Addis Ababa University

A Graduate project report

On

Feed – Forward Networks

Submitted to Addis Ababa University, Faculty of Computer and Mathematical Sciences, Department of Mathematics in Partial fulfillment of the requirements for the Master of Science in Mathematics

By: Dereje Kifle

Advisor: Birhanu Bekele (PhD)

January , 2011

Addis Ababa, Ethiopia

List of Figures

Declaration

I declare that this project has been composed by me and that no part of the project has formed the basis for the award of any degree, diploma, associate ship, fellowship or any other similar title to me.

Dereje Kifle

Signature …… Date ……..

Permission

This is to certify that this project is compiled by **Dereje Kifle** in the Department of mathematics, Addis Ababa University, under my supervision. I hereby also confirm that the project can be submitted for evaluation by examiners and eventual defense.

Dr. Birhanu Bekele

Signature …… Date ……..

Table of Contents

Acknowledgement

First and foremost, I would like to thank the Almighty God who helped me with his strength and gave me wisdom to complete this project as well as the graduate program study.

Next, I express my sincere gratitude to my advisor Dr.Birhanu Bekele. for his genuine advice, constructive comments, invaluable suggestions and provision of materials (references) in preparing this project. I also extend my deep appreciation to Mr. Yonas Abebe (M.Sc.) for his thoughtful comments and suggestion, and material support. And also I would like to thank my friends and colleagues for their help and support in finding the necessary materials.

Last but not least, I am grateful to Addis Ababa University, department of Mathematics for its material support and to Gede'o Zone Building Capacity Office for their financial support.

Dereje Kifle

Abstract

The complete set of this project is about feed – forward networks. The realization of Boolean function for linearly separable problem by a single perceptron is possible. The problem is that how the non-linearly separable problems can be realized by a perceptron. This project intended to answer this question particularly the XOR problem. To this end, it is often advantageous to link several perceptrons in order to increase the number of functions that can be represented. So, this project is devoted to multi – layer networks (feed – forward networks) that really solves this problem using one – layer perceptron.We are going to do this with the help of separating hyperplanes.

List of Figures

Glossary of Symbols

N : The set of Natural Numbers.

Z_2 : Integers modulo two.

Z : The set of Integers.

Q : The set of Rational Numbers.

$\mathbb{R}$: The set of Real Numbers.

Abbreviations

- **NN :** Neural Networks
- **ANN:** Artificial Neural networks
- **BNN:** Biological Neural Networks
- **PLA :** Perceptron Learning Algorithms
- **FFN:** Feed – Forward Network
- **BPA:** Back – Propagation Algorithms
- $\boldsymbol{Conv(A)}$: Convex hull of set *A*

Introduction

One of the modern technologies that help the users to communicate to each other and facilitate the communication is a network. When the users are neurons this network is neural network. Whence a neural network is, in essence, an attempt to simulate the brain. Neural network theory revolves around the idea that certain key properties of biological neurons can be extracted and applied to simulations, thus creating a simulated (and very much simplified) brain. To this end, the first artificial neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pitts. They developed a simple but fundamental model which has the capacity to realize the elementary logical functions **NOT, AND, OR.** It is known that the structure and dynamics of biological neuron is very complex this is because the human brain is estimated to have ten to a hundred billion neurons and each neuron on average connected to 10,000 other neurons. That is, the network is relatively sparsely connected. Each neuron receives signals through synapses that control the effects of the signal on the neuron. These synaptic connections are believed to play a key role in the behavior of the brain. Artificial neural network is an attempt to approach the marvelous world of a real neural network: the human brain. The fundamental building block in an artificial neural network is the mathematical model of a neuron. So, we can easily understand the complexity of biological neuron via artificial neural network. This is because unlike biological neural networks, artificial neural networks do not have more than 1,000 artificial neuron. Among numerous different (artificial) neural network architectures the feed – forward neural network is the one which allow signal to travel one way only; from input to output. So this project focuses only on this type of network.

In chapter one, the perceptron (a simple feed – forward network) is introduced as a simple model of a nerve cell. It is shown that perceptrons are closely linked with the problem of linear separation. The perceptron learning algorithm will be described, and the issue of optimal separation will be studied. It will be shown that an optimal linear separation of finite disjoints sets can be achieved by solving a linear quadratic

optimization problem. In the second chapter, the structure of feed – forward network will be shown. To realize any Boolean functions, multi – layer perceptron will be studied and the issue of back – propagation algorithm will be described.

CHAPTER ONE: PERCEPTRONS

1.1 Formal Neuron

Definition of Key terms

We begin by noticing the following definitions:

Network: A network is a set of points, called **nodes**, and a set of curves called **branches** (or **arcs** or **links**), that connect certain pairs of nodes.

***Neuron*:** It is a functional unit of nervous system.

Neural Network (NN): It is powerful data modeling tool that is able to capture and represent complex input-output relationships. NN technology performs "intelligent" tasks similar to those performed by the human brain. It acquires knowledge through learning and then stores that knowledge within interneuron connection strengths known as synaptic weights. In NN model simple nodes, which can be called variously "neurons," "neurodes," "processing elements," or "units" are connected together to form a network of nodes –hence the term "NN."

Structure of Neuron

The neuron contains all structures of an animal cell. However, the structure and dynamics of biological neuron is very complex. Structurally the neuron can be divided into three major parts

i. the cell body(soma)
ii. the dendrites; and
iii.the axon

Consider the figure shown below

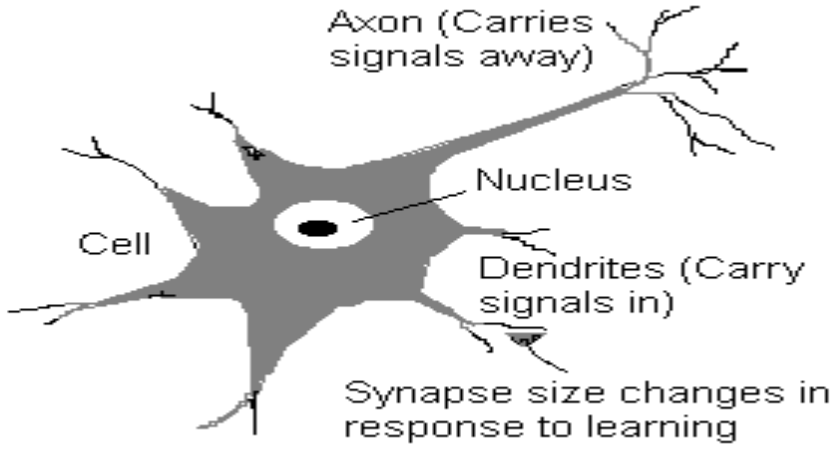

Fig1.1: *Structure of Neuron*

The cell body contains the organells of the neuron and also the dendrites are orginating there. Input connection are made from the axon of other cells to the dendrites or directly to the body of the cell. These are known as axosomatic synapses.

Traditionally, the term **NN** had been used to refer to a network or circuit of biological neuron; the modern usage of the term often refers to artificial neural networks (ANN), which are composed of artificial neurons or nodes.

Biological neural networks (BNN) are made up of real biological neurons that are connected or functionally related in the peripheral nervous system or the central nervous system. **ANN** are made up of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). ANN may either be used to gain an understanding of BNN, or for solving artificial intelligence problems without necessarily creating a model of a real biological system.

An artificial neuron: It is a device with many inputs and one output.

To achieve a simple and easily manageable models of artificial neurons (as shown below),it is necessarily to comprehend the essential functional characteristics of a neurons in a drastically simplified form. Such models are referred to as ***formal neurons.***

We conduct these neural networks by first trying to deduce the essential features of neurons and their interconnections. We then typically program a computer to simulate these features. However because our knowledge of neurons is incomplete and our computing power is limited, our models are necessarily gross idealizations of real networks of neurons.

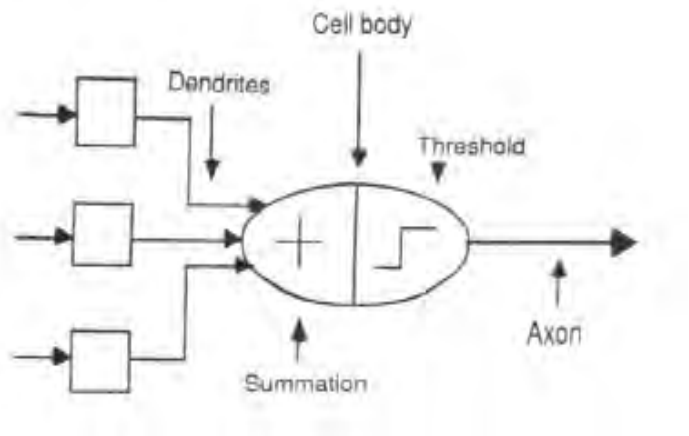

Fig1.2.*The artificial neuron model*

Before going to define what a formal neuron mean let us define the following functions:

A function $\sigma : \mathbb{R} \to Y$, where $Y \subseteq \mathbb{R}$ is called **Output map** & Y is an output value set.

A function s : $X \to \mathbb{R}$, where $X \subseteq \mathbb{R}^n$ (for some positive integer n) is called an **activation map** & X is an input value sets.

Definition 1.1: Let σ and s be as defined above. Then a data quadruple **$(X, Y, \sigma,$ s)** ,where n is number of inputs is called **formal neuron.**

Definition1.2: A function $f : X \to Y$, where $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}$ is called **transfer function** or **input – output map**. Thus a formal neuron sometimes will be identified with its transfer function.

Definition1.3: An affine function is any function of the form $f(w) = \alpha w + \beta$, where α and β are constants.

