



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY
BY EVERYONE'S SMART UNIVERSITY

Mathematical Foundations For Data Science and AI

Lecture 05

AI102

Calculus; Optimization in Neural Networks – Part 1

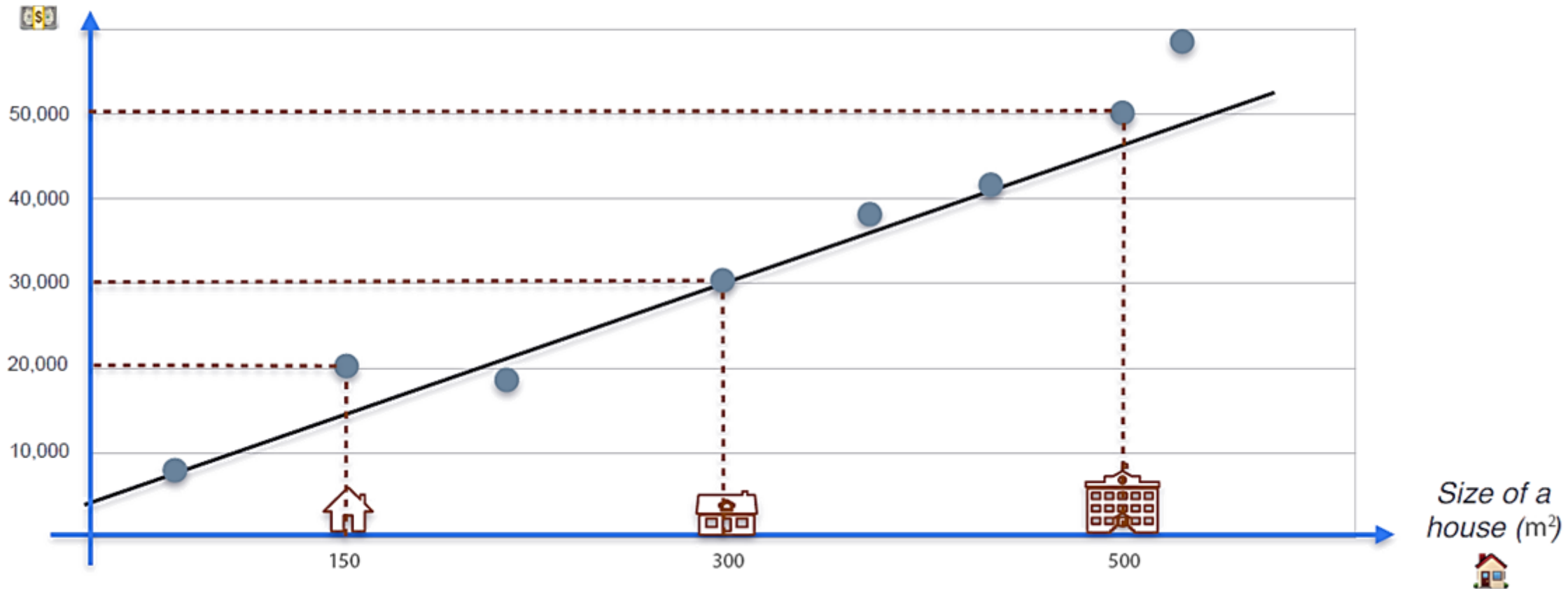
By
Dr. Khaled ELMenshawy

www.midocean.ae

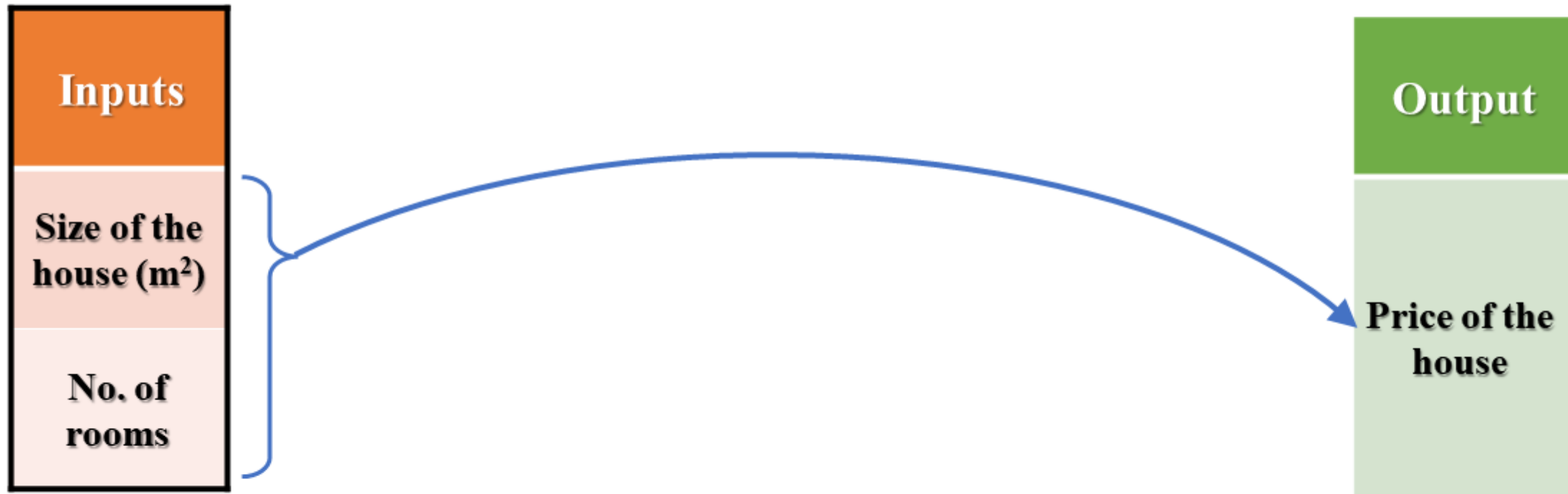
LEARNING OUTCOMES

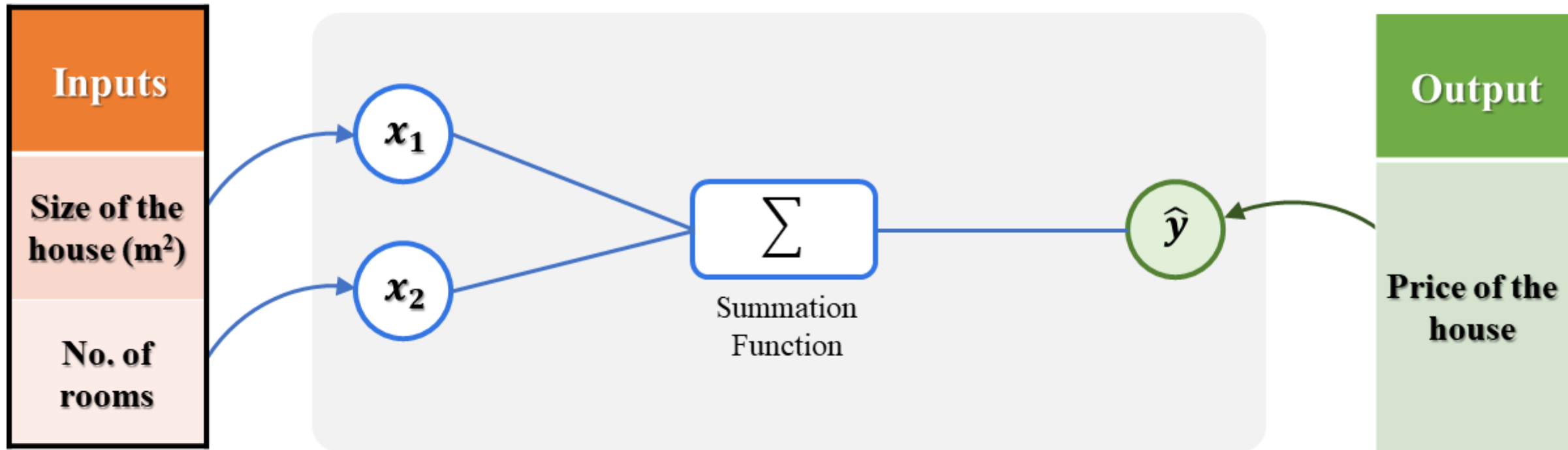
- Understanding Loss Function.
- Understanding Activation Function.
- Understanding Sigmoid Function
- Understanding how to use the Log-Loss Function in optimization.
- Understanding Optimization in Neural Network.

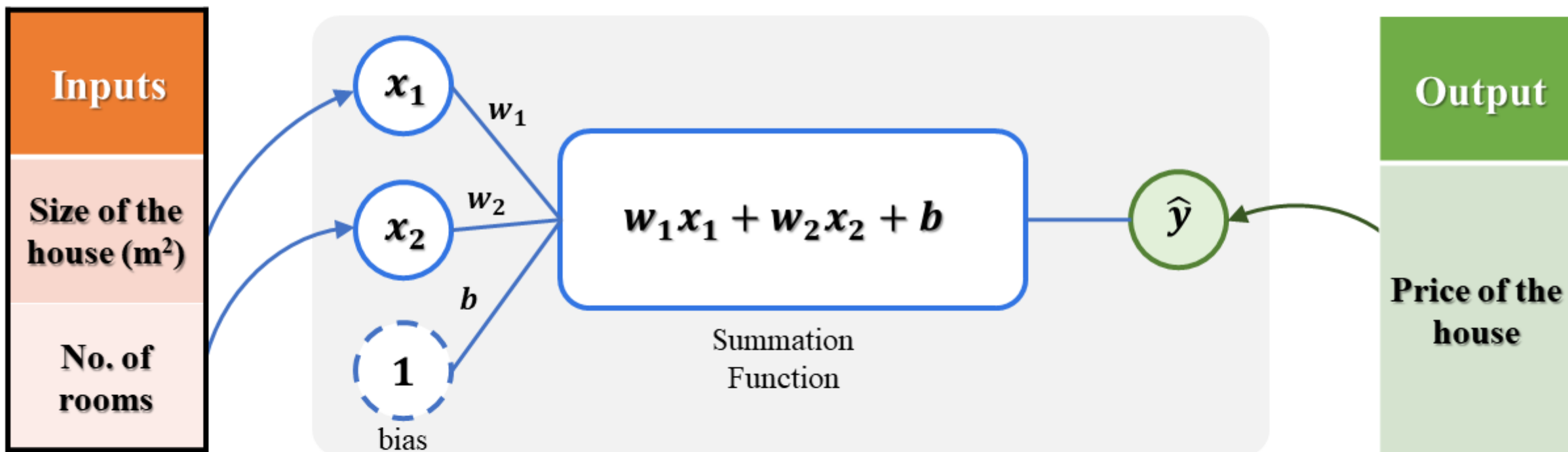
Price of
house (Pound x 100)

