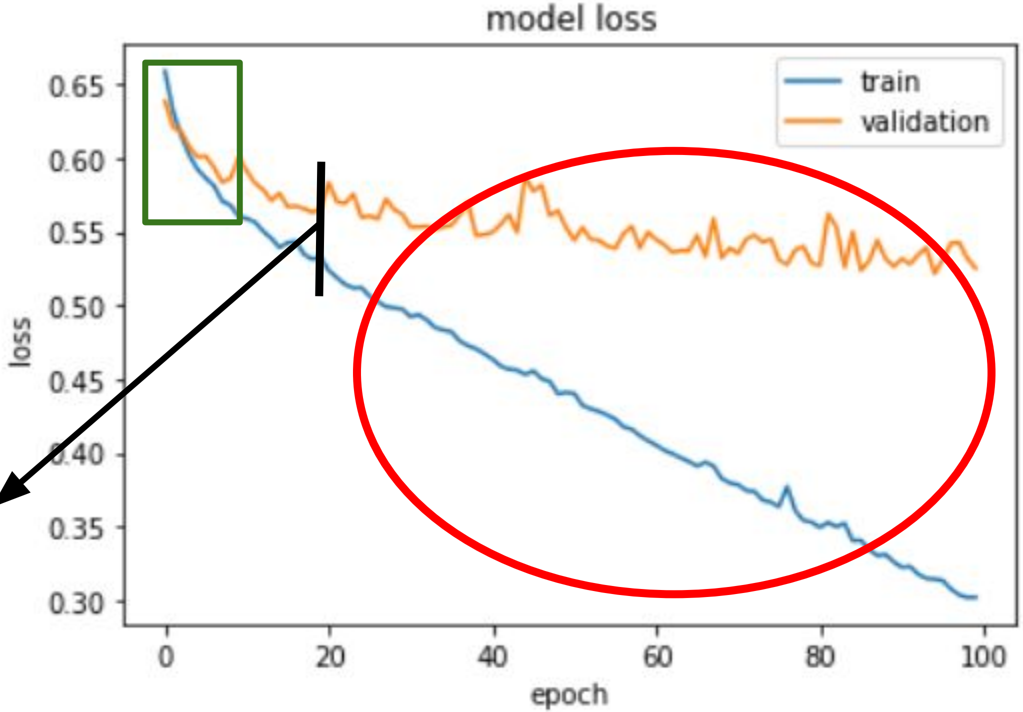
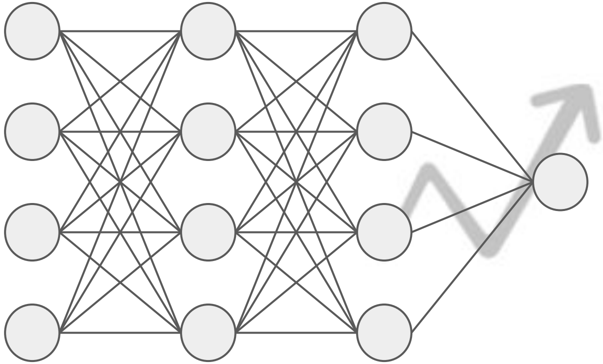
Improving your Neural Network

Solution 1: Early Stopping

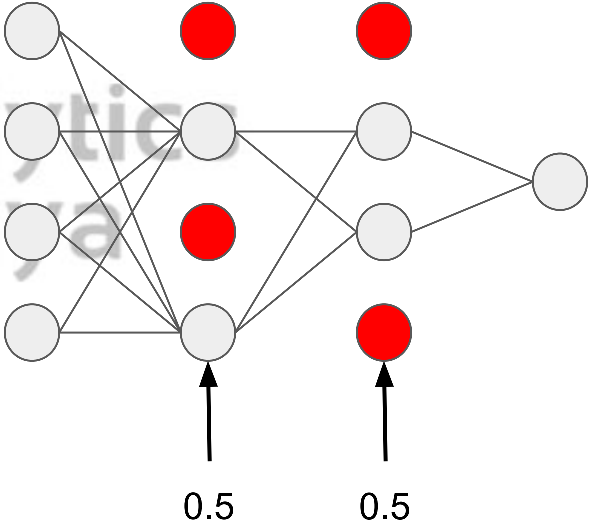


Stop the training here

Solution 2: Dropout Regularization



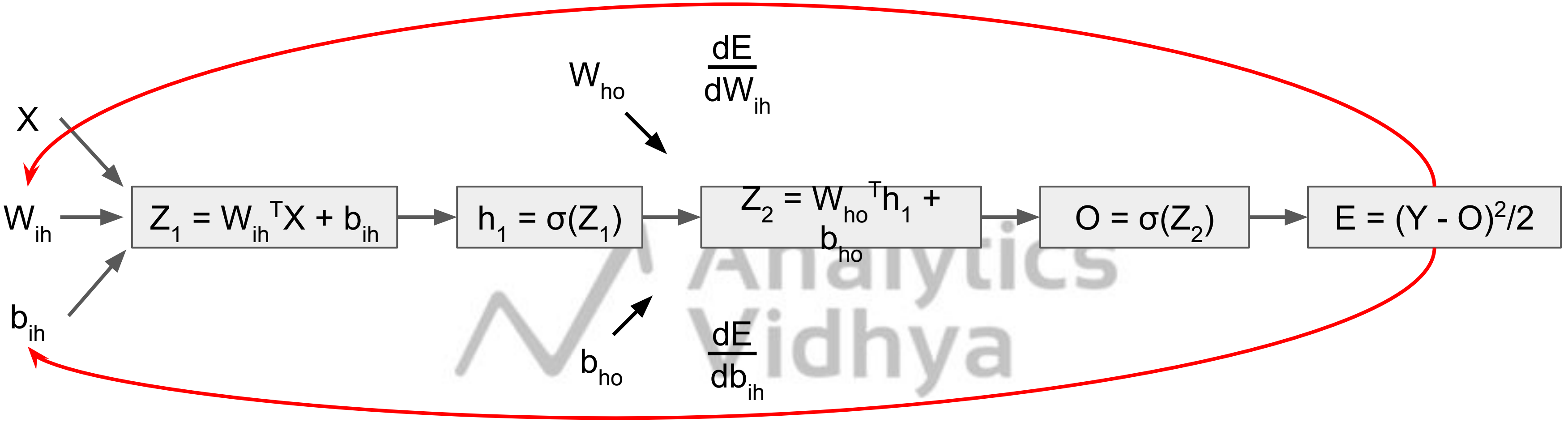
Without Dropout



With Dropout

Problem: Vanishing / Exploding Gradients

Problem: Vanishing / Exploding Gradients



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$\frac{dE}{dW_{ih}}$

$\frac{dE}{db_{ih}}$

$Z_2 = W_{ho}^T h_1 + b_{ho}$

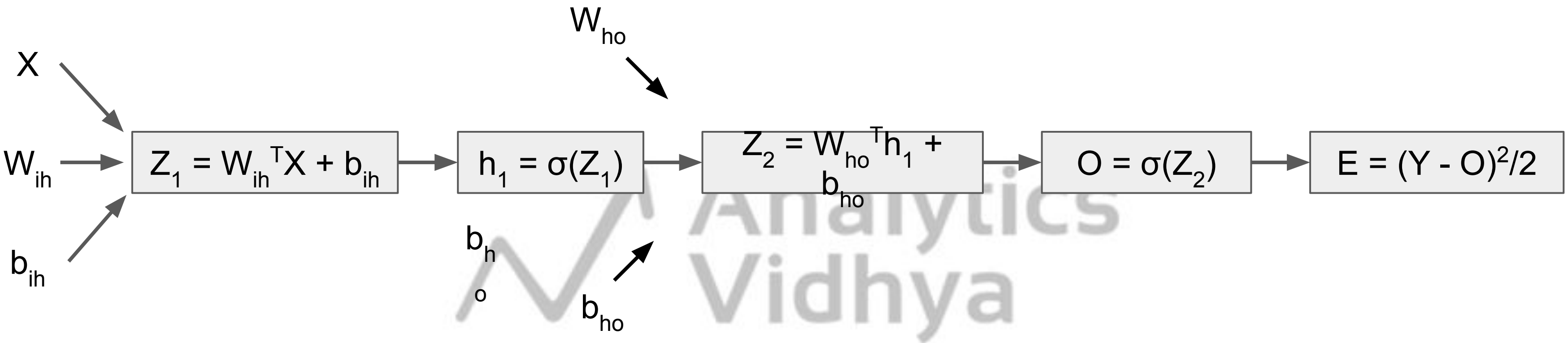
$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

$$\frac{dE}{db_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{db_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * 1$$

Problem: Vanishing / Exploding Gradients



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_h

o

b_{ho}

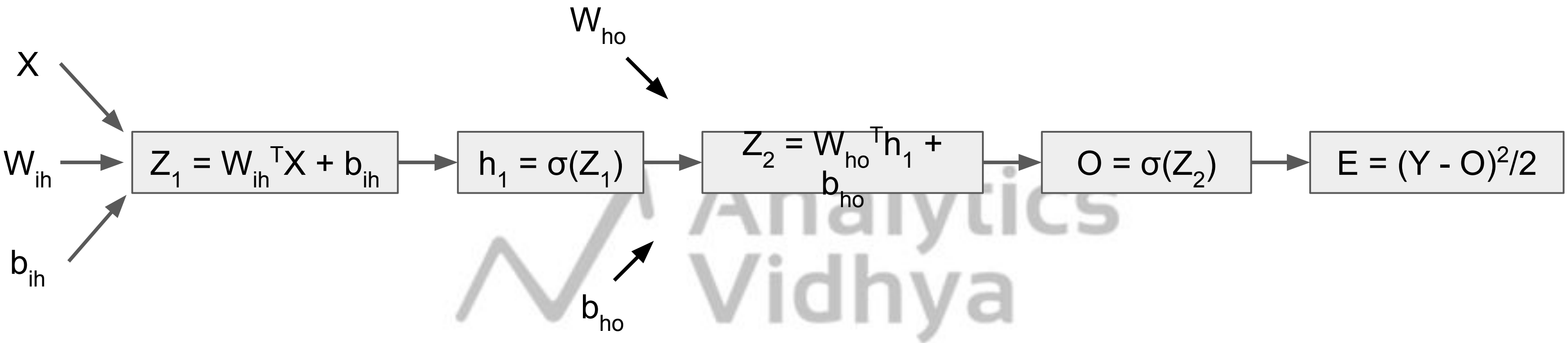
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Problem: Vanishing / Exploding Gradients



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

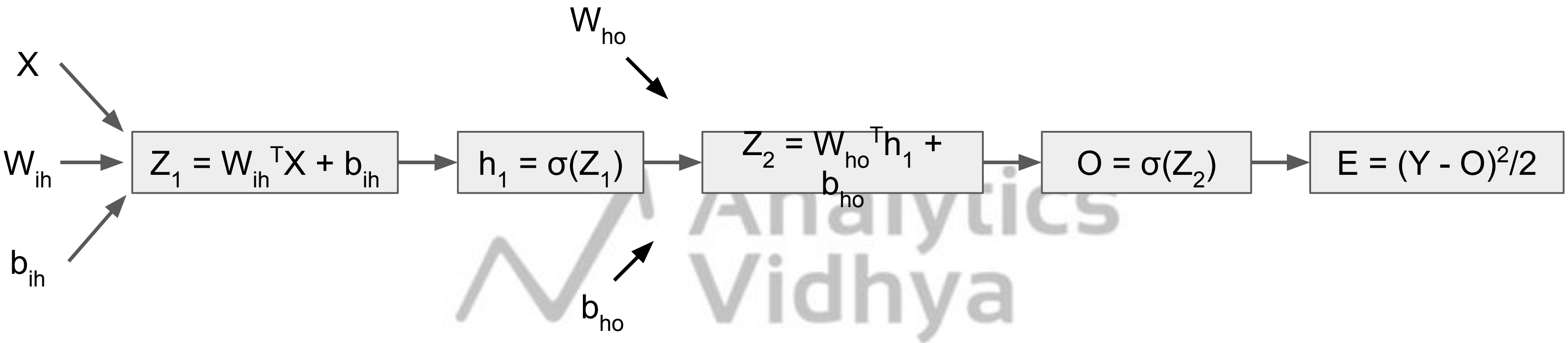
$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Small

Problem: Vanishing / Exploding Gradients



X, W_{ih}, b_{ih} → $Z_1 = W_{ih}^T X + b_{ih}$ → $h_1 = \sigma(Z_1)$ → $Z_2 = W_{ho}^T h_1 + b_{ho}$ (with W_{ho}, b_{ho}) → $O = \sigma(Z_2)$ → $E = (Y - O)^2/2$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Small

Vanishing Gradient

Problem: Vanishing Gradients

$$W_{ih} = W_{ih} - \alpha * \frac{dE}{dW_{ih}}$$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Small

Vanishing Gradient

Problem: Vanishing Gradients

$$W_{ih} = W_{ih} - \boxed{\alpha * \frac{dE}{dW_{ih}}} \longrightarrow \approx 0$$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Small