In 1943, **Warren McCulloch and Walter Pitts** developed a simple but fundamental model which has the capacity to realize the elementary logical functions **NOT**, **AND, OR**.

From biological point of view, the input signals can be viewed to stem from receptors of connected neurons .The signals are modified at the synapses and condensed into a single signal at the soma. If the quantity surpasses a certain threshold, the neurons fires (excited). The simplest way to model such a neuron therefore uses $Y = \{0, 1\}$ as its output value set.

Thus, Output $= \begin{cases} 1 \text{ if the neuron fires} \\ 0 \qquad otherwise \end{cases}$

And the affine linear map

$$s(x_1, x_2, \ldots, x_n) = \sum_j^n w_j\, x_j - \theta \tag{1.1}$$

as an activation function. The weights $w_j \in \mathbb{R}$ model the influence of the synapses on the signal x_j and $\theta \in \mathbb{R}$ represents the threshold (the smallest detectible sensation).

A function Sat**:** $\mathbb{R} \rightarrow \{0, 1\}$, which is defined by

$$\text{Sat}(z) = \begin{cases} 1 \text{ if } z \geq 0 \\ 0 \text{ othewise} \end{cases}$$

is called **Heaviside function.**

If this map is used as the output map, such a formal neuron transmits the signal **1** if and only if $\sum_j^n \boldsymbol{w_j\, x_j} - \boldsymbol{\theta} \geq \mathbf{0}$. Such a formal neurons are called **McCulloch – Pitts neurons or perceptrons.**

Definition 1.4: A function $\sigma : \mathbb{R} \rightarrow [0,1]$ with

$$\lim_{z \rightarrow \infty} \sigma(z) = 1 \;\&\; \lim_{z \rightarrow -\infty} \sigma(z) = 0$$

is called **sigmoid function.**

Common examples of sigmoid functions are:

1. Heaviside function Sat

Since it is a monotone function with $\lim_{z\to\infty} \mathrm{Sat}(z) = 1$ and $\lim_{z\to-\infty} \mathrm{Sat}(z) = 0$.

2. Ramp function, where for $\alpha > 0$

$$\sigma(z) = \begin{cases} 0 & \text{if } z \le -\alpha \\ \frac{1}{2}\left(\frac{z}{\alpha} + 1\right) & \text{if } -\alpha < z < \alpha \\ 1 & \text{if } \alpha \le z \end{cases}$$

3. The Fermi function $\sigma(z) = \frac{1}{1+e^{-z}}$ and its scaled version is $\sigma(z) = \frac{1}{1+e^{-\alpha z}}$ where $\alpha \ge 0$.

4. The modified hyperbolic tangent

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

In summary, we arrive at the following definition

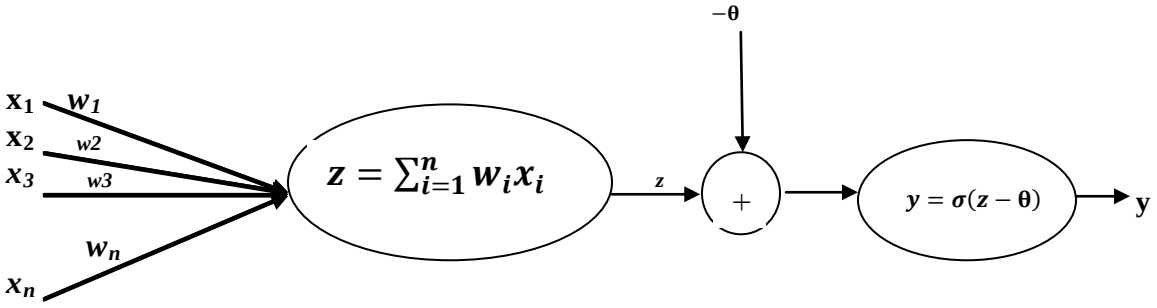

Figure 1.3: *Representation of a formal neuron with an affine activation function. For the perceptron, the Heaviside function* Sat *is used as the out–put function* σ.

Definition1.5: Let $s(x_1, x_2, \ldots, x_n) = \sum_j^n w_j x_j - \theta \quad w_j, \theta \in \mathbb{R},\ i = 1, \ldots, n$. A formal neuron (X, Y, σ, s) is called a σ **−Perceptron** and if $\sigma \equiv \mathrm{Sat}$, then a formal neuron **(*X*,*Y* , Sat , s)** is called a perceptron or McCulloch – Pitts neuron, where $Y = \{0, 1\}$.

Note: If a formal neuron **(X, Y, σ, s)** is a σ −perceptron it is easily shown that it is perceptron.

We first focus on McCulloch – Pitts neuron with $X = \{0,1\}^n$.
If only binary inputs and out puts are used, a perceptron's transfer function is a Boolean or Switching function.

Definition1.6: A transfer function $f: \{0,1\}^n \to \{0,1\}$ is called **Boolean** or **switching** function. This function is of the form $f(x_1, x_2, \dots, x_n) = \text{Sat}\,(\sum_i^n w_i x_i - \theta)$, with fixed $w = (w_1, , \dots, w_n)^T$ and a fixed threshold $\theta \in \mathbb{R}$.

Since $f =$ Sat o s we have

$$f(x_1, x_2, \dots, x_n) = (\text{Sat o s})\,(x_1, x_2, \dots, x_n)$$
$$= \text{Sat}\,(s(x_1, x_2, \dots, x_n))$$
$$= \text{Sat}\,(\sum_i^n w_i x_i - \theta)$$

And $\sum_i^n w_i x_i - \theta = \langle w, x \rangle$. Hence, $f(x_1, x_2, \dots, x_n) = \text{Sat}(\langle w, x \rangle - \theta)$.
Thus various switching function $f: \{0,1\}^n \to \{0,1\}$ can be realized by a perceptron, that is, we can find a perceptron whose transfer function equals *f*.

Examples:

1. $f =$ NOT

Define $f: \{0,1\} \to \{0,1\}$ by $f(x) = \text{Sat}\,(-x + 0.5)$ that is f can be realized by a perceptron.

To see this, $f\,(0) = \text{Sat}\,(0 + 0.5) = 1$, $f\,(1) = \text{Sat}\,(-1 + 0.5) = 0$ by definition of Sat

That is,

x	$f(x)$
0	1
1	0

2. $f =$ AND

Define $f: \{0,1\}^2 \to \{0,1\}$ by $f(x_1, x_2) = \text{Sat}\,(x_1 + x_2 - 1.5)$. This is an appropriate perceptron. $\theta \in (1,2]$

x_1	x_2	$f(x_1, x_2)$
0	0	0
0	1	0
1	0	0
1	1	1

3. $f =$ OR

Define $f: \{0,1\}^2 \to \{0,1\}$ by $f(x_1, x_2) = \text{Sat}\,(x_1 + x_2 - 0.5)$. This is an appropriate perceptron. $\theta \in (0,1]$

x_1	x_2	$f(x_1, x_2)$
0	0	0
0	1	1
1	0	1
1	1	1

Since any Boolean function can be written in disjunctive form, using only AND, OR, and NOT operations this implies that all Boolean functions can be represented by a **suitable network** consisting of perceptrons. This doe sn't infer, how ever, th at a ll Boolean/switching functions can actually be realized by a single perceptron.

Example: Consider the exclusive OR, which coincides with the addition $\oplus$ *on* Z_2

Define $f: \{0,1\}^2 \to \{0,1\}$ by $f(x_1, x_2) = x_1 \oplus x_2$

x_1	x_2	$f(x_1, x_2)$
0	0	0
0	1	1
1	0	1
1	1	0

Note: A **XOR** B = A or B and not A and B i.e., A **XOR** B = (A $\vee B$) $\wedge \neg$**(**A $\wedge$ B**)**

Lemma1.1: There is no perceptron that represents the XOR function.

Proof: Assume that for all $(x_1, x_2) \in \{0,1\}^2$

$f(x) = f: \{0,1\}^2 \to \{0,1\}$ is defined by $f(x_1, x_2) = x_1 \oplus x_2$

$$= \text{Sat}(w_1\, x_1 + w_2\, x_2 - \theta)$$

Then we have

i. $f(0,0) = 0 \oplus 0 = 0 = \text{Sat}(0 + 0 - \theta) \Longrightarrow \theta > 0$

ii. $f(1,0) = 1 \oplus 0 = 1 = \text{Sat}(w_1 - \theta) \Longrightarrow w_1 - \theta \geq 0$

iii. $f(0,1) = 0 \oplus 1 = 1 = \text{Sat}(w_2 - \theta) \Longrightarrow w_2 - \theta \geq 0$

iv. $f(1,1) = 1 \oplus 1 = 0 = \text{Sat}(w_1 + w_2 - \theta)\,) \Longrightarrow w_1 + w_2 - \theta < 0$

Add ii and iii. yields $w_1 + w_2 - 2\theta \geq 0$ $(\boldsymbol{\alpha})$

Add i and iv. yields $w_1 + w_2 - 2\theta < 0$, contradiction to the inequality in $(\boldsymbol{\alpha})$ ■

1.2. Affine Separation

The realization of Boolean functions by a perceptron is closely linked to the affine separation of sets. This yields a criterion for deciding whether a given Boolean function can be realized by a perceptron. In fact it will turn out the number of realizable Boolean functions is actually quite small.

Definition1.7: A set $A \subset \mathbb{R}^n$ is called **affinely separable** from $B \subset \mathbb{R}^n$ if there exists $(w,\theta) \in R^{n+1}$ with

$$\langle w, x\rangle - \theta \begin{cases} \geq 0 \text{ for x} \in A \\ < 0 \; for \;\; x \in B \end{cases} \qquad (1.2)$$

Then the set $\boldsymbol{H} = \{x \in R^n : \langle w, x\rangle = \theta\}$ is called a **separating hyperplane.** If $A \subset B$,then clearly they are affinely separable.