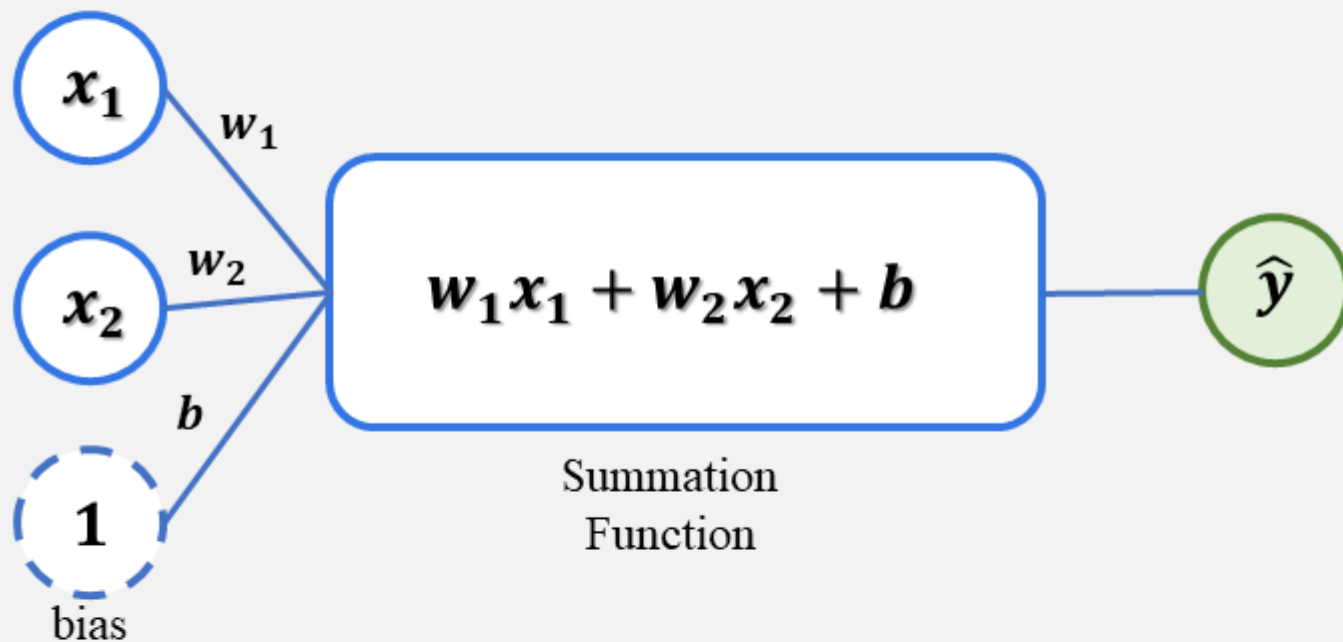


	Size of the house (m ²)	No. of rooms	Price of the house (× 100 Pounds)
	250	3	20,000
	500	5	30,000
	800	8	50,000







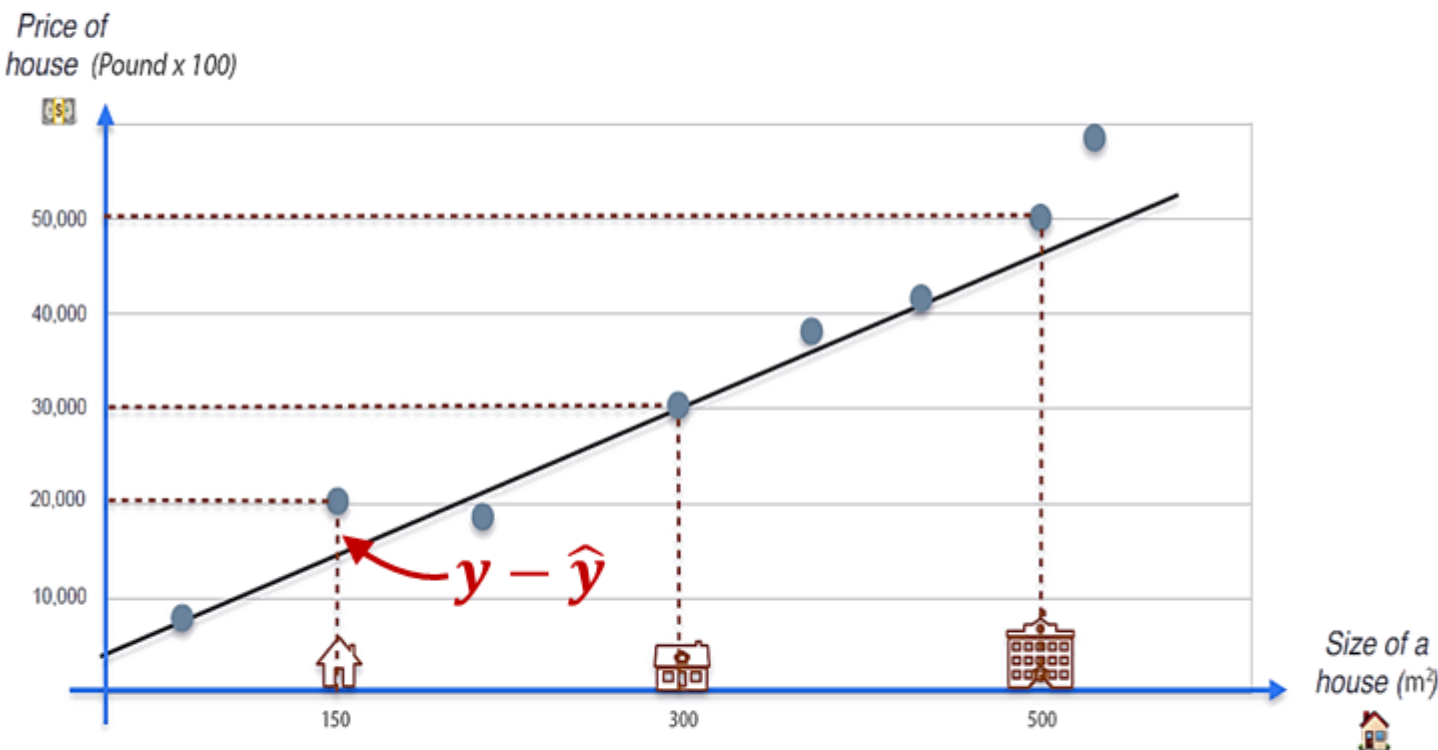


$$\hat{y} = w_1x_1 + w_2x_2 + b$$

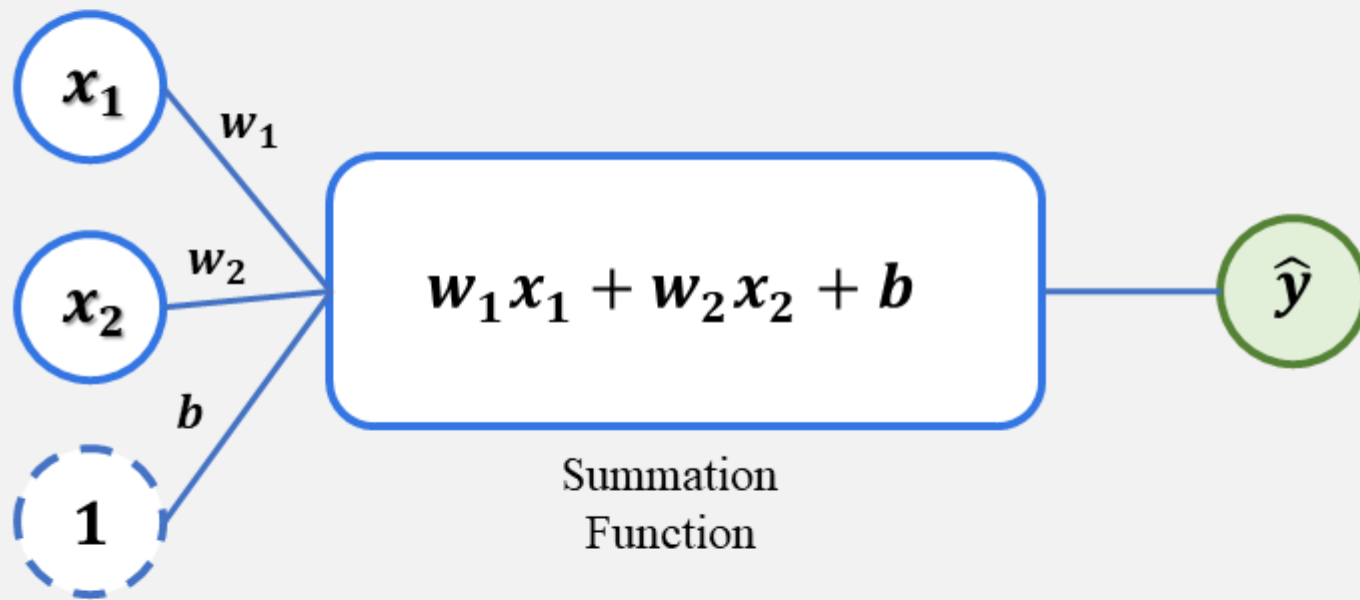
Goal: Find weights and bias that
Reduce the errors in the predictions.

The Loss Function

*A **loss function** is a mathematical function that quantifies the difference between predicted and actual values in a machine learning model.*



	y	$\hat{y}$	$\frac{1}{2} (y - \hat{y})^2$
	20,000	15,000	Error
	30,000	30,000	-
	45,000	50,000	Error



$$\hat{y} = w_1x_1 + w_2x_2 + b$$

The Loss Function:

$$L(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

Goal: Find w_1 , w_2 , and b that give $\hat{y}$ with the least error.

