

Insert Title Here

by

PARBHU Varun

Project Supervisor: Professor (Dr) BHURUTH MUDDUN OSK

Project submitted to the Department of Mathematics,

Faculty of Science, University of Mauritius,

as a partial fulfillment of the requirement

for the degree of

M.Sc. Mathematical and Scientific Computing (Part time)

December 2022

Contents

List of Figures	iii
List of Tables	iv
Acknowledgement	v
Abstract	vi
Terms and Definitions	vii
1 Introduction	1
2 Deep Learning Networks	2
2.1 Introduction	2
2.2 Perceptron	3
2.3 Perceptron Convergence Theorem	7
2.4 Solving Gate problem	11
2.4.1 AND Function	12
2.4.2 OR Function	13
2.4.3 XOR Function	15
2.5 Feedforward Network	18
2.6 Backpropagation	25
3 Optimization	32
3.1 Introduction	32
3.2 Optimization algorithms	34
3.2.1 Gradient Descent	35
3.2.2 Stochastic Gradient Descent	36
3.2.3 Mini-Batch Stochastic Gradient Descent	37

3.2.4	Newton's Method	38
3.2.5	Momentum	39
3.2.6	RMSProp	39
3.2.7	Adam	40
4	Sentiment Analysis	42
4.1	Introduction	42

List of Figures

2.1	Rosenblatt perceptron	3
2.2	Caption	4
2.3	Caption	4
2.4	!!!TO BE CHANGED!!!	5
2.5	!!!!TO BE CHANGED!!!!	6
2.6	AND Function with separating line	12
2.7	AND Function with separating line	13
2.8	OR function	15
2.9	XOR approximate solution	18
2.10	f_1	19
2.11	f_1	20
2.12	FFN hidden layer	21
2.13	XOR layer	22
2.14	Feedforward network for XOR problem	23
2.15	!!!TO BE CHANGED!!!	25
2.16	!!!TO BE CHANGED!!!	26
3.1	!!!TO BE CHANGED!!! Min max	33
3.2	Network Error	35

List of Tables

2.1	Truth Table for f_1^*	19
2.2	Truth Table for f_2^*	20
2.3	Truth Table for f^*	21
2.4	Truth Table from hidden layer	22

Acknowledgement

Placeholder

Abstract

Placeholder

Terms and Definitions

α	Learning Rate
η	Reduced Learning Rate
I	Identity Matrix
Placeholder	
	Placeholder

Chapter 1

Introduction

- Discussion around Cryptocurrency and frequency of changes
- Discussion around Twitter and the API (amount of data)
- Discussion around the advancement Compute processing speed and Deep Learning algorithms

Chapter 2

Deep Learning Networks

2.1 Introduction

Deep learning is a multidisciplinary field ranging from linear algebra, calculus and probability theory. With major advancement made in compute processing power; deep learning is commonly used in the industry now.

McCullough and Walter Pitts (MCP) Neuron

In 1943, McCulloch and Walter Pitts demonstrated that simple binary threshold units wired up as logical gates could be used to build a digital computer (add-ref). The McCulloch-Pitts (MCP) neuron is a simple mathematical model of a biological neuron. It was the earliest mathematical model of a neural network and had only three types of weights; excitatory (1), inhibitory (-1) and inactive (0). The model had an activation function which had a value of 1 if the weighted sum was greater or equal to a given threshold, else 0. Using the MCP neuron, one of the first digital computers that contained stored programs was built. However, the MCP neuron was very restrictive.

Rosenblatt's Perceptron Algorithm

In 1958, Frank Rosenblatt proposed the perceptron and an algorithm to adjust weights of the MCP neuron, extending the work done by McCulloch and Pitts. The MCP neuron had some limitations that the Perceptron managed to solve, for exam-

ple the input was not restricted to boolean values but expanded to real numbers. Rosenblatt proved that if the data used to train the perceptron are linearly separable classes, then the perceptron algorithm converges and separates the two classes by a hyperplane.

2.2 Perceptron

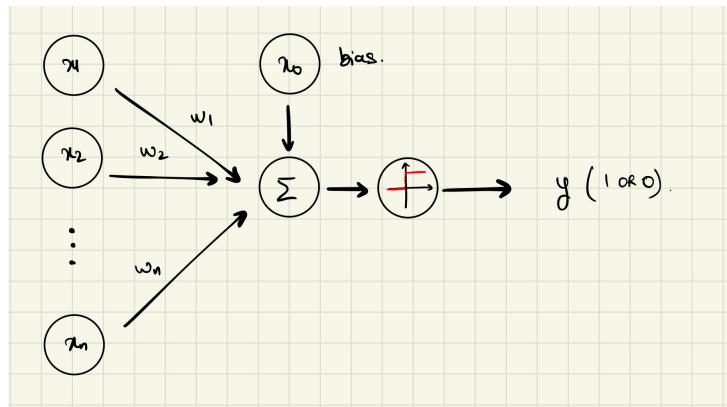


Figure 2.1: Rosenblatt perceptron

Let \mathbf{x} be a vector of inputs where each $x_i \in \mathbb{Z}$ and \mathbf{w} be a vector of weights corresponding to the input signals where each $w_i \in \mathbb{Z}$.

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$$

Then, the mathematical definition of the perceptron is given by

$$\phi(z) = \begin{cases} 1 & \text{if } \sum_{i=0}^n w_i x_i \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

where $\phi(z)$ is known as the hard delimiter.

Consider the two sets of point **A** and **B**

$$\mathbf{A} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\} \quad \mathbf{B} = \left\{ \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\} \quad (2.2)$$

and fitting the perceptron function (2.1) to learn how to classify points in each respective set. Let set **A** and **B** be class 1 and 0 respectively.

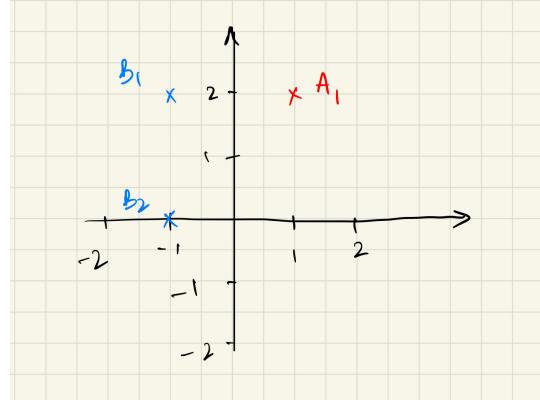


Figure 2.2: Caption

From 2.2, we can separate the two classes by a single line. Consider the random separating line passing through the origin (this will set the value of x_0 in (2.1) to 0).

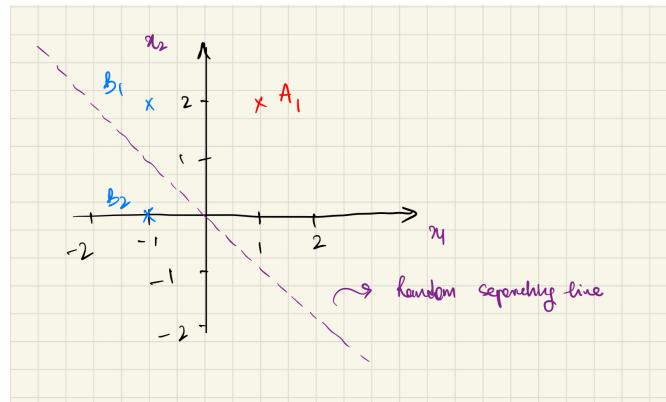


Figure 2.3: Caption

The equation of the random line is given by

$$x_1 = -x_2 \quad (2.3)$$

Comparing to the equation (2.1); we can deduce that

$$w_1 = 1 \quad w_2 = 1 \quad (2.4)$$

Visually, we can already see that the separating line does not split the points properly. The computation of the binary classification using the random line is given by

$$\phi(A_1) = \phi \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right) = \phi(3) = 1 \quad (2.5)$$

$$\phi(B_1) = f \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right) = \phi(1) = 1 \quad (2.6)$$

$$\phi(B_2) = f \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right) = \phi(-1) = 0 \quad (2.7)$$

The computation confirms the visual representation and groups \mathbf{A}_1 and \mathbf{B}_1 together and \mathbf{B}_2 separated which is incorrect. We can either alter the separating line (by moving the weight vector) with respect to \mathbf{A}_1 and/or \mathbf{B}_1 .

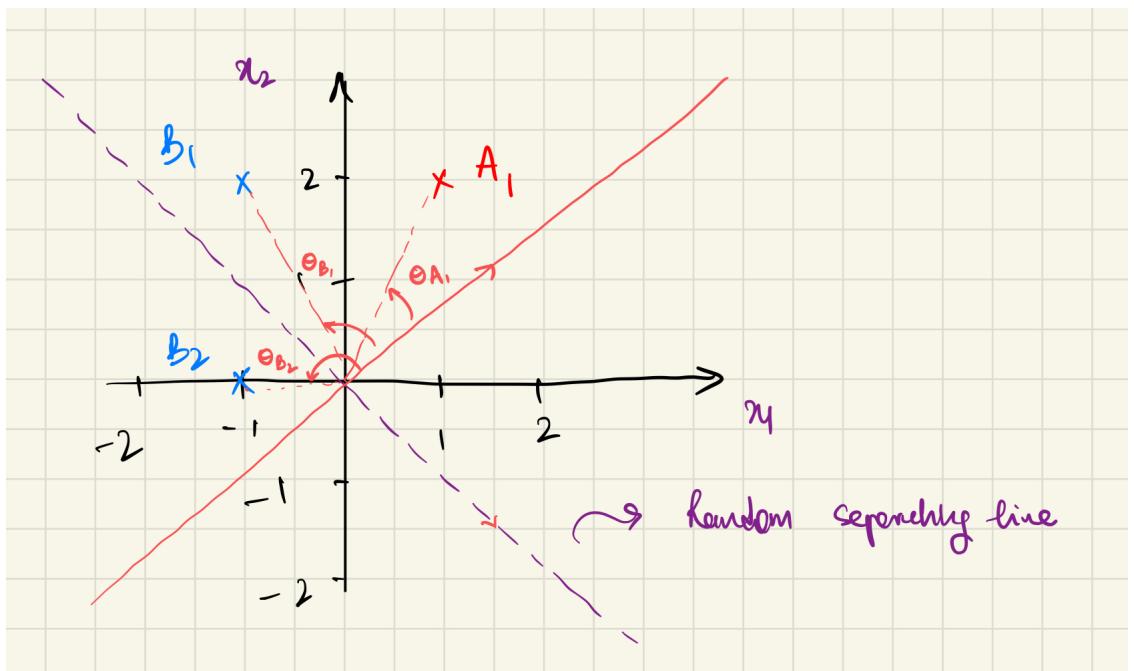


Figure 2.4: !!!TO BE CHANGED!!!