Note: 1. If the inequalities in (1.2) are both strict, we say that A and B are strictly affinely separable (from each other).

2. If $\theta = 0$, we say that A is linearly separable from B, or A and B are strictly linearly separable, respectively.

Thus the geometrical interpretation of the **XOR** problem is there exists no affine hyperplane $\langle w, x\rangle = \theta$ that separates $\{(1,0),(0,1)\}$ from $\{(0,0),(1,1)\}$ which implies these two sets are not affinely separable. However, the separating hyperplane that separates $\{(0,0),(1,0),(0,1)\}$ from $\{(1,1)\}$ is $x_1 + x_2 = 0.5$ (by definition), hence the

two sets are affinely separable since there exists $(1,1,\frac{1}{2}) \in \mathbb{R}^3$ satisfying the condition in (1.2).

Remark: Let A be a set and B be a compact set. If A is affinely separable from B, then A and B are linearly separable.

Now we formulate a necessary and sufficient condition for $\{0,1\}-$valued functions to be realizable by McCulloch-Pitts neurons.

Theorem1.2: Let $X \subseteq \mathbb{R}^n$ (where X is an arbitrary subset of $\mathbb{R}^n$). A function $f: X \to \{0,1\}$ can be represented by a perceptron if and only if X_+ is affinely separable from X_-, where

$$X_+ = f^{-1}(1) \subseteq X \text{ and } X_- = f^{-1}(0) \subseteq X$$

Proof: By hypothesis we have

$$f(x) = \text{Sat}(\langle w, x\rangle - \theta)$$

For some $(w,\theta) \in R^{n+1}$ if and only if

$$\langle w, x\rangle - \theta \begin{cases} \geq 0 & \text{for } x \in X_+ \\ < 0 & \text{for } x \in X_- \end{cases}$$

if and only if X_+ is affinely separable from X_-. This yields the desired result. ■

Definition1.8: The degree of difficulty in separating arbitrary finite disjoint sets is measured by means of the so-called **capacity.**

In order to decide whether two given sets are affinely separable, the concept of convexity plays important role.

Definition1.9: A set $A \subset \mathbb{R}^n$ is called **convex** if for any $x, y \in A$, the complete line segment between x and y is also contained in A, that is, $x, y \in A$ and $0 \leq \lambda \leq 1 \Longrightarrow \lambda x + (1-\lambda)y \in A$.

Example1: Trivially empty set and a set containing a single element is convex.

Example2. Consider the equation of plane in $\mathbb{R}^3$ given below:

$$S = \{(x, y, z): x + 2y - z = 4\}$$

Clearly S is convex set.To see this, let x = $(x_1\ y_1, z_1)$, $y = (x_2, y_2, z_2) \in S$ and $0 \leq \lambda \leq 1$

$$\lambda x + (1 - \lambda)y = \lambda(x_1\ y_1, z_1) + (1 - \lambda)\ (x_2, y_2, z_2)$$
$$= (\lambda\ x_1, \lambda\ y_1, \lambda\ z_1 + ((1 - \lambda)\ x_2, (1 - \lambda)y_2\ , (1 - \lambda)\ z_2)$$
$$= (\lambda\ x_1 + (1 - \lambda)\ x_2, \lambda\ y_1 + (1 - \lambda)\ y_2\ , \lambda\ z_1 + (1 - \lambda)z_2\ )$$

Now

$$\lambda\ x_1 + (1 - \lambda)\ x_2\ + 2(\lambda\ y_1 + (1 - \lambda)\ y_2)) - (\lambda\ z_1 + (1 - \lambda)\ z_2\ )$$
$$= \lambda\ x_1 + 2\lambda\ y_1 - \lambda\ z_1 + (1 - \lambda)x_2 + 2(1 - \lambda)\ x_2 - (1 - \lambda)\ x_2$$
$$= \lambda x_1 + 2y_1 - z_1 + (1 - \lambda)\ (x_2 + 2y_2 - z_2)$$
$$= \lambda\ (4) + (1 - \lambda)\ (4)$$
$$= 4$$

$\Longrightarrow\ \lambda x + (1 - \lambda)y \in S \Longrightarrow S$ is convex set. ■

Example3: Consider the set $B = \{(x, y): 0 \leq y < 1\} \cup \{(0, 1)\}$ this set is also convex.

Definition1.10: a. Let A be a subset in $\mathbb{R}^n$, and let $x \in A$. Then, x is called an interior point of A if there is an ε −neighborhood of x that is contained in A, that is, if there exists an $\varepsilon > 0$ such that $\|y - x\| \leq \varepsilon$ implies that $y \in A$. The set of all such points is called the interior of A and is denoted by $int\ A$ and A is called open if $A = intA$.

b. Let A be a subset in $\mathbb{R}^n$. The closure of A, denoted by cl A,is the set of all points that are arbitrarily close to A. In particular, $x \in$ cl A if for each $\varepsilon > 0$, $A \cap N_\varepsilon(x) \neq \emptyset$,where $N_\varepsilon(x) = \{y: \|y - x\| \leq \varepsilon\}$. The set A is said to be closed if $A =$ cl A and A is compact if it is closed and bounded.

Theorem 1.3 given below gives sufficient condition for separability of convex sets.

Theorem 1.3: Let A and B be two non – empty, disjoint convex subsets of $\mathbb{R}^n$

a. If B is open, then A is affinely separable from B
b. If one of the sets is closed and the other is compact then, they are strictly affinely separable

Proof: b. Let $A \subset \mathbb{R}^n$ be a non – empty, closed, convex set. Let $z \in \mathbb{R}^n$.

Claim 1: A possesses exactly one element of minimal norm, i.e. there is exactly one x_0 such that

$$\|x_0\| = \min_{x\in A} \|x\|.$$

To see this, let $\{x_n\}_{n=1}^{\infty}$ be a sequence in A with $\|x_n\| \to \inf_{x\in A} \|x\| = \beta$

Now using the parallelogram law of equation

$$\left\|\frac{x_n - x_m}{2}\right\|^2 = \frac{\|x_n\|^2}{2} + \frac{\|x_m\|^2}{2} - \left\|\frac{x_n + x_m}{2}\right\|^2$$

By the minimality property of β we have

$\|x_n\|^2 \geq \beta^2$, $\|x_m\|^2 \geq \beta^2$.

Again since A is convex $\frac{x_1+x_2}{2} \in A$ implies $\left\|\frac{x_n+x_m}{2}\right\|^2 \geq \beta^2$.

Now for each $\varepsilon > 0$ and for sufficiently great $n, m \in N$, we have

$\|x_n\|^2 \leq \beta^2 + \varepsilon$, $\|x_m\|^2 \leq \beta^2 + \varepsilon$, $-\left\|\frac{x_n+x_m}{2}\right\|^2 \leq -\beta^2$.

Then we get

$\left\|\frac{x_n - x_m}{2}\right\|^2 \leq \frac{\beta^2+\varepsilon}{2} + \frac{\beta^2+\varepsilon}{2} - \beta^2 = \varepsilon.$

$\Rightarrow \{x_n\}_{n=1}^{\infty}$ is a Cauchy sequence and therefore it is convergent, i.e. it has a limit point $x_0 \in \mathbb{R}^n$.

Since A is closed we get $x_0 \in A$. So we have

$\beta = \lim_{n\to\infty}\|x_n\| = \|x_0\| = \inf_{x\in A} \|x\| = \min_{x\in A} \|x\|.$

Now we prove the uniqueness,

Let $x_0, y_0 \in A$ with $\beta = \|y_0\| = \|x_0\|$

Then $\frac{x_0+y_0}{2} \in A$ (since A is convex). Hence it follows

$$\beta^2 \leq \left\|\frac{x_0+y_0}{2}\right\|^2 = \frac{\|x_0\|^2}{2} + \frac{\|y_0\|^2}{2} - \left\|\frac{x_0-y_0}{2}\right\|^2 = \beta^2 - \left\|\frac{x_0-y_0}{2}\right\|^2.$$

$\Rightarrow \left\|\frac{x_0-y_0}{2}\right\|^2 \leq 0 \Rightarrow \|x_0 - y_0\| = 0 \Rightarrow x_0 = y_0$

Claim2: If $0 \notin A$, then we show that $\{0\}$ and A can be strictly affinely separable and moreover, if $a \in A$ is the element of minimal norm, then $\langle a, x\rangle \geq \|a\|^2 > 0$ for all $x \in A$.

To see this, by claim1 the element of minimal norm exists. Let this element be $a \in A$.

Since by assumption $0 \notin A$ implies that $a \neq 0$.Then since A is convex, we have

$tx + (1-t)a = t(x-a) + a \in A$ for all $t \in [0,1]$, for all $x \in A$.

So we have (using the minimal property of a)

$0 \leq \|a\|^2 \leq \|t(x-a) + a\|^2 = \langle t(x-a) + a, t(x-a) + a\rangle$

$$= \|a\|^2 - 2t\|a\|^2 + 2t\langle a, x\rangle + t^2\|x - a\|^2$$

$\Rightarrow 2t\|a\|^2 - t^2\|x - a\|^2 \leq 2t\langle a, x\rangle$

$\Rightarrow \|a\|^2 - \frac{t\|x-a\|^2}{2} \leq \langle a, x\rangle$ for all $t > 0$. In particular, for $t \to 0$ we get $\|a\|^2 \leq \langle a, x\rangle$

$\Rightarrow \langle a, 0\rangle = 0 < \|a\|^2 \leq \langle a, x\rangle$ for all $x \in A$.

