Neural Network 101

Rafael Geurgas

02/07/25

PURDUE UNIVERSITY

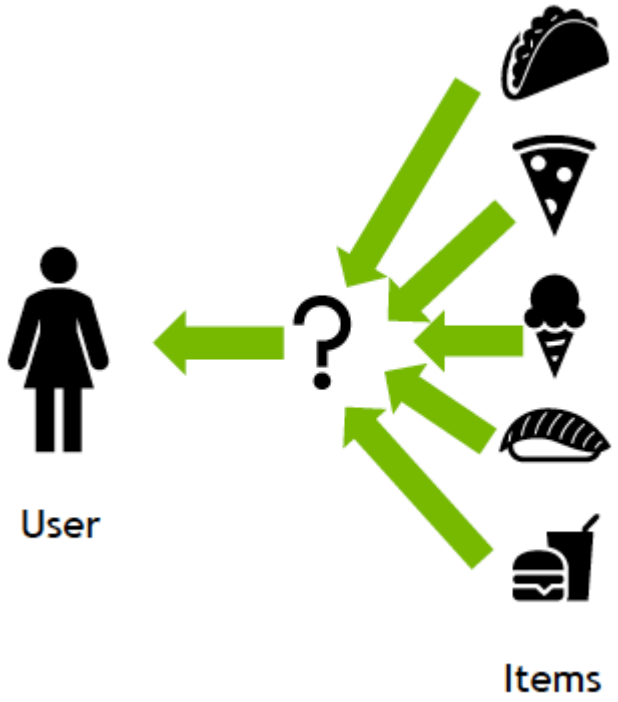
SCAN ME

Artificial Intelligence

- Artificial intelligence (AI): imitating human intelligence or behavioral patterns or any other living entity;

- Machine Learning (ML) is a branch of AI that focuses on the development of algorithms and statistical models that enable computers to perform tasks without being explicitly programmed for those tasks.

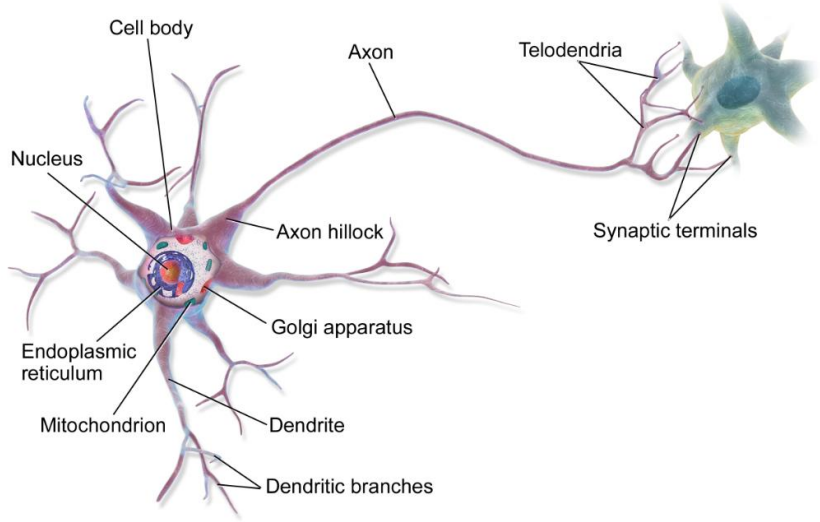
- ML is commonly used in problems where the answer is easily obtainable, but explaining it explicitly is not.
  - Pattern Recognition;
  - Recommendation Systems.

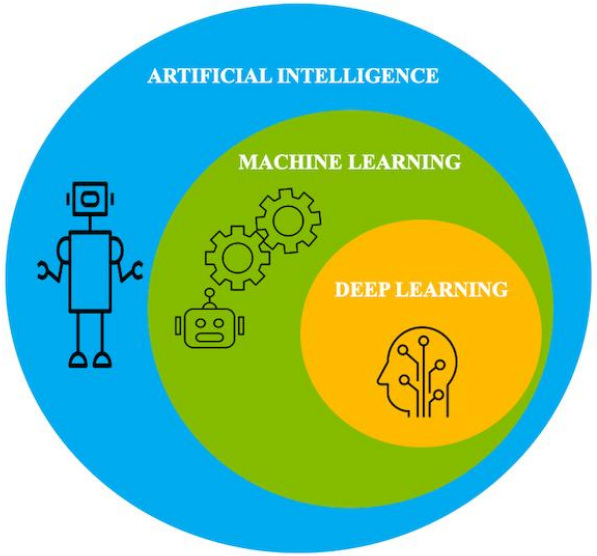


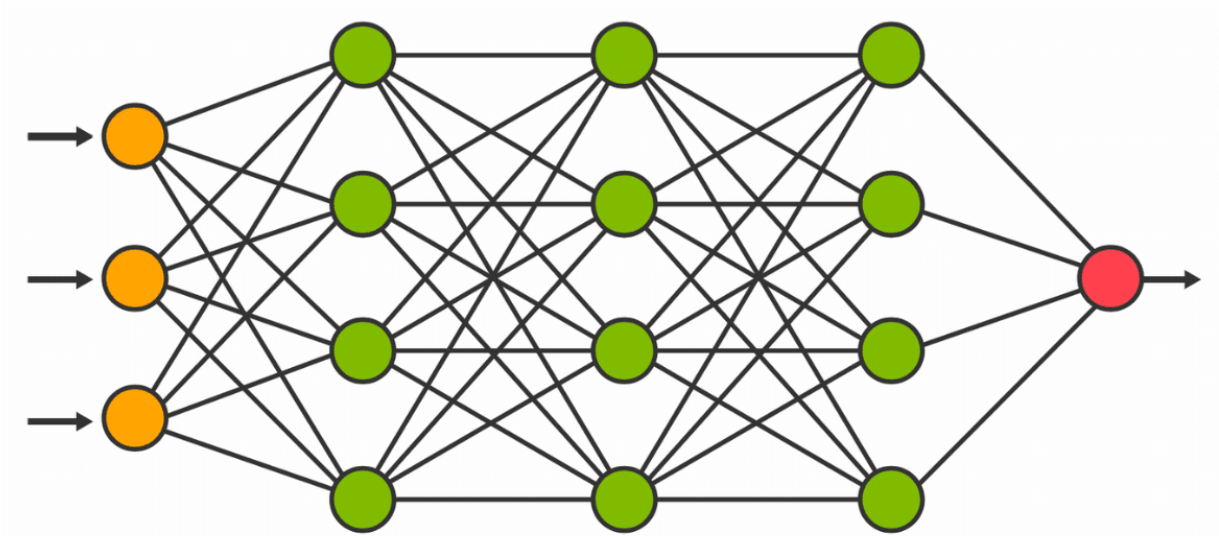
PURDUE UNIVERSITY

Deep Learning

Deep learning is a subset of AI and machine learning (ML) that aims to mimic the human brain's structure and function through **artificial neural networks**.







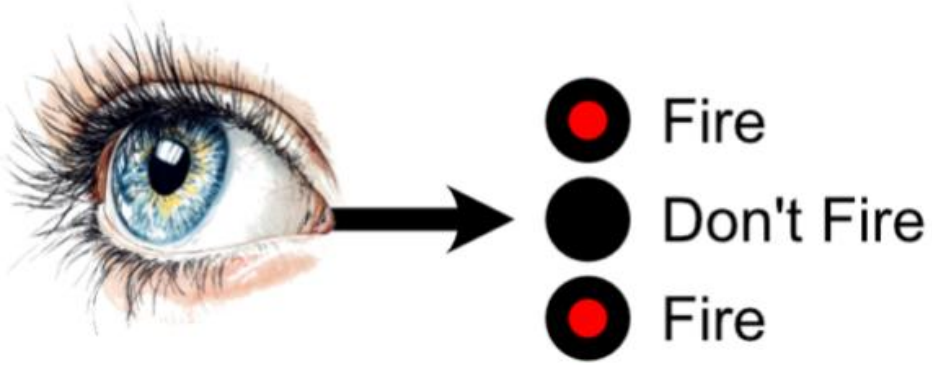
INPUT LAYER: input Data

HIDDEN LAYERS: process information

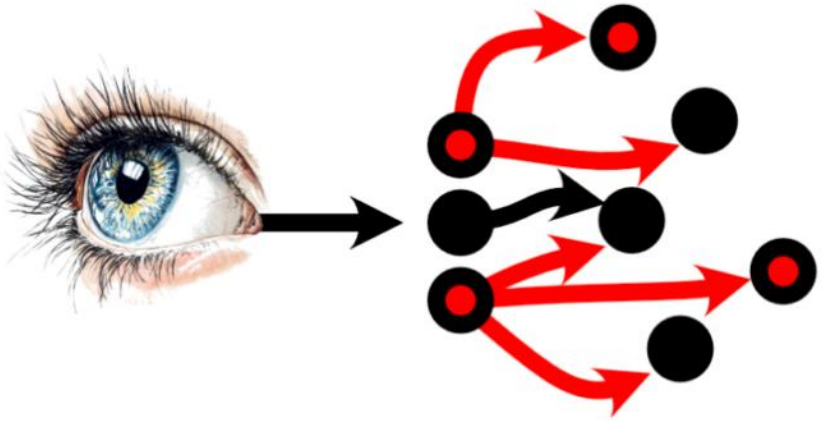
OUTPUT LAYER: result

PURDUE UNIVERSITY

How the Brain Works



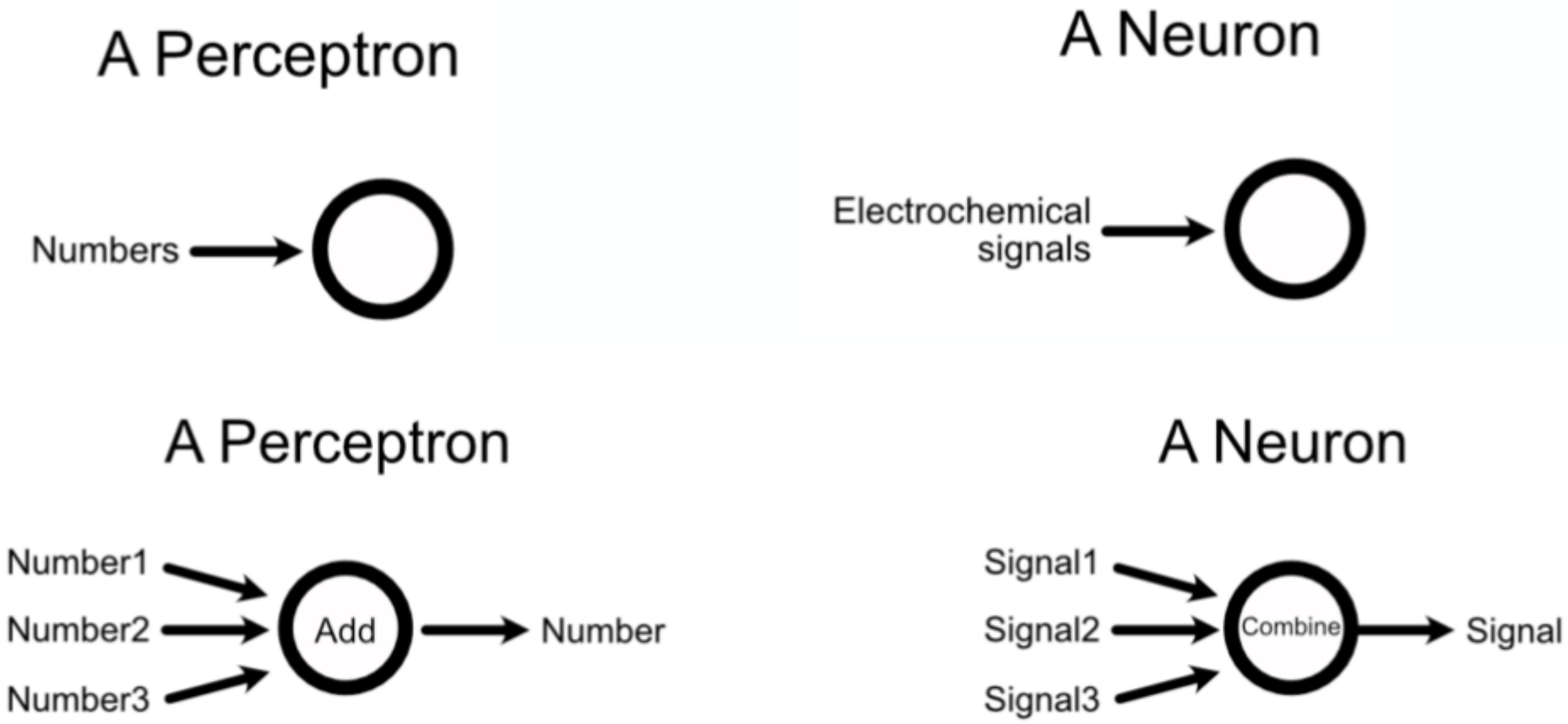
Imagine the signal from the eye directly feeds into three neurons, and two decide to fire.