Since \mathbf{B}_1 is incorrectly classified, using the property of subtraction of vectors; we move the weight vector (\mathbf{w}) away from \mathbf{B}_1 and check the classification again.

$$\mathbf{w}^{(1)} = \mathbf{w} - \mathbf{B}_1 \quad (2.8)$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} \quad (2.9)$$

$$= \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad (2.10)$$

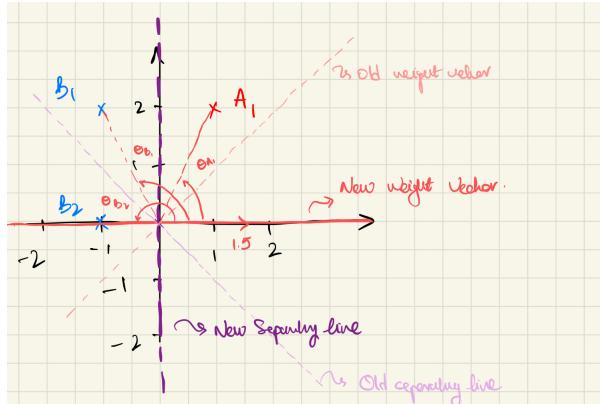


Figure 2.5: !!!!TO BE CHANGED!!!!

Based on the new separating line, we can observe visually that the points are correctly classified. The computation of the binary classification using the new separating line $\mathbf{w}^{(1)}$ is given by

$$\phi(A_1) = \phi \left(\begin{bmatrix} 2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right) = \phi(0) = 1 \quad (2.11)$$

$$\phi(B_1) = \phi \left(\begin{bmatrix} 2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right) = \phi(-4) = 0 \quad (2.12)$$

$$\phi(B_2) = \phi \left(\begin{bmatrix} 2 \\ -1 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right) = \phi(-2) = 0 \quad (2.13)$$

Thus, the perceptron learned how to classify the two sets of points. The function

$f(\mathbf{x})$ could be re-used to classify new points added to the sets.

Perceptron Algorithm

The perceptron algorithm adjusts the weights and as a result, the hard delimiter as well in order to linearly separate a set of binary labelled input.

Algorithm 1 Perceptron Algorithm

Require:

- 1: Let $t = 0$ and $\mathbf{w} = [0 \ 0 \ \dots \ 0]^T$
 - 2: Consider the training set D s.t $D = C_1 \cup C_2$
 - 3: $C_1 \leftarrow$ input with label 1
 - 4: $C_2 \leftarrow$ input with label 0
 - 5: **Start**
 - 6: **while** !convergence **do**
 - 7: Select a random input \mathbf{x}
 - 8: **if** $\mathbf{x} \in A$ and $\mathbf{w} \cdot \mathbf{x} < 0$ **then**
 - 9: $\mathbf{w} = \mathbf{w} + \mathbf{x}$
 - 10: **end if**
 - 11: **if** $\mathbf{x} \in B$ and $\mathbf{w} \cdot \mathbf{x} \geq 0$ **then**
 - 12: $\mathbf{w} = \mathbf{w} - \mathbf{x}$
 - 13: **end if**
 - 14: **end while**
-

If the input data are binary and linearly separable; then algorithm (1) converges. The proof of convergence of the algorithm is known as the **perceptron convergence theorem**.

2.3 Perceptron Convergence Theorem

Theorem

Consider algorithm (1) and let D be a set of training vectors which are linearly separable. Let \mathbf{w}^* be the weight vectors which defines the separating line with $\|\mathbf{w}^*\| = 1$. Then the number of mistakes m made by the perceptron algorithm satisfies

$$m \leq \frac{1}{\gamma^2} \quad \text{where} \quad \gamma = \min_{\mathbf{x} \in D} \frac{|\mathbf{w}^{*T} \mathbf{x}|}{\|\mathbf{x}\|_2} \quad (2.14)$$

Note

1. Since, $\|\mathbf{w}^*\| = 1 \implies \cos(\theta) = \frac{\mathbf{w}^{*T}\mathbf{x}}{\|\mathbf{x}\|_2}$
2. If $\|\mathbf{x}\|_2 = 1$, that is, we scale all the training examples to have unit norm (which has no effect their orientation), then $\gamma = \min_{\mathbf{x} \in D} |\mathbf{w}^{*T}\mathbf{x}|$ is the minimum distance from any example $\mathbf{x} \in D$ to the separating line.

Proof

We first prove the inequality given by

$$(\mathbf{w}^{(t+1)})^T \mathbf{w}^* \geq (\mathbf{w}^{(t)})^T \mathbf{w}^* + \gamma \quad (2.15)$$

State 1

In state 1, let us consider \mathbf{x} being positive ($\mathbf{x} \in C_1$) and incorrectly classified then

$$(\mathbf{w}^{(t+1)})^T \mathbf{w}^* = (\mathbf{w}^{(t)} + \mathbf{x})^T \mathbf{w}^* = (\mathbf{w}^{(t)})^T \mathbf{w}^* + \mathbf{x}^T \mathbf{w}^*$$

Assuming that all $\|\mathbf{x}\|_2 = 1$ then $\mathbf{x}^T \mathbf{w}^* \geq \gamma$ since γ is the minimum.

$$(\mathbf{w}^{(t+1)})^T \mathbf{w}^* \geq (\mathbf{w}^{(t)})^T \mathbf{w}^* + \gamma \quad (2.16)$$

Thus, we have proved (2.15) under **State 1**.

State 2

In state 2, let us consider \mathbf{x} being negative ($\mathbf{x} \in C_2$) and incorrectly classified then

$$(\mathbf{w}^{(t+1)})^T \mathbf{w}^* = (\mathbf{w}^{(t)} - \mathbf{x})^T \mathbf{w}^* = (\mathbf{w}^{(t)})^T \mathbf{w}^* - \mathbf{x}^T \mathbf{w}^*$$

Since $\mathbf{x} \in C_2 \implies |\mathbf{w}^{*T}\mathbf{x}| = -\mathbf{x}^T \mathbf{w}^* \geq 0$ and $|\mathbf{x}^T \mathbf{w}^*| \geq \gamma$

$$\begin{aligned} (\mathbf{w}^{(t+1)})^T \mathbf{w}^* &= (\mathbf{w}^{(t)})^T \mathbf{w}^* + |\mathbf{w}^{*T}\mathbf{x}| \\ (\mathbf{w}^{(t+1)})^T \mathbf{w}^* &\geq (\mathbf{w}^{(t)})^T \mathbf{w}^* + \gamma \end{aligned} \quad (2.17)$$

Thus, we have proved (2.15) under **State 2**.

From (2.16) and (2.17), we have shown that $\forall \mathbf{x} \in D$ that inequality (2.15) holds.

Assume that the inequality (2.15) holds for an arbitrary integer value m and $m - 1$

$$(\mathbf{w}^{(m)})^T \mathbf{w}^* \geq (\mathbf{w}^{(m-1)})^T \mathbf{w}^* + \gamma \quad (2.18)$$

$$(\mathbf{w}^{(m-1)})^T \mathbf{w}^* \geq (\mathbf{w}^{(m-2)})^T \mathbf{w}^* + \gamma \quad (2.19)$$

Merging inequality (2.18) and (2.19) gives us

$$(\mathbf{w}^{(m)})^T \mathbf{w}^* \geq (\mathbf{w}^{(m-2)})^T \mathbf{w}^* + 2\gamma \quad (2.20)$$

Hence, by induction, after M mistakes, inequality (2.20) becomes

$$(\mathbf{w}^{(m)})^T \mathbf{w}^* \geq (\mathbf{w}^{(0)})^T \mathbf{w}^* + m\gamma \quad (2.21)$$

Since $\mathbf{w}^{(0)} = \mathbf{0}$ (the zero vector), we have

$$(\mathbf{w}^{(m)})^T \mathbf{w}^* \geq m\gamma \quad (2.22)$$

Next, we show that

$$\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 1 \quad (2.23)$$

State 1

Consider the same state 1 as before, thus we have

$$(\mathbf{w}^{(t)})^T \mathbf{x} \leq 0 \quad \text{and} \quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \mathbf{x} \quad (2.24)$$

then

$$\begin{aligned} \|\mathbf{w}^{(t+1)}\|_2^2 &= (\mathbf{w}^{(t)} + \mathbf{x})^T (\mathbf{w}^{(t)} + \mathbf{x}) \\ &= (\mathbf{w}^{(t)})^T \mathbf{w}^{(t)} + (\mathbf{w}^{(t)})^T \mathbf{x} + \mathbf{x}^T \mathbf{w}^{(t)} + \mathbf{x}^T \mathbf{x} \\ &= (\mathbf{w}^{(t)})^T \mathbf{w}^{(t)} + \mathbf{x}^T \mathbf{w}^{(t)} + (\mathbf{w}^{(t)})^T \mathbf{x} + \mathbf{x}^T \mathbf{x} \\ &= (\mathbf{w}^{(t)})^T \mathbf{w}^{(t)} + 2(\mathbf{w}^{(t)})^T \mathbf{x} + \mathbf{x}^T \mathbf{x} \end{aligned} \quad (2.25)$$