This implies

$$\|a\|^2 = \inf_{x \in A} \langle a, x\rangle > \sup_{y=0} \langle a, y\rangle$$

Now suppose $z \notin A$.

Claim3: $\{z\}$ and A can be strictly affinely separable.

To see this, let $\notin A$, $C = A - \{z\} = \{a - z : a \in A\}$. Clearly C is closed set. Let α be the element of minimal norm of C. Then there exists $a \in A$ such that $\alpha = a - z$.

Obviously, $0 \notin C$ otherwise $z \in A$, this is contradiction. Using claim 2 with

$\alpha = a - z$ and $x = a' - z$. We have

$\langle \alpha, x\rangle = \langle \alpha, a' - z\rangle \geq \|\alpha\|^2 > 0$ for all $x \in C$ i.e. for all $x \in A$.Then we get

$\langle \alpha, a'\rangle \geq \langle \alpha, z\rangle + \|\alpha\|^2 > \langle \alpha, z\rangle$ for all $x \in A$.

This implies

$\langle \alpha, a'\rangle > \langle \alpha, z\rangle$ for all $a' \in A$.

$\Rightarrow \inf_{a' \in A} \langle \alpha, a'\rangle > \langle \alpha, z\rangle = \sup_{y=z} \langle \alpha, y\rangle$

By definition $\{z\}$ and A are strictly affinely separable and

Now suppose the set B is closed and A is compact such that $A \cap B = \emptyset$.

Claim4: A and B are strictly affinely separable.

Since $A \cap B = \emptyset$, every $x \in A$ is not in B Now let $C = B - A$. Clearly C is a closed set. $C = \{b - a : a \in A, b \in B\}$

Clearly $0 \notin C$ otherwise $A \cap B \neq \emptyset$ which is contradiction. Hence by claim2

A and B are strictly affinely separable. ■

Proof a: If A and B are not affinely separable then (by definition) there doesn't exist a separating hyperplane implies the set A and B have point in common implies B contains a boundary point which is implies the set B is not open. ■

Example: Let $A = (2,4) \subseteq \mathbb{R}$, and $B = [4,6] \subseteq \mathbb{R}$, $C = (-\infty, 3] \subseteq \mathbb{R}$,

It is easily seen that set A, B and C are convex, and set A is open and set B is compact and $A \cap B = \emptyset$ which implies A is affinely separable from B (by theorem1.3)

And since the set C is closed and $B \cap C = \emptyset$, then (by theorem1.3) the two sets are strictly affinely separable.

For arbitrary sets A, it is useful to consider their convex super sets, the smallest of which is called the **convex hull of *A***.

Definition1.11: Let $A \subset \mathbb{R}^n$ be an arbitrary set. The **convex hull,** denoted $Conv(A)$, of A is defined as the intersection of all convex subsets of $\mathbb{R}^n$ that contains A. In other words, $x \in \text{Conv(A)} \Leftrightarrow x$ can be represented as $x = \sum_{i=1}^{k} \lambda_i x_i \,, \sum_{i=1}^{k} \lambda_i \,, \,, \lambda_i \geq 0$ $i = 1, \dots, k,$ where k is positive integer and $x_1, \dots, , x_k \in A$. Thus

$$Conv(A) = \{\sum_{i=1}^{k} \lambda_i x_i : k \in N, x_i \in A, \sum_{i=1}^{k} \lambda_i \,, \,, \lambda_i \geq 0\} \quad (1.3)$$

Remark: 1. If a set X is convex, then $Conv\,(X) = X$.

2. For arbitrary sets A, B with $A \subset B$ we have $Conv\,(A) \subset Conv\,(B)$.
3. For any set A, B we have $Conv\,(A - B) = Conv\,(A) - Conv\,(B)$.
4. Convex hull of an open set in $\mathbb{R}^n$ is again open.
5. Convex hull of a compact set in $\mathbb{R}^n$ is again compact.
6. Convex hull of a closed set in $\mathbb{R}^n$ is need not be closed.

Counter example: Consider the set $A = \{(x, 0): x \in \mathbb{R}\} \cup \{(0, 1)\} \subset \mathbb{R}^2$

Clearly the set A is closed, however, $Conv\,(A) = \{(x, y): 0 \leq y < 1\} \cup \{(0, 1)\}$ is not closed.

A necessary and sufficient criterion for the separability of non – convex sets is illustrated by the following corollary.

Corollary1.4:

1. Let A, $B \subset R^n$ be non – empty, and let B be open. Then, A is affinely separable from B if and only if $Conv(A) \cap Conv(B) = \emptyset$.

2. Let $A, B \subset \mathbb{R}^n$ be non – empty, with A closed and B compact. Let $\overline{Conv(A)}$ denote the closure of Conv(A). Then, A and B are strictly separable from B if and only if $\overline{Conv(A)} \cap Conv(B) = \emptyset$.

Proof: 1. Suppose A is affinely separable from B

Then there are half spaces

$H^+ = \{x \in \mathbb{R}^n : \langle w, x\rangle \geq \theta\}$ and $H^- = \{x \in \mathbb{R}^n : \langle w, x\rangle < \theta\}$ with $A \subset H^+$ and $B \subset H^-$

Clearly H^+ and H^- are convex sets. To see this, let $x, y \in H^+$ then $\langle w, x\rangle \geq \theta, \langle w, y\rangle \geq \theta$

Now for $0 \leq \lambda \leq 1, \langle w,\ \lambda x + (1-\lambda)y\rangle = \langle w, \lambda x\rangle + \langle w, (1-\lambda)y\rangle$

$$= \lambda\langle w, x\rangle + (1-\lambda)\langle w, y\rangle$$
$$\geq \lambda\,\theta + (1-\lambda)\theta$$
$$= \theta$$

$\Rightarrow\ \lambda x + (1-\lambda)y \in H^+$. Thus H^+ is convex set. Dually we can show H^- is convex set.

and $Conv(A) \subset H^+$ (since $A \subset H^+$) , $Conv(B) \subset H^-$ (since $B \subset H^-$)

Since $H^+ \cap H^- = \emptyset \Rightarrow Conv\,(A) \cap Conv\,(B) = \emptyset$.

Suppose $Conv\,(A) \cap Conv\,(B) = \emptyset$. We show that A is affinely separable from B.

Since the set B is open by the remark 4 above, $Conv\,(A)$ is open and hence by theorem 1.3 $Conv\,(A)$ and a re af finely s eparable. H owever, $A \subseteq Conv\,(A)$ and $B \subseteq Conv\,(B)$ implies A and B are affinely separable this complete the proof of (1). ■

Proof 2. As pa rt i n (1), w e ha ve a c losed a nd an ope n ha lf-spaces H^+ and H^-, respectively, such that $Conv\,(A) \subset H^+$ and $Conv\,(B) \subset H^-$.

Now $Conv\,(A) \subset H^+$ implies $\overline{Conv(A)} \subset \overline{H^+} = H^+$ (since is H^+ closed set)

And hence $\overline{Conv(A)} \cap Conv(B) = \emptyset$.

Suppose ($\Leftarrow$) holds since $Conv(B)$is compact , by remark 5 and $\overline{Conv(A)}$ is convex set hence the result by Theorem 1.3. ■

1.2.1 Separation of Finite sets

Finite s ets and t heir c onvex hul ls a re c ompact. H ence, i n vi ew of corollary1.4, t he condition $Conv\,(A) \cap Conv\,(B) = \emptyset$ is equivalent to the strict affine separability of the non empty sets A and B. Now for convenience, instead of functions $f: X \to \{0, 1\}$ we consider mappings $f: X \to \{-1, 1\}$, where $\{x_1, x_2, \ldots, x_N\} \subset \mathbb{R}^n$. Let $X_{\pm} = f^{-1}(\pm 1)$. Suppose $X_- = f^{-1}(-1)$ and $X_+ = f^{-1}(1)$ are bot h non -empty. L et $y_i = f(x_i)$ for $1 \leq i \leq N$. Then the sets X_+ and X_- are (strictly) affinely separable if and only if

$y_i\ (\langle w, x\rangle - \theta) > 0$ for $1 \le i \le N$ for some $\in \mathbb{R}^n$, $\theta \in \mathbb{R}$. Equivalently,

$$f(x) = \text{Sign}(\langle w, x\rangle - \theta) \qquad \text{for all } x \in X.$$

i.e., the function f is realizable by a perceptron with a modified Heaviside or sign function

$$\text{Sign}\ (z) = \begin{cases} 1 & if\ z \ge 0 \\ -1 & if\ z < 0 \end{cases}$$

Note: For any two non – empty sets $A, B \subset \mathbb{R}^n$,we define $A - B = \{\ a - b\colon a \in A, b \in B\}$

Theorem 1.5: X_+ and X_- are affinely separable if and only if $0 \notin C = Conv(\ X_+ - X_-)$.

Proof : We have $Conv(\ X_+ - X_-)\ = Conv(\ X_+) - Conv(\ X_-)$ and

$0 \notin C$ is equivalent to $Conv(\ X_+) \cap Conv(\ X_-)\ = \emptyset$. ■

Note: Equivalently, X_+ and X_- are affinely separable if and only if the associated polyhedron.

$$\Gamma = \{\ (w,\theta) \in \mathbb{R}^{n+1}\colon\ y_i\ (\langle w, x\rangle - \theta) > 0\} \ne \emptyset$$

Clearly, Γ consists of all (w,θ) that define hyperplane which separates X_+ from X_-. In the following two sections, we treat the problem of finding an arbitrary element of Γ (in a finite number of computations steps), and we address the problem of selecting an optimal separating hyper plane.