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

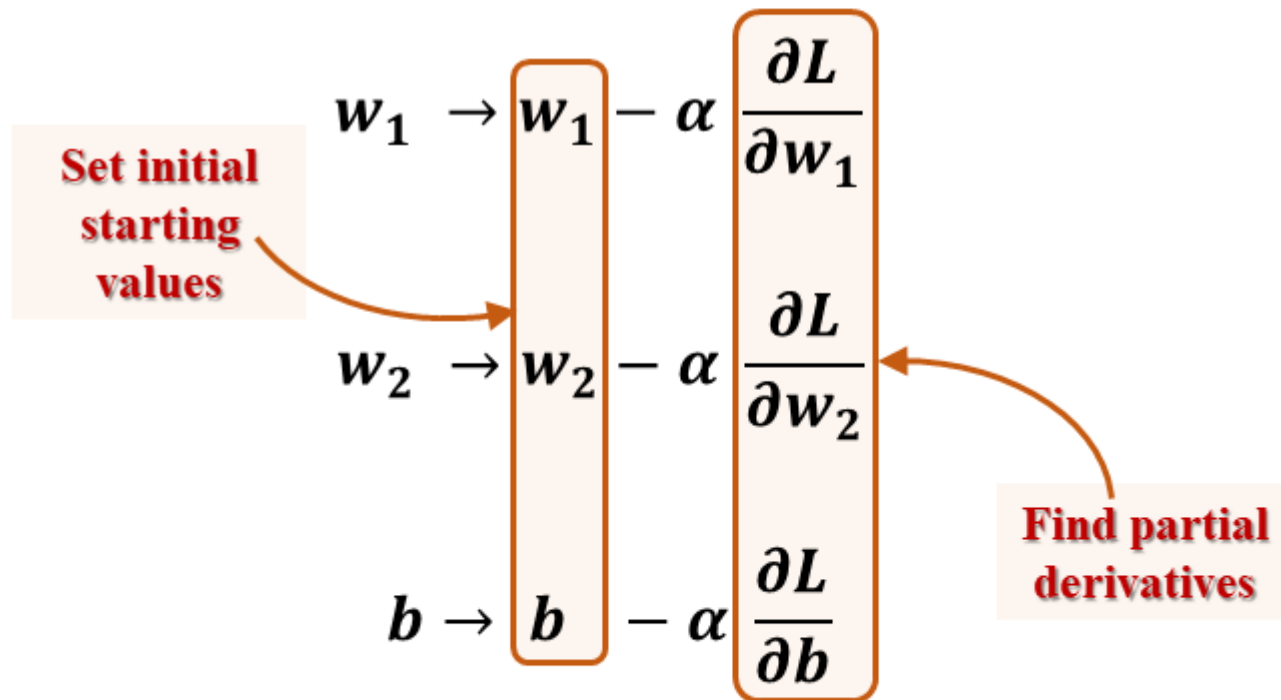
The Loss Function:

$$L(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

Goal: Find w_1 , w_2 , and b that give $\hat{y}$ with the least error.

To find optimal values for w_1 , w_2 , and b

Using *gradient descent* and calculate:



Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

The Loss Function:

$$L(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

Using Chain Rule:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b}$$

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

The Loss Function:

$$L(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

$\frac{\partial L}{\partial \hat{y}}$	$\frac{1}{2} (y - \hat{y})^2$	$-(y - \hat{y})$
$\frac{\partial \hat{y}}{\partial w_1}$	$w_1 x_1 + w_2 x_2 + b$	x_1
$\frac{\partial \hat{y}}{\partial w_2}$	$w_1 x_1 + w_2 x_2 + b$	x_2
$\frac{\partial \hat{y}}{\partial b}$	$w_1 x_1 + w_2 x_2 + b$	1

Prediction Function:

$$\hat{y} = w_1 x_1 + w_2 x_2 + b$$

The Loss Function:

$$L(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$$

$\frac{\partial L}{\partial w_1}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1}$	$-x_1(y - \hat{y})$
$\frac{\partial L}{\partial w_2}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2}$	$-x_2(y - \hat{y})$
$\frac{\partial L}{\partial b}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b}$	$-(y - \hat{y})$

Goal: Find the optimal values for w_1 , w_2 , and b that give $\hat{y}$ with the least error.

Perform Gradient Descent

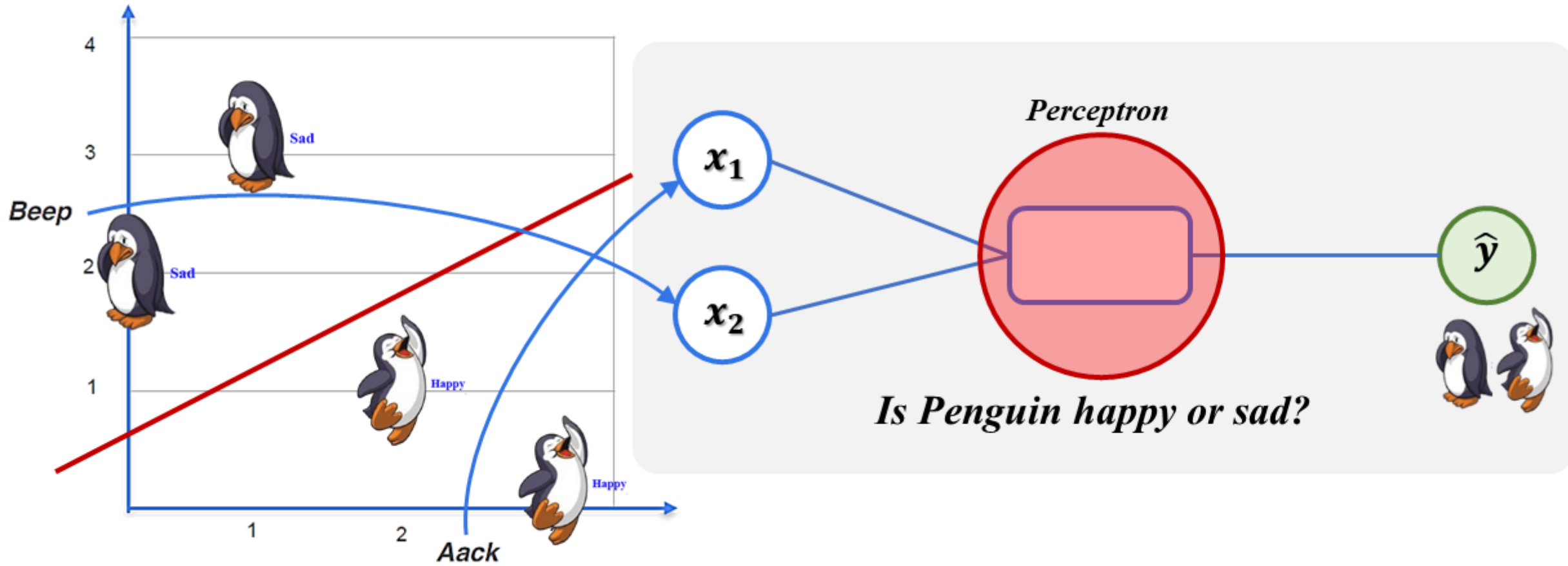
$$\begin{aligned}w_1 &= w_1 - \alpha \frac{\partial L}{\partial w_1} \\ &= w_1 - \alpha (-x_1(y - \hat{y}))\end{aligned}$$

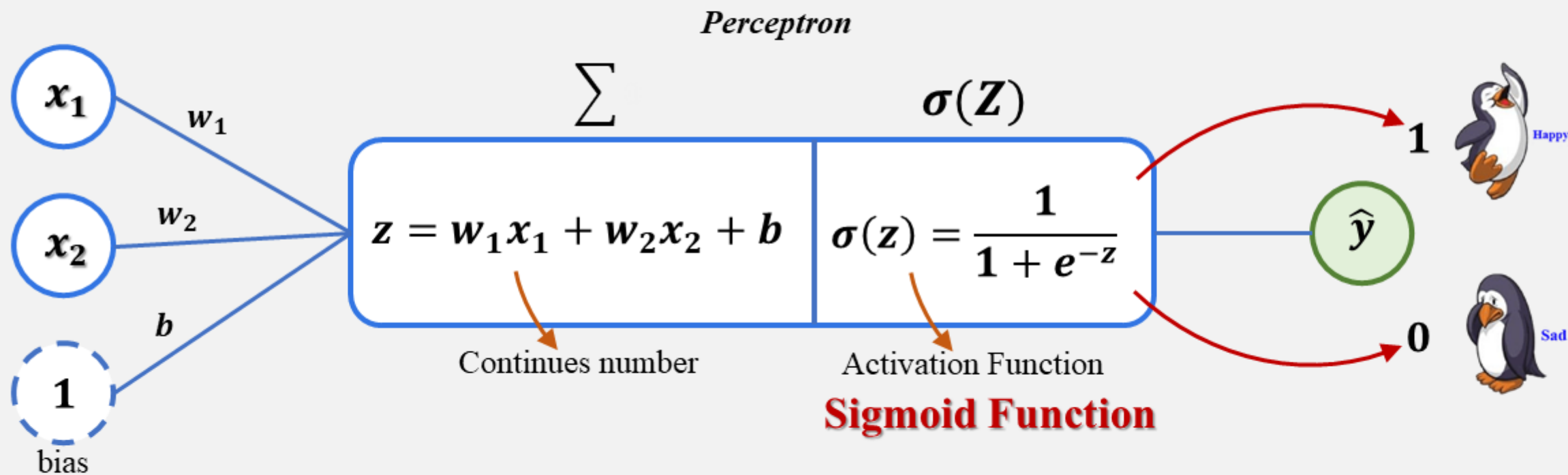
$$\begin{aligned}w_2 &= w_2 - \alpha \frac{\partial L}{\partial w_2} \\ &= w_2 - \alpha (-x_2(y - \hat{y}))\end{aligned}$$

$$\begin{aligned}b &= b - \alpha \frac{\partial L}{\partial b} \\ &= b - \alpha (-(y - \hat{y}))\end{aligned}$$

Activation Function

Sound of Penguin	Aack	Beep	Mood
Aack Aack Aack	3	0	 Happy
Beep Beep	0	2	 Sad
Aack Beep Beep Beep	1	3	 Sad
Aack Beep Aack	2	1	 Happy