Since (2.24) and $\mathbf{x}^T \mathbf{x} = 1$, equation (2.25) becomes

$$\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 1 \quad (2.26)$$

State 2

Consider the same state 2 as before, thus we have

$$(\mathbf{w}^{(t)})^T \mathbf{x} > 0 \quad \text{and} \quad \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \mathbf{x} \quad (2.27)$$

then

$$\begin{aligned} \|\mathbf{w}^{(t+1)}\|_2^2 &= (\mathbf{w}^{(t)} - \mathbf{x})^T (\mathbf{w}^{(t)} - \mathbf{x}) \\ &= (\mathbf{w}^{(t)} - \mathbf{x})^T \mathbf{w}^{(t)} - (\mathbf{w}^{(t)} - \mathbf{x})^T \mathbf{x} \\ &= (\mathbf{w}^{(t)})^T \mathbf{w}^{(t)} - \mathbf{x}^T \mathbf{w}^{(t)} - (\mathbf{w}^{(t)})^T \mathbf{x} + \mathbf{x}^T \mathbf{x} \\ &= (\mathbf{w}^{(t)})^T \mathbf{w}^{(t)} - 2(\mathbf{w}^{(t)})^T \mathbf{x} + \mathbf{x}^T \mathbf{x} \end{aligned} \quad (2.28)$$

Since (2.27) and $\mathbf{x}^T \mathbf{x} = 1$, equation (2.28) becomes

$$\|\mathbf{w}^{(t+1)}\|_2^2 \leq \|\mathbf{w}^{(t)}\|_2^2 + 1 \quad (2.29)$$

Thus, we have proved (2.23) under **State 2**.

By the same induction defined in (2.21), we have

$$\|\mathbf{w}^{(m)}\|_2^2 \leq m \quad (2.30)$$

Finally, we merge result (2.22) and (2.30) using the Cauchy-Schwarz (CS) inequality
Recall that the CS inequality is given by

$$|\mathbf{w}^{*T} \mathbf{x}| \leq \|\mathbf{w}^*\|_2 \|\mathbf{x}\|_2 \quad (2.31)$$

Hence, using (2.31) we get

$$|(\mathbf{w}^m)^T \mathbf{w}^*| \leq \|(\mathbf{w}^m)^T\|_2 \|\mathbf{w}^*\|_2 \quad (2.32)$$

From the result (2.22), (2.32) becomes

$$m\gamma \leq |(\mathbf{w}^m)^T \mathbf{w}^*| \leq \|(\mathbf{w}^m)^T\|_2 \|\mathbf{w}^*\|_2 \quad (2.33)$$

Using the result (2.30) and the fact that $\|\mathbf{w}^*\|_2 = 1$, we get

$$\begin{aligned} m\gamma &\leq \|(\mathbf{w}^m)^T\|_2 \\ m^2\gamma^2 &\leq \|(\mathbf{w}^m)^T\|_2^2 \end{aligned} \quad (2.34)$$

$$m^2\gamma^2 \leq m \quad (2.35)$$

$$m \leq \frac{1}{\gamma^2} \quad (2.36)$$

This proves the Perceptron Convergence Theorem that the number of mistakes is at most $\frac{1}{\gamma^2}$, where γ is the margin.

2.4 Solving Gate problem

The perceptron alone is not always a good model to solve binary classification problem. It has certain limitation by nature of its definition. Consider the logic functions AND, OR and XOR across these 4 points:

$$\mathbf{p}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{p}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{p}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mathbf{p}_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (2.37)$$

2.4.1 AND Function

The AND function returns a value of 1 if both of the inputs is 1. Let f be the AND function such that

$$f(\mathbf{p}_1) = 0 \quad f(\mathbf{p}_2) = 0 \quad f(\mathbf{p}_3) = 1 \quad f(\mathbf{p}_4) = 0$$

Consider the AND function with the separating line below

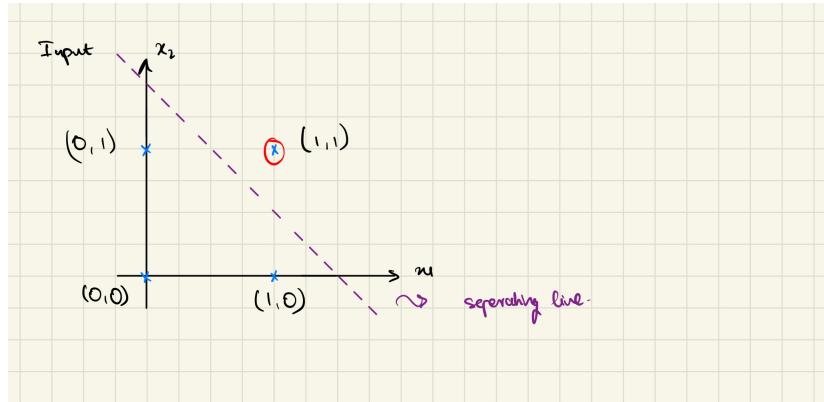


Figure 2.6: AND Function with separating line

The equation of the separating line in figure 2.6 is given by

$$x_1 + x_2 - 1.5 = 0 \quad (2.38)$$

We can transform the separating line in the form of $\mathbf{w}^T \mathbf{x} = 0$ to fit the perceptron.

Let

$$\mathbf{w} = \begin{bmatrix} -1.5 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad (2.39)$$

Applying function (2.1) on each point to check whether they are correctly classified.

$$\phi(p_1) = \phi \left(\begin{bmatrix} -1.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) = \phi(-1.5) = 0$$

$$\phi(p_2) = \phi \left(\begin{bmatrix} -1.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right) = \phi(-0.5) = 0$$

$$\phi(p_3) = \phi \left(\begin{bmatrix} -1.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) = \phi(0.5) = 1$$

$$\phi(p_4) = \phi \left(\begin{bmatrix} -1.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right) = \phi(-0.5) = 0$$

Thus, the perceptron is able to solve the AND function.

2.4.2 OR Function

The OR function returns a value of 1 if either of the inputs is 1. Let f be the OR function such that

$$f(\mathbf{p}_1) = 0 \quad f(\mathbf{p}_2) = 1 \quad f(\mathbf{p}_3) = 1 \quad f(\mathbf{p}_4) = 1$$

Consider the OR function with the separating line below

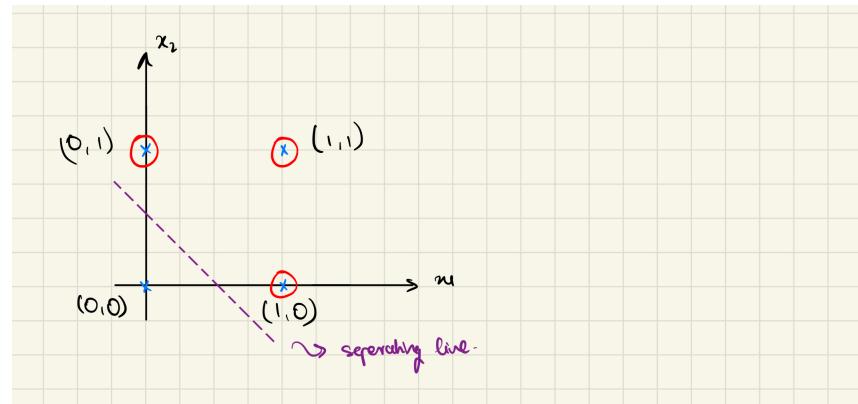


Figure 2.7: AND Function with separating line

The equation of the separating line in figure 2.7 is given by

$$x_1 + x_2 - 0.5 = 0 \quad (2.40)$$

We can transform the separating line in the form of $\mathbf{w}^T \mathbf{x} = 0$ to fit the perceptron.

Let

$$\mathbf{w} = \begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad (2.41)$$

Applying function (2.1) on each point to check whether they are correctly classified.

$$\phi(p_1) = \phi \left(\begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) = \phi(-0.5) = 0$$

$$\phi(p_2) = \phi \left(\begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right) = \phi(0.5) = 1$$

$$\phi(p_3) = \phi \left(\begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) = \phi(1.5) = 1$$

$$\phi(p_4) = \phi \left(\begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \right) = \phi(0.5) = 1$$

Thus, the perceptron is able to solve the OR function.

2.4.3 XOR Function

The XOR function returns a value of 1 if either of the inputs is 1 but not both. Let f be the XOR function such that

$$f(\mathbf{p}_1) = 0, \quad f(\mathbf{p}_2) = 1, \quad f(\mathbf{p}_3) = 0, \quad f(\mathbf{p}_4) = 1 \quad (2.42)$$

Consider a plot of the XOR function below

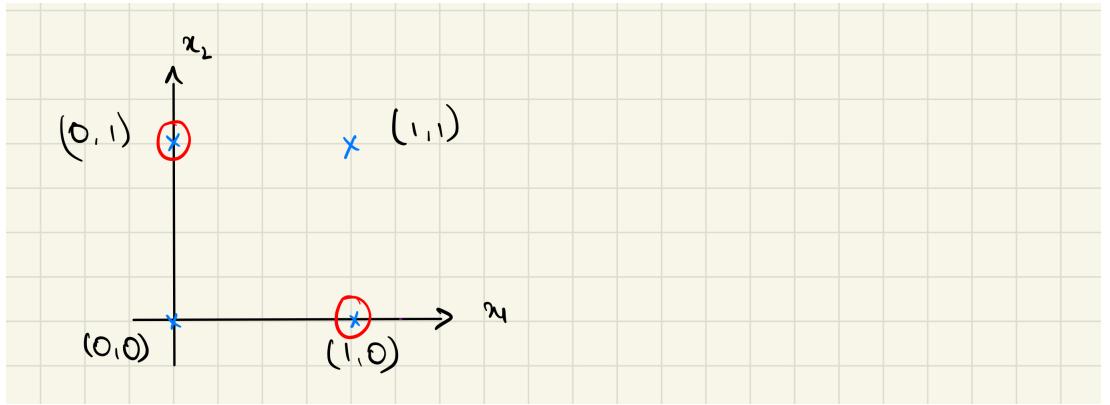


Figure 2.8: OR function

From figure 2.8; we can observe that no separating line alone could group the two different classes together. We attempt to find a combination of weight for the perceptron which reduces the error (margin) as much as possible.