1.3 Perceptron Learning Algorithms

If the finite sets X_+ and X_- are to be separated, we are interested in finding a concrete separating hyperplane. A method of constructing such a hyperplane is given by the perceptron Learning Algorithm (PLA).

Note: The strict separability of the sets X_+ and X_- is assumed throughout this section. We start by reformulating the task.

1. PL problem: Two finite sets X_+ , $X_- \subset \mathbb{R}^n$ with $Conv(\ X_+) \cap Conv(\ X_-)\ = \emptyset$ shall be affinely separated by a perceptron. In other words, find $(w,\theta) \in \mathbb{R}^{n+1}$ such that

$$\text{Sign}(\langle w, x\rangle - \theta) = \begin{cases} 1\ if\ \ x \in X_+ \\ -1\ if\ x \in X_- \end{cases}$$

It is convenient to eliminate the threshold θ : replace (x,w) by $(\hat{x},\hat{w})$,where $\hat{x}=\begin{bmatrix}x\\1\end{bmatrix}$ and $\hat{w}=\begin{bmatrix}w\\-\theta\end{bmatrix}$. Hence, the new input domain is $\mathbb{R}^n\times\{1\}$ and

$$\text{Sign}(\langle\hat{w},\hat{x}\rangle)=\text{Sign}(\langle w,x\rangle-\theta).$$

Instead of the affine separation of X_+ and X_- in $\mathbb{R}^n$, we now face the problem of linear separation of $\hat{X}_+$and $\hat{X}_-$ in $\mathbb{R}^{n+1}$. Let further $m:=n+1$ and

$$\hat{X}=\hat{X}_+\cup\{-x: x\in\hat{X}_-\}\subset\mathbb{R}^m$$

Hence, PL problem is equivalent to the following

2. Reformulated PL problem:

Find a vector $\hat{w}\in\mathbb{R}^{n+1}$ with $\langle\hat{w},\xi\rangle>0$ for all $\xi\in\hat{X}$. (1.4)

If $\hat{X}=\{\xi_1,\xi_2,\dots,\xi_N\}$, de fine $A:=[\xi_1,\xi_2,\dots,\xi_N]^T\in\mathbb{R}^{N\times m}$. Then (1.4) reads $A\hat{w}>0$ where the inequality is to be understood component – wise. As will be illustrated, the PLA yields such a separating vector $\hat{w}$ after finitely many steps.

Note: A: $\mathbb{R}^m\longrightarrow\mathbb{R}^N$ where $A=\begin{bmatrix}\xi_{11}&\cdots&\xi_{1m}\\\vdots&\ddots&\vdots\\\xi_{N1}&\cdots&\xi_{Nm}\end{bmatrix}$ and $\hat{w}=\begin{bmatrix}w_1\\\vdots\\w_m\end{bmatrix}^T$

Definition 1.12: Let $\hat{X}\subset\mathbb{R}^m$ be a finite set. A sequence $u:N\longrightarrow\hat{X}$ is called a **training sequence** for $\hat{X}$ if for all $\xi\in\hat{X}$ and all $k_0\in N$, there is $k\geq k_0$ such that $u(k)=\xi$. That is, each pattern $\xi\in\hat{X}$ appears infinitely often in a training sequence u for $\hat{X}$.

Example: $u=\{x_1,x_2,\dots,x_N,x_1,\dots,x_N,\dots\}$ i.e. after every k_o elements we have to get the element again for $\hat{X}=\{x_1,x_2,\dots,x_N\}$

PL Algorithm

Let $A:=[\xi_1,\xi_2,\dots,\xi_N]^T\in\mathbb{R}^{N\times m}$ be given. Let u be a training sequence for $\{\xi_1,\xi_2,\dots,\xi_N\}$, and let $\psi:N\longrightarrow R$ be a sequence with $0<\inf_{k\in N}\psi(k)\leq\sup_{k\in N}\psi(k)<\infty$

Step 0**:** Choose $w(0)\in\mathbb{R}^m$ and put $k:=0$

Step 1**:** Set

$$w(0)=\begin{cases}w(k) & if\ u(k)^Tw(k)>0\\ w(k)+\psi(k)u(k) & if\ u(k)^Tw(k)\leq 0\end{cases} \quad (1.5)$$

Step 3**:** If $Aw(k+1) > 0$ applies component-wise, an admissible separation has been found and the algorithm stops. Otherwise, augment k by one and go to step 1.

***Theorem 1.6:* (*Perceptron convergence theorem*)** Let a matrix $A := [\xi_1, \xi_2, \dots, \xi_N]^T \in \mathbb{R}^{N\times m}$ be given and suppose that there exists $w^* \in \mathbb{R}^m$ such that $Aw^* > 0$ is fulfilled component-wise. Let u and ψ be as described above. Let $w : N \longrightarrow \mathbb{R}^m$ be the (infinite) sequence of weight vectors that is determined by the recursion (1.5) from an arbitrary initial value $w(0) \in \mathbb{R}^m$.

Then *w* becomes stationary, i.e., there is a $k_L \in N$ such that for all $k \geq k_L$:

$w(k) = w(k_L) =: \hat{w}$ and $A\hat{w} > 0$ component – wise.

This signifies that the sequence of weights produced according to (1.5) converges to a solution of the PL problem, in finitely many steps. In view of Step2 above, this implies that the PL Algorithm stops after finitely many iterations.

Proof:

By assumption,$Aw^* > 0$, hence

$$\alpha := \inf_{k\in N} \psi(k) \cdot \min_{j\in\{1,\dots,N\}} (\xi_j{}^T w^*) > 0 \text{ where } \xi_j{}^T = [\xi_{1j}, \dots, \xi_{Nj}]$$

Define

$$\beta := \sup_{k\in N} \psi(k) \cdot \max_{j\in\{1,\dots,N\}} \|\xi_j\|_2 \text{ and } \bar{w} = \frac{\beta^2}{\alpha} w^*.$$

Consider an iteration step in which the weight vector is properly updated, i.e. suppose

$w(k+1) \neq w(k)$

and thus $u(k)^T w(k) \leq 0$. Then we have

$$\begin{aligned} \|w(k+1) - \bar{w}\|_2^2 &= \|w(k) + \psi(k)u(k) - \bar{w}\|_2^2 \\ &= \|w(k) - \bar{w} + \psi(k)u(k)\|_2^2 \\ &= \|w(k) - \bar{w}\|_2^2 + 2\,\psi(k)u(k)^T w(k) - \bar{w} + \psi(k)^2 \|u(k)\|_2^2 \\ &\leq \|w(k) - \bar{w}\|_2^2 - 2\,\psi(k)u(k)^T \bar{w} + \beta^2 \qquad (*) \end{aligned}$$

(since $u(k)^T w(k) \leq 0$ and $\psi(k)^2 \|u(k)\|_2^2 \leq \beta^2$)

To see this, $\beta := \sup_{k\in N} \psi(k) \cdot \max_{j\in\{1,\dots,N\}} \|\xi_j\|_2$ and $u : N \longrightarrow \{\xi_1, \xi_2, \dots, \xi_N\}$

$\Rightarrow \beta \geq \psi(k)\|\xi_j\|_2$ for all $j \in \{1, \dots, N\}$.

$\Rightarrow \beta^2 \geq \psi(k)^2 \|u(k)\|_2^2$ (since $u(k) \in \{\xi_1, \xi_2, \ldots, \xi_N\}$)

Furthermore, $0 < \alpha \leq \psi(k)u(k)^T w^* \Rightarrow 1 \leq \frac{\psi(k)u(k)^T w^*}{\alpha}$

and $\psi(k)u(k)^T \bar{w} = \beta^2 \frac{\psi(k)u(k)^T w^*}{\alpha} \geq \beta^2$ implies $-2\psi(k)u(k)^T \bar{w} \leq -2\beta^2$

Thus (*) become

$\|w(k+1) - \bar{w}\|_2^2 \leq \|w(k) - \bar{w}\|_2^2 + \beta^2$

Set $k_1 := 0$ and let $0 < k_2 < k_3 < \cdots$ be the sequence of integers with

$w(k_{i-1}) \neq w(k_i)$

Let $(w(k_1), w(k_2), \ldots)$ be the corresponding subsequence of w. Then

$\|w(k_1) - \bar{w}\|_2^2 \leq \|w(0) - \bar{w}\|_2^2 - \beta^2$

$\|w(k_2) - \bar{w}\|_2^2 \leq \|w(1) - \bar{w}\|_2^2 - \beta^2 \leq \|w(0) - \bar{w}\|_2^2 - 2\beta^2$

By induction, we have

$0 \leq \|w(k_{j+1}) - \bar{w}\|_2^2 \leq \|w(0) - \bar{w}\|_2^2 - j\beta^2$

and hence

$j \leq \frac{\|w(0) - \bar{w}\|_2^2}{\beta^2}$

We conclude that the length of the subsequence, and hence the number of proper updating steps is bounded. Let L denote the length of the subsequence. Then

$w(k) = w(k_L)$ for all $k \geq k_L$.