Neurons receive stimulus from previous neurons and then choose whether or not to fire based on the magnitude of the stimulus.

PURDUE UNIVERSITY

Similarity between Neural Networks and the Brain



Perceptrons output numbers, while neurons output electrochemical signals.

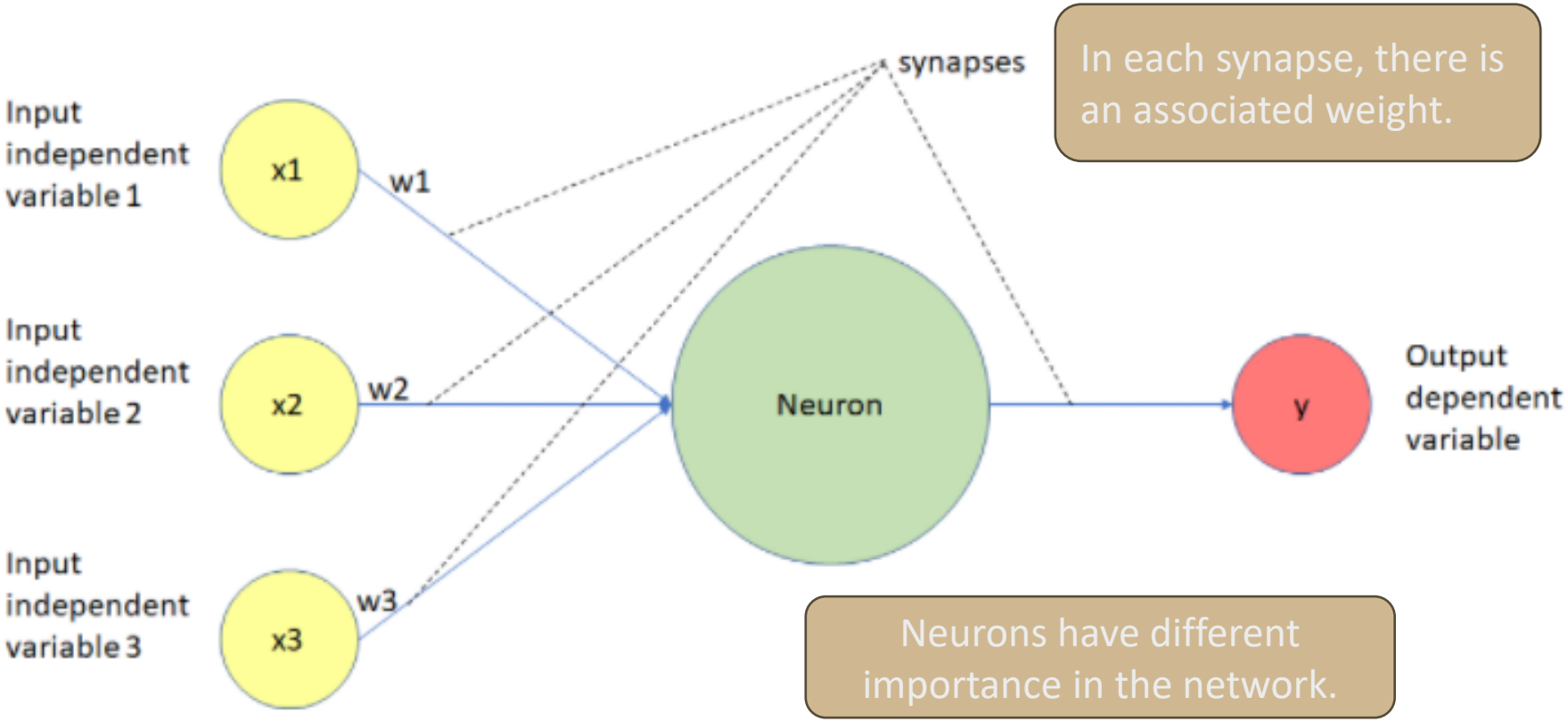
PURDUE UNIVERSITY

Similarity between Neural Networks and the Brain

Perceptrons output a continuous range of numbers, while Neurons either fire or they don't.

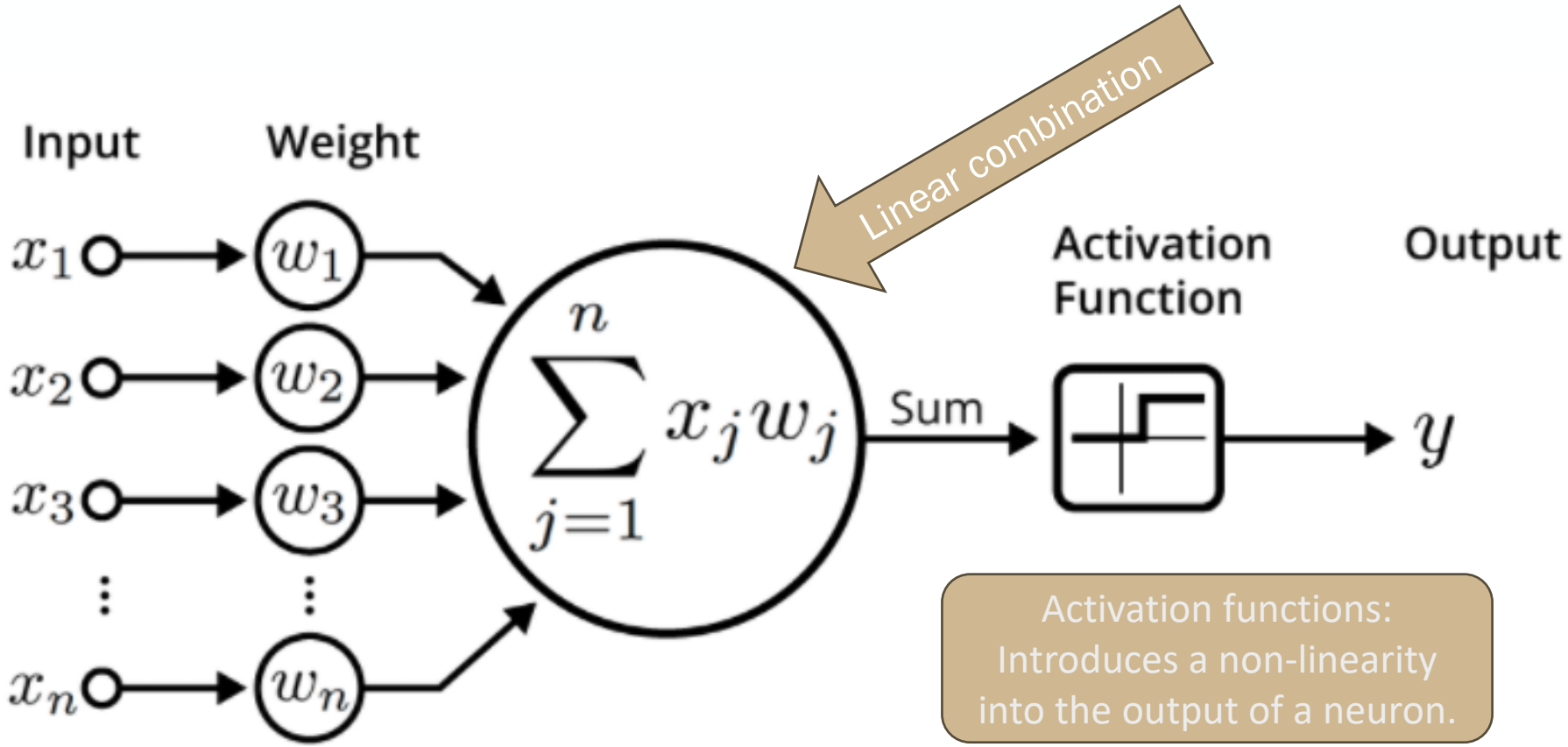
Neural Network: Neuron (Percepton)

Node where information and computational calculations are passed.