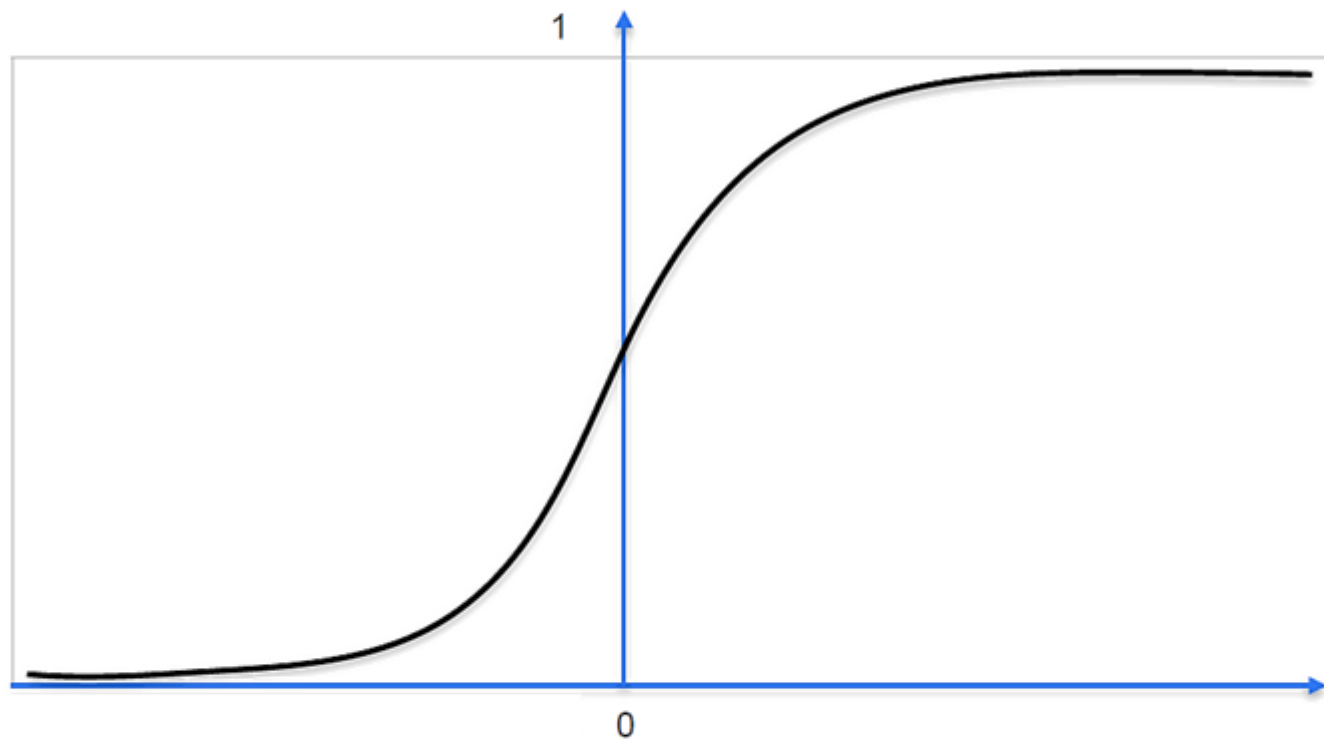


The sigmoid function takes the entire number line and crunches into the interval zero, one, which is what we want for the output.

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

Or;

$$\sigma(z) = (1 + e^{-z})^{-1}$$



$$\sigma(z) = (1 + e^{-z})^{-1}$$

$$\frac{d}{dz} \sigma(z) = -1 (1 + e^{-z})^{-1-1} \times \left(\frac{d}{dz} (1 + e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \times \left(\frac{d}{dz} (1) + \frac{d}{dz} (e^{-z}) \right)$$

$$= -1 (1 + e^{-z})^{-2} \times \left((0) + e^{-z} \frac{d}{dz} (-z) \right)$$

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

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

Sigmoid Function

$$\frac{d}{dz} \sigma(z) = (1 + e^{-z})^{-2} (e^{-z})$$

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

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

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

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

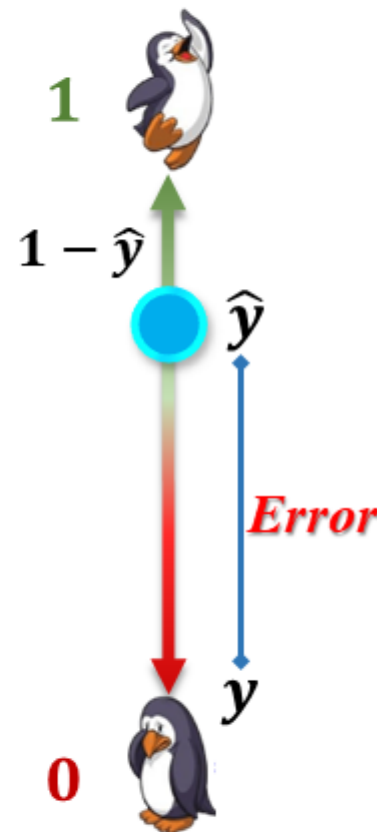
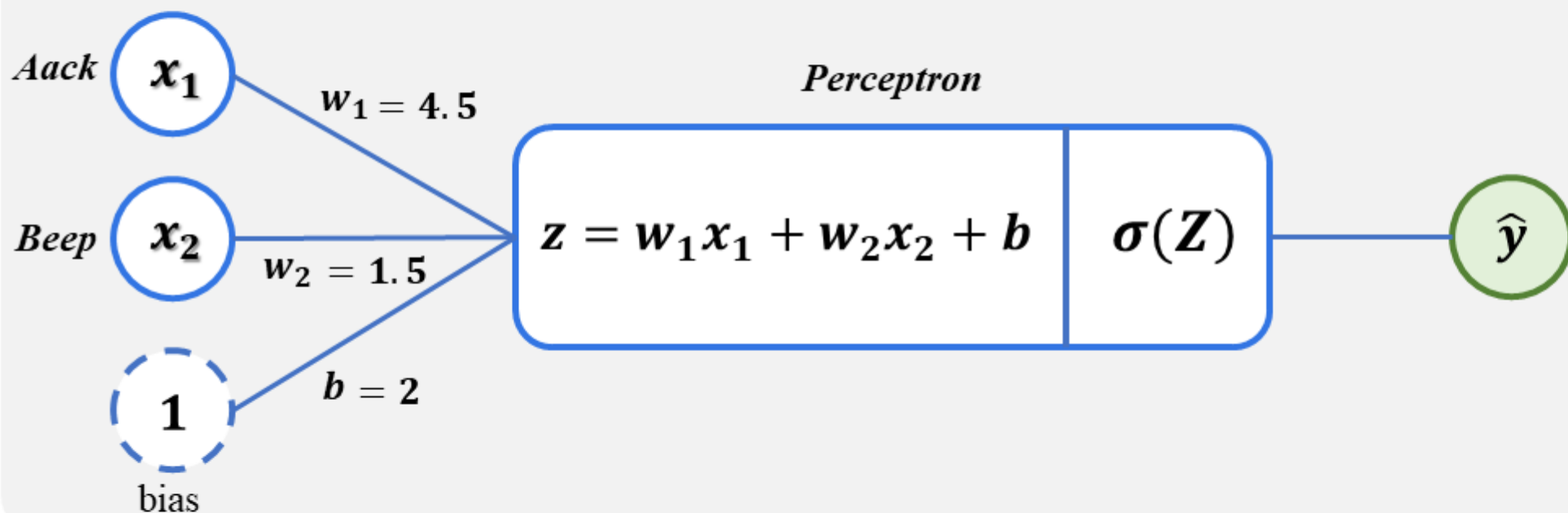
Sigmoid Function

$$\frac{d}{dz} \sigma(z) = \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1 + e^{-z}} \right)$$

Since $\sigma(z) = \frac{1}{1 + e^{-z}}$

$$\frac{d}{dz} \sigma(z) = \sigma(z)(1 - \sigma(z))$$

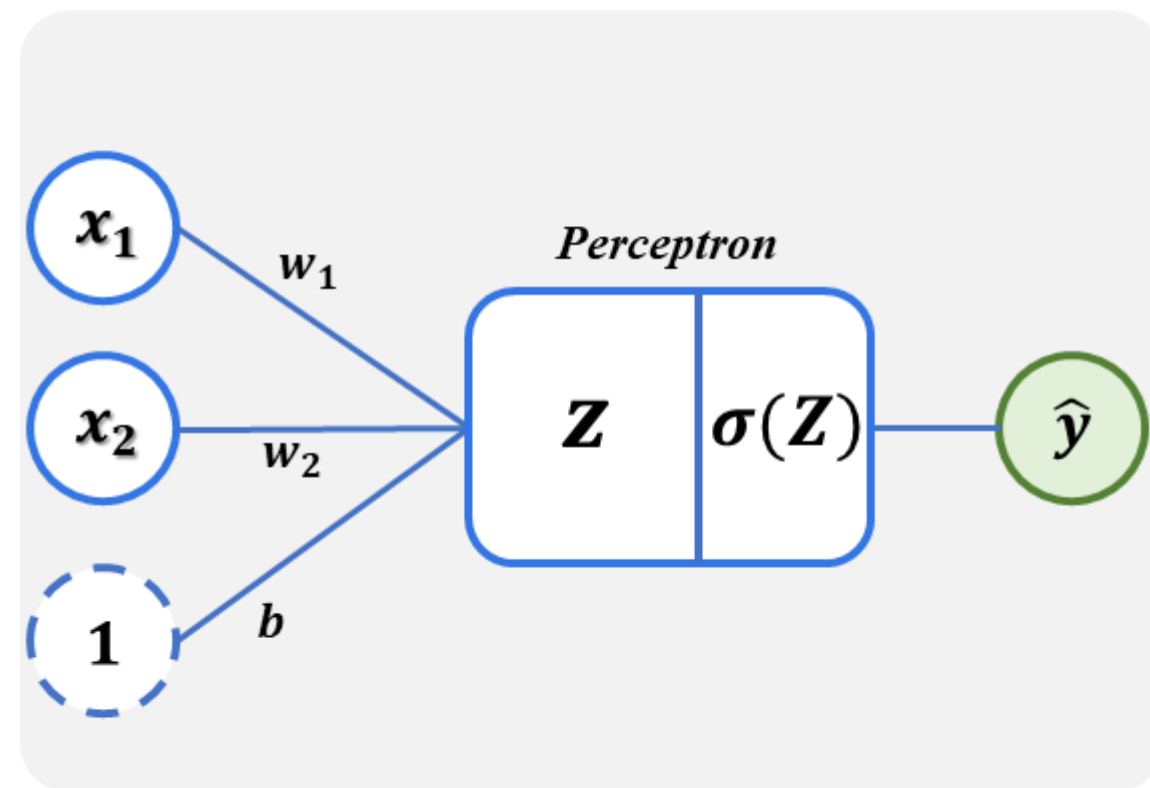
Aack Beep Beep Beep



Apply Log-Loss Function $L(y, \hat{y})$

Log-loss Function is indicative of how close the prediction probability is to the corresponding actual/true value