Vanishing Gradient

Problem: Vanishing Gradients

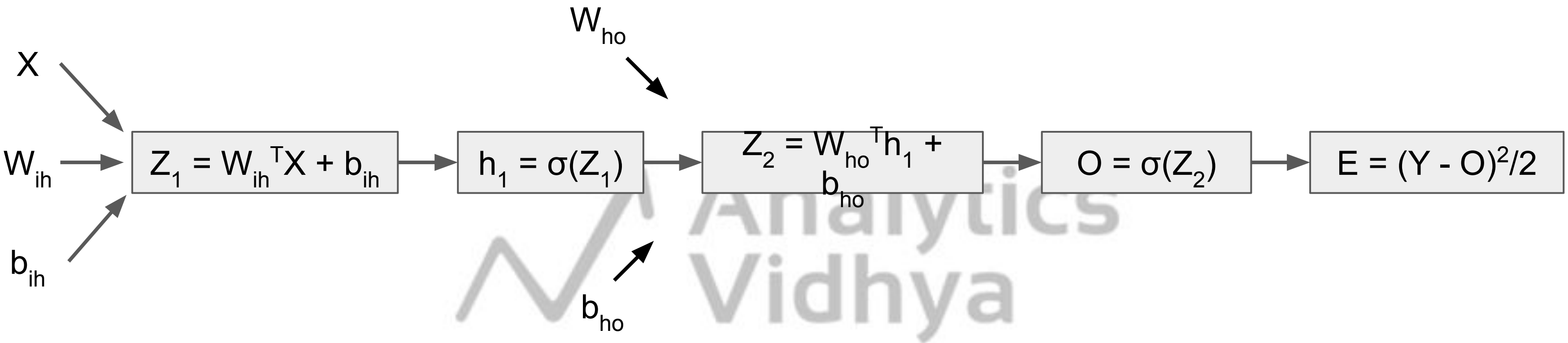
$$W_{ih} = W_{ih} \boxed{- \alpha * \frac{dE}{dW_{ih}}} \longrightarrow \approx 0 \qquad\qquad W_{ih} \approx W_{ih}$$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Small

Vanishing Gradient

Problem: Vanishing / Exploding Gradients



W_{ho}

X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

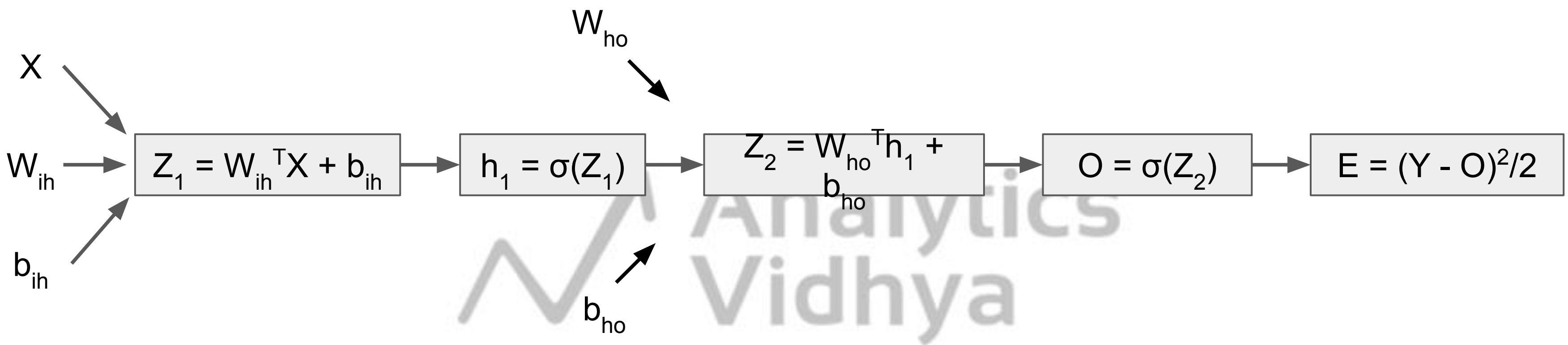
$E = (Y - O)^2/2$

b_{ho}

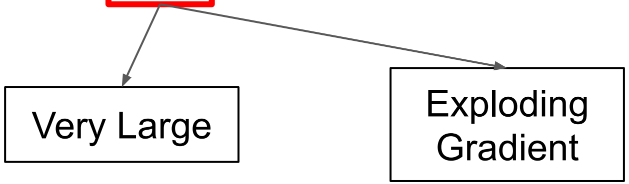
$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Large

Problem: Vanishing / Exploding Gradients



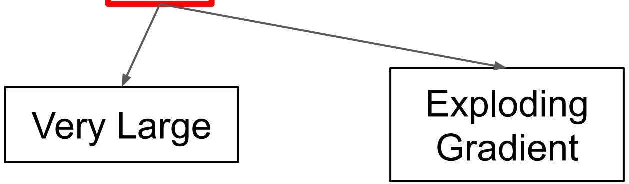
$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$



Problem: Exploding Gradients

$$W_{ih} = W_{ih} - \alpha * \frac{dE}{dW_{ih}}$$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$



Very Large

Exploding Gradient

Problem: Exploding Gradients

$$W_{ih} = W_{ih} \boxed{- \alpha * \frac{dE}{dW_{ih}}} \rightarrow \text{Very Large}$$

$$\boxed{\frac{dE}{dW_{ih}}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Very Large

Exploding Gradient

Problem: Vanishing / Exploding Gradients

1

Problem: Vanishing / Exploding Gradients

1	1.0002

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1	12	-10	8	-6	14	-12

Exploding Gradient

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1	12

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1	12	-10

Problem: Vanishing / Exploding Gradients

1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1	12	-10	8	-6	14	-12

Problem: Vanishing / Exploding Gradients

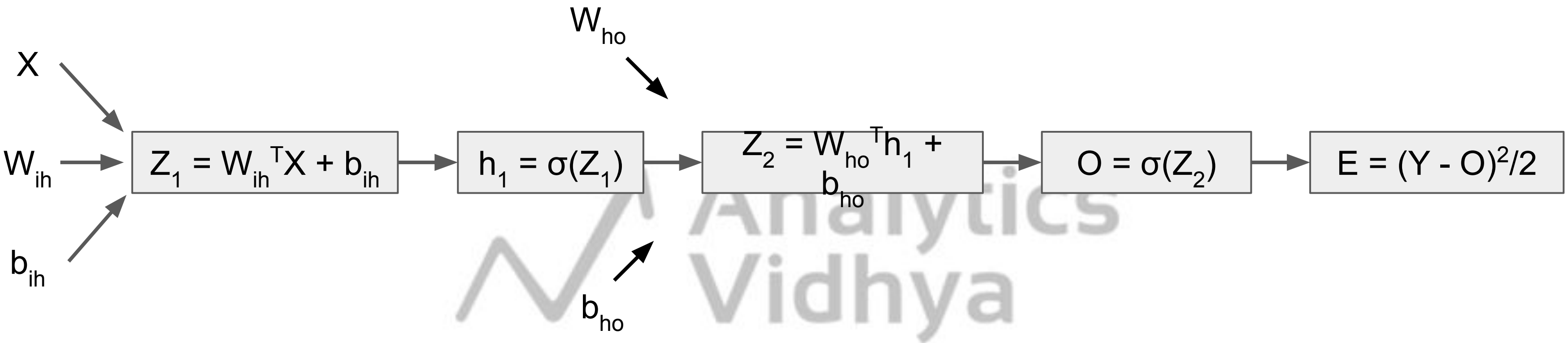
1	1.0002	1.0003	1.0004	1.0005	1.0006	1.0007

Vanishing Gradient

1	12	-10	8	-6	14	-12

Exploding Gradient

Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

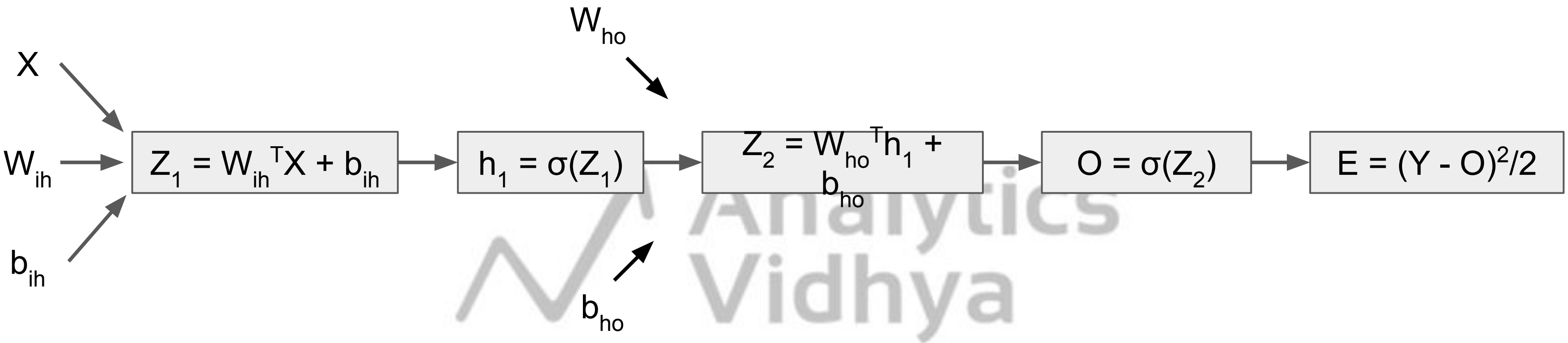
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

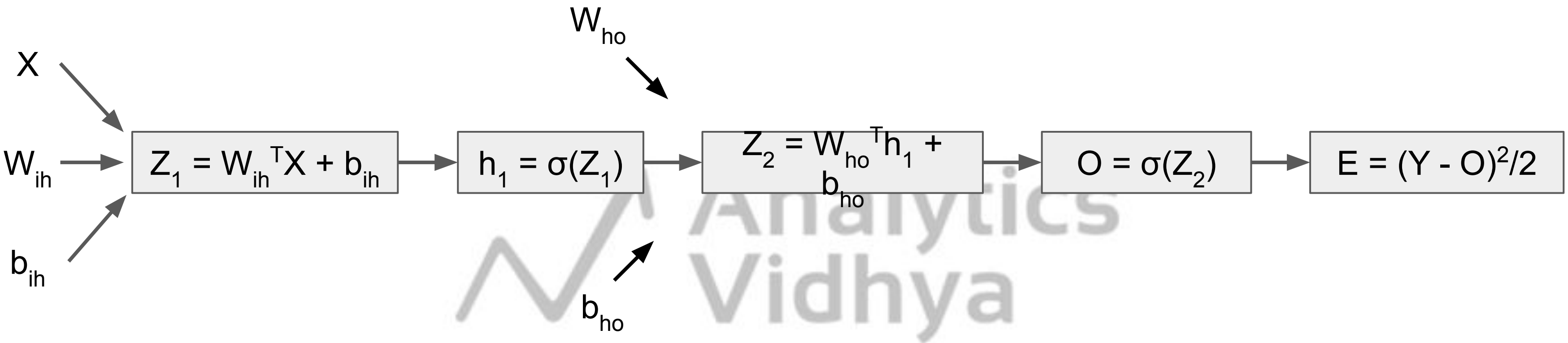
$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Vanishing Gradients happen?



$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \boxed{\frac{dZ_2}{dh_1}} * \boxed{\frac{dh_1}{dZ_1}} * \boxed{\frac{dZ_1}{dW_{ih}}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

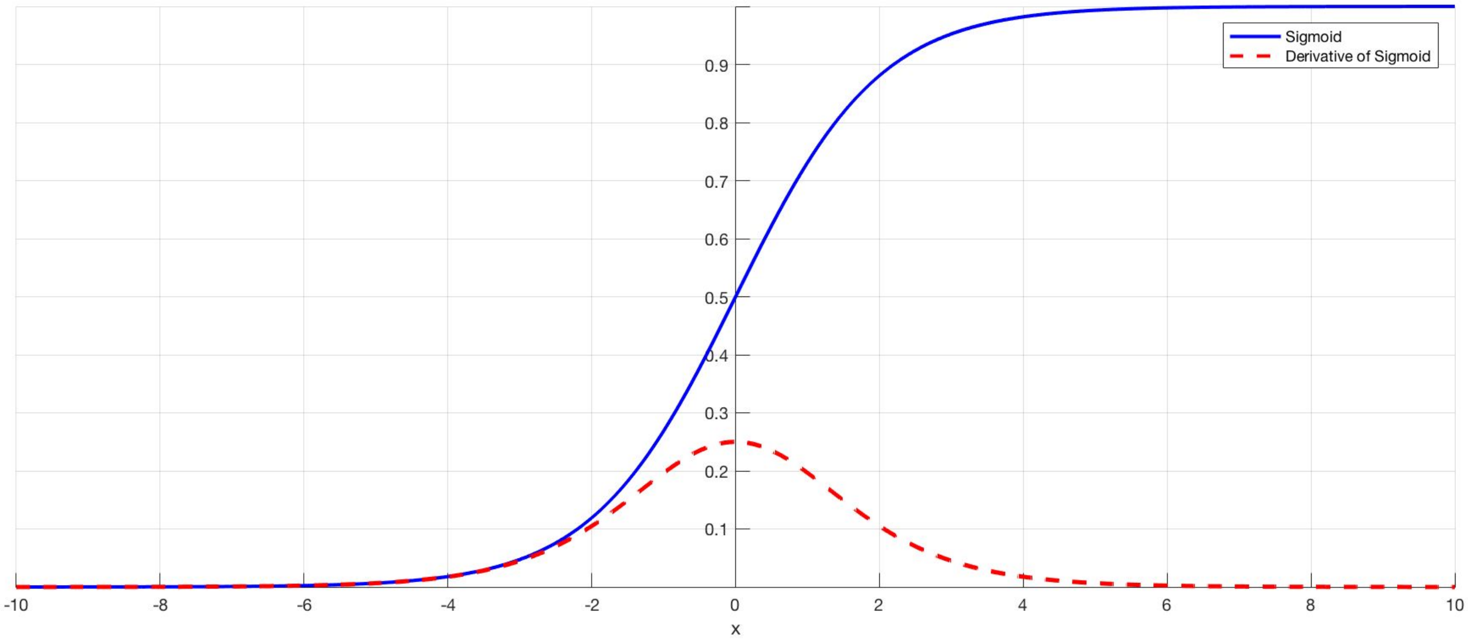
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