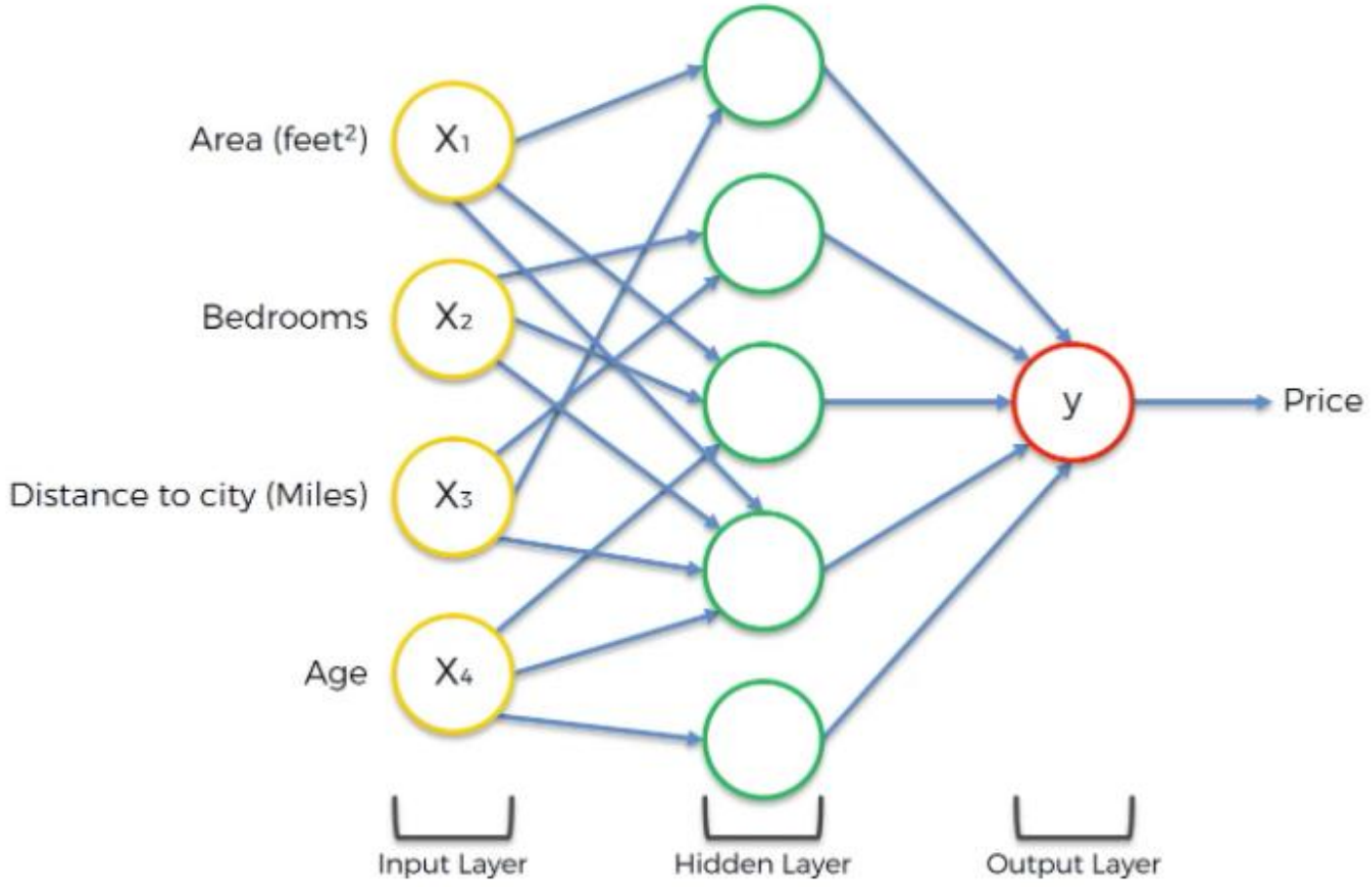
Weights

The progressive adjustment of weights is the primary way in which a neural network learns.



Simple Example

Simple Example



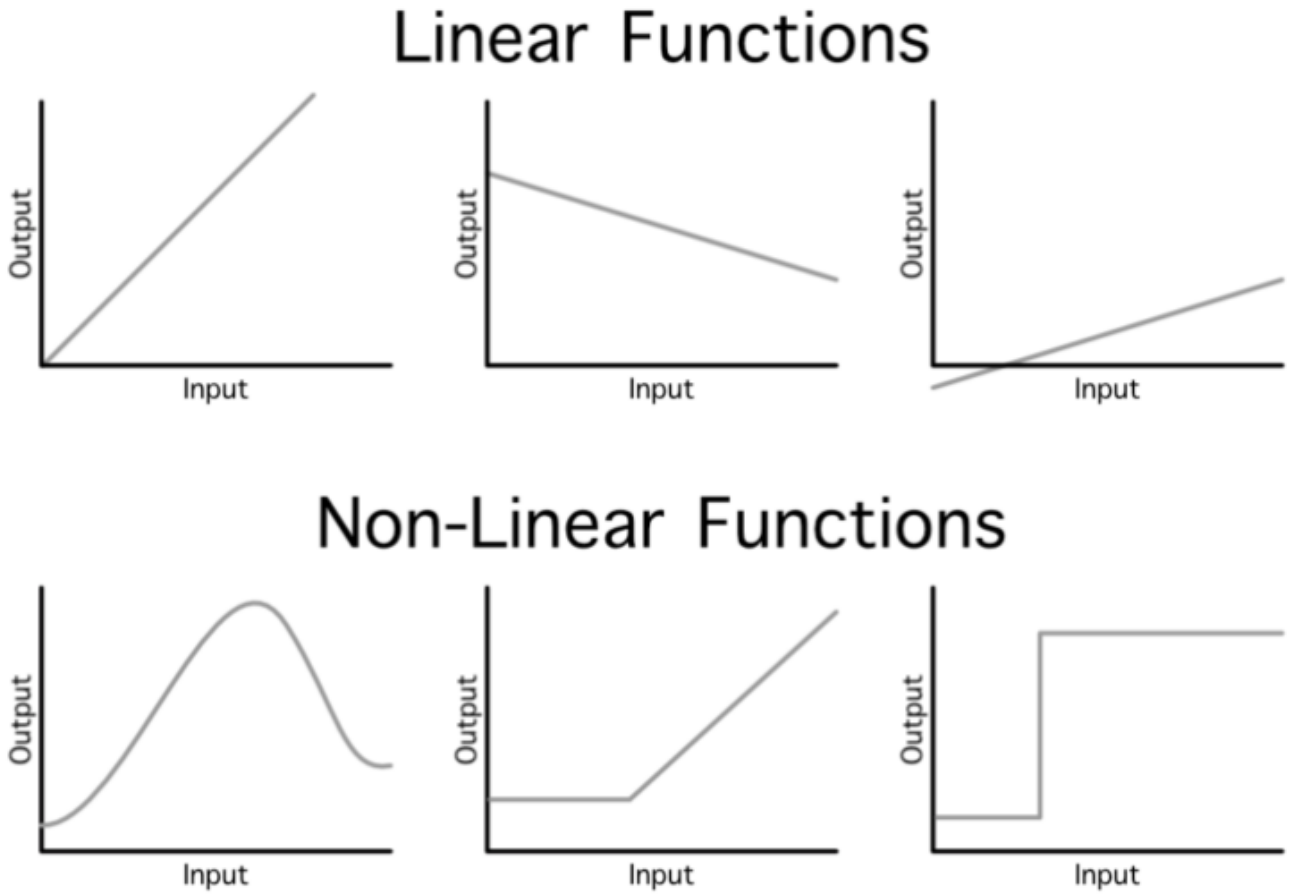
- Adding a Hidden layer adds more parameters (weights) to the network.
- From 4 to 16.
- Adding an activation function can give more freedom to the output

Bias


Area (feet²)
X1
W1
Bedrooms
X2
W2
Distance to city (Miles)
X3
W3
Age
X4
W4
y
Price = w1*x1+ w2*x2+ w3*x3+ w4*x4 +b
Input Layer
Output Layer

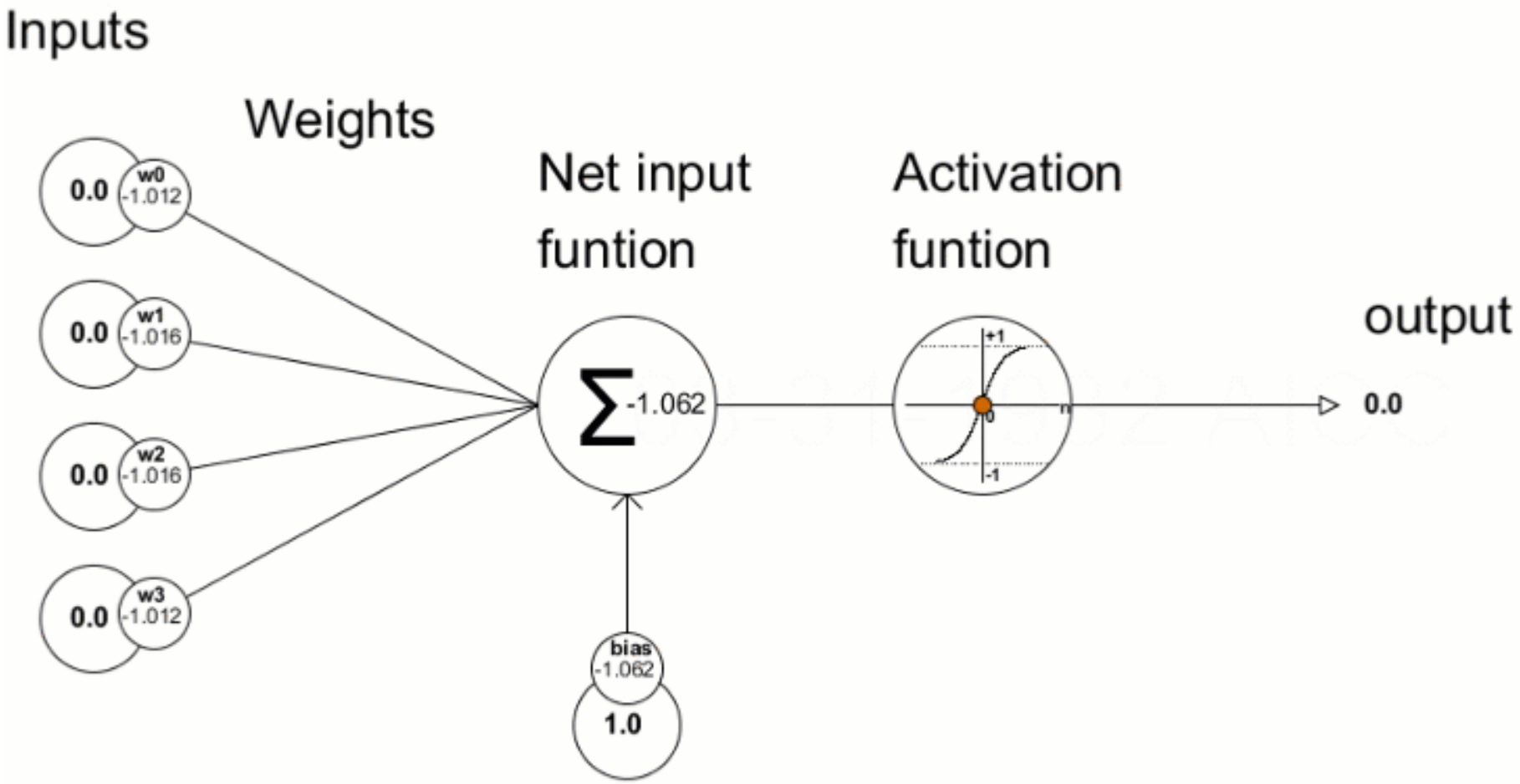
Activation Function

- This is the biggest difference between Neural Network and Linear Regression



PURDUE UNIVERSITY

Neuron


Inputs
Weights
Net input funtion
Activation funtion
output
0.0
w0
-1.012
0.0
w1
-1.016
0.0
w2
-1.016
0.0
w3
-1.012
-1.062
+1
0
-1
0.0
bias
-1.062
1.0