Using the Log-Loss Function

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

Log-Loss Function:

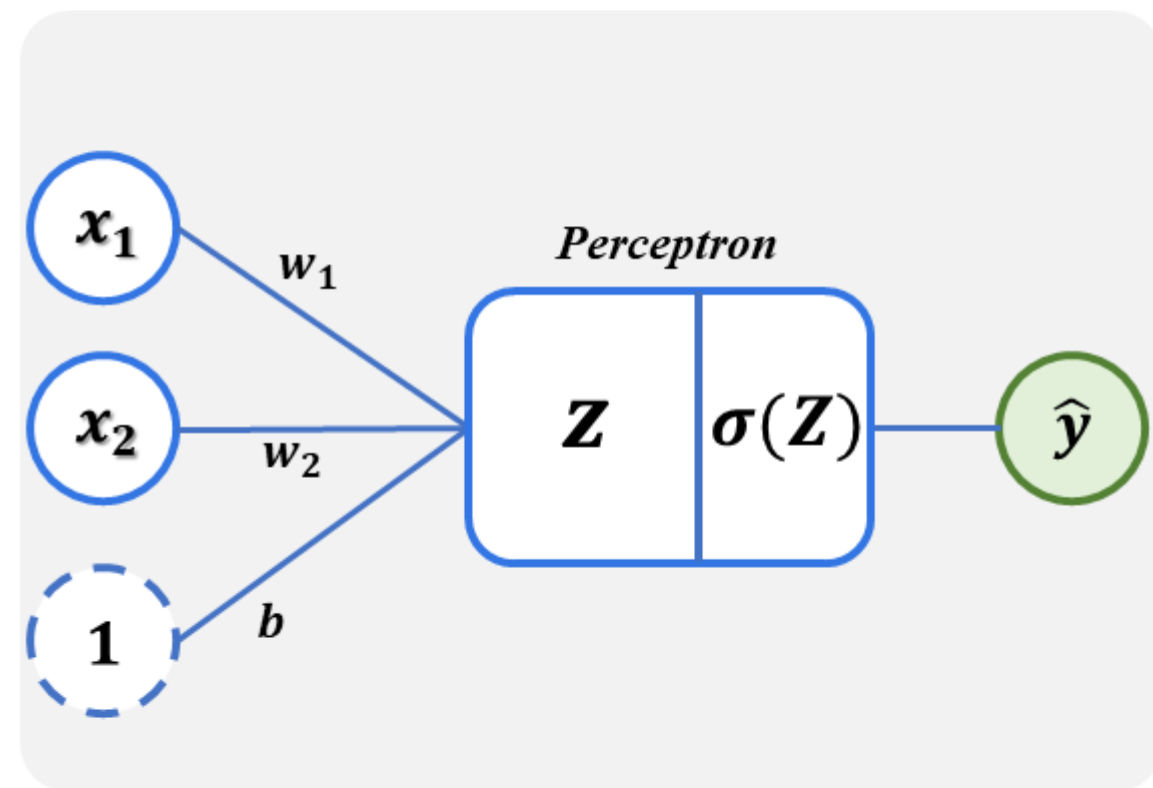
$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

If $L(y, \hat{y})$ is large number, the difference between y and $\hat{y}$ is large, and If $L(y, \hat{y})$ is small number, the difference between y and $\hat{y}$ is small.

Goal:

Find the optimal values for w_1 , w_2 , and b that give $\hat{y}$ with the least error.

Using the Log-Loss Function



1

To find optimal values for w_1 , w_2 , and b

Using gradient descent and calculate:

Set initial
starting values $L(y, \hat{y})$ 

0

$$\begin{aligned}
 w_1 &\rightarrow w_1 - \alpha \frac{\partial L}{\partial w_1} \\
 w_2 &\rightarrow w_2 - \alpha \frac{\partial L}{\partial w_2} \\
 b &\rightarrow b - \alpha \frac{\partial L}{\partial b}
 \end{aligned}$$

Find partial
derivatives

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Using Chain Rule:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b}$$

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y}{\hat{y}} - \frac{(1 - y)}{(1 - \hat{y})}$$

$$= \frac{-y + y\hat{y} + \hat{y} - y\hat{y}}{\hat{y}(1 - \hat{y})}$$

$$= \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$\frac{d}{dz} \sigma(z)$	$= \sigma(z)(1 - \sigma(z))$ $= \hat{y}(1 - \hat{y})$	
$\frac{\partial \hat{y}}{\partial w_1}$	$\sigma(w_1x_1 + w_2x_2 + b)$	$\hat{y}(1 - \hat{y})x_1$
$\frac{\partial \hat{y}}{\partial w_2}$	$\sigma(w_1x_1 + w_2x_2 + b)$	$\hat{y}(1 - \hat{y})x_2$
$\frac{\partial \hat{y}}{\partial b}$	$\sigma(w_1x_1 + w_2x_2 + b)$	$\hat{y}(1 - \hat{y})$

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})x_1$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})x_2$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b} = \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})$$

Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$\frac{\partial L}{\partial w_1}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1}$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})x_1$
$\frac{\partial L}{\partial w_2}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2}$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})x_2$
$\frac{\partial L}{\partial b}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b}$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})} \times \hat{y}(1 - \hat{y})$

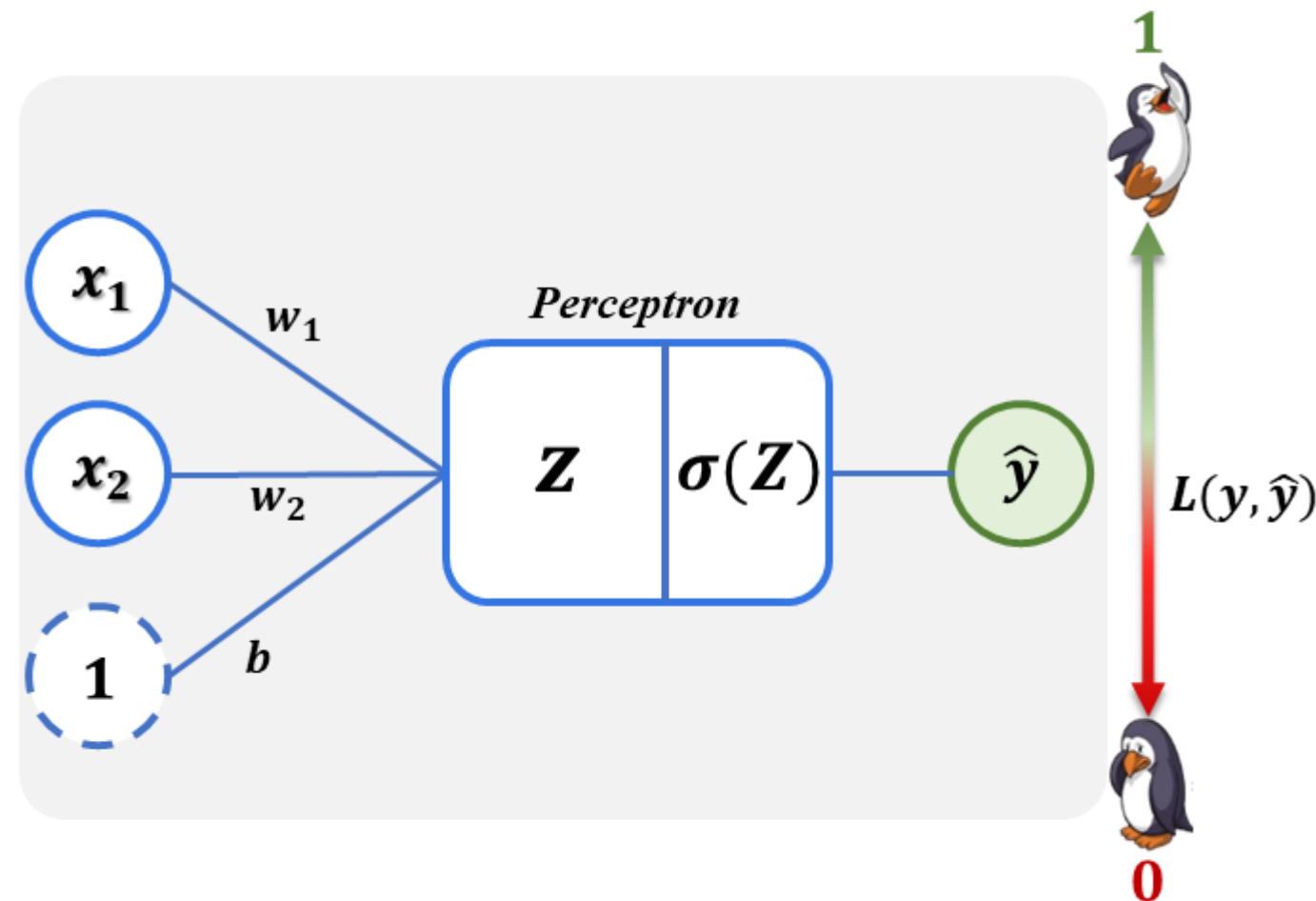
Prediction Function:

$$\hat{y} = \sigma(w_1x_1 + w_2x_2 + b)$$

The Log-Loss Function:

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

$\frac{\partial L}{\partial w_1}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_1}$	$-(y - \hat{y})x_1$
$\frac{\partial L}{\partial w_2}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_2}$	$-(y - \hat{y})x_2$
$\frac{\partial L}{\partial b}$	$= \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial b}$	$-(y - \hat{y})$