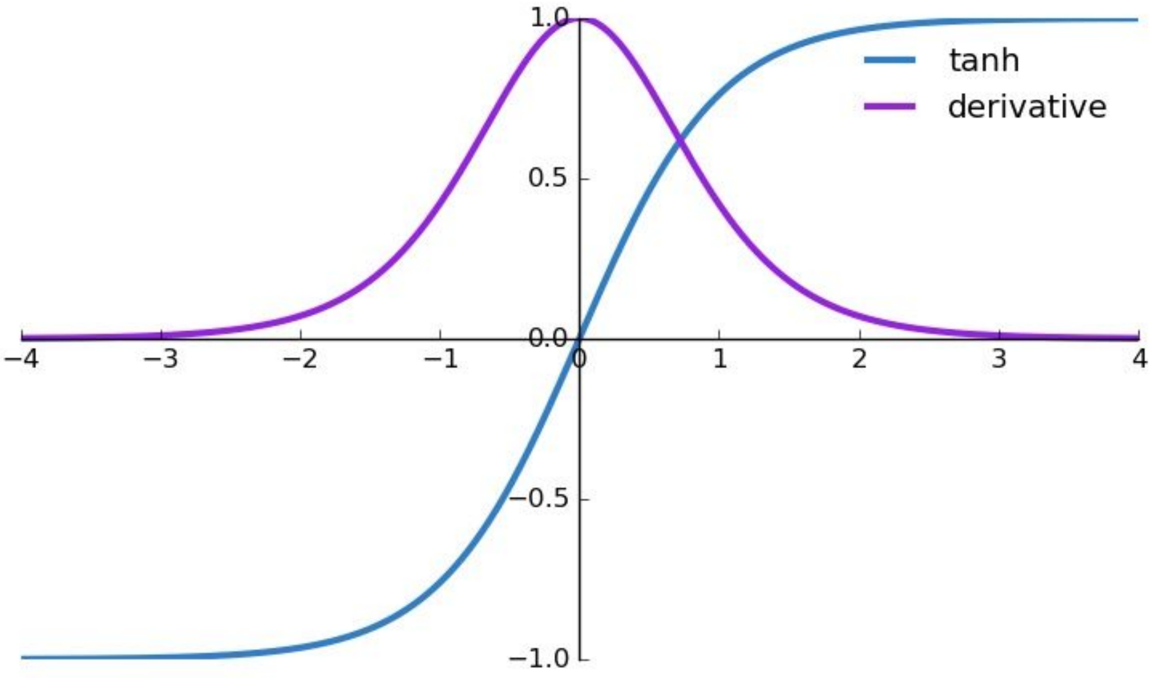
$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

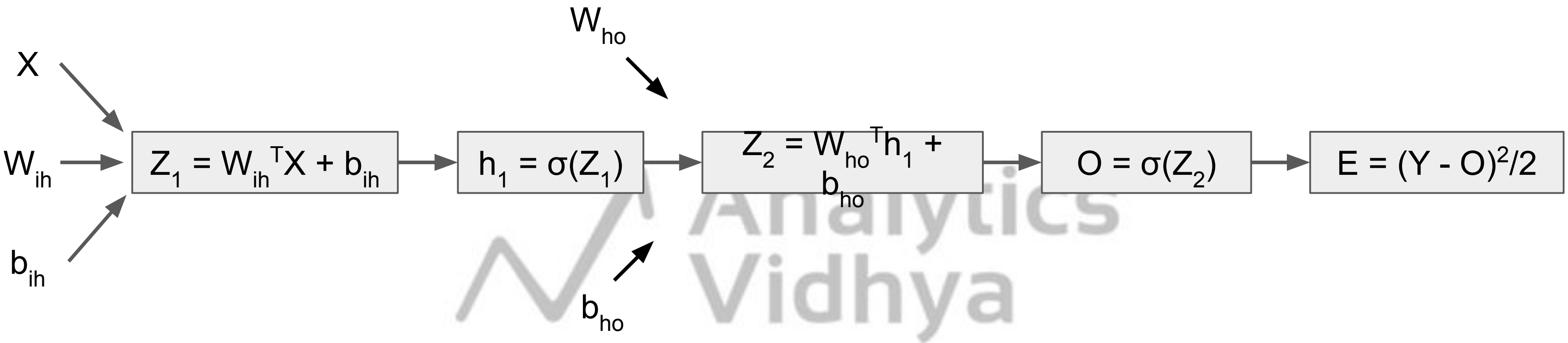
Why does Vanishing Gradients happen?



Why does Vanishing Gradients happen?



Why does Vanishing Gradients happen?

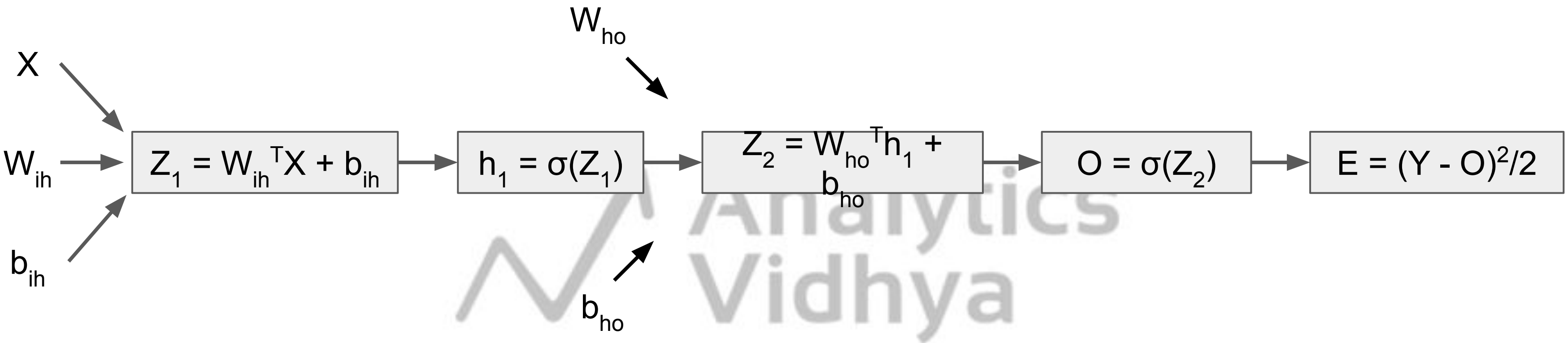


X, W_{ih}, b_{ih} → $Z_1 = W_{ih}^T X + b_{ih}$ → $h_1 = \sigma(Z_1)$ → $Z_2 = W_{ho}^T h_1 + b_{ho}$ → $O = \sigma(Z_2)$ → $E = (Y - O)^2/2$

W_{ho}, b_{ho} → $Z_2 = W_{ho}^T h_1 + b_{ho}$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

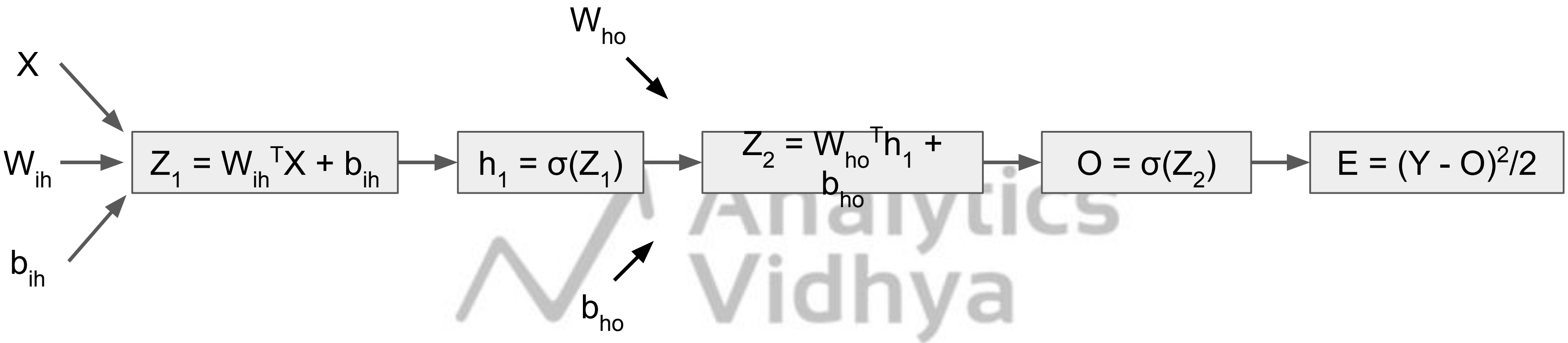
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \underset{0.2}{\boxed{\frac{dO}{dZ_2}}} * \frac{dZ_2}{dh_1} * \underset{0.2}{\boxed{\frac{dh_1}{dZ_1}}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

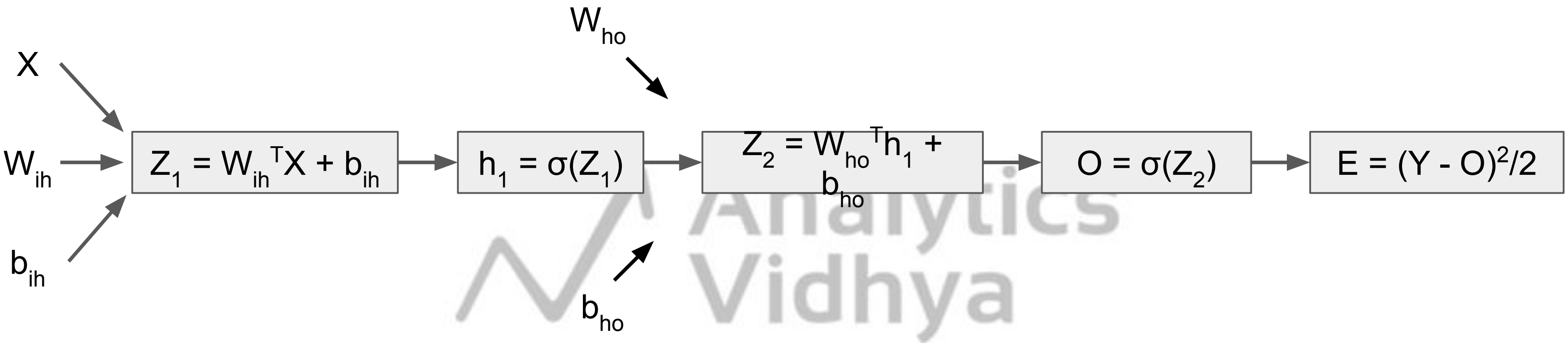
$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

0.2 0.2

$$(0.2)^2 (O\text{-}Y) * W_{ho} * X$$

Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

b_{ho}

$O = \sigma(Z_2)$

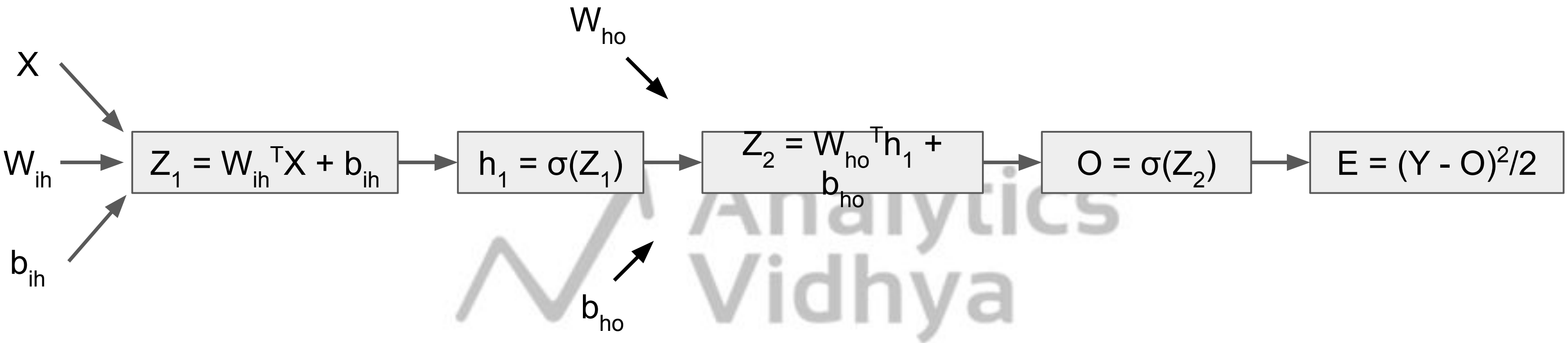
$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

$(0.2)^{10}$ (O-Y) * ….. * W_{ho} * X

Layers = 10

Solution: Vanishing Gradients



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

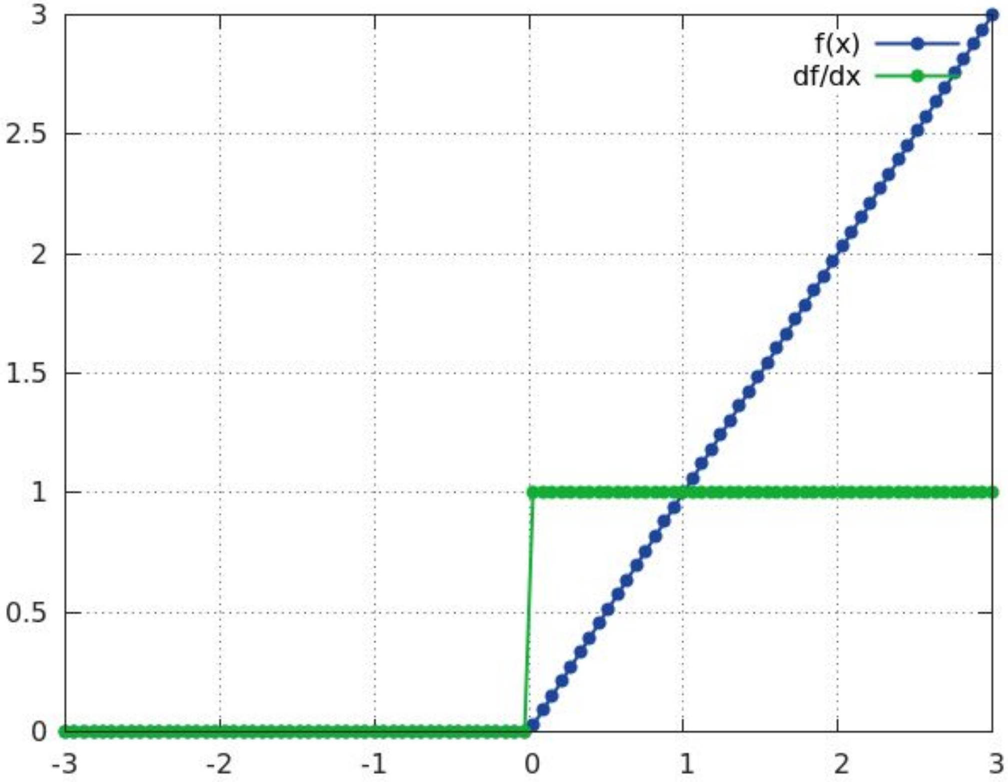
$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