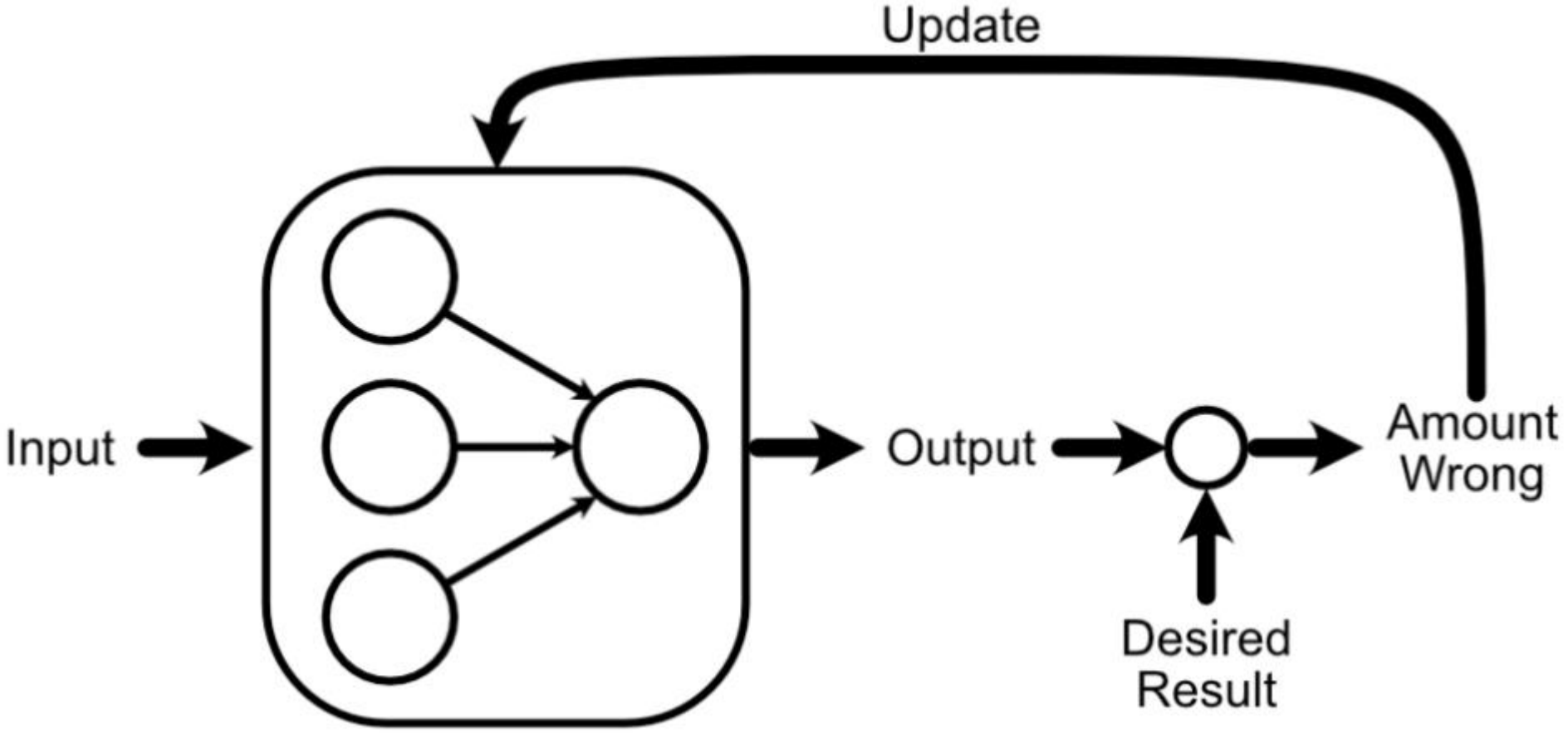
Classification

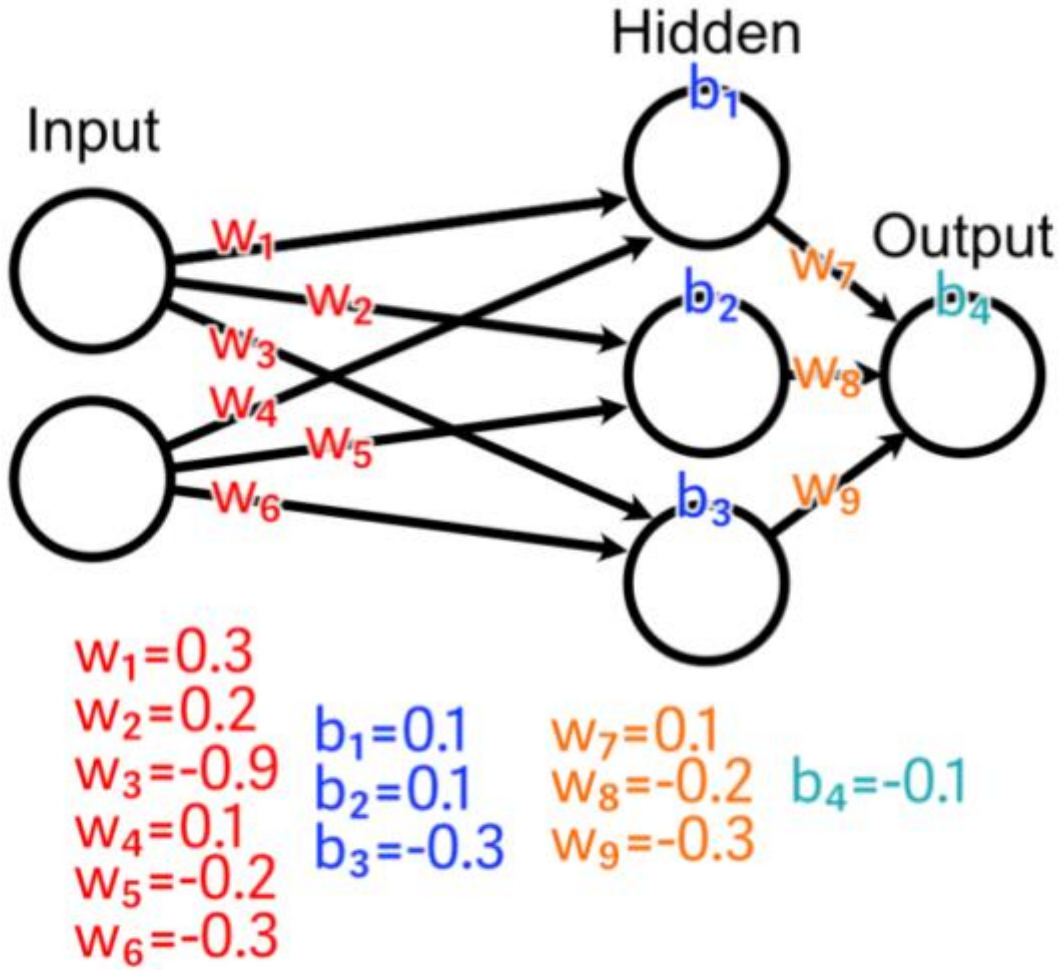
- Unsupervised: There is no information about classification in the database.
  - Clustering;
  - Recommendation Algorithms.

- Supervised: There is information about classification in the database.
  - Classification;
  - Prediction.

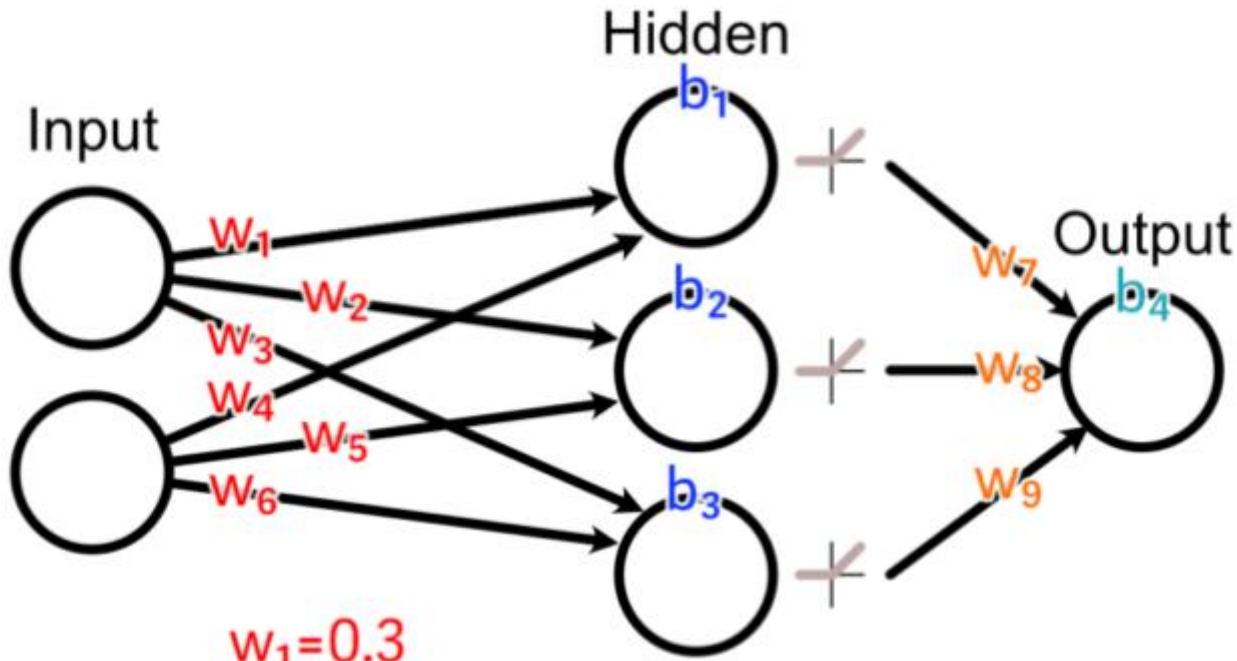
- Semi-supervised: There is information about classification in a fraction of data in the database.

Training Process





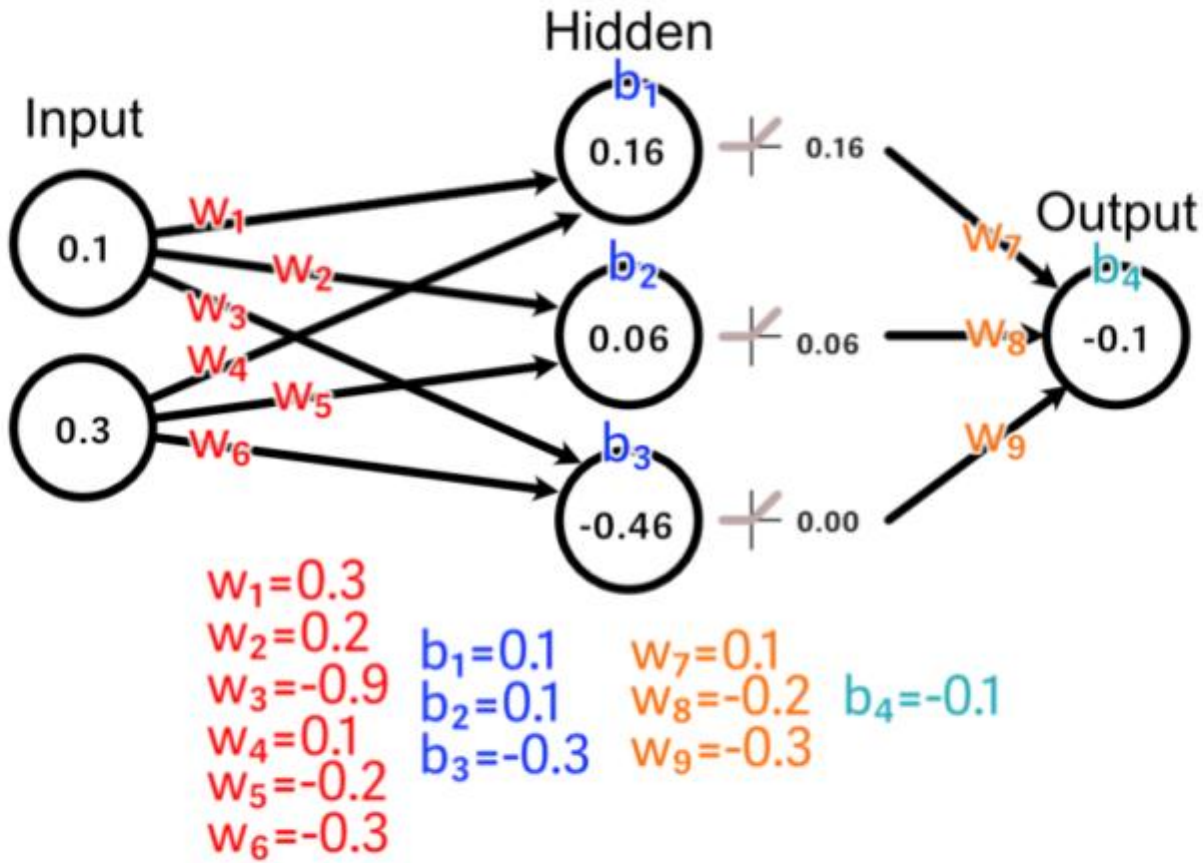
The neural network, with random weights and biases defined.



