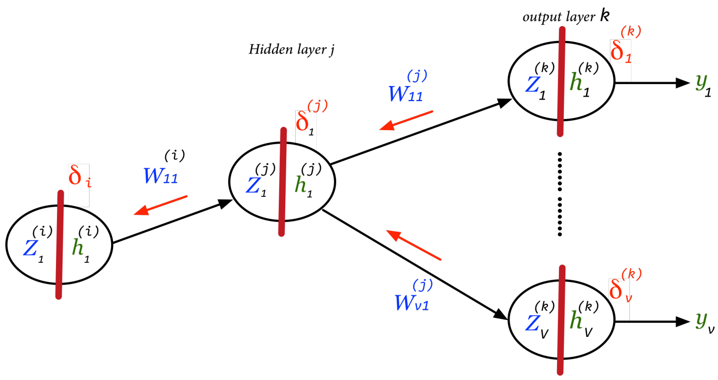
$$\frac{\partial MSE}{\partial w_{12}^{(2)}} = \delta_1^{(3)} h_2^{(2)} \ with \ \ \delta_1^{(3)} = \sigma'(Z_1^{(3)})(h_1^{(3)} - t_1)$$

$$h_2^{(2)} = 0.525$$

$$\frac{\partial MSE}{\partial w_{12}^{(2)}} = (-0.1311)(0.525) = \mathbf{-0.0688275}$$

$$\begin{pmatrix} \frac{\partial MSE}{\partial w_{12}^{(2)}} \\ \\ \frac{\partial MSE}{\partial w_{12}^{(2)}} \end{pmatrix} = \begin{pmatrix} \mathbf{-0.0435252} \\ \\ \mathbf{-0.0688275} \end{pmatrix}$$

# Backward for hidden Layer




$$\frac{\partial E}{\partial w_{ij}^{(l)}} = h_j^{(l)} \delta_i^{(l+1)}$$

$$\delta_j^{(l)} = \sum_{i=1}^{s_{(l+1)}} w_{ij}^{(l)} \quad \delta_i^{(l+1)} \quad \sigma'(Z_j^{(l)})$$

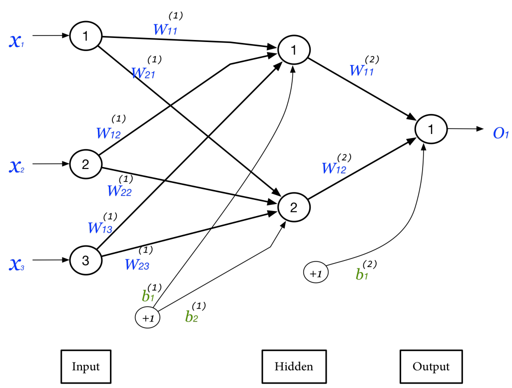
$$\delta_j^{(l)} = \sigma'(Z_j^{(l)}) \sum_{i=1}^{s_{(l+1)}} w_{ij}^{(l)} \quad \delta_i^{(l+1)}$$

$$\frac{\partial MSE}{\partial w_{11}^{(i)}} = h_1^{(i)} \delta_1^{(j)}$$

$$\delta_1^{(j)} = \sigma'(Z_1^{(j)}) \sum_{i=1}^{(k)} w_{i1}^{(j)} \quad \delta_i^{(k)}$$

$$\boxed{\begin{aligned} &\frac{\partial E}{\partial w_{ij}^{(l)}} = h_j^{(l)} \delta_i^{(l+1)} \quad with \quad \delta_j^{(l)} = \sigma'(Z_j^{(l)}) \sum_{i=1}^{s_{(l+1)}} w_{ij}^{(l)} \quad \delta_i^{(l+1)} \\ &\qquad \frac{\partial E}{\partial b_i^{(l)}} = \delta_i^{(l+1)} \end{aligned}}$$

For our example we have only one hidden layer :




$$\frac{\partial MSE}{\partial w_{11}^{(1)}} = h_1^{(1)}\delta_1^{(2)}$$

$$\delta_1^{(2)} = \sigma^{'}(Z_1^{(2)})w_{11}^{(2)} \quad \delta_1^{(3)}$$

$$Z^2 = X^T W^1 + b^1 = \begin{pmatrix} -0.7 & 0.1 \end{pmatrix}$$

$$\delta_1^{(3)} = (0.474)(1 - 0.474)(-0.526) = -0.1311$$

$$\sigma^{'}(Z_1^{(2)}) = \sigma(Z_1^{(2)})(1 - \sigma(Z_1^{(2)})) = \sigma(-0.7)(1 - \sigma(-0.7)$$

$$\sigma^{'}(Z_1^{(2)}) = (0.332)(1 - 0.332)$$

$$w_{11}^{(2)} = -0.3$$

$$\delta_1^{(2)} = (0.332)(1 - 0.332)(-0.1311)(-0.3) = 0.00872245008$$

$$\frac{\partial MSE}{\partial w_{11}^{(1)}} = h_1^{(1)} * 0.00872245008$$

$$\frac{\partial MSE}{\partial w_{21}^{(1)}} = h_1^{(1)}\delta_2^{(2)} \quad ; \quad \delta_2^{(2)} = \sigma^{'}(Z_2^{(2)})w_{12}^{(2)} \quad \delta_1^{(3)}$$

$$\delta_2^{(2)} = \sigma(0.1)(1 - \sigma(0.1))(-0.1311)(-0.2) = 0.0065386125$$

$\frac{\partial MSE}{\partial w_{11}^{(1)}} = (h_1^{(1)}) * \delta_1^{(2)} = 1 * (0.00872245008)$

$\frac{\partial MSE}{\partial w_{21}^{(1)}} = (h_1^{(1)}) * \delta_2^{(2)} = 1 * (0.0065386125)$