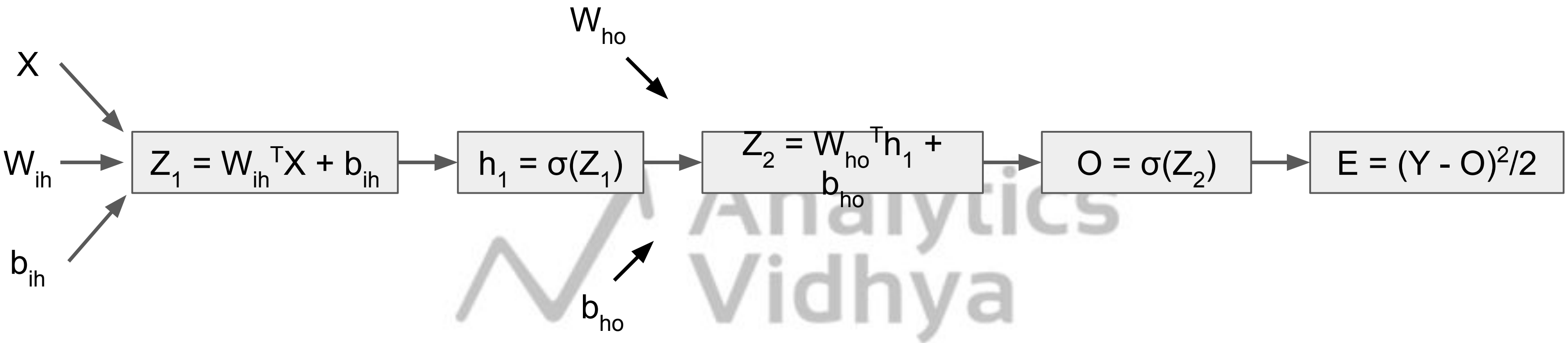
$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

ReLU activation function

Solution: Vanishing Gradients



Why does Vanishing Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

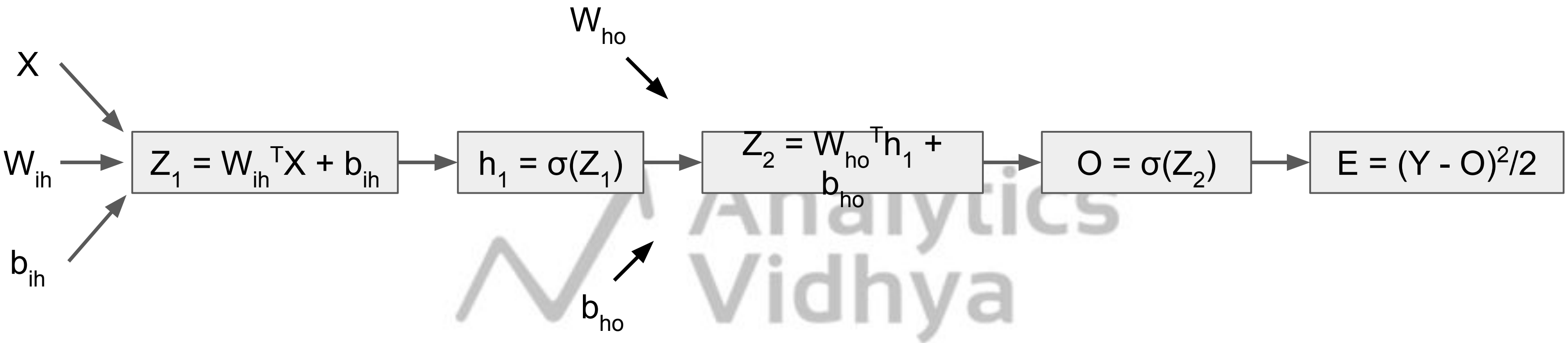
b_{ho}

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \boxed{\frac{dZ_2}{dh_1}} * \frac{dh_1}{dZ_1} * \boxed{\frac{dZ_1}{dW_{ih}}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Exploding Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

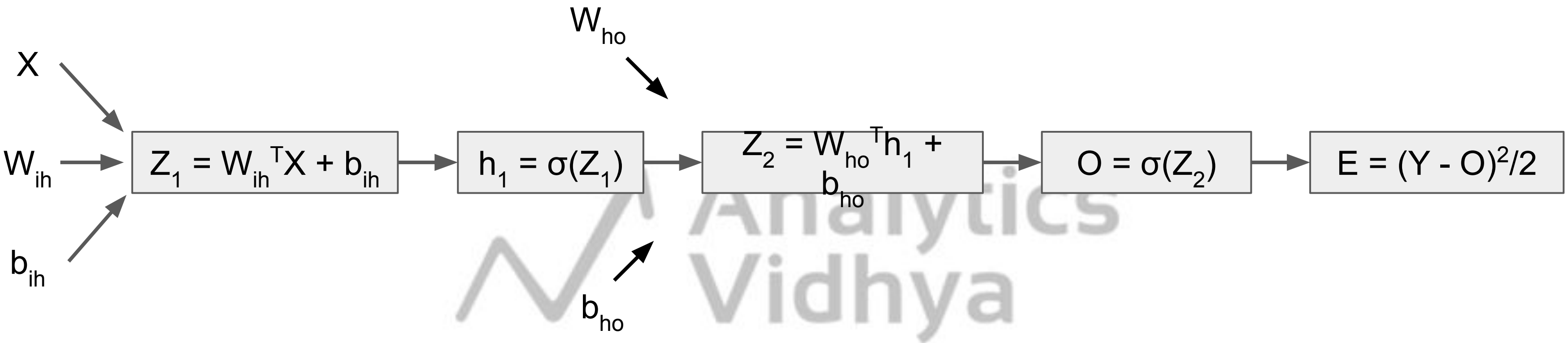
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

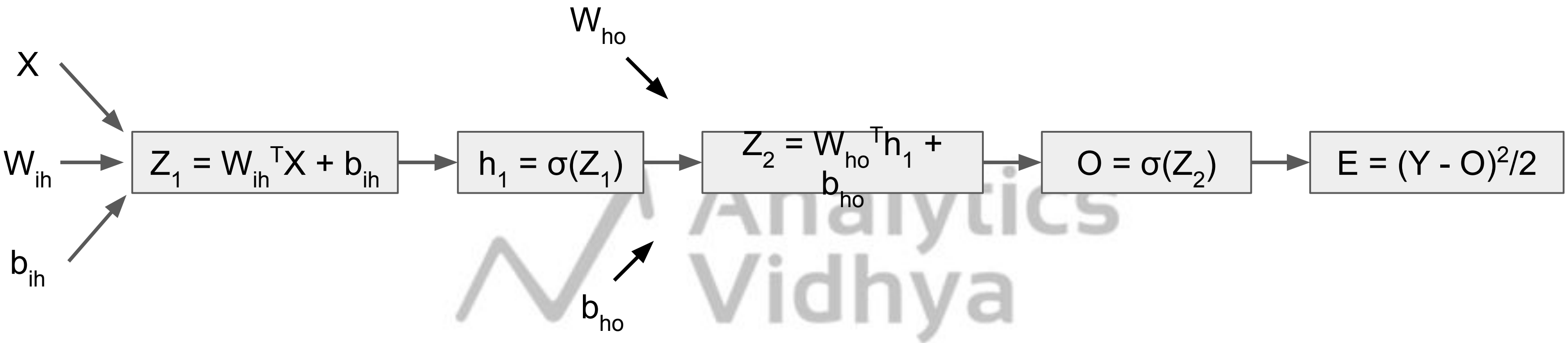
$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Exploding Gradients happen?



$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \boxed{\frac{dZ_2}{dh_1}} * \boxed{\frac{dh_1}{dZ_1}} * \boxed{\frac{dZ_1}{dW_{ih}}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Exploding Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

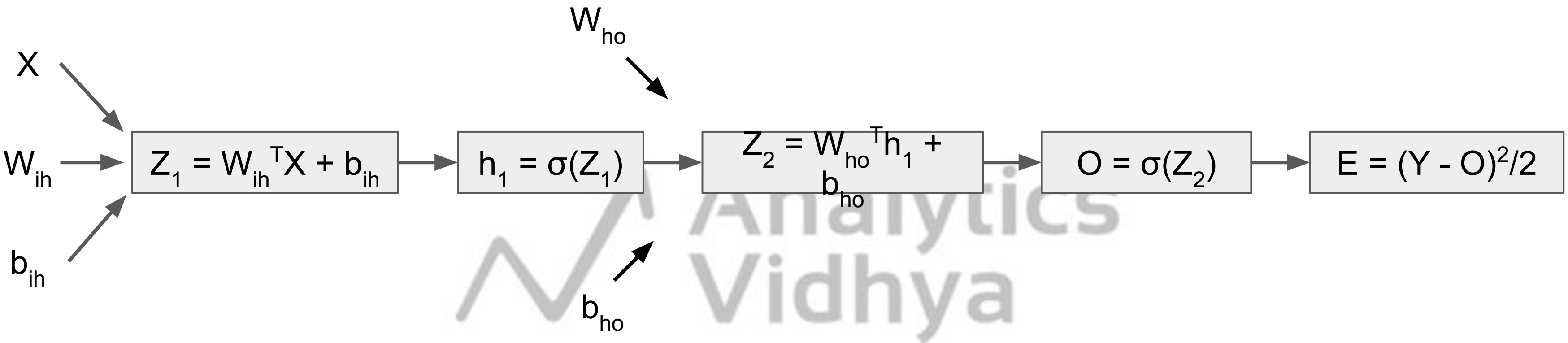
$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

1,5

1.5

Why does Exploding Gradients happen?



X, W_{ih}, b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}, b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

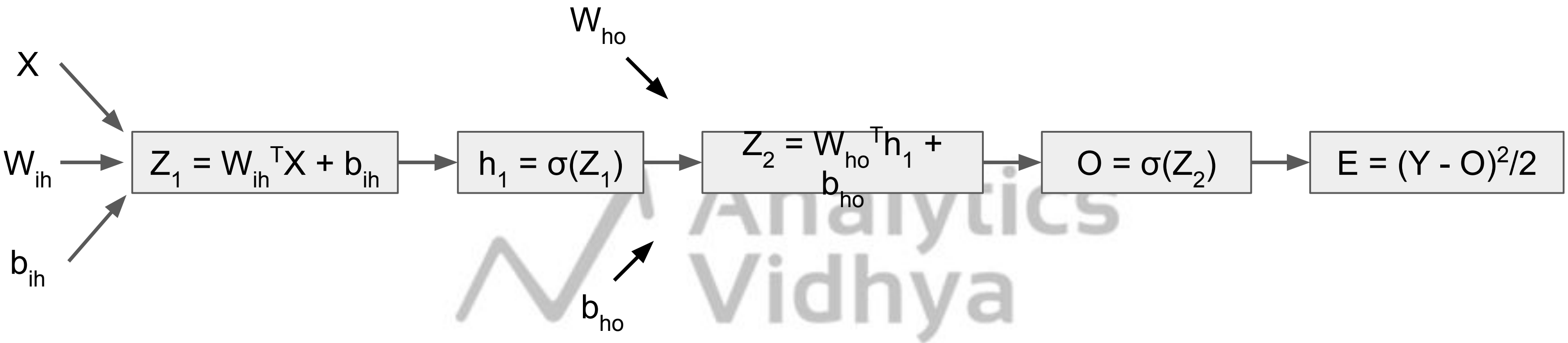
$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

1,5   1.5 $(1.5)^2\,(O\text{-}Y) * W_{ho} * X$

Why does Exploding Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

b_{ho}

$O = \sigma(Z_2)$

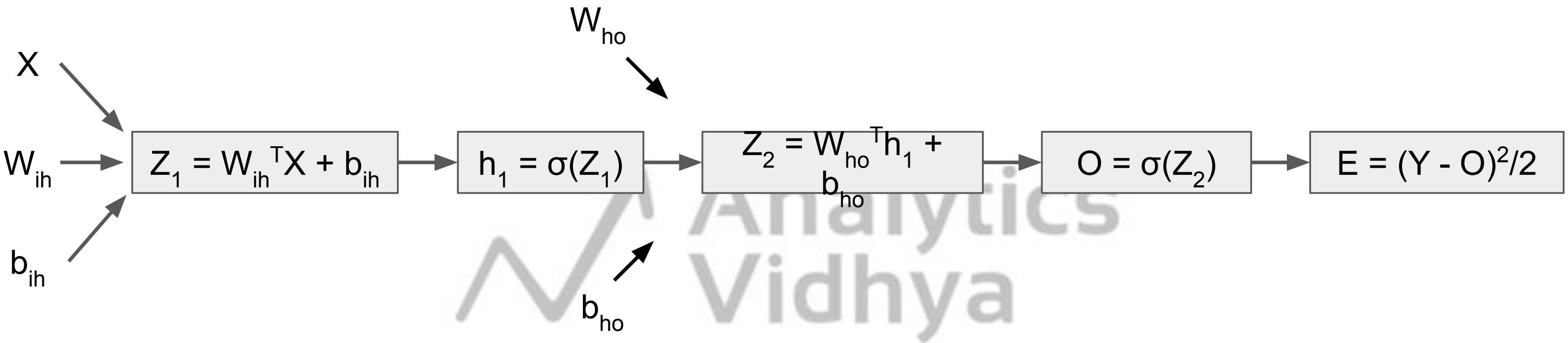
$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \boxed{\frac{dO}{dZ_2}} * \frac{dZ_2}{dh_1} * \boxed{\frac{dh_1}{dZ_1}} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