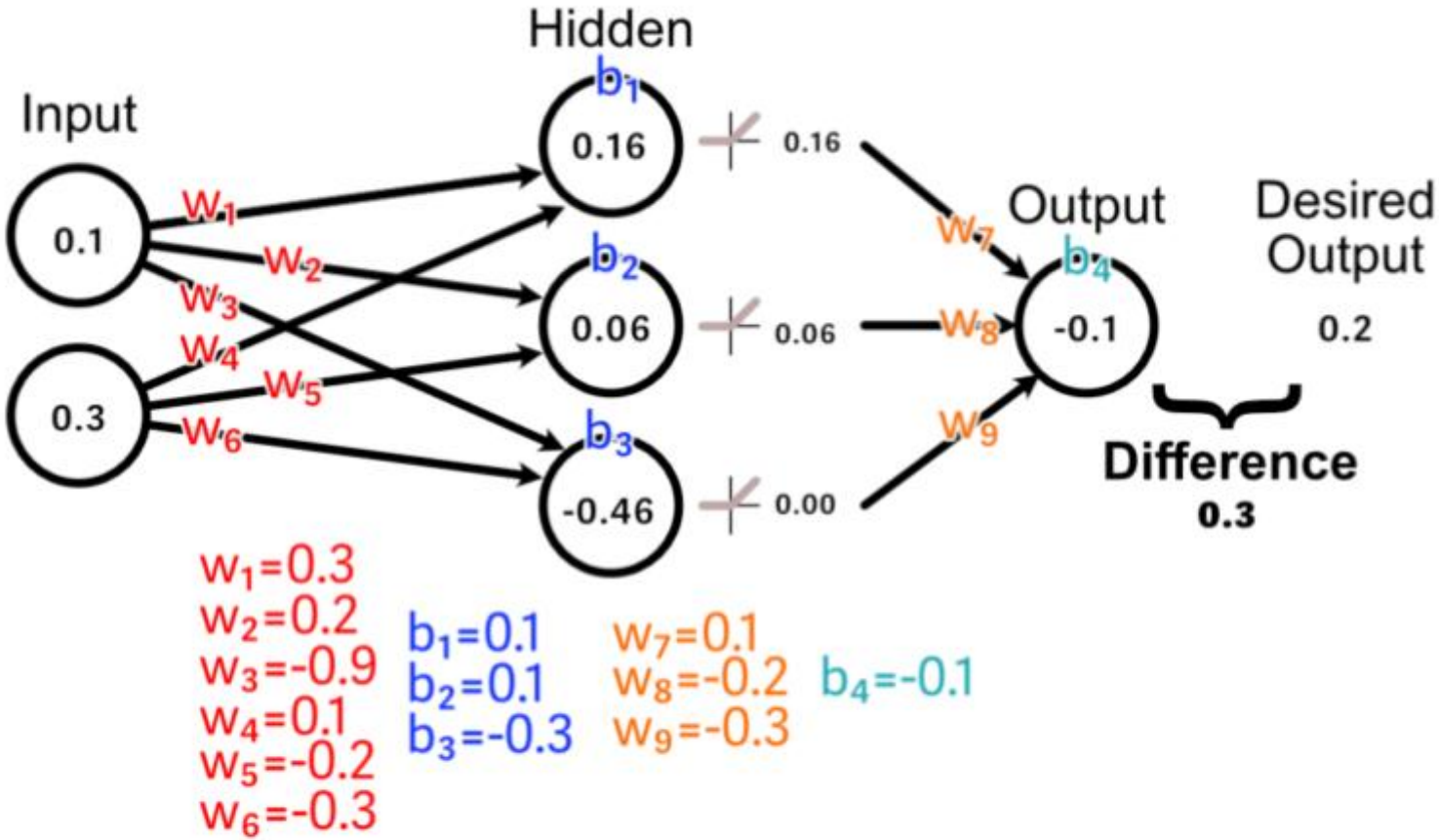
We'll apply the ReLU activation function to the value of our hidden perceptrons.

Input	Desired Output
0.1, 0.3	0.2
0.1, -0.1	0.0
-0.8, 0.2	-0.3
⋮	⋮

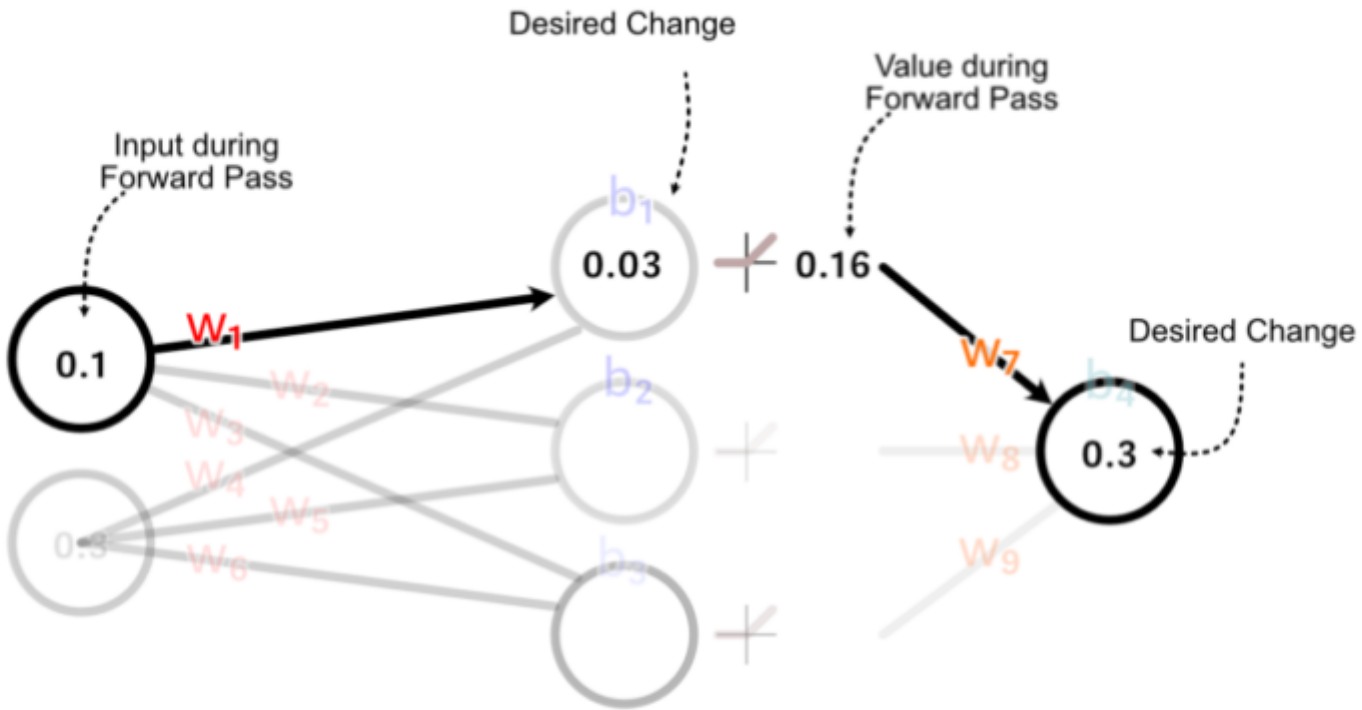
An example of the data that we'll be training off of.



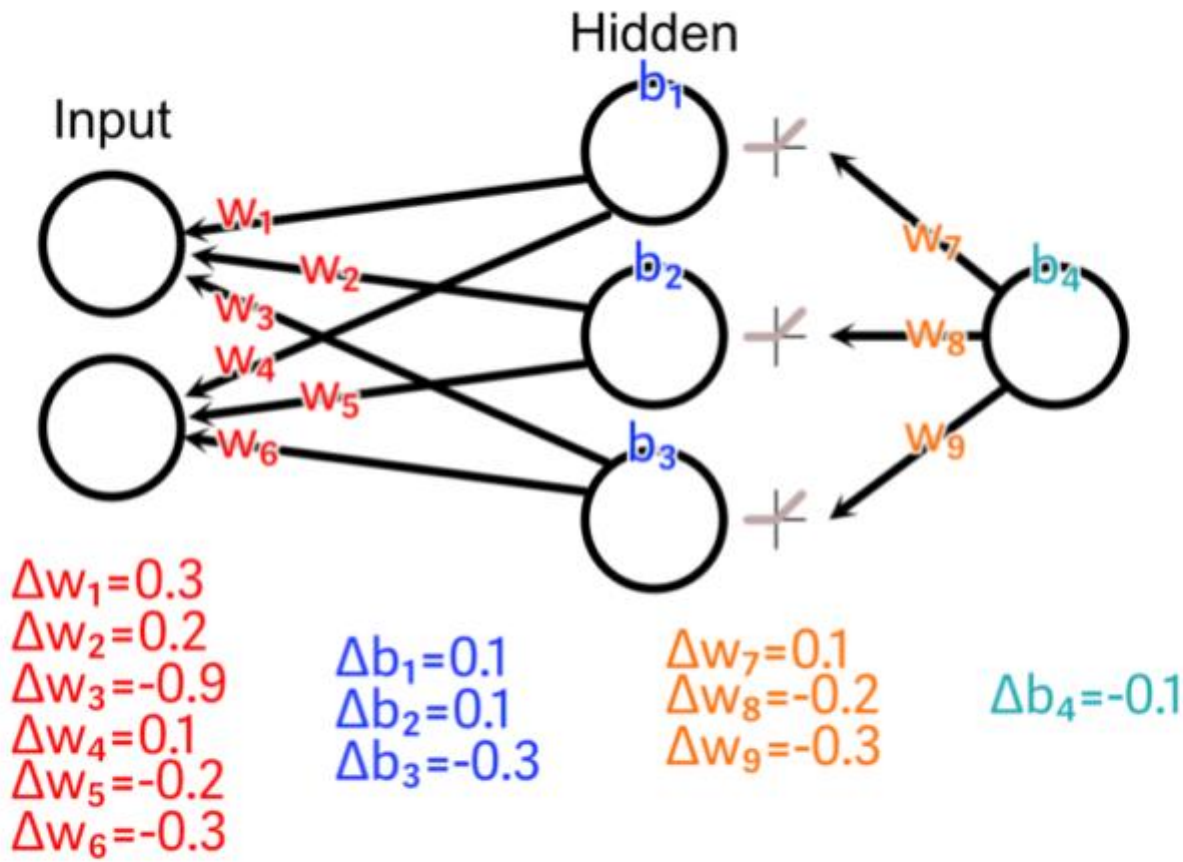
Calculating the value of the hidden layer and output based on the input, including all major intermediary steps.