Let f^* be the exact function and $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b$, where $\mathbf{w} = [w_1 \ w_2]^T$, be the approximate function. Consider the error function (the difference between the exact and approximate value)

$$\mathbb{E}(\mathbf{w}, b) = \frac{1}{4} \sum_{x \in D} (f^*(\mathbf{x}) - \mathbf{w}^T \mathbf{x} - b)^2 \quad (2.43)$$

we minimise the error function \mathbb{E} by setting the partial derivatives to 0. The partial derivative with respect to the bias is given by

$$\begin{aligned} \frac{\partial \mathbb{E}}{\partial b} &= -\frac{1}{2} \sum_{x \in D} (f^*(\mathbf{x}) - \mathbf{w}^T \mathbf{x} - b) \\ \frac{\partial \mathbb{E}}{\partial b} = 0 \rightarrow \sum_{x \in D} f^*(\mathbf{x}) &= \sum_{x \in D} \mathbf{w}^T \mathbf{x} - \sum_{x \in D} b \end{aligned} \quad (2.44)$$

The partial derivative with respect to the weight vector is given by

$$\begin{aligned}
\frac{\partial \mathbb{E}}{\partial \mathbf{x}} &= \frac{1}{2} \sum_{x \in D} (f^*(\mathbf{x}) - \mathbf{w}^T \mathbf{x} - b) \frac{\partial (-\mathbf{w}^T \mathbf{x})}{\partial \mathbf{w}} \\
&= -\frac{1}{2} \sum_{x \in D} (f^*(\mathbf{x}) - \mathbf{w}^T \mathbf{x} - b) \mathbf{x} \\
\frac{\partial \mathbb{E}}{\partial \mathbf{w}} &= 0 \rightarrow \sum_{x \in D} \mathbf{x} f^*(\mathbf{x}) = \sum_{x \in D} (\mathbf{w}^T \mathbf{x}) \mathbf{x} + b \sum_{x \in D} \mathbf{x}
\end{aligned} \tag{2.45}$$

Substituting for values in equation (2.44)

$$\begin{aligned}
\sum_{x \in D} f^*(\mathbf{x}) &= f^*(\mathbf{p}_1) + f^*(\mathbf{p}_2) + f^*(\mathbf{p}_3) + f^*(\mathbf{p}_4) \\
&= 0 + 1 + 1 + 0 \\
&= 2
\end{aligned} \tag{2.46}$$

$$\begin{aligned}
\sum_{x \in D} \mathbf{w}^T \mathbf{x} &= \mathbf{w}^T \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \mathbf{w}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \mathbf{w}^T \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \mathbf{w}^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= 2w_1 + 2w_2
\end{aligned} \tag{2.47}$$

$$\begin{aligned}
\sum_{x \in D} b &= b \sum_{x \in D} 1 \\
&= 4b
\end{aligned} \tag{2.48}$$

Thus, we get

$$\begin{aligned}
2 &= 2w_1 + 2w_2 + 4b \\
w_1 + w_2 + 2b &= 1
\end{aligned} \tag{2.49}$$

Substituting for values in equation (2.45)

$$\begin{aligned} \sum_{x \in D} \mathbf{x} f^*(\mathbf{x}) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} (0) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (1) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} (0) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (1) \\ &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned} \quad (2.50)$$

$$\begin{aligned} \sum_{x \in D} (\mathbf{w}^T \mathbf{x}) \mathbf{x} &= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &\quad + \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 2w_1 + w_2 \\ w_1 + 2w_2 \end{bmatrix} \end{aligned} \quad (2.51)$$

$$\begin{aligned} b \sum_{x \in D} \mathbf{x} &= b \begin{bmatrix} 0 \\ 0 \end{bmatrix} + b \begin{bmatrix} 1 \\ 0 \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= b \begin{bmatrix} 2 \\ 2 \end{bmatrix} \end{aligned} \quad (2.52)$$

Thus, we get

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2w_1 + w_2 \\ w_1 + 2w_2 \end{bmatrix} + b \begin{bmatrix} 2 \\ 2 \end{bmatrix} \quad (2.53)$$

From equations (2.49) and (2.53), we get the set of equations that forms the approximate function with the minimum error

$$w_1 + w_2 + 2b = 1$$

$$2w_1 + w_2 + 2b = 1$$

$$w_1 + 2w_2 + 2b = 1$$

Solving the system of linear equations we get

$$w_1 = 0 \quad w_2 = 0 \quad b = \frac{1}{2} \quad (2.54)$$

Hence, the approximate function $f(\mathbf{x}, \mathbf{w})$ is given by

$$f(\mathbf{x}, \mathbf{w}) = \frac{1}{2} \quad (2.55)$$

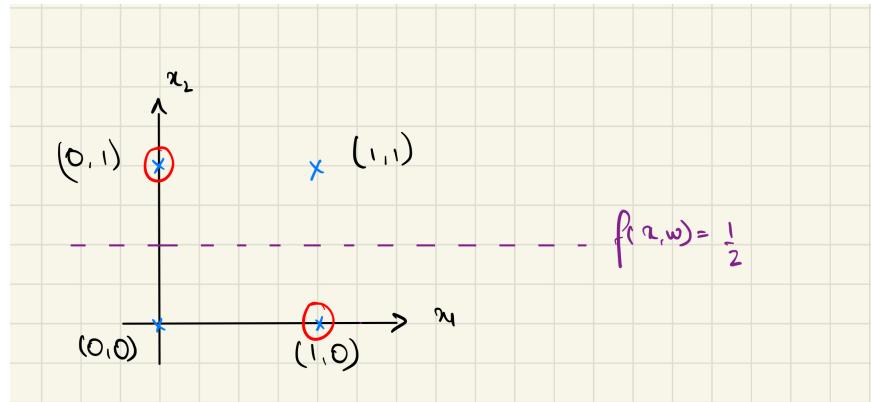


Figure 2.9: XOR approximate solution

Clearly from figure 2.9, the perceptron model cannot solve the XOR logic function. The XOR problem is actually not linearly separable. Thus, we introduce **feedforward networks** to overcome this problem.

2.5 Feedforward Network

We have seen that the single perceptron cannot solve the XOR problem. In order to solve the problem, we use a set of perceptrons that learns two different functions within a connected network.

First, let's consider two NOT boolean function $f_1^* = x_1 \text{ AND } (\text{NOT } x_2)$ and $f_2^* = (\text{NOT } x_1) \text{ AND } x_2$.

Consider f_1^* which returns 1 if x_1 and (not x_2) are equal to 1.

x_1	x_2	NOT (x_2)	x_1 AND (NOT x_2)
0	0	1	0
1	0	1	1
1	1	0	0
0	1	0	0

Table 2.1: Truth Table for f_1^*

The function f_1^* can be separated using the perceptron.

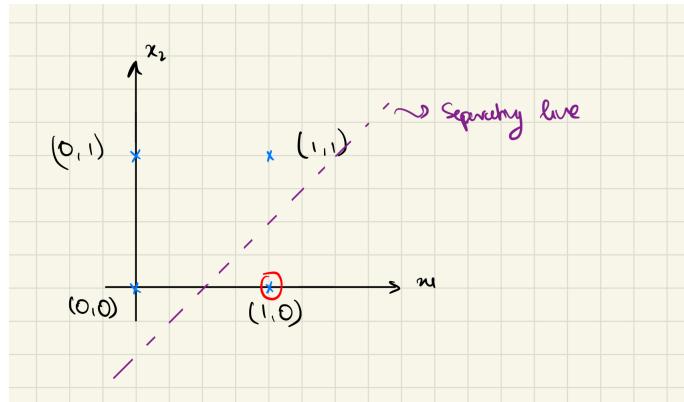


Figure 2.10: f_1

The equation of the separating line in figure 2.10 is given by

$$x_1 - x_2 - \frac{1}{2} = 0 \quad (2.56)$$

We can transform the separating line in the form of $\mathbf{w}^T \mathbf{x} = 0$ to fit the perceptron.

$$\mathbf{w} = \begin{bmatrix} -0.5 \\ 1 \\ -1 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad (2.57)$$

Similarly, consider f_2^* which returns 1 if (not x_1) and x_2 are equal to 1.

x_1	x_2	NOT (x_1)	(NOT x_1) AND x_2
0	0	1	0
1	0	0	0
1	1	0	0
0	1	1	1

Table 2.2: Truth Table for f_2^*

The function f_2^* can be separated using the perceptron.