$$(1.5)^{10} (O\text{-}Y) * \ldots.. * W_{ho} * X$$

Layers = 10

Why does Exploding Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

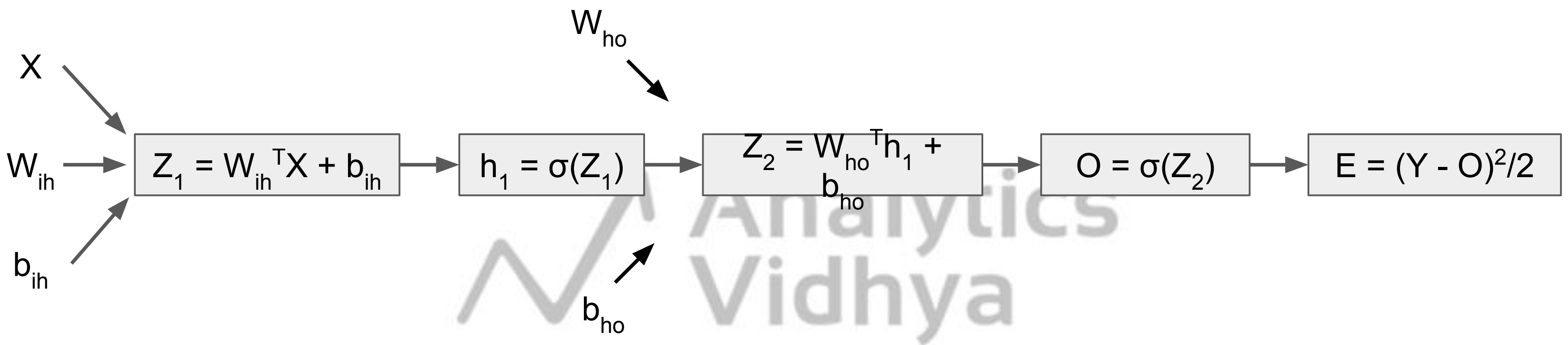
$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \boxed{\frac{dZ_2}{dh_1}} * \frac{dh_1}{dZ_1} * \boxed{\frac{dZ_1}{dW_{ih}}} = (O-Y) * O(1-O) * W_{ho} * h_1(1-h_1) * X$$

Why does Exploding Gradients happen?



X

W_{ih}

b_{ih}

$Z_1 = W_{ih}^T X + b_{ih}$

$h_1 = \sigma(Z_1)$

W_{ho}

b_{ho}

$Z_2 = W_{ho}^T h_1 + b_{ho}$

$O = \sigma(Z_2)$

$E = (Y - O)^2/2$

$$\frac{dE}{dW_{ih}} = \frac{dE}{dO} * \frac{dO}{dZ_2} * \frac{dZ_2}{dh_1} * \frac{dh_1}{dZ_1} * \frac{dZ_1}{dW_{ih}} = (O\text{-}Y) * O(1-O) * \underset{\begin{array}{c}\downarrow\\ \boxed{\text{High}}\end{array}}{\boxed{W_{ho} *}}\; h_1(1-h_1) * X$$

Gradient Clipping

Gradient Clipping

- Clips the derivatives or gradients

Gradient Clipping

- Clips the derivatives or gradients
- Define a threshold, clipvalue

Gradient Clipping

- Clips the derivatives or gradients

- Define a threshold: clipvalue

clipvalue = 1

Gradient Clipping

- Clips the derivatives or gradients

- Define a threshold: clipvalue

If gradient value is less than -1, it will be clipped to -1

clipvalue = 1

Gradient Clipping

- Clips the derivatives or gradients
- Define a threshold: clipvalue

clipvalue = 1

If gradient value is less than -1, it will be clipped to -1

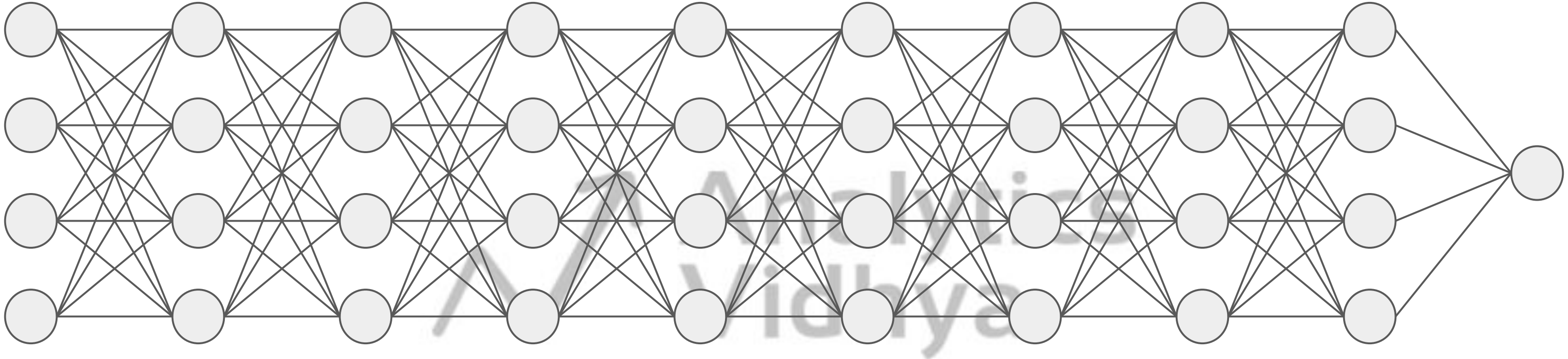
If gradient value is more than 1, it will be clipped to 1

Steps to solve emergency vs non-emergency vehicle classification problem by applying Gradient Clipping

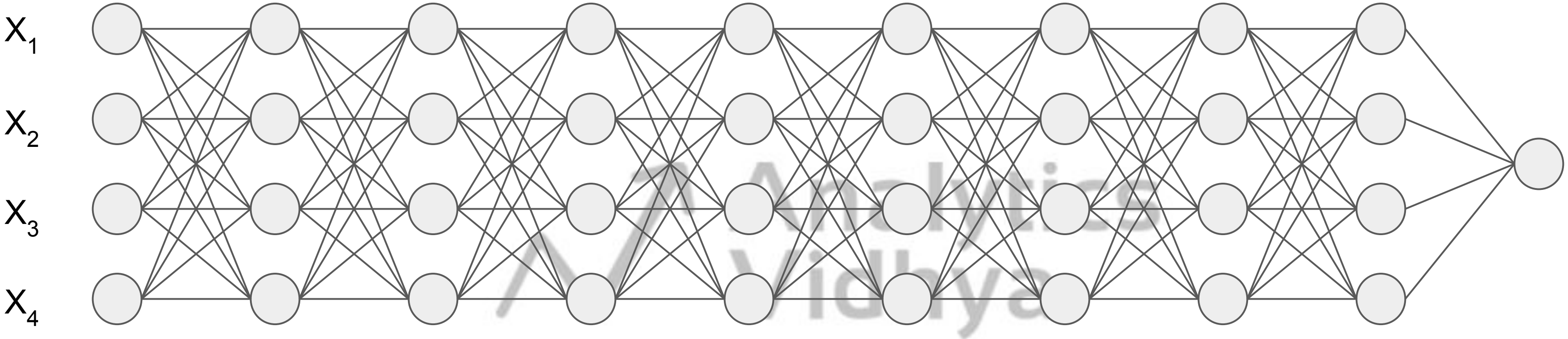
1. Loading the dataset
2. Pre-processing the data
3. Creating training and validation set
4. Defining the model architecture
5. Compiling the model
    - Define clipvalue while defining the optimizer
6. Training the model
7. Evaluating model performance

Thank You

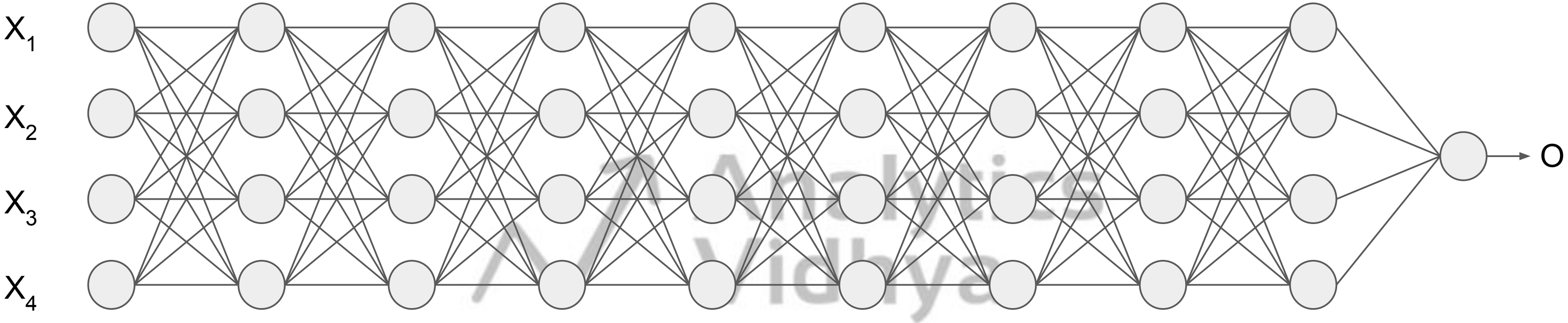
Problem: Vanishing / Exploding Gradients



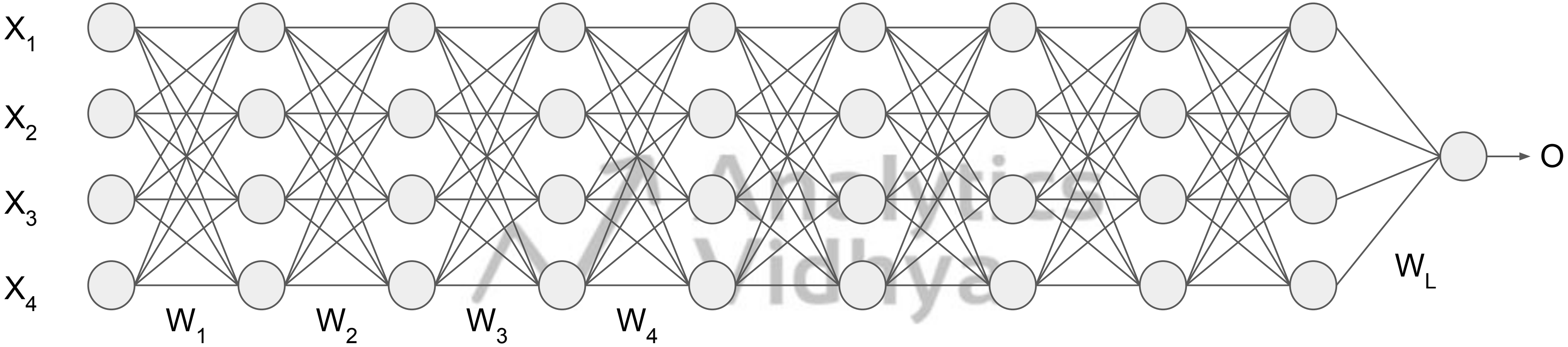
Problem: Vanishing / Exploding Gradients



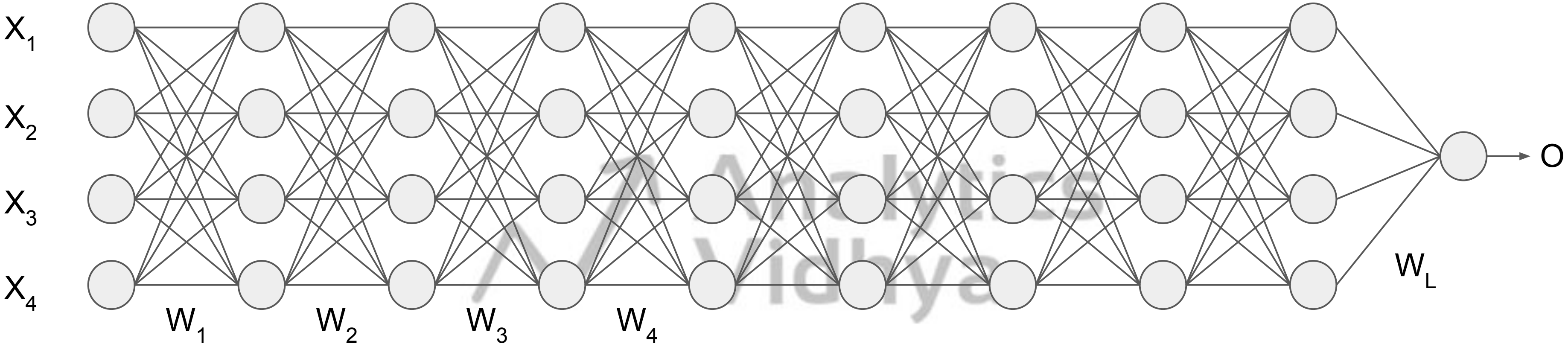
Problem: Vanishing / Exploding Gradients