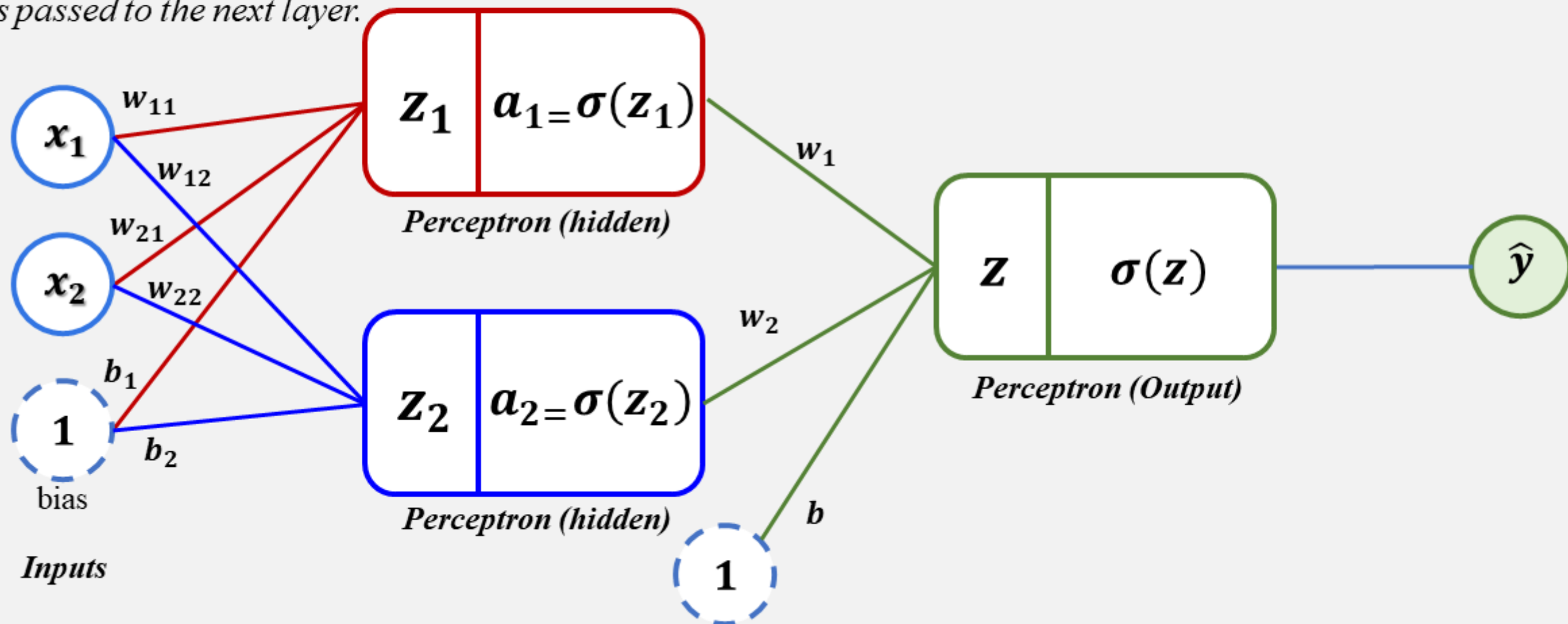
Perform Gradient Descent

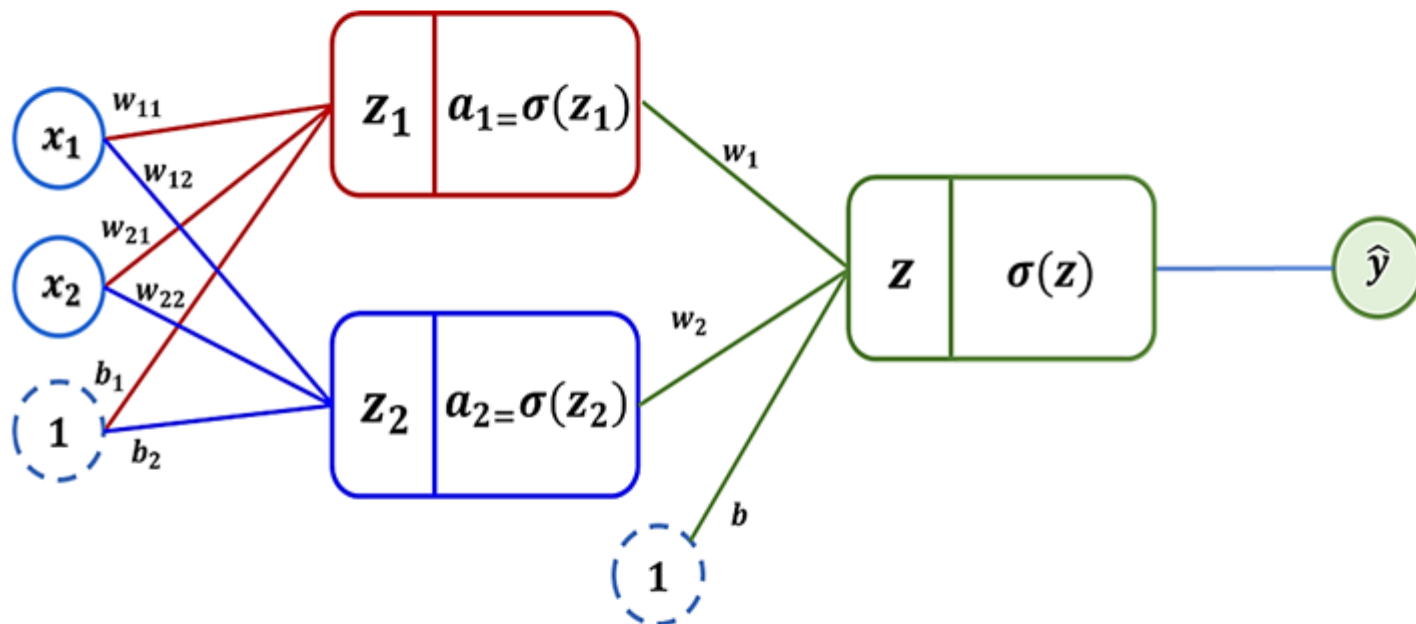
$$\begin{aligned} w_1 &= w_1 - \alpha \frac{\partial L}{\partial w_1} \\ &= w_1 - \alpha (-x_1 (y - \hat{y})) \end{aligned}$$

$$\begin{aligned} w_2 &= w_2 - \alpha \frac{\partial L}{\partial w_2} \\ &= w_2 - \alpha (-x_2 (y - \hat{y})) \end{aligned}$$

$$\begin{aligned} b &= b - \alpha \frac{\partial L}{\partial b} \\ &= b - \alpha (-(y - \hat{y})) \end{aligned}$$

A neural network is just a bunch of Perceptrons organized in layers where the information of these perceptions is passed to the next layer.