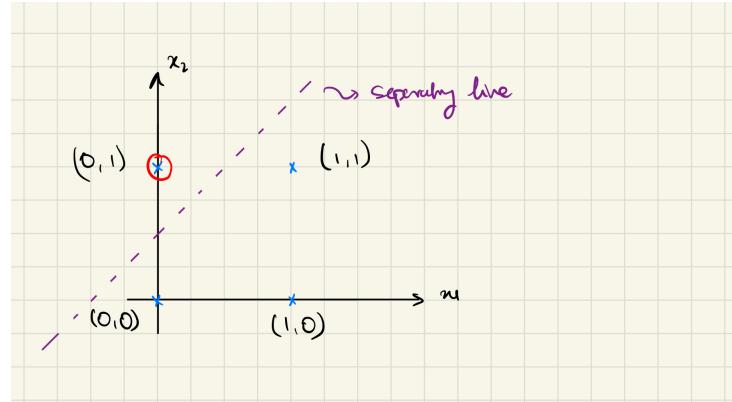


Figure 2.11: f_1

The equation of the separating line in figure 2.11 is given by

$$-x_1 + x_2 - \frac{1}{2} = 0 \quad (2.58)$$

We can transform the separating line in the form of $\mathbf{w}^T \mathbf{x} = 0$ to fit the perceptron.

$$\mathbf{w} = \begin{bmatrix} -0.5 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad (2.59)$$

Consequently, the XOR function defined in (2.42) can be expressed as f^* such that

$$f^*(x_1, x_2) = f_1^*(x_1, x_2) \text{ OR } f_2^*(x_1, x_2)$$

x_1	x_2	$f_1^*(x_1, x_2)$	$f_2^*(x_1, x_2)$	$f_2^*(x_1, x_2)$
0	0	0	0	0
1	0	1	0	1
1	1	0	0	0
0	1	0	1	1

Table 2.3: Truth Table for f^*

Consider a feedforward network with a hidden layer with two units h_1 and h_2 such that h_1 learns $f_1^*(x_1, x_2)$ and h_2 learns $f_2^*(x_1, x_2)$.

$$\begin{aligned} h_1 &= f_1^*(x_1, x_2) \\ &= \phi(z_1) \end{aligned}$$

where $z_1 = w_{11}x_1 + w_{12}x_2 + b_1$ and $w_{11} = 1$, $w_{12} = -1$ and $b_1 = -0.5$ from separating line in (2.57).

$$\begin{aligned} h_2 &= f_2^*(x_1, x_2) \\ &= \phi(z_2) \end{aligned}$$

where $z_2 = w_{21}x_1 + w_{22}x_2 + b_2$ and $w_{21} = -1$, $w_{22} = 1$ and $b_2 = -0.5$ from separating line in (2.59).

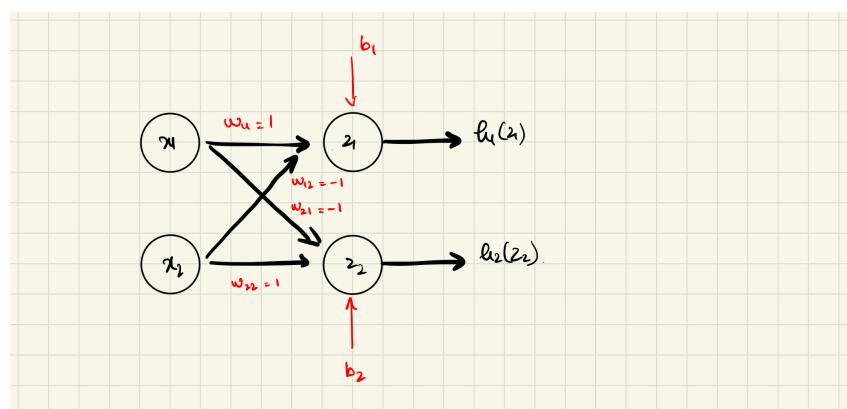


Figure 2.12: FFN hidden layer

Thus we obtain the system below for the hidden layer.

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} b_1 & w_{11} & w_{12} \\ b_1 & w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad (2.60)$$

We then feed forward the information from the hidden layer to another layer.

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \phi \left(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right) \quad (2.61)$$

At the hidden layer, we are left with the following truth table.

x_1	x_2	XOR	
0	0	0	
1	0	1	
0	0	0	redundant
0	1	1	

Table 2.4: Truth Table from hidden layer

The XOR problem can not be separated from using another perceptron after the hidden layer.

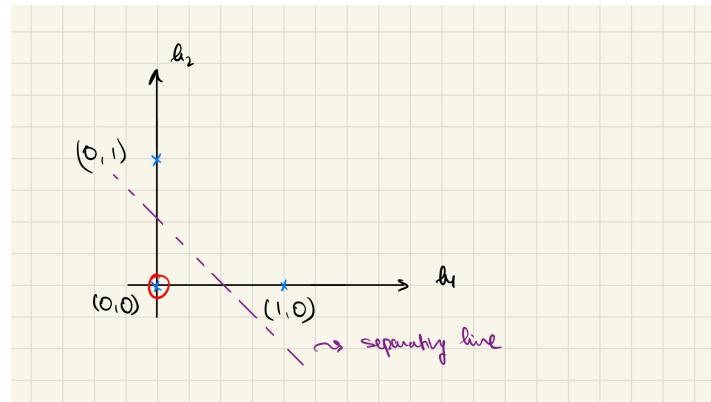


Figure 2.13: XOR layer

The equation of the separation line is in figure (2.13) is given binary

$$h_1 + h_2 - \frac{1}{2} = 0 \quad (2.62)$$

We can now transform the separating line in the form of $\mathbf{w}^T \mathbf{x} = 0$ to fit the perceptron.

$$\mathbf{w} = \begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{h} = \begin{bmatrix} 1 \\ h_1 \\ h_2 \end{bmatrix} \quad (2.63)$$

Finally, the XOR function can be represented using the set of transform defined below.

$$f(x_1, x_2) = \phi' \left(\mathbf{w}_2^T \left(\phi \left(\mathbf{w}_1^T \mathbf{x} \right) \right) \right) \quad (2.64)$$

where

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad \mathbf{w}_1 = \begin{bmatrix} -0.5 & -0.5 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \quad \mathbf{w}_2 = \begin{bmatrix} -0.5 \\ 1 \\ 1 \end{bmatrix} \quad (2.65)$$

ϕ' taken an augmented input from the calculation of $\phi(\mathbf{w}_1^T \mathbf{x})$ to accommodate for the bias.

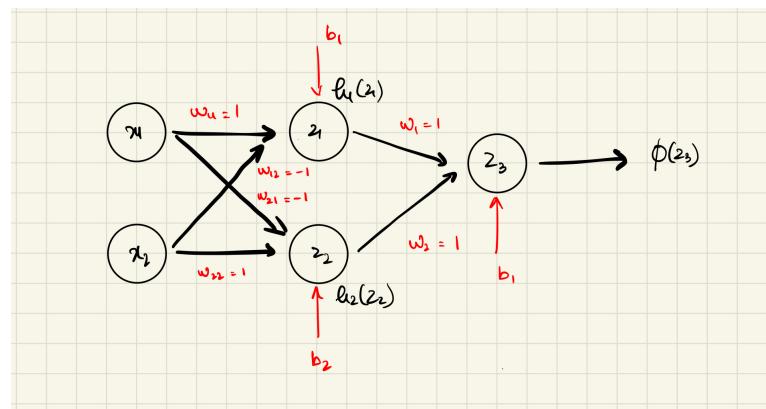


Figure 2.14: Feedforward network for XOR problem

Checking for $\mathbf{x} = [1 \ 0 \ 0]^T$

$$\mathbf{w}_1^T \mathbf{x} = \begin{bmatrix} -0.5 & 1 & -1 \\ -0.5 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}$$

$$\phi(\mathbf{w}_1^T \mathbf{x}) = \phi\left(\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\phi'(\mathbf{w}_2^T (\phi(\mathbf{w}_1^T \mathbf{x}))) = \phi'\left(\begin{bmatrix} -0.5 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = \phi'(-0.5)$$

Thus, for $\mathbf{p}_1 = [0, 0]^T$

$$f(x_1, x_2) = \phi'(-0.5) = 0$$

Checking for $\mathbf{x} = [1 \ 0 \ 1]^T$

$$\mathbf{w}_1^T \mathbf{x} = \begin{bmatrix} -0.5 & 1 & -1 \\ -0.5 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1.5 \\ 0.5 \end{bmatrix}$$

$$\phi(\mathbf{w}_1^T \mathbf{x}) = \phi\left(\begin{bmatrix} -1.5 \\ 0.5 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\phi'(\mathbf{w}_2^T (\phi(\mathbf{w}_1^T \mathbf{x}))) = \phi'\left(\begin{bmatrix} -0.5 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}\right) = \phi'(0.5)$$

Thus, for $\mathbf{p}_4 = [0, 0]^T$

$$f(x_1, x_2) = \phi'(0.5) = 1$$

2.6 Backpropagation

We have seen how information flows in a feedforward neural network. Backpropagation algorithm is the process of spreading the error at the output layer throughout the other layers in the network. The goal is to understand how does the error changes with respect to the weight and biases. Consider the neural network below with N number of layers (including the input and output layer).