Problem: Vanishing / Exploding Gradients

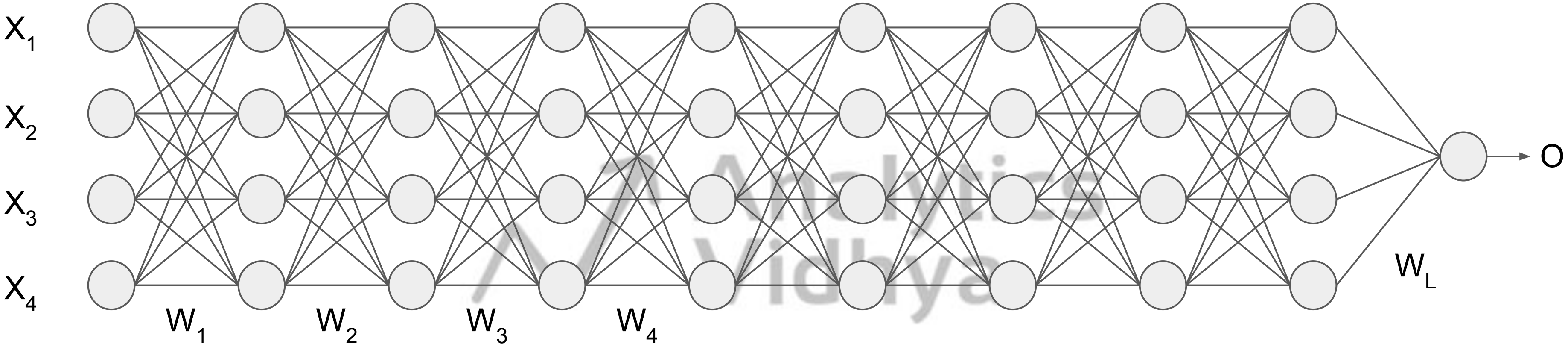


Problem: Vanishing / Exploding Gradients



L = Number of layers

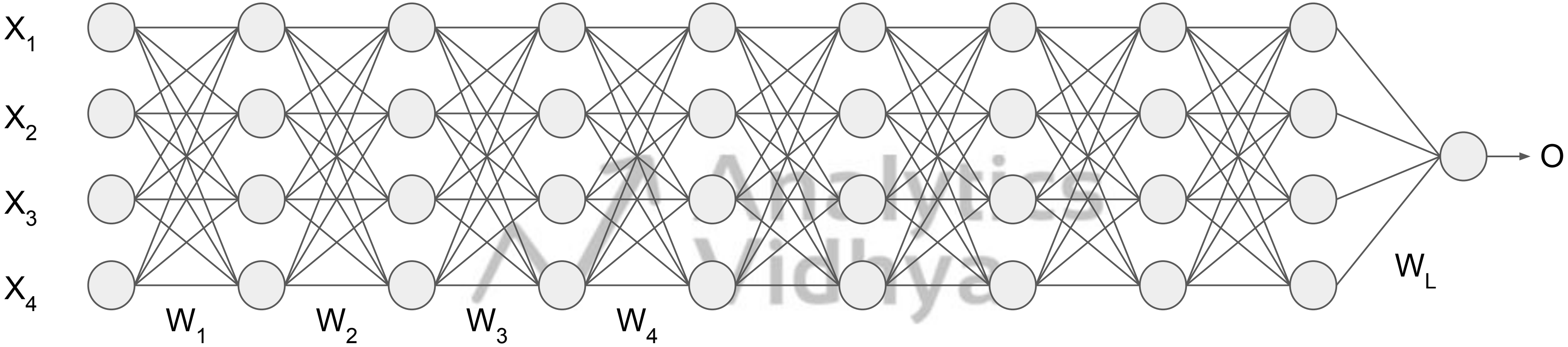
Problem: Vanishing / Exploding Gradients