As u is a training sequence for the rows of A, this implies $Aw(k_L) > 0$. ■

Example: An affine separation of the sets $X_+ = \{(0,0), (0,1), (1,0)(-1,1)\}$ and $X_- = \{(-2,1), (-2,0), (-1,0)(-1,-1)\}$. After eliminating the threshold and reflecting the set X_- at the origin, we obtain the set $\hat{X}$ as in equation (#) below, i.e.,

$\hat{X}_+ = \begin{bmatrix} X_+ \\ 1 \end{bmatrix} = \{(0,0,1), (0,1,1), (1,0,1)(-1,1,1)\}$

$\hat{X}_- = \begin{bmatrix} X_- \\ 1 \end{bmatrix} = \{(-2,1,1), (-2,0,1), (-1,0,1)(-1,-1,1)\}$

$\hat{X} = \hat{X}_+ \cup \{-x : x \in \hat{X}_-\}$

$= \{(0,0,1),(0,1,1),(1,0,1),(-1,1,1)\} \cup$

$\{(2,-1,-1),(2,0,-1),(1,0,-1)(1,1,-1)\}$

$= \{(0,0,1),(0,1,1),(1,0,1),(-1,1,1),(2,-1,-1),(2,0,-1),(1,0,-1)(1,1,-1)\}$ (#)

By reformulated PL problem: we need to find a vector $w \in R^3$ with $\langle w,\xi\rangle > 0$ for all $\xi \in \hat{X} = \{\xi_1,\xi_2,\dots,\xi_8\}$ and $H = \{\xi \in \mathbb{R}^3 : \langle w,\xi\rangle = 0\}$ is a separating hyperplane.

Now a weight vector $w \in \mathbb{R}^3$ is sought with $\langle w,\xi\rangle > 0$ for all $\xi \in \hat{X}$. We have a matrix $A := [\xi_1,\xi_2,\dots,\xi_8]^T \in \mathbb{R}^{8\times 3}$, that is,

$$A = \begin{bmatrix} \xi_{11} & \cdots & \xi_{18} \\ \xi_{21} & \cdots & \xi_{28} \\ \xi_{31} & \cdots & \xi_{38} \end{bmatrix}^T \Rightarrow A = \begin{bmatrix} 0 & 0 & 1 & -1 & 2 & 2 & 1 & 1 \\ 0 & 1 & 0 & 1 & -1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}^T \in \mathbb{R}^{8\times 3}$$

For this, perceptron algorithm is started with the randomly chosen weight vector $w(0) = (0,2,1)$ and for simplicity a sequence in $\mathbb{R}$, $\psi : N \longrightarrow \mathbb{R}$ is defined by $\psi(k) = 1$ for all $k \in N$ i.e. $\psi \equiv 1$. The hyperplane $H = \{\xi \in \mathbb{R}^3 : \langle w(0),\xi\rangle = 0\}$ which is orthogonal to $w(0)$ separates the correctly classified points from the falsely classified points.

Falsely classified points are found by the relation $A\,w(0) \le 0$ applying component – wise.

$$A\,w(0) = \begin{bmatrix} 0 & 0 & 1 & -1 & 2 & 2 & 1 & 1 \\ 0 & 1 & 0 & 1 & -1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{bmatrix}^T \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$$

$= [1 \quad 3 \quad 1 \quad 3 \quad -3 \quad -1 \quad -1 \quad 1]^T \Rightarrow$ falsely classified points are $(2,-1,-1),(2,0,-1)$, and $(1,0,-1)$ others are correctly classified.

Consider a training sequence $u : N \longrightarrow \hat{X}$. If a falsely classified point emerges in the training sequence (e.g. $u(0) = (2,0,-1)$) by step 1 of PL algorithm, we have the weight vector corrected accordingly: $w(1) = w(0) + u(0) = (0,2,1) + (2,0,-1) = (2,2,0)$ i.e. by step 1, we have

$$w(1) = \begin{cases} w(0) & if\ u(0)^T w(0) > 0 \\ w(0) + u(0) & if\ u(0)^T w(0) \le 0 \end{cases} \text{ (since } \psi \equiv 1 \text{)}$$

$= w(0) + u(0)$ (since $u(0)^T w(0) \le 0$) $= (2,2,0)$

Now $A\,w(1) = \begin{bmatrix} \underset{\uparrow}{0} & 2 & 2 & \underset{\uparrow}{0} & 2 & 4 & 2 & 4 \end{bmatrix}^T$. Arrows s how that th e point t hat is falsely c lassified p osition. H ence, t hey ar e $(0,0,1)$ and $(-1,1,1)$. W e c ontinued the process because the algorithm only stops if w found with $\langle w, \xi \rangle > 0$ for all $\xi \in \widehat{X}$. If these poits are emerges in the training sequence; take $u(1) = (-1,1,1)$, then the weight vector is corrected accordingly by step 1 of PL algorithm.

$w(2) = w(1) + u(1)$ (since $0 = u(1)^T w(1) \leq 0$) $= (2,2,0) + (-1,1,1) = (1,3,1)$

$\Longrightarrow A\,w(2) = \begin{bmatrix} 1 & 4 & 2 & 3 & \underset{\uparrow}{-2} & 1 & \underset{\uparrow}{0} & 3 \end{bmatrix}^T$.

In similar way we take $u(2) = (2,-1,-1)$ and hence the weight vector corrected accordingly:

$w(3) = w(2) + u(2) = (1,3,1) + (2,-1,-1) = (3,2,0)$

$\Longrightarrow A\,w(3) = \begin{bmatrix} \underset{\uparrow}{0} & 2 & 3 & \underset{\uparrow}{-1} & 4 & 6 & 3 & 5 \end{bmatrix}^T$. Take $u(3) = (0,0,1)$, then we have

$w(4) = w(3) + u(3) = (3,2,0) + (0,0,1) = (3,2,1)$

$\Longrightarrow A\,w(4) = \begin{bmatrix} 1 & 3 & 4 & \underset{\uparrow}{0} & 3 & 5 & 2 & 4 \end{bmatrix}^T$. Now take $u(4) = (-1,1,1)$, t hen t he weight vector is: $w(5) = w(4) + u(4) = (3,2,1) + (-1,1,1) = (2,3,2)$

$\Longrightarrow A\,w(5) = \begin{bmatrix} 2 & 5 & 4 & 3 & \underset{\uparrow}{-1} & 2 & \underset{\uparrow}{0} & 3 \end{bmatrix}^T$. Put $u(5) = (2,-1,-1)$,then

$w(6) = w(5) + u(5) = (2,3,2) + (2,-1,-1) = (4,2,1)$

$\Longrightarrow A\,w(6) = \begin{bmatrix} 1 & 3 & 5 & \underset{\uparrow}{-1} & 6 & 7 & 3 & 5 \end{bmatrix}^T$. Put $u(6) = (-1,1,1)$,then

$w(7) = w(6) + u(6) = (4,2,1) + (-1,1,1) = (3,3,2)$

$\Longrightarrow A\,w(7) = [2 \quad 5 \quad 5 \quad 2 \quad 1 \quad 4 \quad 1 \quad 4]^T$ and there is no falsely classified point.

$\Longrightarrow \langle w(7), \xi \rangle > 0$ for a ll $\xi \in \hat{X}$ i.e. $w \equiv (3,3,2)$, t hen t he c orresponding a ffine separation of

X_+ and X_- is $\langle w(7), \xi \rangle = 0 \Longrightarrow 3\xi_1 + 3\xi_2 + 2 = 0$ and t he c orresponding p erceptron has the weights $w = (3,3)^T$ and the threshold $\theta = -2$. ■

1.4 Optimal Separation

Suppose that X_+ and X_- are finite non-empty subsets of $\mathbb{R}^n$ that are affinely separable. Let $X = X_+ \cup X_- = \{x_1, x_2, \dots, x_N\}$ and let $y_i = \pm 1$ for $x_i \in X_\pm$. The set of separating hyperplanes is uniquely determined by the weight polyhedron.

$\Gamma = \{ (w,\theta) \in \mathbb{R}^{n+1} : y_i (\langle w, x_i \rangle - \theta) > 0 \text{ for } 1 \le i \le N\} \neq \emptyset.$ (1.6)

We define a functional that measures the quality of separation achieved by a particular separating hyperplane. Intuitively, this depends on the distance between $X = X_+ \cup X_-$ and the separating hyperplane. If this distance is large, we may expect a certain robustness of the separation with respect to noisy data.

Definition1.13: For $(w,\theta) \in \mathbb{R}^{n+1}$, $w \neq 0$, let $\rho(w,\theta) = \min_{1 \le i \le N} \rho_i (w,\theta)$, where

$\rho_i (w,\theta) = \frac{y_i (\langle w, x_i \rangle - \theta)}{\|w\|}$ be the separation margin of (w,θ). (1.7)

This notion has the following geometrical interpretation: Let $x_i \in X$ and let $H = \{x : \langle w, x \rangle = \theta\}$ be the hyperplane defined by $(w,\theta) \in \mathbb{R}^{n+1}$. The distance of x_i from H is

$$dist(x_i, H) = \min_{x \in H} \|x_i - x\|.$$

We can easily see that $dist(x_i, H) = |\rho_i (w,\theta)|$ (*)

Recall that H separates X_+ from X_- if and only if all the $\rho_i (w,\theta)$ are positive.

In that case by (*), $\rho(w,\theta) = \min_{1 \le i \le N} dis(x_i, H) = dist(X, H)$ $(since x_i \in X)$

i.e. $\rho(w,\theta)$ equals the distance of X from H, and thus measures the quality of separation.

On the other hand, if H doesn't separate X_+ and X_-, then at least one of the $\rho_i (w,\theta)$ is non-positive. This yields the inequality $\rho(w,\theta) \le dist(X, H)$ which is true for all $(w,\theta) \in \mathbb{R}^{n+1}$, and equality holds if and only if $(w,\theta) \in \Gamma$ i.e., If H is a separating hyperplane. Therefore, the optimal separation problem is to maximize $\rho(w,\theta)$ by choice of $(w,\theta) \in \mathbb{R}^{n+1}$.