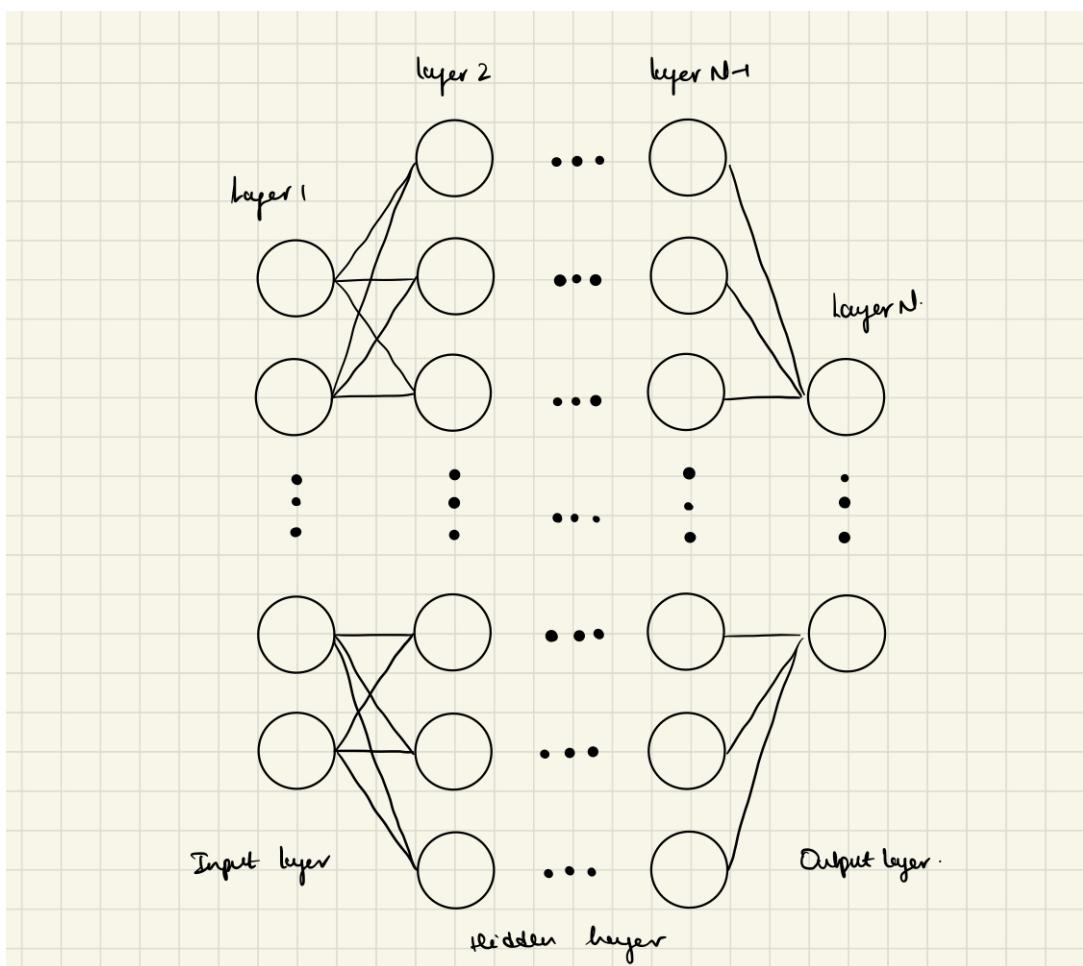


Figure 2.15: !!!TO BE CHANGED!!!

The annotation between two consecutive layers, l and $l + 1$, in the network is given by the following set of indexed values.

- $b_j^{l+1} \rightarrow$ Bias input at the j^{th} neuron in the $l + 1^{th}$ layer
 $w_{jk}^{l+1} \rightarrow$ Weight between the k^{th} and j^{th} neuron in the l^{th} and $(l + 1)^{th}$ layer respectively
 $z_j^{l+1} \rightarrow$ Input of the activation function at the j^{th} neuron in the $(l + 1)^{th}$ layer
 $a_j^l \rightarrow$ Output from the activation function at the j^{th} neuron in the l^{th} layer

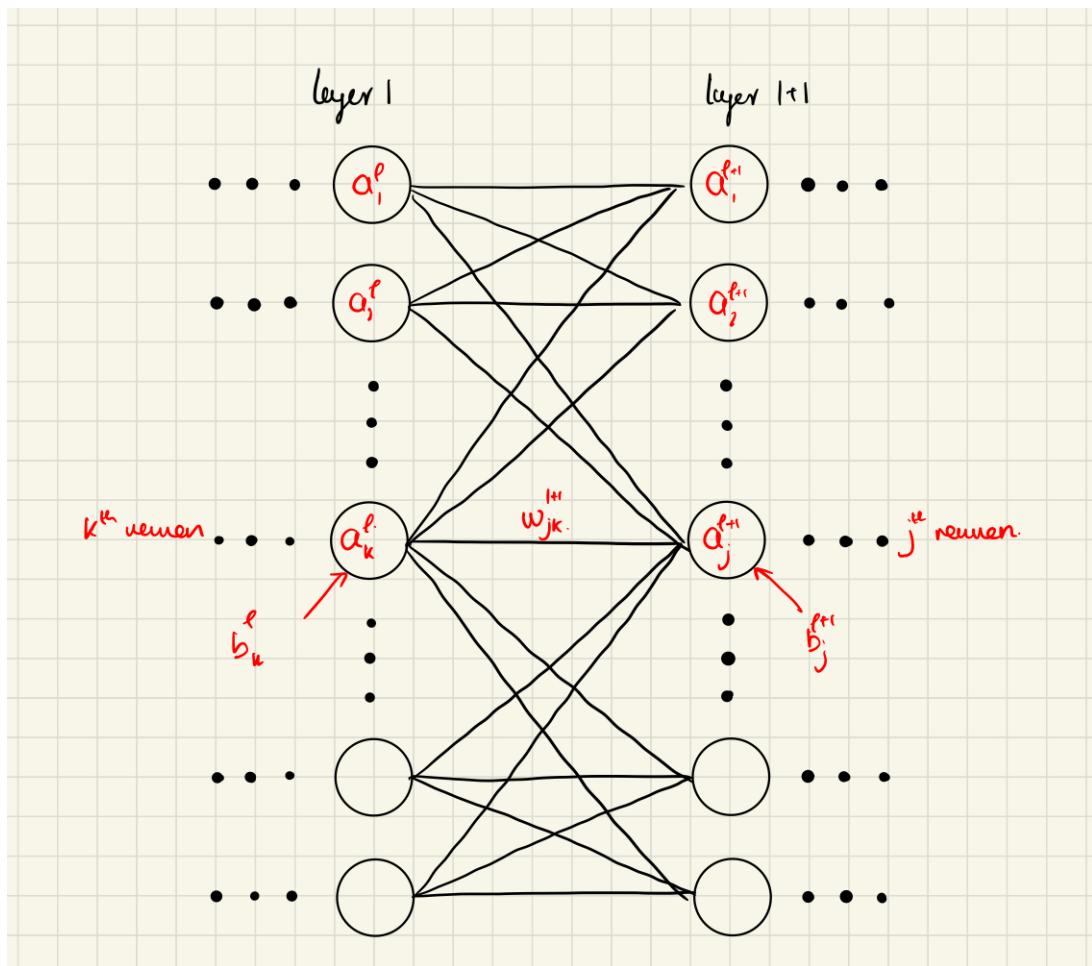


Figure 2.16: !!!TO BE CHANGED!!!

Let the activation function be denoted as σ such that

$$a_j^{(l+1)} = \sigma(z_j^{(l+1)}) \quad (2.66)$$

$$z_j^{(l+1)} = \sum_k w_{jk}^{(l+1)} a_k^{(l)} + b_j^{(l+1)} \quad (2.67)$$

We start at the end of the network by calculating the value of the cost/error function.

Let C be the cost function which takes as input the output layer

$$\begin{aligned} C &= C(a_j^L) \\ &= \frac{1}{2} \sum_j (y_j - a_j^L)^2 \end{aligned} \tag{2.68}$$

where y_j is the desired exact value that we want the network to predict. Also, let the change in the cost function with respect to the input of the activation function be given by δ_j^L where

$$\delta_j^L = \frac{\partial C}{\partial z_j^L}$$

Using result (2.66) and the definition (2.68)

$$\begin{aligned} \delta_j^L &= \frac{\partial C(a_j^L)}{\partial z_j^L} \\ &= \frac{\partial C(\sigma(z_j^L))}{\partial z_j^L} \end{aligned}$$

Applying the chain rule

$$\begin{aligned} \delta_j^L &= \frac{\partial C(\sigma(z_j^L))}{\partial z_j^L} \\ &= \frac{\partial C}{\partial a_j^L} \frac{\partial a_j^L}{\partial z_j^L} \end{aligned}$$

From equation (2.66)

$$\begin{aligned} a_j^{(L)} &= \sigma(z_j^{(L)}) \\ \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} &= \sigma'(z_j^{(L)}) \end{aligned}$$

Thus, the change in the cost function with respect to the input of the activation function in the last layer is given by

$$\delta_j^{(L)} = \frac{\partial C}{\partial a_j^{(L)}} \sigma'(z_j^{(L)}) \quad (2.69)$$

Next, we find how to back-propagate the changes from an upper layer $l + 1$ to the previous layer l . Consider the change in the cost function based at the l^{th} layer

$$\delta_j^{(l)} = \frac{\partial C}{\partial z_j^{(l)}} \quad (2.70)$$

From figure (2.16); an element in the l^{th} layer influences all the neurons in the $(l + 1)^{th}$ layer. Thus, when applying the chain rule we sum over all the neurons which gives

$$\delta_j^l = \sum_k \frac{\partial C}{\partial z_k^{(l+1)}} \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} \quad (2.71)$$

From equations (2.66) and (2.67)

$$\begin{aligned} \frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} &= \frac{\partial}{\partial z_j^{(l)}} \left(\sum_s w_{ks}^{(l+1)} \sigma(z_s^{(l)}) + b_k^{(l+1)} \right) \\ &= \sum_s w_{ks}^{(l+1)} \frac{\partial \sigma(z_s^{(l)})}{\partial z_j^{(l)}} \\ &= \sum_s w_{ks}^{(l+1)} \sigma'(z_s^{(l)}) \frac{\partial z_s^{(l)}}{\partial z_j^{(l)}} \end{aligned}$$

The partial derivative only exist and is 1 when $s = j$; else the value is 0.

$$\frac{\partial z_k^{(l+1)}}{\partial z_j^{(l)}} = w_{kj}^{(l+1)} \sigma'(z_j^{(l)}) \quad (2.72)$$

Substituting (2.72) and (2.70) into (2.71)

$$\delta_j^l = \sum_k \delta_k^{(l+1)} w_{kj}^{(l+1)} \sigma'(z_j^{(l)}) \quad (2.73)$$