L = Number of layers

Bias = 0

Problem: Vanishing / Exploding Gradients

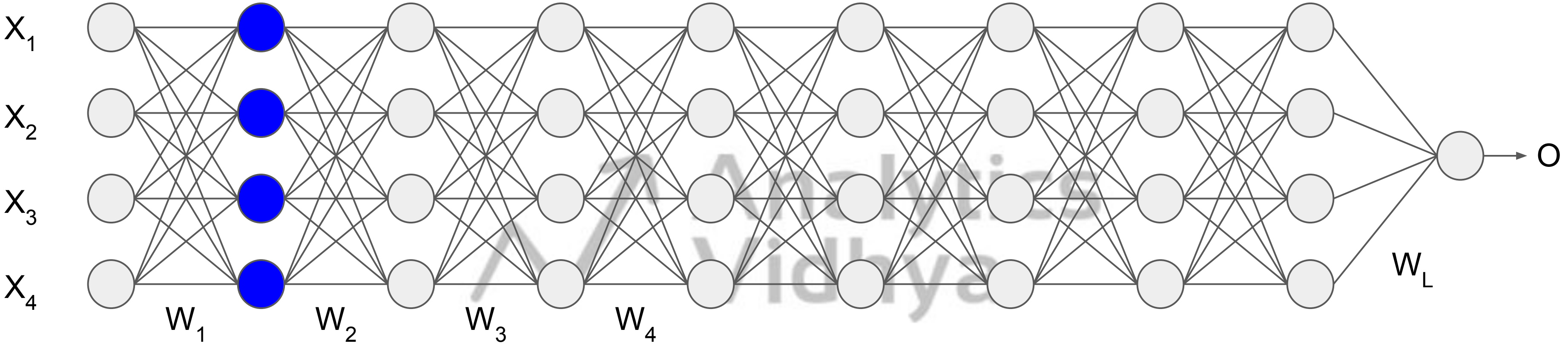


L = Number of layers

Bias = 0

Activation Function: Linear

Problem: Vanishing / Exploding Gradients

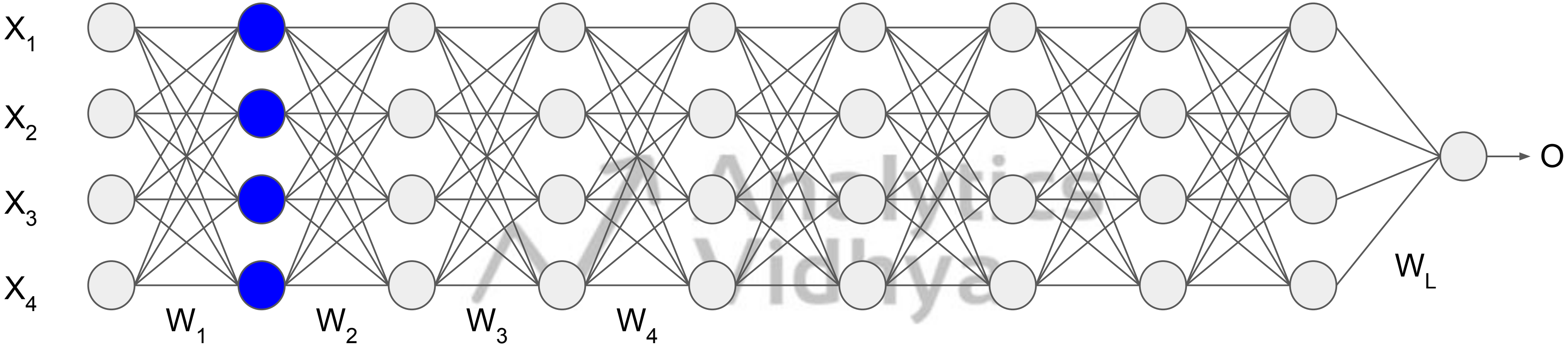


L = Number of layers

Bias = 0

Activation Function: Linear

Problem: Vanishing / Exploding Gradients



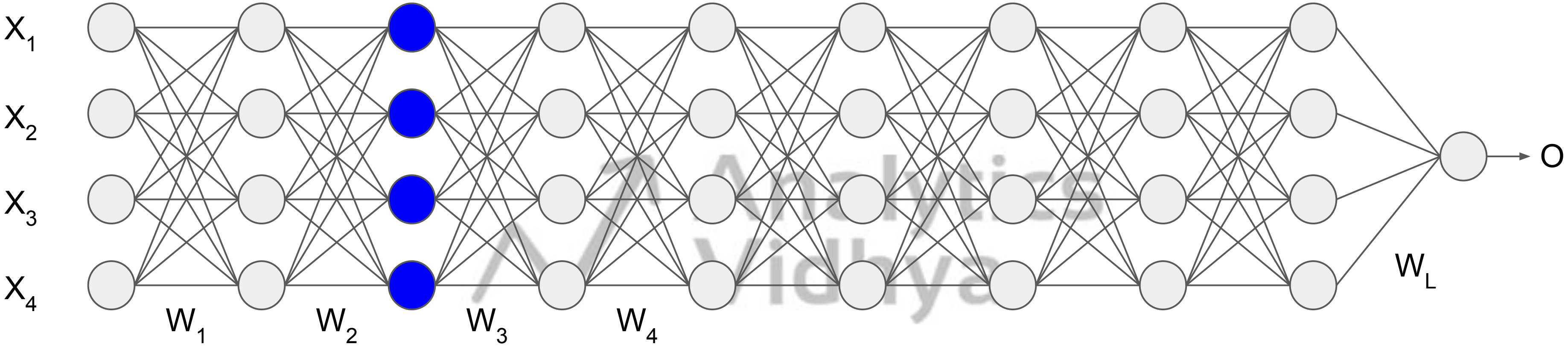
$h_1 = W_1X$

L = Number of layers

Bias = 0

Activation Function: Linear

Problem: Vanishing / Exploding Gradients



$h_1 = W_1X$

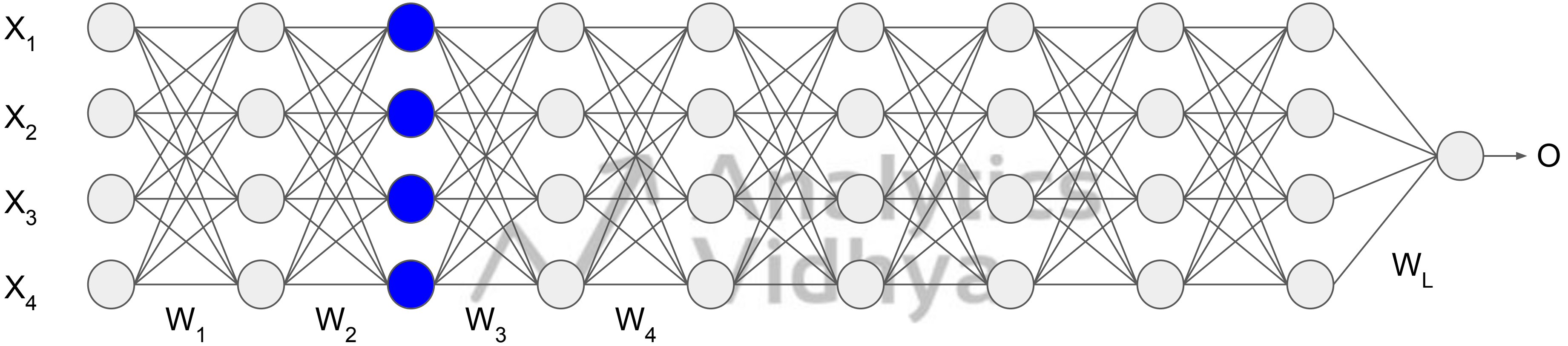
$h_2 = W_2h_1$

L = Number of layers

Bias = 0

Activation Function: Linear

Problem: Vanishing / Exploding Gradients



L = Number of layers

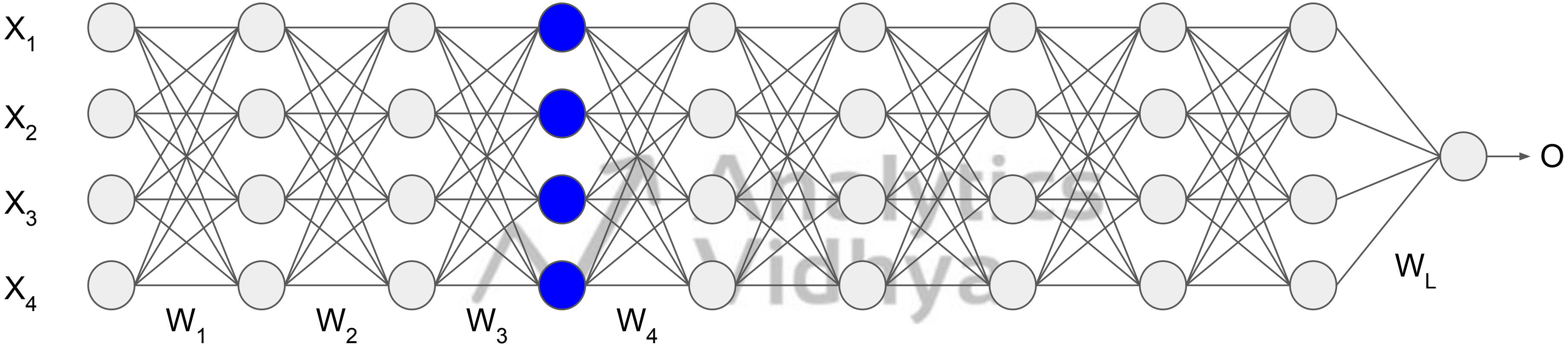
Bias = 0

Activation Function: Linear

$h_1 = W_1X$

$h_2 = W_2h_1 = W_1W_2X$

Problem: Vanishing / Exploding Gradients



$h_1 = W_1X$

$h_2 = W_2h_1 = W_1W_2X$

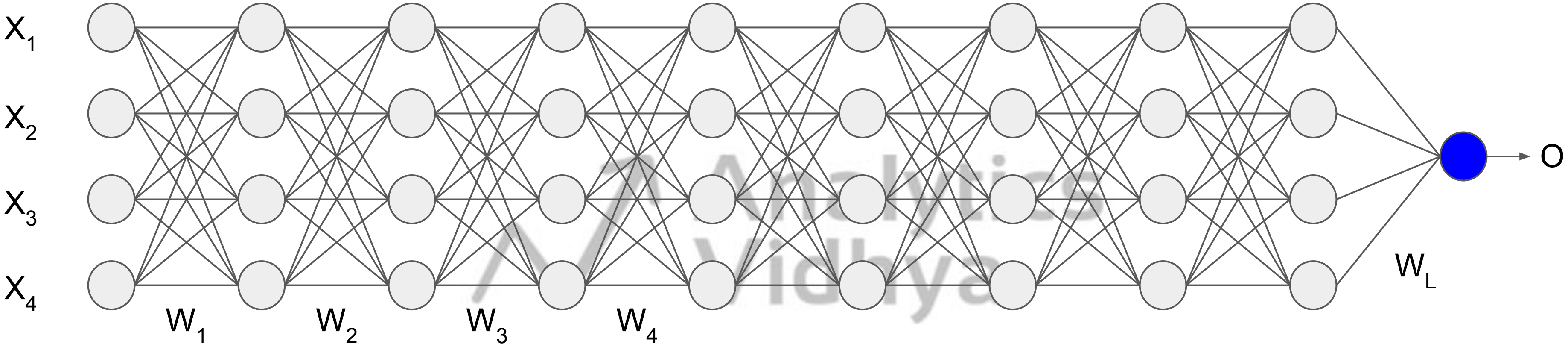
$h_3 = W_3h_2 = W_1W_2W_3X$

L = Number of layers

Bias = 0

Activation Function: Linear

Problem: Vanishing / Exploding Gradients



L = Number of layers

Bias = 0

Activation Function: Linear

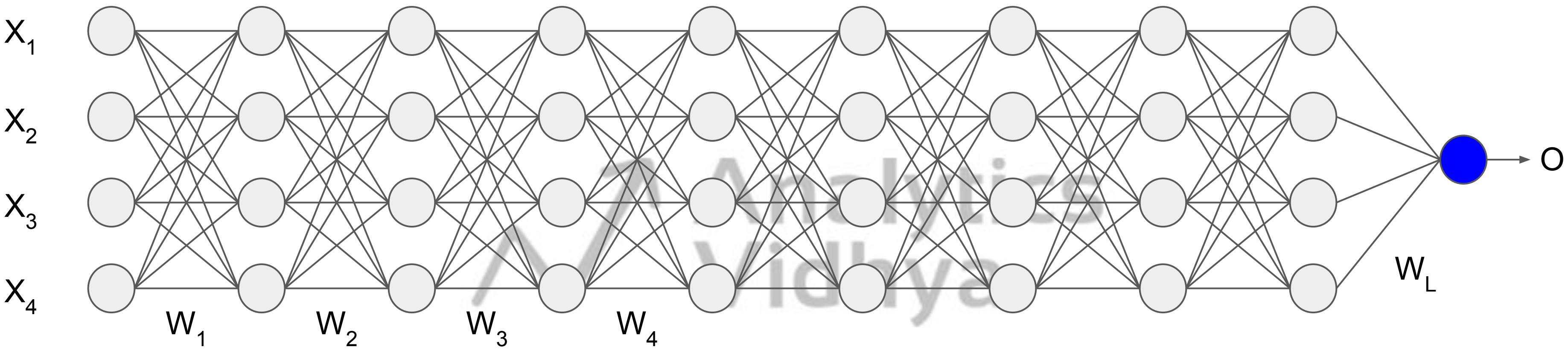
$h_1 = W_1X$

$h_2 = W_2h_1 = W_1W_2X$

$h_3 = W_3h_2 = W_1W_2W_3X$

$O = W_Lh_{L-1} =$
$W_1W_2W_3...W_LX$

Problem: Vanishing / Exploding Gradients



$$W_1 = W_1 = W_{l-1} = \begin{array}{|c|c|c|c|}\hline 1.5 & 0 & 0 & 0 \\ \hline 0 & 1.5 & 0 & 0 \\ \hline 0 & 0 & 1.5 & 0 \\ \hline 0 & 0 & 0 & 1.5 \\ \hline \end{array}$$

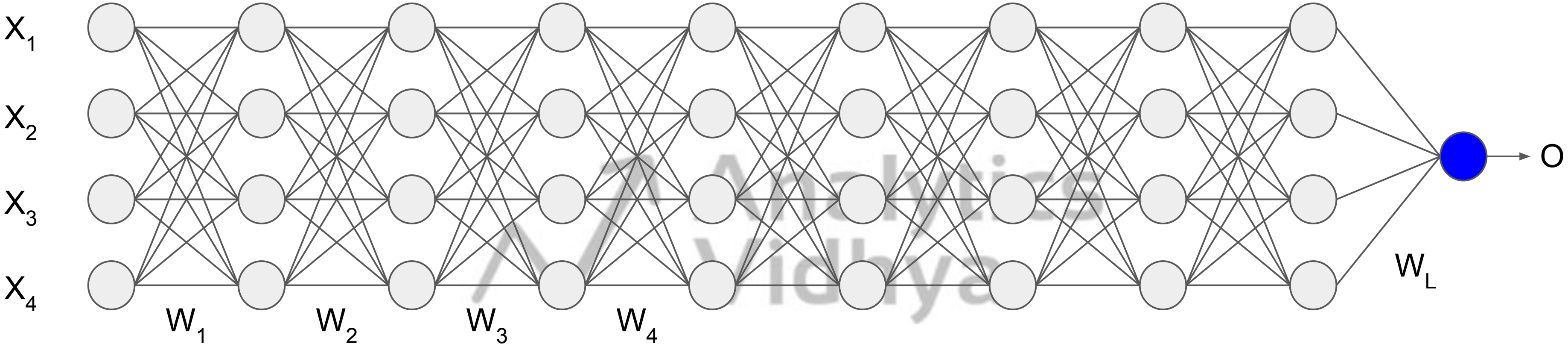
$h_1 = W_1X$

$h_2 = W_2h_1 = W_1W_2X$

$h_3 = W_3h_2 = W_1W_2W_3X$

$O = W_Lh_{L-1} =$
$W_1W_2W_3 \ldots W_LX$

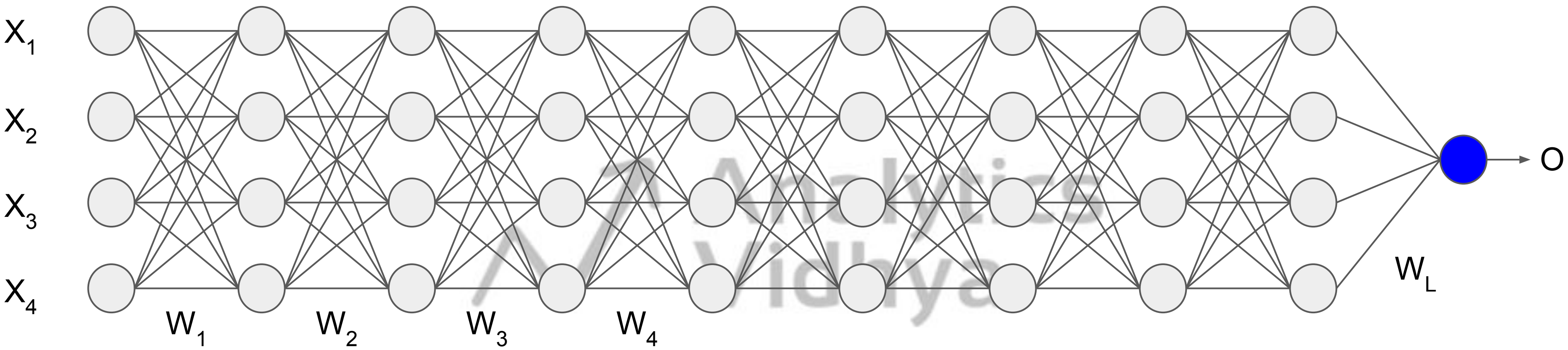
Problem: Vanishing / Exploding Gradients



$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}$$

$$O = W_1 W_2 ... W_{L-1} W_L X$$

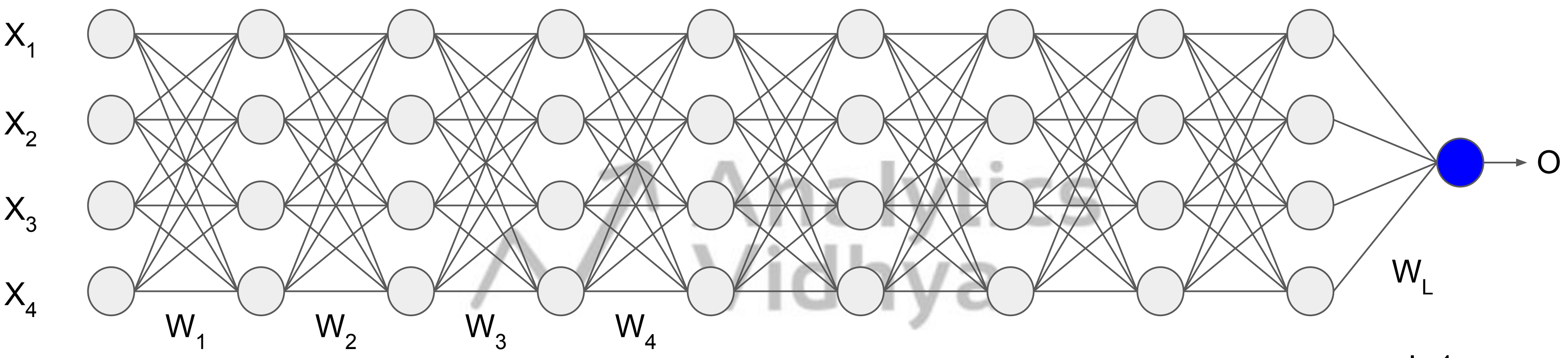
Problem: Vanishing / Exploding Gradients



$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}$$

$$O = W_1W_2...W_{L-1}W_LX = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}^{L-1} W_LX$$

Problem: Vanishing / Exploding Gradients

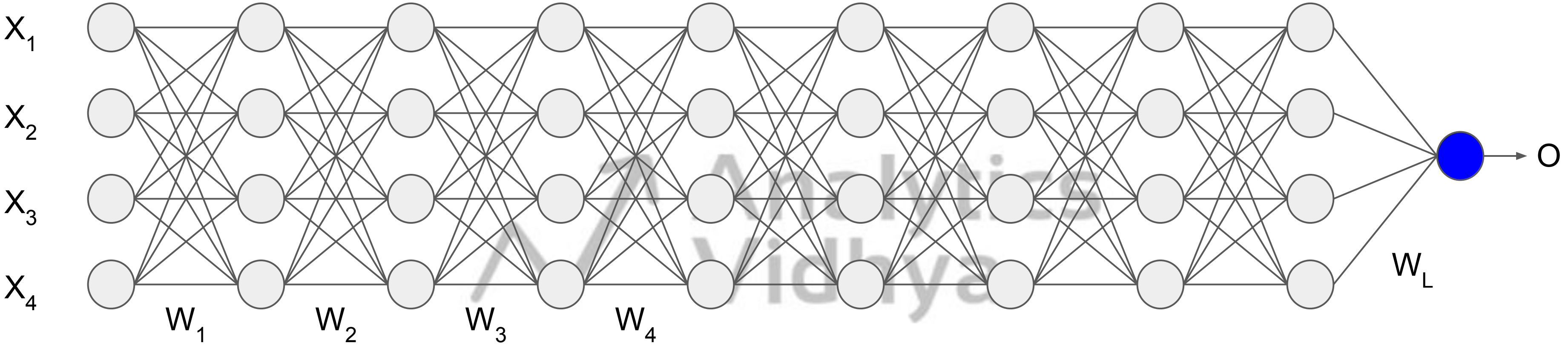


$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}$$

$$L = 10$$

$$O = W_1W_2...W_{L-1}W_LX = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}^{L-1} W_L X$$

Problem: Vanishing / Exploding Gradients

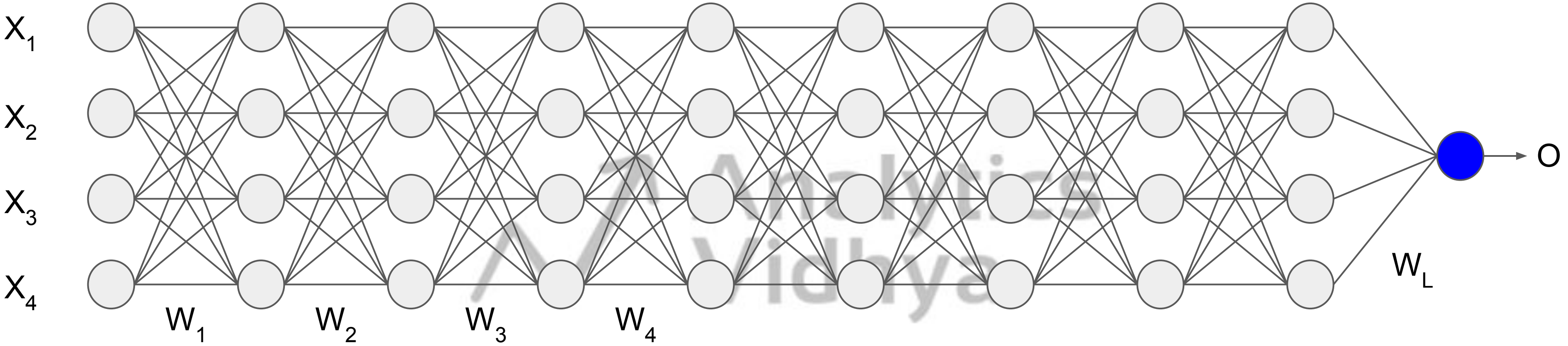


L = 10

$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.5 & 0 & 0 \\ 0 & 0 & 1.5 & 0 \\ 0 & 0 & 0 & 1.5 \end{bmatrix}$$

$$O = W_1W_2 \ldots W_{L-1}W_LX = \begin{bmatrix} 38 & 0 & 0 & 0 \\ 0 & 38 & 0 & 0 \\ 0 & 0 & 38 & 0 \\ 0 & 0 & 0 & 38 \end{bmatrix} W_LX$$