$$a_1 = \sigma(z_1)$$

$$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$$

$$a_2 = \sigma(z_2)$$

$$z_2 = x_1 w_{12} + x_2 w_{22} + b_2$$

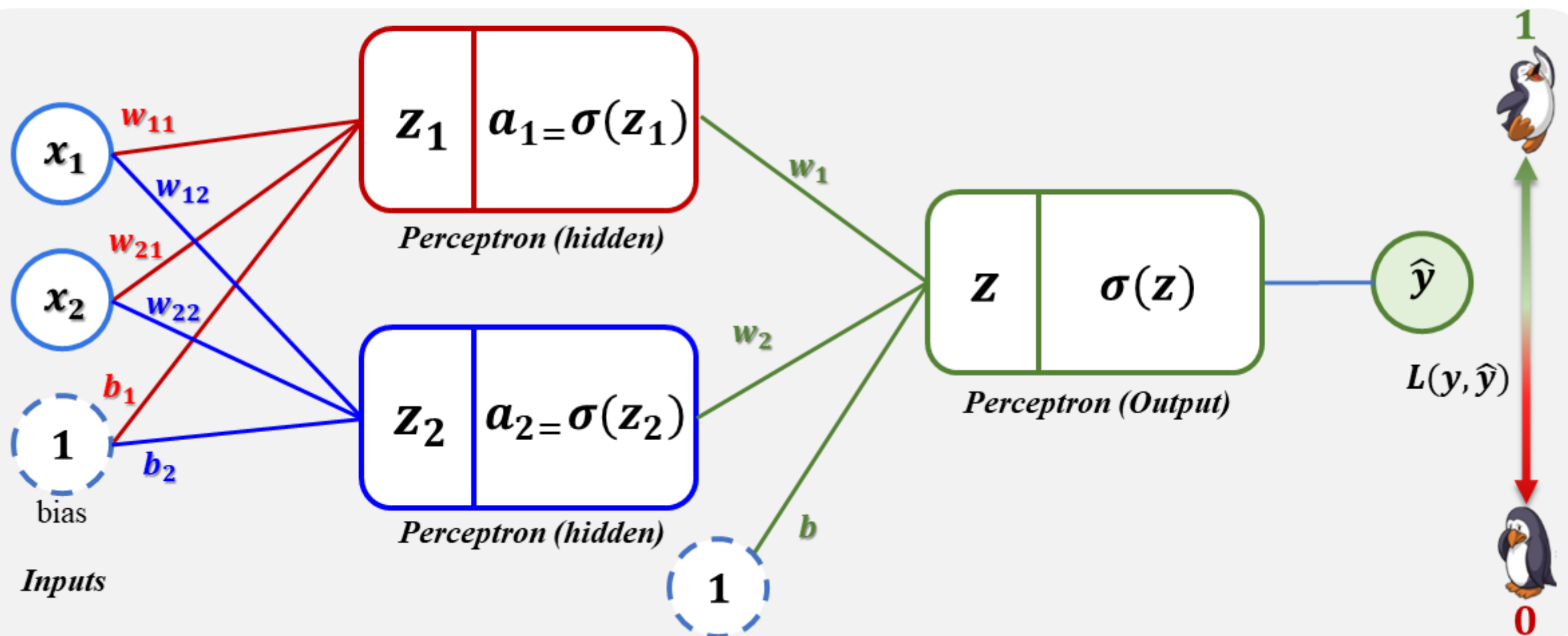
$$\hat{y} = \sigma(z)$$

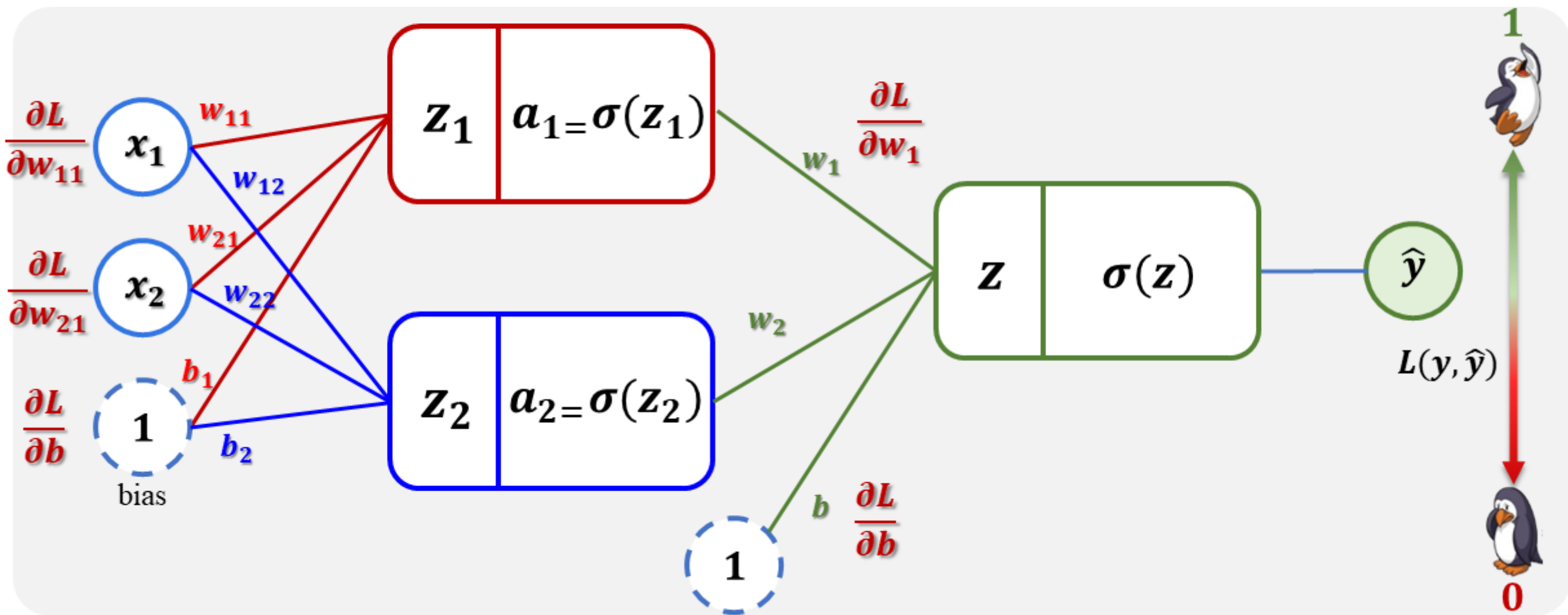
$$z = a_1 w_1 + a_2 w_2 + b$$

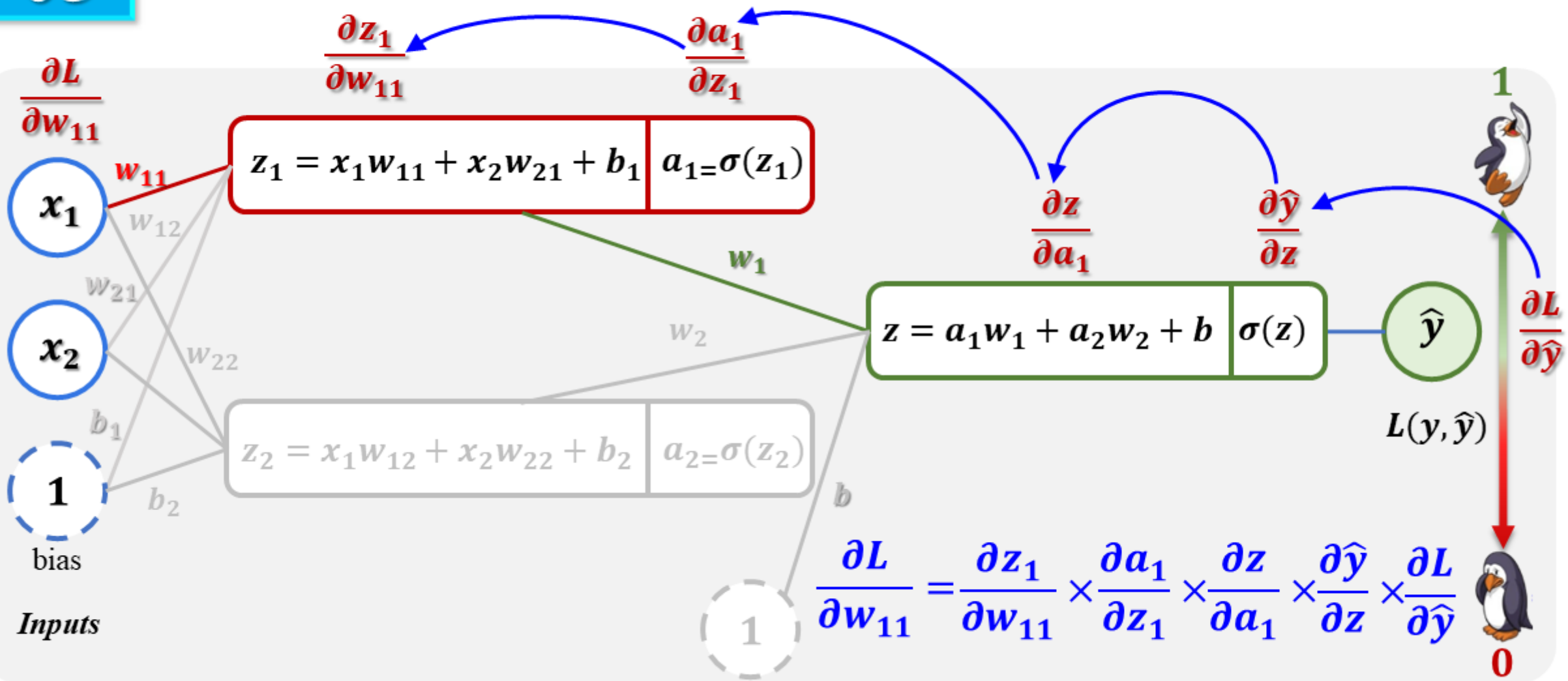
Log-Loss Function

$$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$$

Goal: Adjust every weight ($\mathbf{w}$) and bias ($\mathbf{b}$) in order to reduce the loss function $L(\mathbf{y}, \hat{\mathbf{y}})$







$\frac{\partial z_1}{\partial w_{11}}$	$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$	x_1
$\frac{\partial a_1}{\partial z_1}$	$a_1 = \sigma(z_1)$	$a_1(1 - a_1)$
$\frac{\partial z}{\partial a_1}$	$z_1 = a_1 w_1 + a_2 w_2 + b$	w_1
$\frac{\partial \hat{y}}{\partial z}$	$\hat{y} = \sigma(z)$	$\hat{y}(1 - \hat{y})$
$\frac{\partial L}{\partial \hat{y}}$	$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$

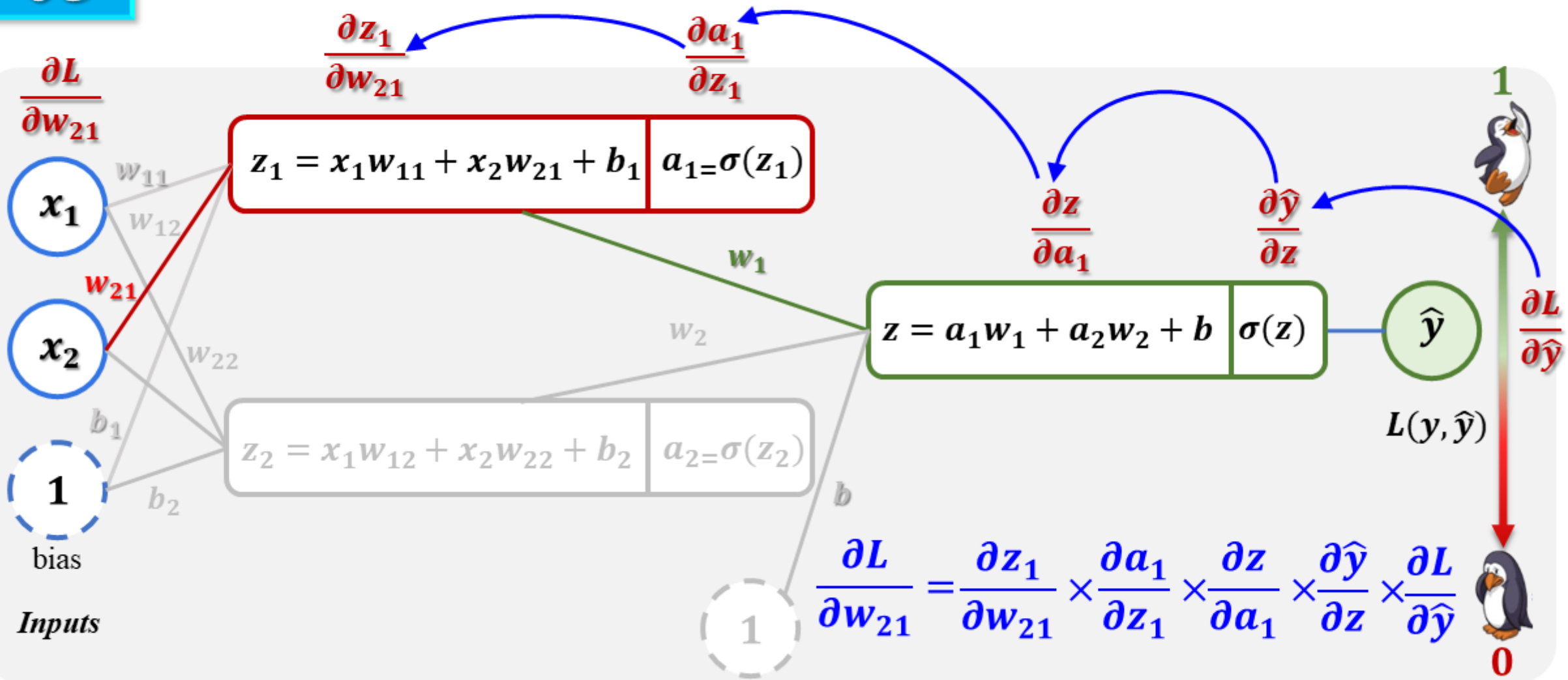
$$\frac{\partial L}{\partial w_{11}} = \frac{\partial z_1}{\partial w_{11}} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z}{\partial a_1} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{11}} = x_1 * a_1(1 - a_1) * w_1 * \hat{y}(1 - \hat{y}) * \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_{11}} = -x_1 w_1 a_1(1 - a_1)(y - \hat{y})$$