The training data has an input of 0.1 and 0.3, and the desired output (the average of the input) is 0.2. The prediction from the model was -0.1. Thus, the difference between the output and the desired output is 0.3.

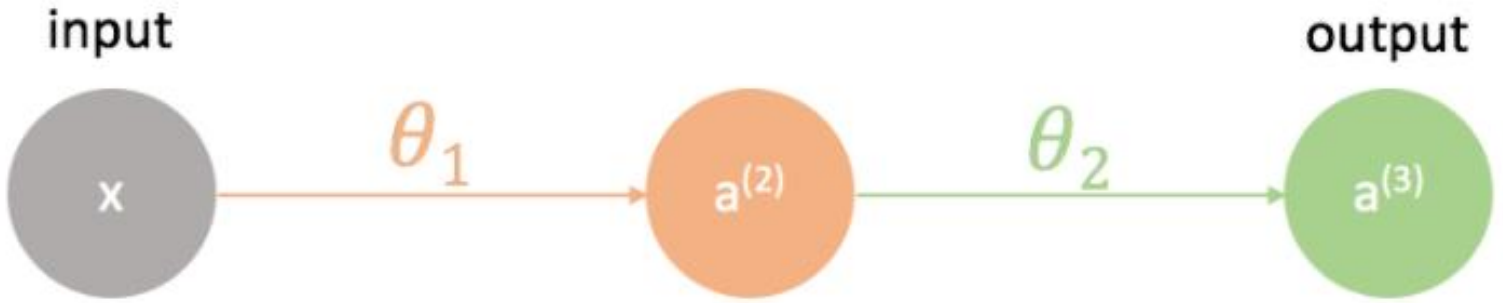


Now that we've calculated how the first hidden neuron should change, we can calculate how we should update w_1 the same way we calculated how w_7 should be updated previously.

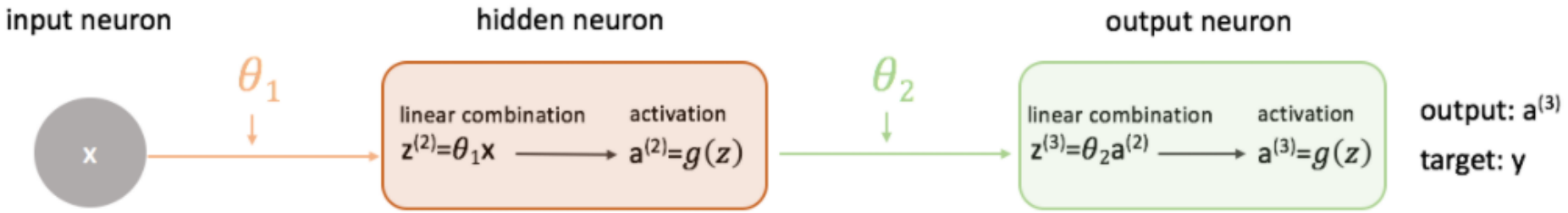


By back propagating through the model, using a combination of values from the forward passes and desired changes from the backward pass at various points of the model, we can calculate how all parameters should

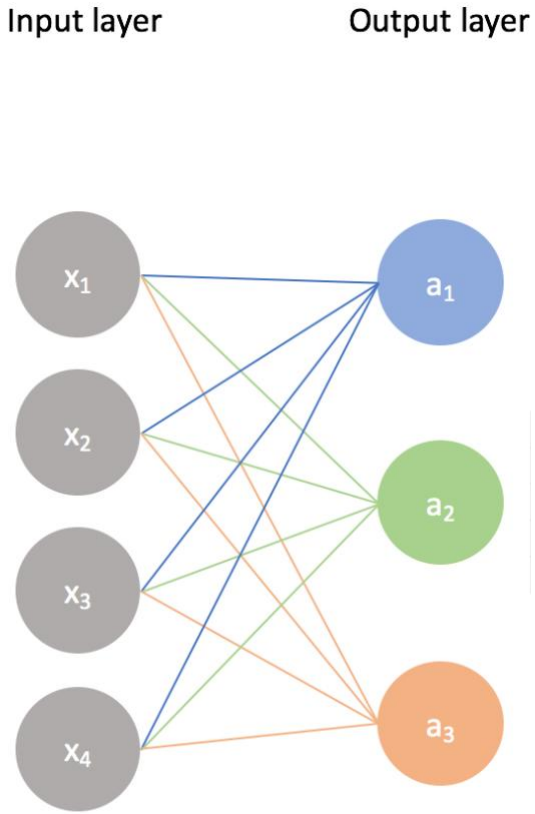
Forward Propagation



$$a^{(2)} = g(\theta_1 x) \qquad a^{(3)} = g(\theta_2 a^{(2)})$$



Matrix Notation



$$\begin{bmatrix} w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \\ w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b \end{bmatrix} \xrightarrow{activation} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

$$A = \sigma(WX + B)$$

Cost Function (J)

- Used to measure the error in a network prediction.

- Training a network: minimize the cost function.

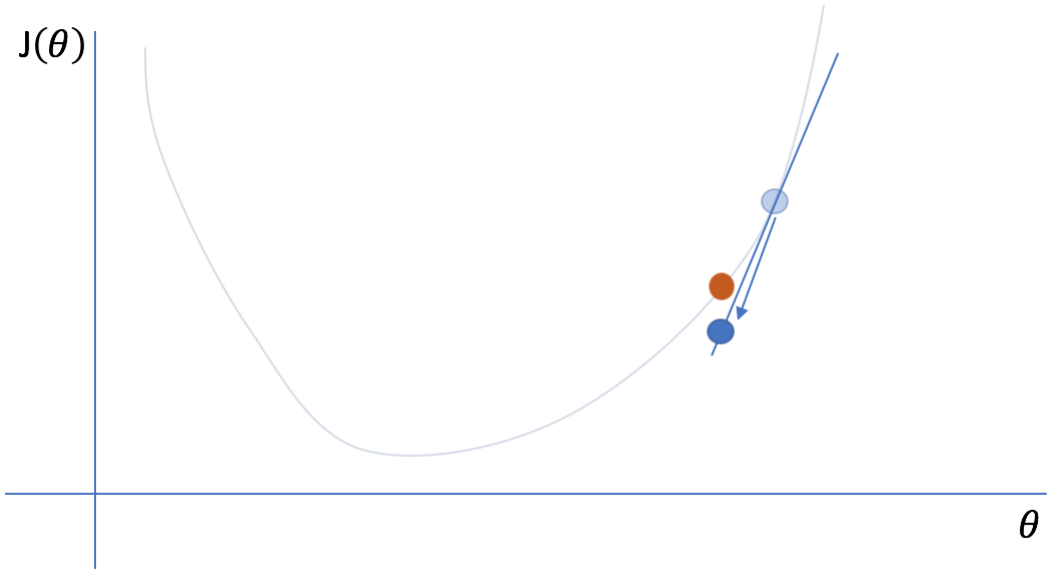
- Examples:
- Regression;
- Mean Squared Error;
- Mean Absolute Error;
- Cross Entropy.

Gradient Descent

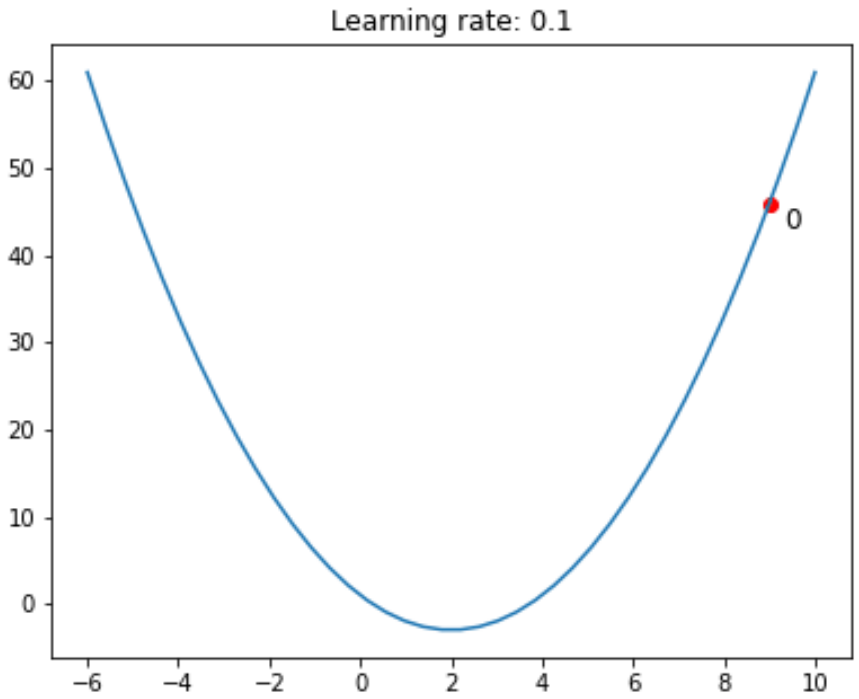
- Gradient: a vector that indicates the direction and magnitude in which, by displacement from the specified point, the greatest possible increase is obtained in the value of a quantity from which a scalar field for the space under consideration is defined.

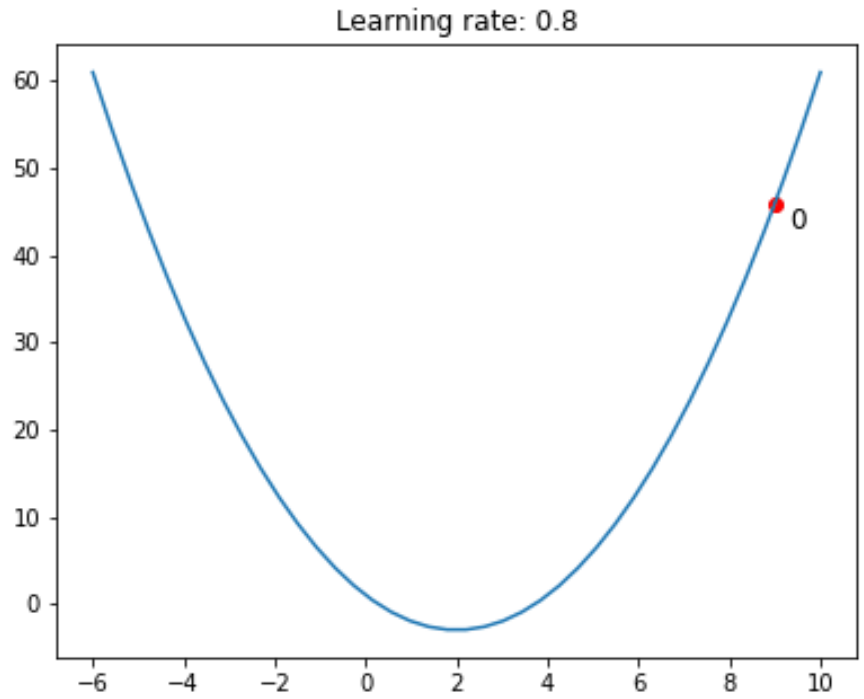
$$\theta_i := \theta_i + \Delta\theta_i$$