Theorem 1.8: Suppose that the non – empty finite sets X_+, X_- are affinely separable. Then the optimal separation margin $\rho := sup\{ \rho(w,\theta) : (w,\theta) \in \mathbb{R}^{n+1}, w \neq 0\}$ is finite,

and the supremum is achieved at some (w^*, θ^*). Moreover, the corresponding optimal hyperplane $H = \{x \in \mathbb{R}^n : \langle w^*, x\rangle = \theta^*\}$ is uniquely determined, and we have

$$\rho = \frac{1}{2} dist(C_+, C_-) \text{ ,where } \quad C_\pm = \text{conv}\,(X_\pm).$$

To prove Theorem1.8, several preliminary steps are needed. We start by characterizing the distance between the compact convex sets C_+ and C_-

Lemma 1.9: Let $X_+, X_-, and\ C_+, C_-$ be as in theorem1.8 and let $C := Conv(X_+ - X_-) = C_+ - C_-$. Due to the separablity assumption, we have $0 \notin C$ according to theorem1.5. Then

$$dist(C_+ - C_-) = dist(C, 0) = \max_{\|v\| = 1} \min_{x \in C} \langle v, z\rangle$$

Proof: Foe any v with $\|v\| = 1$,we have $\langle v, z\rangle \le |\langle v, z\rangle| \le \|v\|\|z\| = \|z\|$ due to Cauchy Schwarz – inequality implies $\langle v, z\rangle \le \|z\|$ and hence

$\min_{x\in C} \langle v, z\rangle \le \min_{x\in C} \|z\| = dist(C, 0).$

$\Rightarrow \max_{\|v\|=1} \min_{x\in C} \langle v, z\rangle \le \max_{\|v\|=1} dist(C, 0) = dist(C, 0)$

$$\max_{\|v\|=1} \min_{x\in C} \langle v, z\rangle \le dist(C, 0). \qquad (\alpha)$$

For the converse direction, since C is compact, and $0 \notin C$. Hence there exists $0 \neq z^* \in C$ with $\|z^*\| = \min_{x\in C} \|z\| = dist(C, 0)$. Since C is convex set, then for any $z \in C$ and $\lambda \in [0,1]$, we have

$\lambda z + (1 - \lambda) z^* \in C$ and since $\|z^*\| \le \|z\|$ for all $z \in C$

Thus $\|z^*\|^2 \le \|\lambda z + (1 - \lambda) z^*\|^2$. But

$$\|\lambda z + (1 - \lambda) z^*\|^2 = \|\lambda(z - z^*) + z^*\|^2$$

$$= \|z^*\|^2 + 2\,\lambda\langle z - z^*, z^*\rangle + \lambda^2\|z - z^*\|^2$$

$\Rightarrow 0 \le 2\langle z - z^*, z^*\rangle + \lambda\|z - z^*\|^2$ for all $z \in C$ and $\lambda \in (0,1]$. However, as $\lambda \longrightarrow 0$, so is $\lambda\|z - z^*\|^2$ which implies

$$0 \le 2\langle z - z^*, z^*\rangle \qquad \text{for all } z \in C.$$

Hence,

$$\langle z - z^*, z^*\rangle \ge 0 \text{ for all } z \in C \text{ which also implies } \langle z^*, z^*\rangle \le \langle z, z^*\rangle.$$ Thus

$$\|z^*\| \le \langle \frac{z^*}{\|z^*\|}, z\rangle$$

holds for all $z \in C$ and put $v = \frac{z^*}{\|z^*\|}$. Thus there exists a vector v with $\|v\| = 1$ and

$dist(C,0) \le \langle v, z\rangle$ (since $\|z^*\| = dist(C,0)$)

$\Rightarrow$ $dist(C,0) \le \min_{x\in C} \langle v, z\rangle$. However, this implies

$$dist(C,0) \le \max_{\|v\|=1} \min_{x\in C} \langle v, z\rangle. \qquad (\beta)$$

From (α) & (β) the result is as desired. ■

Lemma 1.10: Let X_+, X_-, be as usual, with $C_\pm = Conv(X_\pm)$ and

$$C = C_+ - C_- = Conv(X_+ - X_-).$$

Let Γ be defined as in **(**1.6**)**. Consider its projection onto the first n components

$$\mathrm{W} = \{ w \in \mathbb{R}^n : \exists \theta \in \mathbb{R}: (w,\theta) \in \Gamma \}$$

Then

1. $\mathrm{W} = \{ w \in \mathbb{R}^n : \langle w, z\rangle > 0 \text{ for all } z \in C\}$

and for any w∈ W

2. $$\theta_w = \frac{1}{2}\left(\min_{x\in C_-} \langle w, x\rangle + \max_{y\in C_+} \langle w, y\rangle \right) \qquad (1.10)$$

is such that $(w, \theta_w) \in \Gamma$. Moreover, for any $0 \neq w \in \mathbb{R}^n$,

$$\rho(w, \theta_w) = \frac{1}{2} \min_{x\in C} \langle \frac{w}{\|w\|}, z\rangle \qquad (1.11)$$

Proof: Of course if $\Gamma \neq \emptyset$ so is W. By definition,

$$w \in \mathrm{W} \iff \exists\theta: \begin{cases} \langle w, x\rangle - \theta > 0 \text{ for all } x \in X_+ \\ \langle w, y\rangle - \theta < 0 \text{ for all } y \in X_- \end{cases}$$

$$\iff \exists\theta: \begin{cases} \langle w, x\rangle - \theta > 0 & \text{for all } x \in X_+ \\ \langle w, -y\rangle + \theta > 0 & \text{for all } -y \in -X_- \end{cases}$$

$$\iff \exists\theta: \langle w, x\rangle - \theta + \langle w, -y\rangle + \theta = \langle w, x-y\rangle > 0 \ (x - y \in X_+ - X_-)$$

Put $z = x - y$ and thus,

$w \in \mathrm{W} \Rightarrow \langle w, z\rangle > 0$ for all $z \in X_+ - X_-$

and then $\langle w, z\rangle > 0$ holds for all $z \in C$, due to convexity.

Now suppose (1), let $w \in \mathbb{R}^n$ and let θ_w be as defined above. We show that $w \in \mathrm{W}$.

For $x \in X_+$ we have

$$\begin{aligned}\langle w,x\rangle - \theta_w &= \langle w,x\rangle - \tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &\geq \min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &= \tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &= \tfrac{1}{2}\min_{z\in C}\langle w,z\rangle. \end{aligned} \tag{1.12}$$

Again for $y \in X_-$,we have

$$\begin{aligned}\langle w,y\rangle - \theta_w &= \langle w,y\rangle - \tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &\leq \max_{x\in C_-}\langle w,y\rangle - \tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle - \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &= -\tfrac{1}{2}\min_{x\in C_+}\langle w,x\rangle + \tfrac{1}{2}\max_{y\in C_-}\langle w,y\rangle \\ &= -\tfrac{1}{2}\min_{z\in C}\langle w,z\rangle. \end{aligned} \tag{1.13}$$

Now if w is such that $\langle w,z\rangle > 0$ for all $z \in C$, we have $\min_{z\in C}\langle w,z\rangle > 0$ due to the compactness of C.

Thus from (1.12) and (1.13), we have

$\langle w,x\rangle - \theta_w > 0$ for all $x \in X_+$ and $\langle w,y\rangle - \theta_w < 0$ for all $y \in X_-$

implies $(w, \theta_w) \in \Gamma$. In particular, $w \in W$.

Note that equality holds in (1.2),(1.3) for at least one $x \in X_+$, $y \in X_-$. This is due to the fact t hat $C_\pm$ are co mpact and co nvex set s. H ence $\langle w,.\rangle$ achieves i ts minimum a nd maximum i n one of t he e xtreme p oints of $C_\pm$, a nd t hese a re c ontained i n t he s ets $X_\pm$,respectively. Thus

$$\min_{x\in X_+}(\langle w,x\rangle - \theta_w) = \min_{y\in X_-}(-\langle w,y\rangle + \theta_w) = \tfrac{1}{2}\min_{z\in C}\langle w,z\rangle.$$

Let $X = X_+ \cup X_- = \{x_1, \dots, x_N\}$ and let $y_i = \pm 1$ for $X_\pm$. For any $0 \neq w \in R^n$,

$$\begin{aligned}\rho(w,\theta_w)\,\|w\| &= \min_{1\leq i\leq N} y_i(\langle w, x_i\rangle - \theta_w) \\ &= \min\left(\min_{x\in X_+}(\langle w,x\rangle - \theta_w),\ \min_{y\in X_-}(-\langle w,y\rangle + \theta_w)\right) \\ &= \tfrac{1}{2}\min_{z\in C}\langle w,z\rangle \text{ (since } C \text{ is compact and convex)} \end{aligned}$$

$$\Rightarrow \rho(w,\theta_w) = \tfrac{1}{2}\min_{z\in C}\langle \tfrac{w}{\|w\|}, z\rangle. \qquad \blacksquare$$

Corollary1.11: The optimal separation margin ρ defined in (1.8) satisfies

$$\rho = \frac{1}{2} dist(C_+, C_-)$$

In particular, it is a finite number.