$\frac{\partial MSE}{\partial w_{12}^{(1)}} = (h_2^{(1)}) * \delta_1^{(2)} = 0 * (0.00872245008)$

$\frac{\partial MSE}{\partial w_{22}^{(1)}} = (h_2^{(1)}) * \delta_2^{(2)} = 0 * (0.0065386125)$

$\frac{\partial MSE}{\partial w_{13}^{(1)}} = (h_3^{(1)}) * \delta_1^{(2)} = 1 * (0.00872245008)$

$\frac{\partial MSE}{\partial w_{23}^{(1)}} = (h_3^{(1)}) * \delta_2^{(2)} = 1 * (0.0065386125)$

Weights update :

$W^{t+1} = W^t - \eta \nabla E(W^t)$

$w_{11}^{(1)} = w_{11}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{11}^{(1)}} = 0.2 - 0.9 * 1 * 0.00872245008 = 0.19214979492$
$w_{21}^{(1)} = w_{21}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{21}^{(1)}} = -0.3 - 0.9 * 1 * 0.0065386125 = -0.30588475125$
$w_{12}^{(1)} = w_{12}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{12}^{(1)}} = 0.4 - 0.9 * 0 * (0.00872245008) = 0.4$
$w_{22}^{(1)} = w_{22}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{22}^{(1)}} = 0.1 - 0.9 * 0 * (0.0065386125) = 0.1$
$w_{13}^{(1)} = w_{13}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{13}^{(1)}} = -0.5 - 0.9 * 1 * (0.00872245008) = -0.50785020507$

$w_{23}^{(1)} = w_{23}^{(1)} - 0.9 * \frac{\partial MSE}{\partial w_{23}^{(1)}} = 0.2 - 0.9 * 1 * (0.0065386125) = 0.19411524875$

$w_{11}^{(2)} = w_{11}^{(2)} - 0.9 * \frac{\partial MSE}{\partial w_{11}^{(2)}} = -0.3 - 0.9 * (-0.0435252) = -0.26082732$

$w_{12}^{(2)} = w_{12}^{(2)} - 0.9 * \frac{\partial MSE}{\partial w_{12}^{(2)}} = -0.2 - 0.9 * (-0.0688275) = -0.13805525$

$b^{t+1} = b^t - \eta \nabla E(b^t)$

$\frac{\partial E}{\partial b_1^{(2)}} = \delta_1^{(3)} = -0.1311$

$b_1^{(2)} = b_1^{(2)} - 0.9 * \delta_1^{(3)} = 0.1 - 0.9 * (-0.1311) = 0.21799$

$\frac{\partial E}{\partial b_1^{(1)}} = \delta_1^{(2)}$

$b_1^{(1)} = b_1^{(1)} - \delta_1^{(2)} * 0.9 = -0.4 - 0.9 * 0.00872245008 = -0.40785020507$

$\frac{\partial E}{\partial b_2^{(1)}} = \delta_2^{(2)}$

$b_2^{(1)} = b_2^{(1)} - 0.9 * \delta_2^{(2)} = 0.2 - 0.9 * 0.0065386125 = 0.19411524875$

## 3.3 Second Forward pass

$$w1 = \begin{pmatrix} w_{11}^{(1)} & w_{21}^{(1)} \\ w_{12}^{(1)} & w_{22}^{(1)} \\ w_{13}^{(1)} & w_{23}^{(1)} \end{pmatrix} = \begin{pmatrix} 0.19214979492 & -0.30588475125 \\ 0.4 & 0.1 \\ -0.50785020507 & 0.19411524875 \end{pmatrix}$$

$$w2 = \begin{pmatrix} w_{11}^{(2)} \\ \\ w_{12}^{(2)} \end{pmatrix} = \begin{pmatrix} -0.26082732 \\ \\ -0.13805525 \end{pmatrix}$$

$$b1 = \begin{pmatrix} b_1^{(1)} & b_2^{(1)} \end{pmatrix} = \begin{pmatrix} -0.40785020507 & 0.19411524875 \end{pmatrix}$$

$$b2 = \begin{pmatrix} b_1^{(2)} \end{pmatrix} = \begin{pmatrix} 0.21799 \end{pmatrix}$$

$$Z^2 = X^T W^1 + b1 = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{pmatrix} + \begin{pmatrix} b_1^{(1)} & b_2^{(1)} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.19214979492 & -0.30588475125 \\ 0.4 & 0.1 \\ -0.50785020507 & 0.19411524875 \end{pmatrix} + \begin{pmatrix} -0.40785020507 & 0.19411524875 \end{pmatrix}$$

$$= \begin{pmatrix} -0.72355062 & 0.08234575 \end{pmatrix}$$

$$h^2 = \sigma(Z^2) = \sigma(\begin{pmatrix} -0.72355062 & 0.08234575 \end{pmatrix}) = \begin{pmatrix} 0.3266115919 & 0.52057481263 \end{pmatrix}$$

$$Z^3 = \begin{pmatrix} 0.3266115919 & 0.52057481263 \end{pmatrix} \begin{pmatrix} -0.26082732 \\ -0.13805525 \end{pmatrix} + \begin{pmatrix} 0.21799 \end{pmatrix} = 0.06093269$$

$$h^3 = \sigma(Z^3) = \sigma(\begin{pmatrix} 0.06093269 \end{pmatrix}) = \begin{pmatrix} \mathbf{0.48477153888} \end{pmatrix}$$

# 4 Comments

old error : $E = \frac{1}{2}(0.474 - 1)^2 = 0.138338$

new error : $E = \frac{1}{2}(0.48477153888 - 1)^2 = 0.132730$

Therefore error has reduced.