$$\Delta\theta_i = -\alpha \frac{\partial J(\theta)}{\partial \theta_i}$$

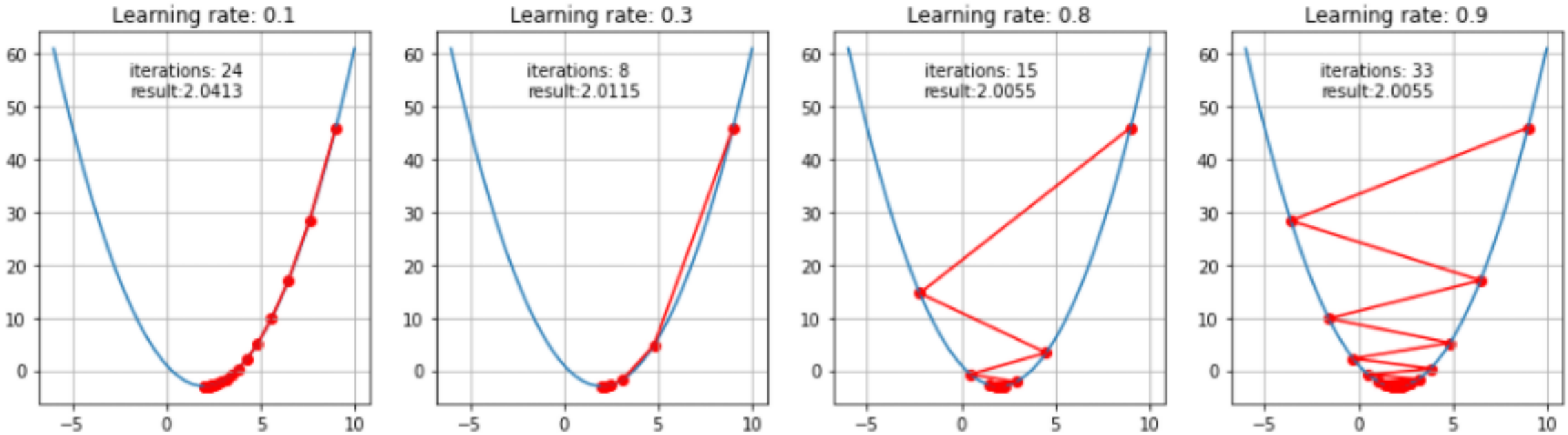


Learning Rate

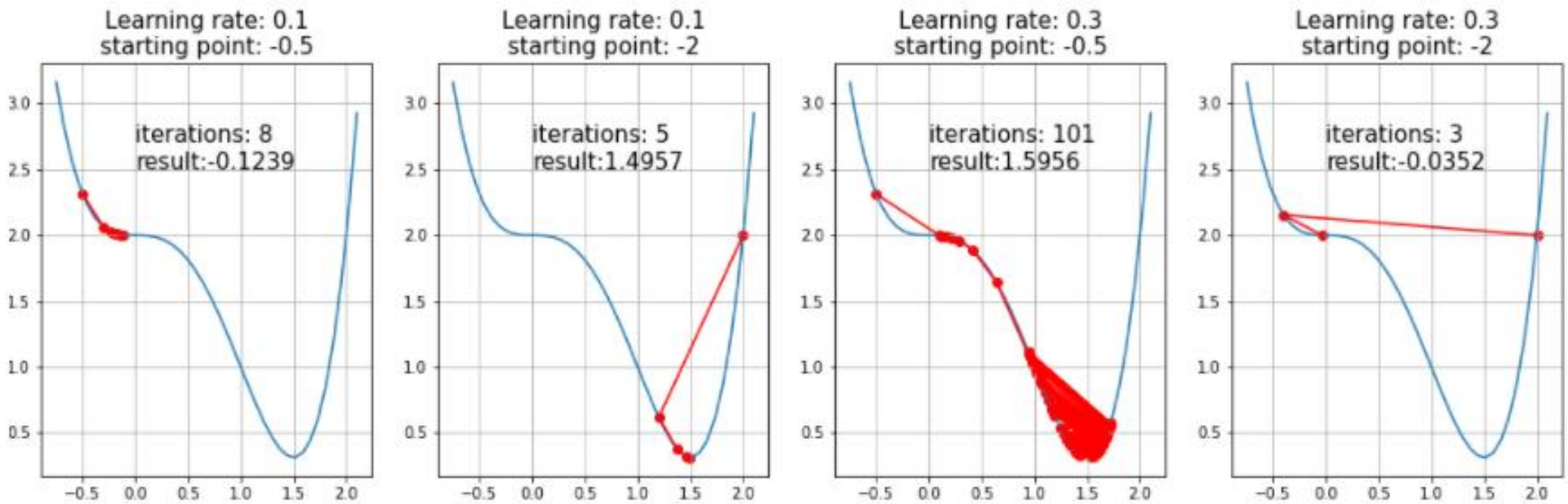




Different Learning rates

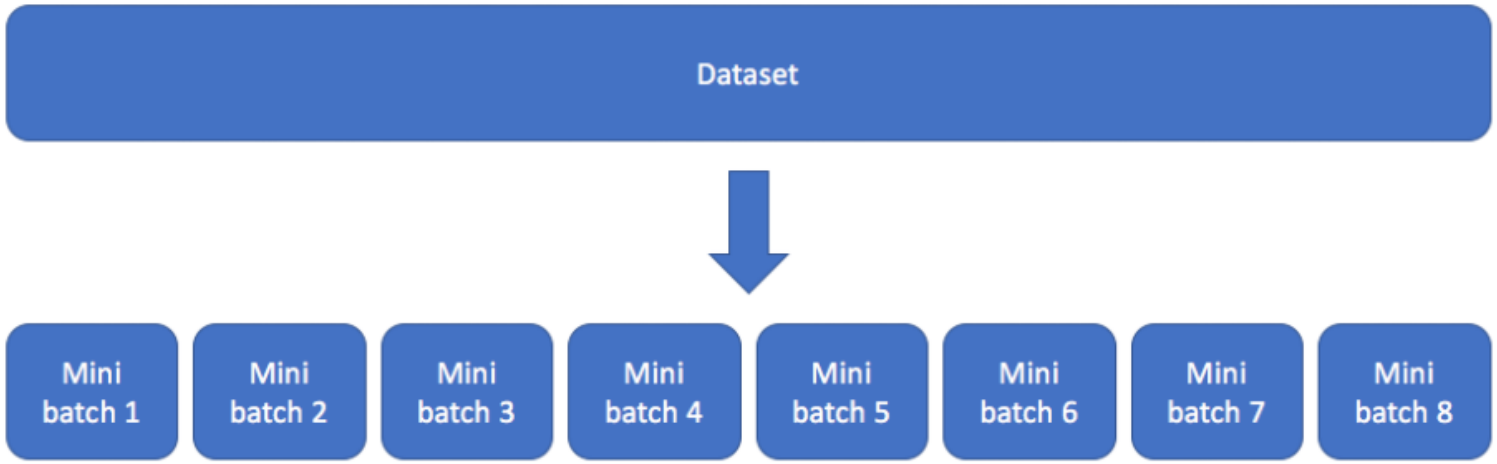


Different Learning rates



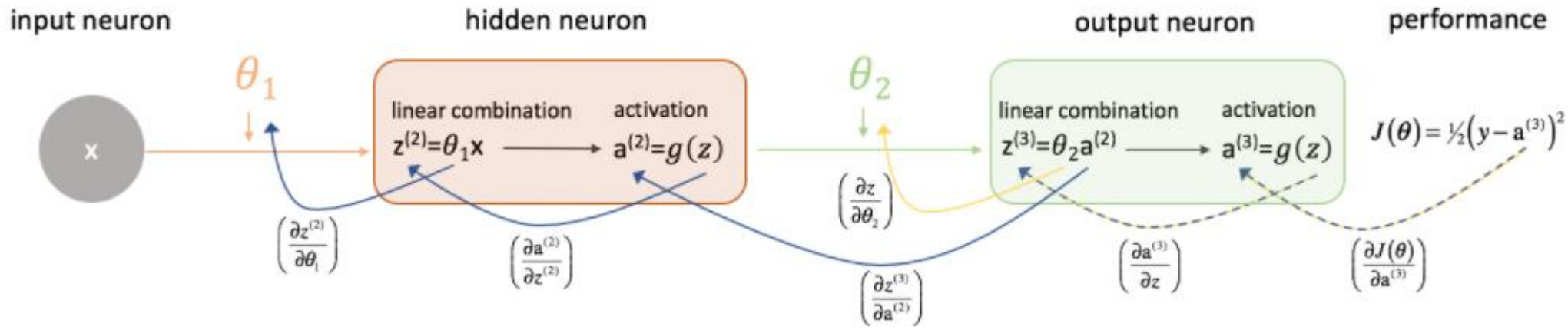
Mini Batch

- Cost function analyzes the loss associated with each training example;
- The sum of the batch cost functions gives the overall cost of the entire dataset;
- Mini-batch: the process of "breaking down" the dataset into smaller chunks to calculate the cost function.



PURDUE UNIVERSITY

Relationship between Weight and Cost Function



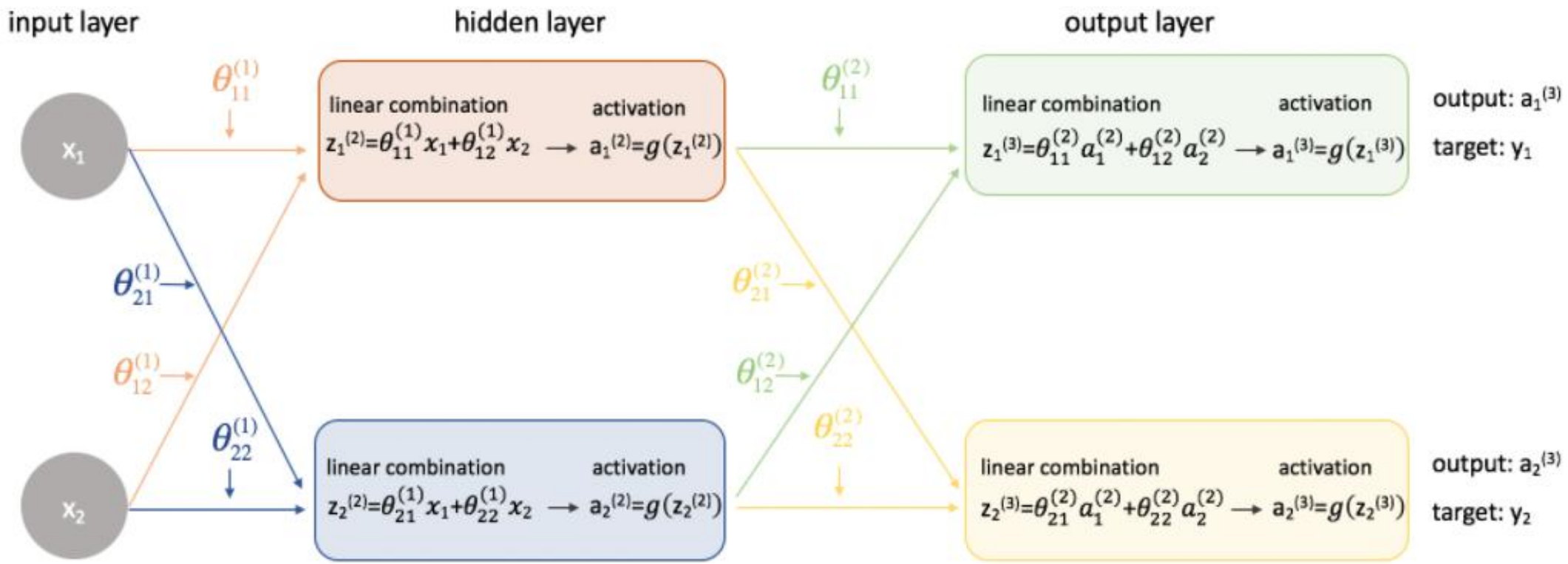
$$\frac{\partial J(\theta)}{\partial \theta_1} = ?$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \left(\frac{\partial J(\theta)}{\partial a^{(3)}}\right)\left(\frac{\partial a^{(3)}}{\partial z^{(3)}}\right)\left(\frac{\partial z^{(3)}}{\partial a^{(2)}}\right)\left(\frac{\partial a^{(2)}}{\partial z^{(2)}}\right)\left(\frac{\partial z^{(2)}}{\partial \theta_1}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = ?$$