Equation (2.73) are the changes in l^{th} layer from the changes that happened in the $(l + 1)^{th}$ layer. Since we do not explicitly adjust the output and input of the activation functions, we derive the changes of the cost function with respect to the bias and weight. Consider the effect of the bias at each layer given by

$$\frac{\partial C}{\partial b_j^{(l)}} = \frac{\partial C}{\partial z_j^{(l)}} \frac{\partial z_j^{(l)}}{\partial b_j^{(l)}}$$

Using equation (2.67)

$$\begin{aligned} \frac{\partial z_j^{(l)}}{\partial b_j^{(l)}} &= \frac{\partial}{\partial b_j^{(l)}} \left(\sum_k w_{jk}^{(l+1)} a_k^{(l)} + b_j^{(l+1)} \right) \\ &= 1 \end{aligned}$$

Thus, the change of in the cost function with respect to the bias at any layer is given by

$$\frac{\partial C}{\partial b_j^{(l)}} = \delta_i^{(j)} \quad (2.74)$$

Finally, consider the rate of change of the cost function with respect to any weight in the network given by

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \frac{\partial C}{\partial z_m^{(l)}} \frac{\partial z_m^{(l)}}{\partial w_{jk}^{(l)}} \quad (2.75)$$

Using equation (2.67)

$$\begin{aligned} \frac{\partial z_m^{(l)}}{\partial w_{jk}^{(l)}} &= \frac{\partial}{\partial w_{jk}^{(l)}} \left(\sum_s w_{ms}^{(l)} a_s^{(l-1)} + b_m^{(l)} \right) \\ &= \sum_s a_s^{(l-1)} \frac{\partial w_{ms}^{(l)}}{\partial w_{jk}^{(l)}} \end{aligned}$$

The partial derivative only exist and is 1 when $s = k$ and $m = j$; else the value is 0.

$$\frac{\partial z_j^{(l)}}{\partial w_{jk}^{(l)}} = a_k^{(l-1)} \quad (2.76)$$

Substituting (2.72) and (2.76) into (2.75)

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)} \quad (2.77)$$

The four main back-propagation equations are given by

$$\delta_j^{(L)} = \frac{\partial C}{\partial a_j^{(L)}} \sigma'(z_j^{(L)}) \quad (2.78)$$

$$\delta_j^l = \sum_k \delta_k^{(l+1)} w_{kj}^{(l+1)} \sigma'(z_j^{(l)}) \quad (2.79)$$

$$\frac{\partial C}{\partial b_j^{(l)}} = \delta_i^{(j)} \quad (2.80)$$

$$\frac{\partial C}{\partial w_{jk}^{(l)}} = \delta_j^{(l)} a_k^{(l-1)} \quad (2.81)$$

We vectorize the above set of equations by using the Hadamard product. Given two matrices of the A and B of the same dimension, $m \times n$, the Hadamard product is given as

$$(A \odot B)_{mn} = (A)_{mn} (B)_{mn}$$

The vectorized form of the back-propagation equations are given by

$$\delta^{(L)} = \nabla_{a^{(L)}} C \odot \sigma'(z^{(L)})$$

$$\delta^{(l)} = (w^{(l+1)})^T \sigma^{(l+1)} \odot \sigma'(z^{(l)})$$

$$\nabla_{b^{(l)}} C = \delta^{(l)}$$

$$\nabla_{w^{(l)}} C = \delta^{(l)} a^{(l-1)}$$

writing a few lines

Algorithm 2 Back-Propagation Algorithm

Require:

1: Let $t = 0$ and $\mathbf{w} = [0 \ 0 \ \dots \ 0]^T$

2:

Chapter 3

Optimization

3.1 Introduction

The optimization problem is a computational problem in which the objective is to find the best of all possible solutions. Deep neural networks is a form of an optimization problem to find the best possible set of weights in order to reduce the error in a network.

A generic form of an optimization problem is given by

$$\begin{aligned} & \underset{x}{\text{minimize/maximize}} && f(x) \\ & \text{subject to} && g_i(x) \leq 0, \quad i = 1, \dots, m \\ & && h_j(x) = 0, \quad j = 1, \dots, p \end{aligned} \tag{3.1}$$

where

$f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective/loss function to be minimised,
 $g_i(x) \leq 0$ are called inequality constraints,
 $h_j(x) = 0$ are called equality constraints, and
 $m \geq 0$ and $p \geq 0$. If $m = p = 0$, the problem is an unconstrained optimization problem.

Minimum/Maximum Point

Let x^* be points of the function $f(x)$ defined in (3.1) and $u \in \mathbb{N}(x^*, \delta)$ where \mathbb{N} is a δ -neighbourhood of x^* . Then, x^* is said to be a

- (i) local minimum if $f(x^*) < f(u) \forall u \in \mathbb{N}$
- (ii) local maximum if $f(x^*) > f(u) \forall u \in \mathbb{N}$
- (iii) global minimum if $f(x^*) < f(u) \forall u \in \mathbb{R}$
- (iv) global maximum if $f(x^*) > f(u) \forall u \in \mathbb{R}$

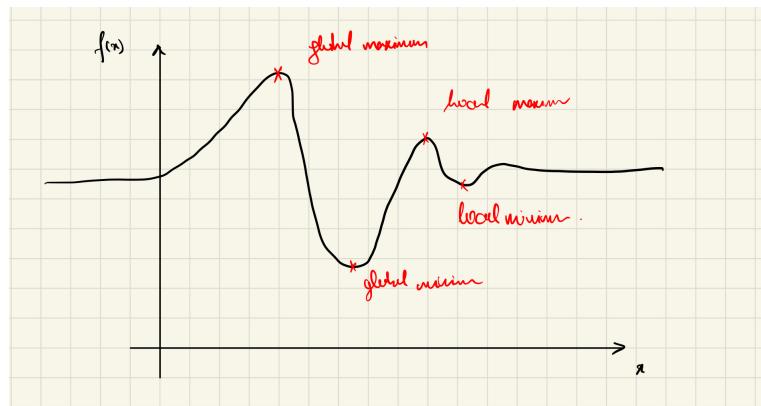


Figure 3.1: !!!TO BE CHANGED!!! Min max

Determining Minimum/Maximum Point

For a function $f(\mathbf{x})$ where $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, the condition for the presence of a stationary point (minimum or maximum point) is given by

$$\mathbf{G} = \nabla f \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right) = \mathbf{0} \quad (3.2)$$

The second derivative of $f(x)$ is given by the Hessian matrix, \mathbf{H}

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (3.3)$$

The nature of the stationary points of $f(x)$ can be determined by studying the positive definiteness of (3.3) using its eigenvalues. If all the eigenvalues of \mathbf{H} are positive, then \mathbf{H} is symmetric positive definite, indicating the presence of a **local minimum**.

3.2 Optimization algorithms

The solutions to the optimization problem are vital in modern machine learning and artificial intelligence algorithms, which includes weight optimization in deep learning. There are a number of popular optimization algorithm currently developed to solve the problem. Hence, choosing the right algorithm can be challenging as well.

Network Error

Let the network error function, e , be given by the squared difference between the predicted value and the true value. The difference is squared to avoid the sum of errors of multiple input vector to be zero which can mislead the network to have perfect predictive power.

$$e = \frac{1}{2} \sum_i^{\text{Total Input}} (\hat{y}_i - y_i)^2 \quad (3.4)$$

The goal of the network is to adjust the weight to reduce the network error as much as possible. The idea of reducing a function to a value is synonymous to (3.1) an optimization problem. The network error in (3.4) is a quadratic equation in this case.

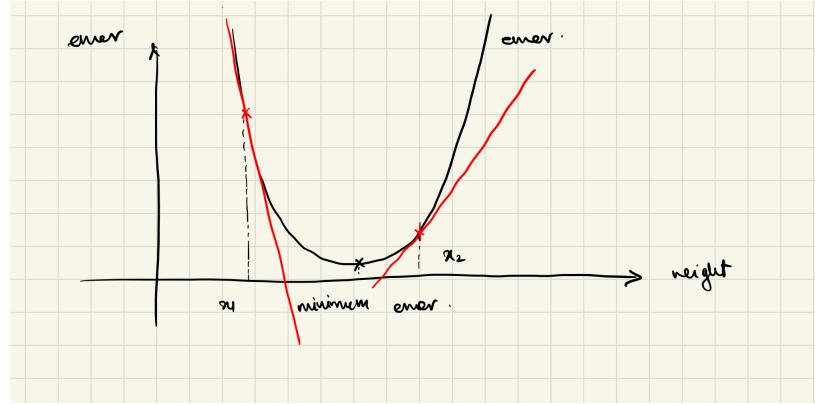


Figure 3.2: Network Error

Figure (3.2) is a graphical representation of the network error function. Our goal is to reach the minimum of the function by adjusting the weight value.

3.2.1 Gradient Descent

In order to reach the bottom of the function; we need to adjust the weight such that the derivative of the error function is 0. The approach of updating the weight based on the gradient of the error function is known as **gradient descent**. The derivative with respect to the weights of the network error (3.4) for a single input is given by

$$\frac{\partial e}{\partial w_{ij}} = (\hat{y}_i - y_i) \frac{\partial \hat{y}_i}{\partial w_{ij}} \quad (3.5)$$

From figure (3.2); we can observe that if the gradient is negative then we have underestimated the predicted value and need to increase the weight to reach the optimal value. On the other hand, if the gradient is positive then we have overestimated the predicted value and need to decrease the weight to reach the optimal value. The equation (3.5) provides us with the opposite direction and amount to adjust the weight. Hence, the update rule of the weights for gradient descent method is given by

$$w_{ij}^{t+1} = w_{ij}^t - \frac{de}{dw_{ij}} \quad (3.6)$$

In the gradient descent method, the network learns from the gradient of the error function and adjust the weights accordingly to reduce the error. However, the gradient alone can be quite large, causing oscillations as we go down the error function. This problem by introducing a **learning rate** α prior to adjusting the weight.