Perform Gradient Descent

$$w_{11} \rightarrow w_{11} - \underbrace{\alpha (-x_1 w_1 a_1(1 - a_1)(y - \hat{y}))}_{\frac{\partial L}{\partial w_{11}}}$$



$\frac{\partial z_1}{\partial w_{21}}$	$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$	x_2
$\frac{\partial a_1}{\partial z_1}$	$a_1 = \sigma(z_1)$	$a_1(1 - a_1)$
$\frac{\partial z}{\partial a_1}$	$z_1 = a_1 w_1 + a_2 w_2 + b$	w_1
$\frac{\partial \hat{y}}{\partial z}$	$\hat{y} = \sigma(z)$	$\hat{y}(1 - \hat{y})$
$\frac{\partial L}{\partial \hat{y}}$	$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$

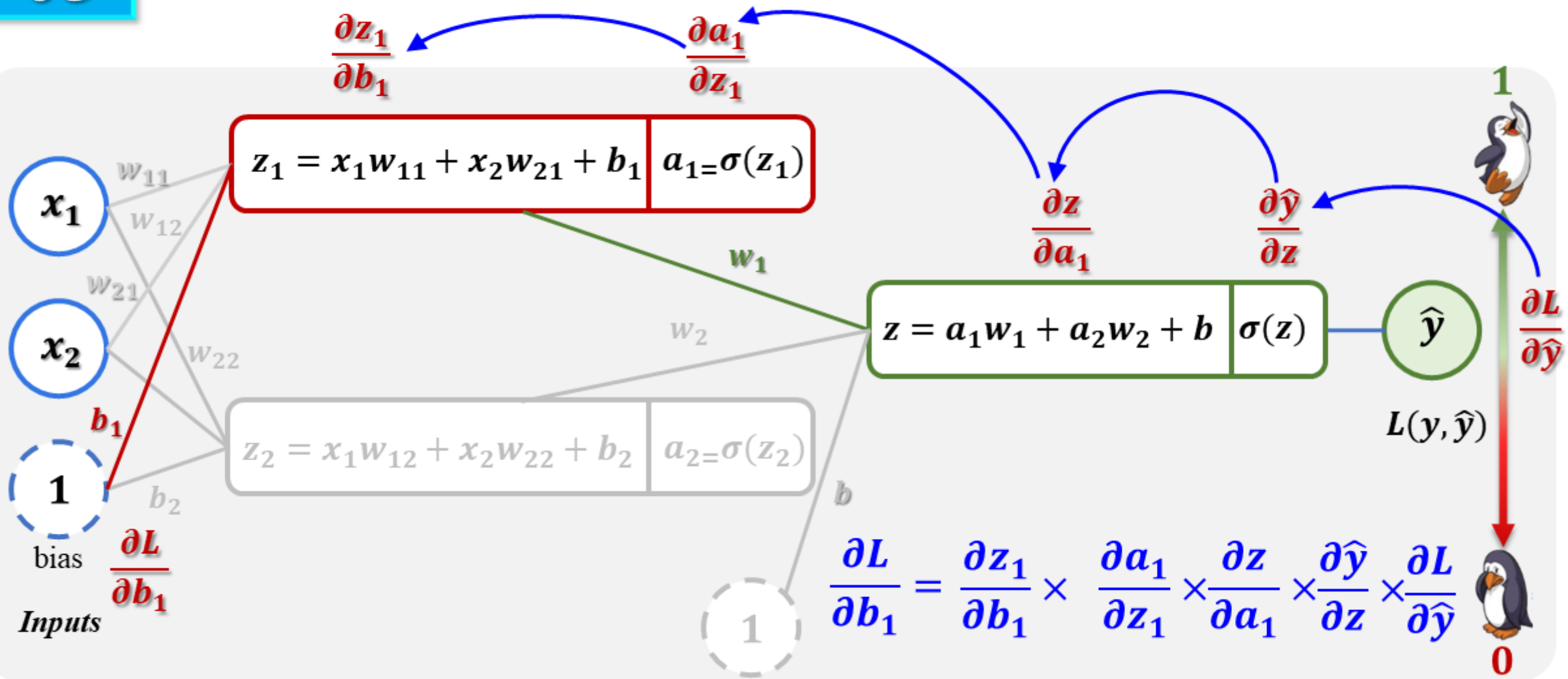
$$\frac{\partial L}{\partial w_{21}} = \frac{\partial z_1}{\partial w_{21}} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z}{\partial a_1} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_{21}} = x_2 * a_1(1 - a_1) * w_1 * \hat{y}(1 - \hat{y}) * \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_{21}} = -x_2 w_1 a_1(1 - a_1)(y - \hat{y})$$

Perform Gradient Descent

$$w_{21} \rightarrow w_{21} - \underbrace{\alpha \left(-x_2 w_1 a_1(1 - a_1)(y - \hat{y}) \right)}_{\frac{\partial L}{\partial w_{21}}}$$



$\frac{\partial z_1}{\partial b_1}$	$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$	1
$\frac{\partial a_1}{\partial z_1}$	$a_1 = \sigma(z_1)$	$a_1(1 - a_1)$
$\frac{\partial z}{\partial a_1}$	$z_1 = a_1 w_1 + a_2 w_2 + b$	w_1
$\frac{\partial \hat{y}}{\partial z}$	$\hat{y} = \sigma(z)$	$\hat{y}(1 - \hat{y})$
$\frac{\partial L}{\partial \hat{y}}$	$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$

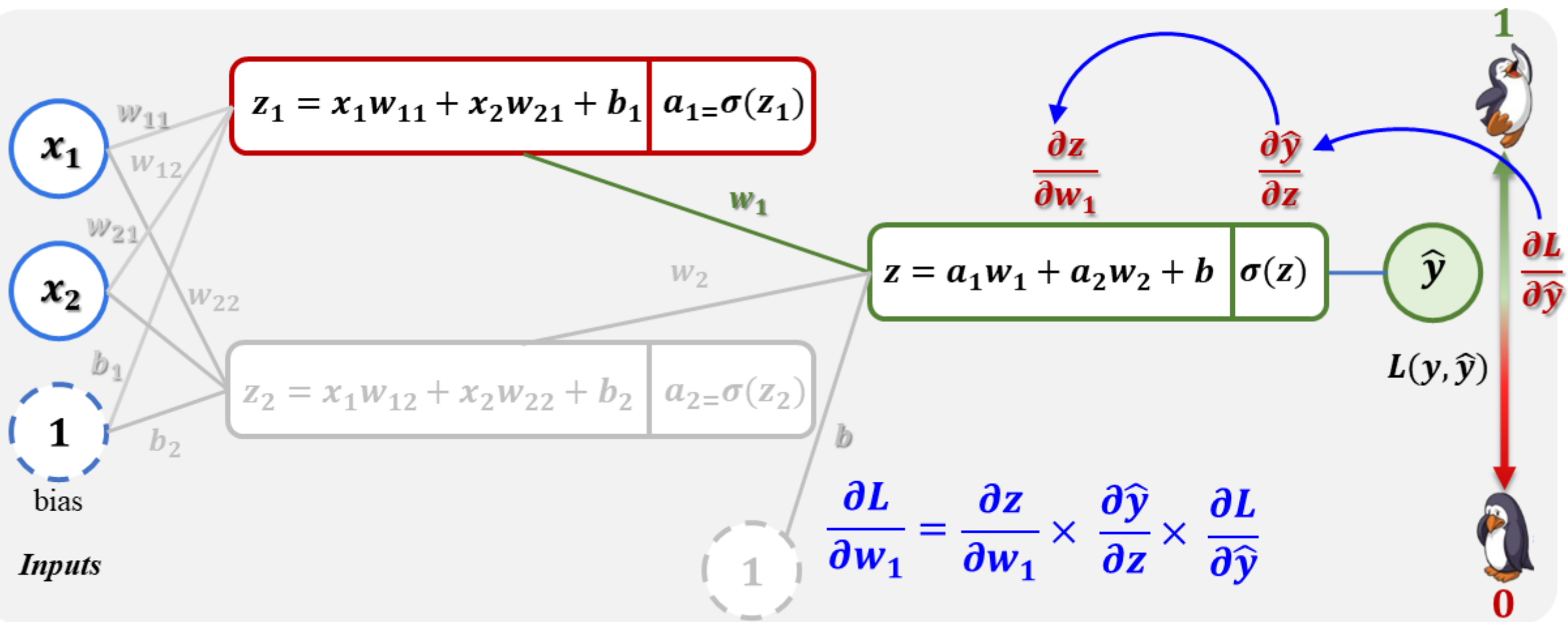
$$\frac{\partial L}{\partial b_1} = \frac{\partial z_1}{\partial b_1} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z}{\partial a_1} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial b_1} = 1 * a_1(1 - a_1) * w_1 * \hat{y}(1 - \hat{y}) * \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial b_1} = -w_1 a_1(1 - a_1)(y - \hat{y})$$

Perform Gradient Descent

$$b_1 \rightarrow b_1 - \alpha \underbrace{(-w_1 a_1(1 - a_1)(y - \hat{y}))}_{\frac{\partial L}{\partial b_1}}$$



$\frac{\partial z}{\partial w_1}$	$z = a_1 w_1 + a_2 w_2 + b$	a_1
$\frac{\partial \hat{y}}{\partial z}$	$\hat{y} = \sigma(z)$	$\hat{y}(1 - \hat{y})$
$\frac{\partial L}{\partial \hat{y}}$	$L(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$	$\frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$

$$\frac{\partial L}{\partial w_1} = \frac{\partial z}{\partial w_1} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial L}{\partial \hat{y}}$$

$$\frac{\partial L}{\partial w_1} = a_1 * \hat{y}(1 - \hat{y}) * \frac{-(y - \hat{y})}{\hat{y}(1 - \hat{y})}$$

$$\frac{\partial L}{\partial w_1} = -a_1 (y - \hat{y})$$

Perform Gradient Descent

$$w_1 \rightarrow w_1 - \underbrace{\alpha (-a_1 (y - \hat{y}))}_{\frac{\partial L}{\partial w_1}}$$

Rule	Logarithm	Equal to
<i>Product Rule</i>	$\log xy$	$\log x + \log y$
<i>Division Rule</i>	$\log \frac{x}{y}$	$\log x - \log y$
<i>Exponential Rule</i>	$\log x^c$	$c \log x$

Method	Rule
The Gradient of $f(x, y)$	$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$
Gradient Descent	$x_{new} = x_{old} - \alpha f'(x_{old})$