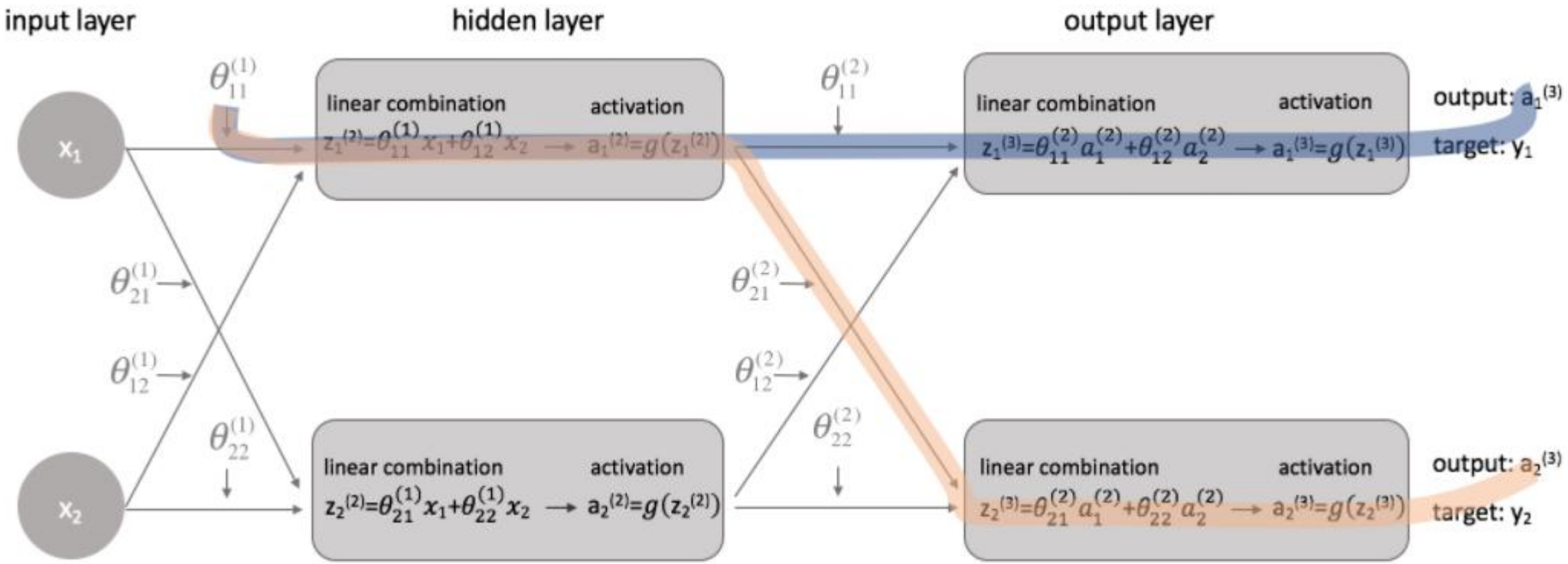
$$\frac{\partial J(\theta)}{\partial \theta_2} = \left(\frac{\partial J(\theta)}{\partial a^{(3)}}\right)\left(\frac{\partial a^{(3)}}{\partial z}\right)\left(\frac{\partial z}{\partial \theta_2}\right)$$

Complicating the Problem: Two Layers and Two Outputs



$$J(\theta) = \frac{1}{2m}\sum\left(y_i - a_i^{(2)}\right)^2$$

Complicating the Problem: Two Layers and Two Outputs



$$\frac{\partial J(\theta)}{\partial \theta_{11}^{(1)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{11}^{(1)}}\right) + \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{11}^{(1)}}\right)$$

Complicating the Problem: Two Layers and Two Outputs

First Layer

$$\frac{\partial J(\theta)}{\partial \theta_{11}^{(1)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{11}^{(1)}}\right) + \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{11}^{(1)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{12}^{(1)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{12}^{(1)}}\right) + \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial a_1^{(2)}}\right)\left(\frac{\partial a_1^{(2)}}{\partial z_1^{(2)}}\right)\left(\frac{\partial z_1^{(2)}}{\partial \theta_{12}^{(1)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{21}^{(1)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial a_2^{(2)}}\right)\left(\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)\left(\frac{\partial z_2^{(2)}}{\partial \theta_{21}^{(1)}}\right) + \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial a_2^{(2)}}\right)\left(\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)\left(\frac{\partial z_2^{(2)}}{\partial \theta_{21}^{(1)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{22}^{(1)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial a_2^{(2)}}\right)\left(\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)\left(\frac{\partial z_2^{(2)}}{\partial \theta_{22}^{(1)}}\right) + \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial a_2^{(2)}}\right)\left(\frac{\partial a_2^{(2)}}{\partial z_2^{(2)}}\right)\left(\frac{\partial z_2^{(2)}}{\partial \theta_{22}^{(1)}}\right)$$

Second Layer

$$\frac{\partial J(\theta)}{\partial \theta_{11}^{(2)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial \theta_{11}^{(2)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{12}^{(2)}} = \left(\frac{\partial J(\theta)}{\partial a_1^{(3)}}\right)\left(\frac{\partial a_1^{(3)}}{\partial z_1^{(3)}}\right)\left(\frac{\partial z_1^{(3)}}{\partial \theta_{12}^{(2)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{21}^{(2)}} = \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial \theta_{21}^{(2)}}\right)$$

$$\frac{\partial J(\theta)}{\partial \theta_{22}^{(2)}} = \left(\frac{\partial J(\theta)}{\partial a_2^{(3)}}\right)\left(\frac{\partial a_2^{(3)}}{\partial z_2^{(3)}}\right)\left(\frac{\partial z_2^{(3)}}{\partial \theta_{22}^{(2)}}\right)$$

Complexity

1 hidden layer and 1 output layer

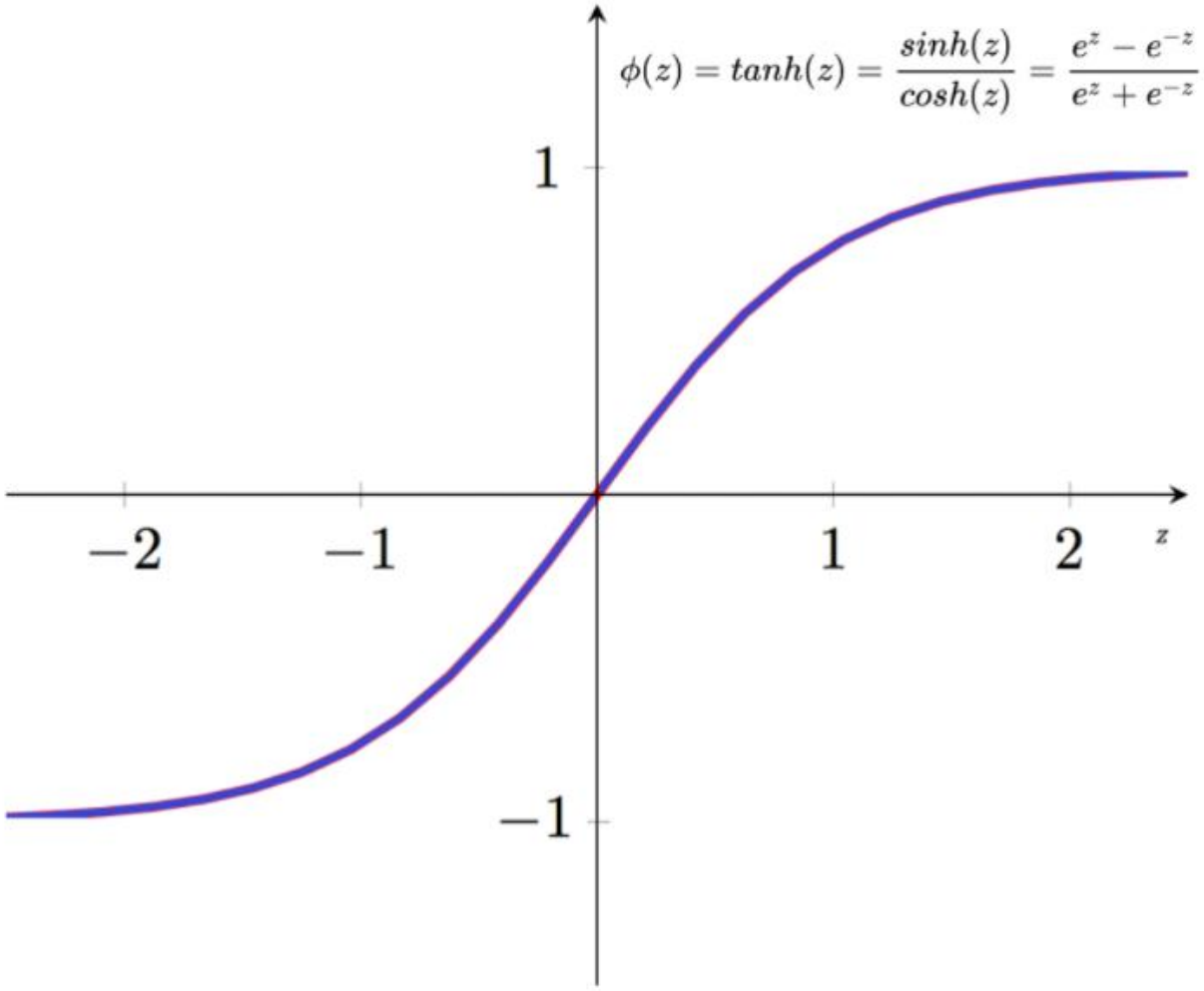
- 2 Parameters;
- 8 Partial Derivatives.

2 hidden layers and 2 output layers

- 8 Parameters;
- 52 Partial Derivatives.

Activation Function

- Examples:
  - Boundary Functions;
  - Sigmoid Functions;
  - Rectified Linear Unit (ReLU) Functions;
  - Hyperbolic Tangent Functions.



Thank You

PURDUE UNIVERSITY®

Bias

- Allows shifting the activation function by adding a constant.