Problem: Vanishing / Exploding Gradients

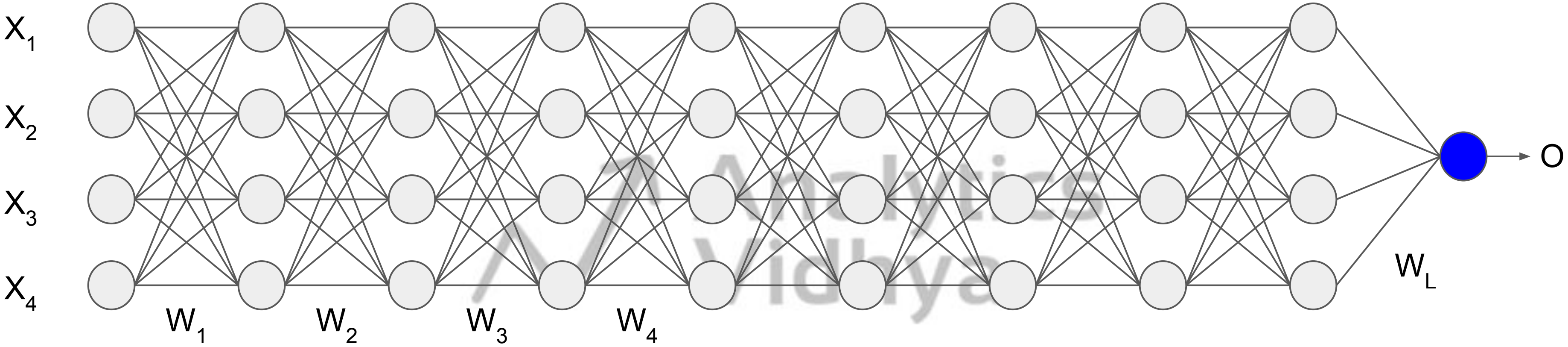


$$W_1 = W_1 = W_{l-1} = \begin{array}{|c|c|c|c|} \hline 1.5 & 0 & 0 & 0 \\ \hline 0 & 1.5 & 0 & 0 \\ \hline 0 & 0 & 1.5 & 0 \\ \hline 0 & 0 & 0 & 1.5 \\ \hline \end{array}$$

Increase Exponentially

$$\boxed{O} = W_1W_2...W_{L-1}W_LX = \begin{array}{|c|c|c|c|} \hline 38 & 0 & 0 & 0 \\ \hline 0 & 38 & 0 & 0 \\ \hline 0 & 0 & 38 & 0 \\ \hline 0 & 0 & 0 & 38 \\ \hline \end{array} \quad W_LX$$

Problem: Vanishing / Exploding Gradients

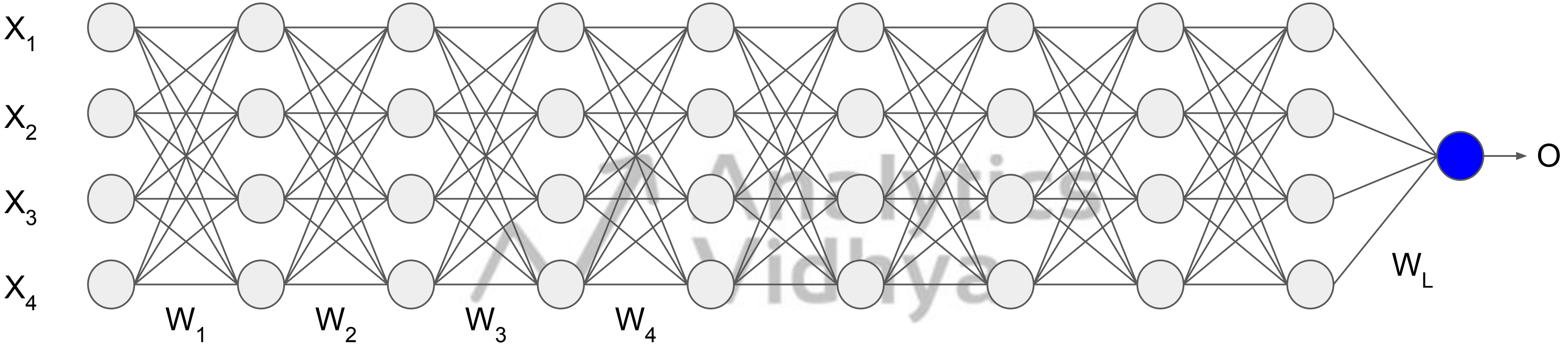


$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

L = 10

$$O = W_1W_2 \ldots W_{L-1}W_LX = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}^{L-1} W_L X$$

Problem: Vanishing / Exploding Gradients

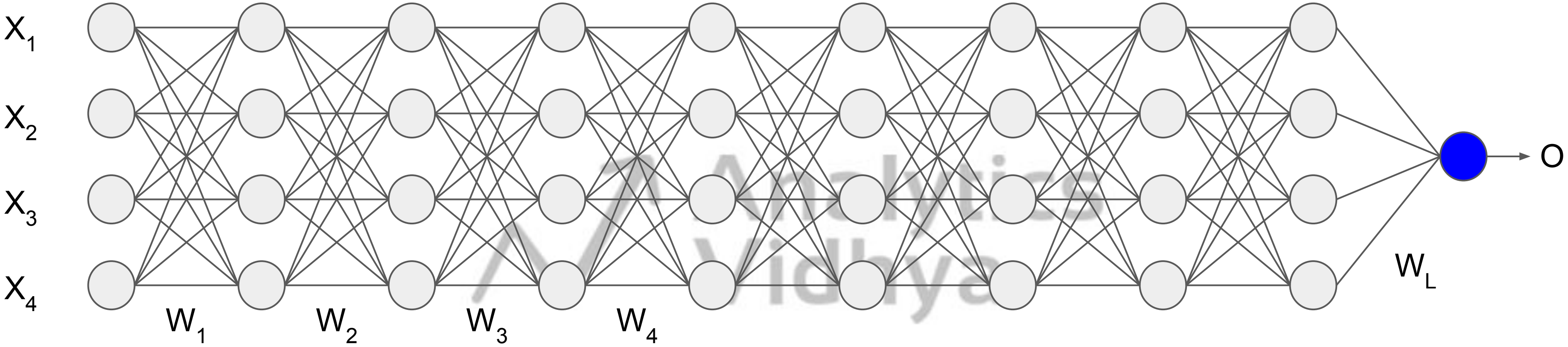


$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

$L = 10$

$$O = W_1 W_2 \ldots W_{L-1} W_L X = \begin{bmatrix} 0.002 & 0 & 0 & 0 \\ 0 & 0.002 & 0 & 0 \\ 0 & 0 & 0.002 & 0 \\ 0 & 0 & 0 & 0.002 \end{bmatrix} W_L X$$

Problem: Vanishing / Exploding Gradients

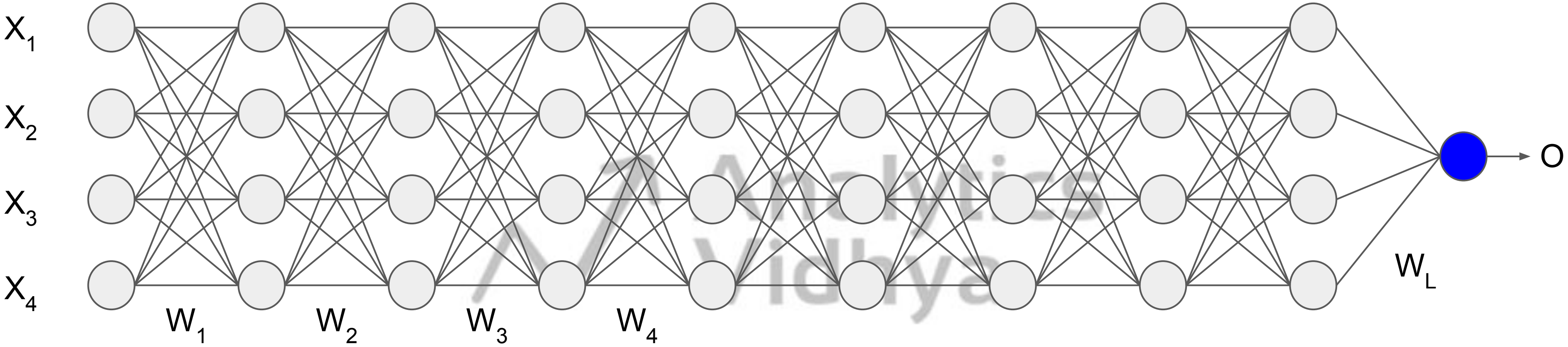


$$W_1 = W_1 = W_{l-1} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

Decrease Exponentially

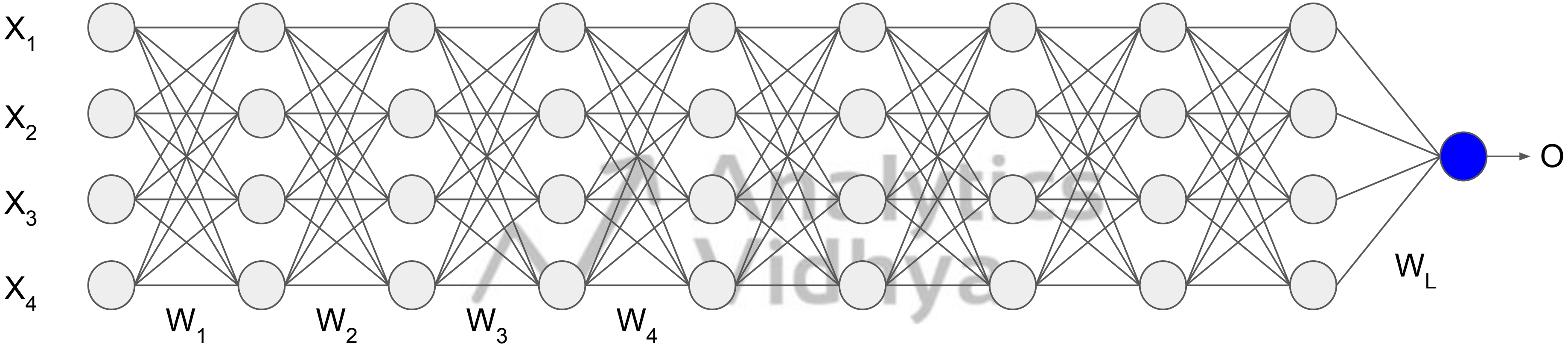
$$\boxed{O} = W_1W_2...W_{L-1}W_LX = \begin{bmatrix} 0.001 & 0 & 0 & 0 \\ 0 & 0.001 & 0 & 0 \\ 0 & 0 & 0.001 & 0 \\ 0 & 0 & 0 & 0.001 \end{bmatrix} W_L X$$

Problem: Vanishing / Exploding Gradients



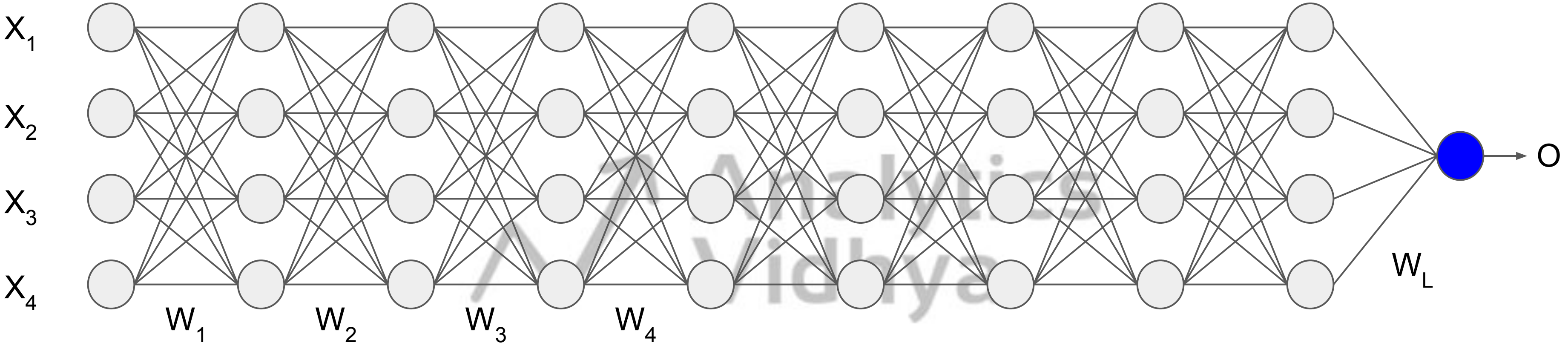
- Similarly: Gradients will also increase / decrease

Problem: Vanishing / Exploding Gradients



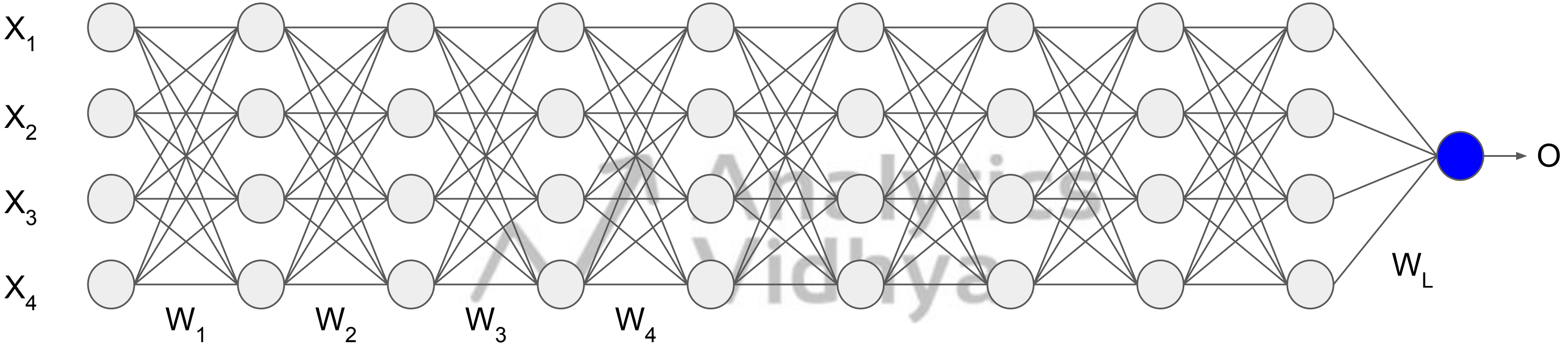
- Similarly: Gradients will also increase / decrease
- **Vanishing Gradients**: Exponentially decreasing gradients

Problem: Vanishing / Exploding Gradients



- Similarly: Gradients will also increase / decrease
- **Vanishing Gradients**: Exponentially decreasing gradients
- **Exploding Gradients**: Exponentially increasing gradients

Problem: Vanishing / Exploding Gradients



- Similarly: Gradients will also increase / decrease
- **Vanishing Gradients**: Exponentially decreasing gradients
- **Exploding Gradients**: Exponentially increasing gradients
- Slows down the training process