Learning Rate

The learning rate is typically denoted by the Greek letter alpha α . It helps the network to control the rate at which the weights are changing. Having a system with a high learning rate may lead to an oscillating network when trying to find the optimal weight and having a slow learning rate increases the number of iterations required when optimizing the network. The adjusted update rule of the weight is given by

$$w_{ij}^{t+1} = w_{ij}^t - \alpha \frac{\partial e}{\partial w_{ij}} \quad (3.7)$$

where $\alpha \in (0, 1)$

3.2.2 Stochastic Gradient Descent

The Stochastic Gradient Descent dates back to the Robbins-Monro(ADD REF) algorithm from the 1950s and is still an important optimization algorithm. Instead of adjusting the weight to the average loss function, we can approximate the gradient by only one random directional derivative of the error. Thus, the number of computation in the gradient descent method can be significantly reduced by a factor of n (where n is the number of direction). The stochastic gradient method (SGM) is given by

$$w^{t+1} = w^t - \alpha_t (\nabla e)_i \quad (3.8)$$

where α_t is the learning rate at the t^{th} iteration which may vary with iterations.

As we move closer to the bottom of the minimum of the function; the solution starts

to oscillate since we are moving the weight in random directions. Thus, the learning rate should be reduced gradually as we iterate. Some commonly used reduction of learning rate is given by

$$\eta(t) = \eta_i \quad \text{if } t_i \leq t \leq t_{i+1} \quad \text{piecewise constant} \quad (3.9)$$

$$\eta(t) = \eta_0 e^{-\lambda t} \quad \text{exponential decay} \quad (3.10)$$

$$\eta(t) = \eta_0 (\beta t + 1)^{-\alpha} \quad \text{polynomial decay} \quad (3.11)$$

where λ , β , and α are known as hyperparameter.

Training dataset can be very large in certain cases which leads to a greater cost of compute at each iteration for the gradient descent, so stochastic gradient descent is preferred in these cases.

3.2.3 Mini-Batch Stochastic Gradient Descent

In order to make use of the features of both gradient descent and stochastic gradient descent; we introduce the mini-batch stochastic gradient descent. In this approach, instead of iterating in the direction of the full gradient or only in one direction of the full gradient, we split the dataset into small batches and compute the gradient of each batch. The formula of the stochastic gradient descent is given by

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \mathbf{g}_t \quad (3.12)$$

$$\mathbf{g}_t = \nabla_{\mathcal{B}_t} e(\mathbf{x}, \mathbf{w})$$

where \mathcal{B}_t is a mini-batch of elements drawn uniformly at random from the training set (the expectation of the gradient remains unchanged).

Since we are using a mini-batch, the updates are closer to the full gradient but at a reduced cost.

3.2.4 Newton's Method

For minimizing $f(x)$, $x \in \mathbb{R}$, we need to solve $g(x) = f'(x) = 0$. Newton's iteration is given by

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} \quad (3.13)$$

$$= x_n - \frac{f'(x_n)}{f''(x_n)} \quad (3.14)$$

For multivariate functions we need to minimise $f(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^n$, that it

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n \quad (3.15)$$

The Newton's iteration for multivariate function is given by

$$\mathbf{x}_{n+1} = \mathbf{x}_n - H(\mathbf{x}_n)^{-1} \nabla f(\mathbf{x}_n) \quad (3.16)$$

where

$H(\mathbf{x}_n)$ is the Hessian matrix of $f(\mathbf{x})$

We can observe from (3.3) that calculating the inverse of the Hessian matrix can be computationally very expensive for higher dimensions. Replacing $H(\mathbf{x}_n)^{-1}$ by $\alpha \mathcal{I}$ where \mathcal{I} is the identity matrix; we get the **method of steepest descent** given by

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha \mathbf{I} \nabla f(\mathbf{x}_n) \quad (3.17)$$

where $\alpha \in (0, 1)$

3.2.5 Momentum

Exponentially Weighted Moving Average

The exponentially weighed moving average (EWMA), also known as exponential moving average (EMA), of a series of data points S_t is given by

$$\mu_t = \beta\mu_{t-1} + (1 - \beta)S_t \quad (3.18)$$

$$\mu_0 = c$$

where $c \in \mathbb{R}$ and $\beta \in (0, 1)$ is known as the smoothing constant. β represents the weightage that is going to be assigned to the past values. The average number of previous reading is approximately given by $n = (1 - \beta)^{-1}$. The higher the value of β the greater the number of points we average over.

SGD with Momentum

The momentum method was mentioned in Rumelhart, Hinton and Williams' paper on back-propagation learning in 1986 (ADD REF). From the SGD, it was observed that the weight update can be very noisy. In order to smoothen the search direction in the SGD, an exponentially moving average is implemented on the gradient. The SGC with momentum can be written in the form

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\boldsymbol{\mu}^t \quad (3.19)$$

where the exponential moving average of the gradient is given by

$$\boldsymbol{\mu}^t = \beta\boldsymbol{\mu}^{t-1} + (1 - \beta)\nabla(e)^t$$

$$\boldsymbol{\mu}_0 = \mathbf{c}$$

3.2.6 RMSProp

Proposed by Geoff Hilton (ADD REF) in lecture 6 of the online course "Neural Network for Machine Learning", he Root Mean Square Propagation (RMSProp)

algorithm alleviates undesirable fluctuations when adjusting the weight during optimization. RMSProp is an adaptive learning rate. The idea is to reduce the step size at very large gradient to avoid fluctuations and increase the step size at smaller gradient to move steadily in the correct direction of the optimal solution; hence an adaptive learning rate at each iteration. The equations for the RMSProp is given by

$$\begin{aligned} S_i^t &= \gamma S_i^{t-1} + (1 - \gamma)(\nabla e^t)_i^2 \\ w_i^{t+1} &= w_i^t - \frac{\eta}{\sqrt{S_i^t} + \epsilon} (\nabla e^t)_i \end{aligned} \quad (3.20)$$

where $\epsilon \approx 10^{-6}$ to ensure that we avoid dividing by zero during iterations.

3.2.7 Adam

The Adam algorithm (Adaptive Moment Estimation algorithm) is a robust update rule for the weight optimization, it was proposed by Kingma and BA in 2014 (ADD REF). The algorithm combines the benefits of momentum (3.19) and RMSProp (3.20). Adam is the most popular generalized algorithm performs very well in many cases. It is considered as a state-of-the-art algorithm for deep neural network optimization. The set of equations for the Adam algorithm is given by

$$\begin{aligned} \mu_i^{(t)} &= \beta_1 \mu_i^{(t-1)} + (1 - \beta_1) \nabla(e^{(t)})_i && \text{Momentum} \\ S_i^{(t)} &= \beta_2 S_i^{(t-1)} + (1 - \beta_2) (\nabla e^{(t)})_i^2 && \text{RMSProp} \end{aligned}$$

To prevent S^t and μ^t from becoming zero during the initial steps, a bias correction is introduced, such that

$$\hat{\mu}^{(t)} = \frac{\mu}{1 - \beta_1^t} \quad \hat{S}^{(t)} = \frac{S^t}{1 - \beta_2^t}$$

Finally, the update rule is given by

$$w_i^{t+1} = w_i^t - \frac{\eta}{\sqrt{S_i^t} + \epsilon} \hat{\mu}_i \quad (3.21)$$

The parameters β_1 , β_2 and ϵ are known as the hyperparameters of the Adam algorithm. A common set of values that works well in literature are $\beta_1 = 0.9$, $\beta_2 = 0.9999$ and $\epsilon = 10^{-8}$.

Chapter 4

Sentiment Analysis

4.1 Introduction

People's opinions, feelings and sentiments towards entities such products, services, other people, events, news, issues, topics, etc. can be in very large volume, complex and difficult to be understood and processed by machines and computers. Thus, sentiment analysis, also known as opinion mining, started to popularised along the rise of social media when large amount of digital text data were suddenly available for mining. Natural language processing (NLP) helps computers process and understand human based language to perform repetitive task. Sentiment analysis is a niche of NLP. It aims at quantifying the positivity, negativity and/or neutrality of implied or expressed in a given text.

Social media have been providing large platforms for people to share their opinion freely and expressed their views on any subject across various geographical and spatial boundaries. They have also allowed people to connect, influence and be influenced by such opinions and views. These interactions have been studied in the 1940s and 1950s among people in organizations by management science researchers. Since 2002, with social media, those studies have been performed at grand scales with the abundance of data. Thus, advanced sentiment analysis research have been performed in field political science, economics, finance and management science as they are heavily dependent on public opinions.

Asur and Huberman (2010) (ADD REF) attempted to solve the revenue predic-

tion problem using both the tweet volume and the tweet sentiment. Same kind of data along with polling results were also used, in Bermingham and Smeaton (2011) (ADD REF), to train a linear regression model to predict election results. Zhang et al. (2010) (ADD REF) leveraged sentiments on Twitter to help predict the movement of stock market indices such as Dow Jones, S&P500 and NASDAQ. It was shown that large negative emotions and opinions caused Dow to go down the next day and lower negative emotions cause Dow to go up on the next day. Bar-Haim et al. (2011) (ADD REF) on one hand leverage sentiments on Twitter but on the other did not treat all Twitter authors equally. Only expert investors were used as features in training stock price movement predictors. Undeniably, modern social media (started early 2000s) have grown into a major influencer of human opinion and sentiments.

Bibliography

AuthorName (Year *a*), Conferencetitle, *in* ‘BookTitle’.

AuthorName (Year *b*), ‘Title of article’, *Journal* .

BookAuthors (Year), *BookTitle*, Publisher.

- McCullough - First Neural Network Paper 1943 - F. Rosenblatt - 1958 - 1962