Proof: For any $0 \neq w \in \mathbb{R}^n$, we have

$$\rho(w, \theta_w)\, \|w\| = \min\left(\min_{x \in X_+} (\langle w, x\rangle - \theta)\ ,\ \min_{y \in X_-} (-\langle w, y\rangle + \theta)\ \right)$$

$$= \min\left(\min_{x \in C_+} (\langle w, x\rangle - \theta)\ ,\ \min_{y \in C_-} (-\langle w, y\rangle + \theta)\ \right)$$

The second equality follows from the fact that $\langle w, . \rangle$ achieves its minimum and maximum in one of the extreme points of $C_{\pm}$ because the minimum values are attained at a point in $X_{\pm}$. For any real numbers a, b, we have

$\min(a, b) \leq \frac{1}{2}(a + b)$, and equality holds only if $a = b$. Using this we obtain

$$\rho(w, \theta_w)\, \|w\| \leq \frac{1}{2} \min_{x \in C_+, y \in C_-} (\langle w, x - y\rangle)$$

and thus

$$\rho(w,\theta) \leq \frac{1}{2} \min_{z \in C} \langle \frac{w}{\|w\|}, z\rangle$$

Equality holds if

$$\min_{x \in C_+} (\langle w, x\rangle - \theta) = \min_{y \in C_-} (-\langle w, y\rangle + \theta) \qquad \text{(a)}$$

Since $\theta_w = \frac{1}{2}\left(\min_{x \in C_-} \langle w, x\rangle + \max_{y \in C_+} \langle w, y\rangle \right)$, by (a) we have $\theta = \theta_w$

To see this,

$$\min_{x \in C_+} (\langle w, x\rangle - \theta) = \min_{y \in C_-} (-\langle w, y\rangle + \theta) = -\max_{y \in C_-} (\langle w, y\rangle + \theta)$$

$$\Rightarrow\ 2\theta = \min_{x \in C_+} \langle w, x\rangle + \max_{y \in C_-} \langle w, y\rangle$$

$$\Rightarrow \theta = \frac{1}{2}\left(\min_{x \in C_+} \langle w, x\rangle + \max_{y \in C_-} \langle w, y\rangle \right) = \theta_w$$

We conclude that for any $0 \neq w \in \mathbb{R}^n$,

$$\sup_{\theta \in \mathbb{R}} \rho(w,\theta) = \rho(w, \theta_w) = \frac{1}{2} \min_{z \in C} \langle \frac{w}{\|w\|}, z\rangle \qquad (1.14)$$

According to lemma 1.9 this yields the desired result, since

$$\rho = \sup_{w \in \mathbb{R}^n} \sup_{\theta \in \mathbb{R}} \rho(w,\theta) = \frac{1}{2} \sup_{w \in \mathbb{R}^n} \min_{z \in C} \langle \frac{w}{\|w\|}, z\rangle = \frac{1}{2} \max_{\|v\|=1} \min_{z \in C} \langle v, z\rangle = \frac{1}{2} dist(C_+, C_-). \qquad \blacksquare$$

So far, we have shown that the optimal separation margin is

$\rho = \frac{1}{2} dist(C, 0) = \frac{1}{2} \min_{z \in C} \|z\|$

where C is a compact convex set. We show that this minimization problem possesses a unique solution. This leads to the construction of the unique optimal separating hyperplane, thus completing the proof of Theorem 1.8.

Lemma1.12: **1.** Let $C \subset \mathbb{R}^n$ be a non – empty closed convex set with $0 \notin C$. Then there exists a unique $z^* \in C$ with

$$\|z^*\| = \min_{z \in C} \|z\| = dist(C, 0)$$

2. Define

$$C^\# = \{ w \in \mathbb{R}^n : \langle w, z \rangle \geq c \text{ for all } z \in C \} \quad (1.15)$$

where $c > 0$. Then $C^\#$ is also non – empty, closed and convex, and $0 \notin C^\#$. Thus there exists a unique $w^* \in C^\#$ with

$$\|w^*\| = \min_{w^* \in C^\#} \|w\| = dist(C^\#, 0)$$

and

$$w^* = \frac{cz^*}{\|z^*\|^2} \text{ and } z^* = \frac{cw^*}{\|w^*\|^2}$$

Proof: Let $z \in C$ be arbitrary. Define $B = \{ x \in \mathbb{R}^n : \|x\| \leq \|z\| \}$. Clearly

$$\inf_{z \in C} \|z\| = \inf_{z \in B \cap C} \|z\|$$

and since $B \cap C$ is compact, the minimum is achieved at some $z^* \in C$. For uniqueness, suppose that z^*_1 and z^*_2 are two solutions. Due to convexity, also $\frac{1}{2}(z^*_1 + z^*_2) \in C$ (i.e., for $\lambda = \frac{1}{2}$). Now for if $z^*_1 \neq z^*_2$, then either $z^*_1 > z^*_2$ or $z^*_1 < z^*_2$. Without loss of generality, assume $z^*_1 > z^*_2$.

Thus we have

$\frac{1}{2}(z^*_1 + z^*_2) < z^*_1 \Rightarrow \left\| \frac{1}{2}(z^*_1 + z^*_2) \right\| < \|z^*_1\| = \|z^*_2\| = \|z^*\| = \min_{z \in C} \|z\|$

$\Rightarrow \frac{1}{2}(z^*_1 + z^*_2) \notin C \Rightarrow C$ is not convex set which is contradiction.

Hence we must have $z^*_1 = z^*_2$

An argument from Lemma1.9 shows that $\|z^*\|^2$

$1 \leq \langle \frac{1z^*}{\|z^*\|^2}, z \rangle$ for all $z \in C$ which implies

$c \leq \langle \frac{cz^*}{\|z^*\|^2}, z \rangle$ for all $z \in C$ and $c > 0 \Rightarrow \frac{cz^*}{\|z^*\|^2} \in C^{\#} \Rightarrow C^{\#} \neq \emptyset$

In particular, $\|w\| \geq c\|z^*\|^{-1}$ for all $\in C^{\#}$. Set $w^* = \frac{cz^*}{\|z^*\|^2}$. To show that C is convex set, let w_1 , w_2 $\in C^{\#}$ and $\lambda \in [0,1]$. Then $\langle w_1, z \rangle \geq$ c , $\langle w_2, z \rangle \geq$ c for all $z \in C$ and $c > 0$. Now

$$\langle \lambda w_1 + (1-\lambda) w_2, z \rangle = \lambda \langle w_1, z \rangle + (1-\lambda) \langle w_2, z \rangle$$
$$\geq \lambda c + (1-\lambda)c \text{ for all } z \in C \text{ and } c > 0$$
$$= c \text{ for all } z \in C \text{ and } c > 0$$

implies $\lambda w_1 + (1-\lambda) w_2 \in C^{\#}$. Hence, $C^{\#}$ is convex set.

Again to show that $0 \notin C^{\#}$. For if $0 \in C^{\#}$, then we have

$0 = \langle 0, z \rangle \geq$ c for all $z \in C$ and $c > 0 \Rightarrow c = 0$ which is impossible. Hence $0 \notin C^{\#}$

This guarantees the unique existence of w^* by (1). Finally, for any $w \in C^{\#}$ and $z \in C$, we have

$\|w\|\|z\| \geq \langle w, z \rangle \geq c$ (the first inequality is due to Cauchy Schwarz)

Since $w^* = \frac{cz^*}{\|z^*\|^2}$

Then $\|w^*\| = c\|z^*\|^{-1} \Rightarrow w^*\|w^*\|^2 = c^2\|z^*\|^{-2}w^* \Rightarrow w^* = \frac{c}{\|z^*\|^2} \cdot \frac{cw^*}{\|w^*\|^2}$

Hence, we get $z^* = \frac{cw^*}{\|w^*\|^2}$ and $\|w^*\| = c\|z^*\|^{-1}$ implies that $\|w^*\|$ is the global minimum. ■

Lemma1.13: Let X_+ , X_-, and $C = Conv(X_+ - X_-)$ be as usual.

Let $C^{\#} = \{ w \in \mathbb{R}^n : \langle w, z \rangle \geq c \; for \; all \; z \in C \}$ where $c > 0$. Let w^* be the unique solution of

$$\|w^*\| = \min_{w^* \in C^{\#}} \|w\|$$

Let $\theta_{w^*} = \frac{1}{2}(\min_{x \in C_-} \langle w^*, x \rangle + \max_{y \in C_+} \langle w^*, y \rangle)$. Then

$H^* = \{ x \in \mathbb{R}^n : \langle w^*, x \rangle - \theta_{w^*} = 0 \}$

is the unique optimal separating hyperplane.

Proof: In view of Lemma 1.10, it is clear that $(w^*, \theta_{w^*}) \in \Gamma$,that is, H^* is a separating hyperplane. Moreover, it is optimal, since

$$\langle w^*, z\rangle \geq c \text{ for all z } \in C$$

and

$$\langle w^*, z^*\rangle = \langle w^*, \frac{cw^*}{\|w^*\|^2}\rangle = c\frac{\langle w^*, w^*\rangle}{\|w^*\|^2} = c$$

Thus

$$\rho(w^*, \theta_{w^*}) = \frac{1}{2}\min_{z\in C}\langle \frac{w^*}{\|w^*\|}, z\rangle = \frac{1}{2}\langle \frac{w^*}{\|w^*\|}, z^*\rangle = \frac{1}{2}\|z^*\| = \rho$$

For uniqueness, let (w, θ) define another separating hyperplane,that is, let $\rho(w, \theta) = \rho$. According to (1.14),

$$\rho(w, \theta) \leq \rho(w, \theta_w) \text{ for all } \theta.$$

and equality holds only if $\theta = \theta_w$. Hence we may assume without loss of generality that $\theta = \theta_w$. Thus we obtain from (1.11)

$$\rho(w, \theta_w) = \frac{1}{2}\min_{z\in C}\langle \frac{w}{\|w\|}, z\rangle = \rho = \frac{1}{2}\min_{z\in C}\|z\| = \frac{1}{2}\|z^*\|$$

Then

$$\|z^*\| \leq \langle \frac{w}{\|w\|}, z^*\rangle \leq \|z^*\|$$

where the second inequality is due to Cauchy-Schwarz. Hence w and z^* are related via

$$w = \alpha z^* \text{ for some } \alpha > 0.$$

We conclude that the (w, θ) for which the separating separation margin ρ is achieved are unique up to scalar positive multiples. ■

1.5 Optimization Techniques for Optimal Separation

In the previous section, we have seen that the optimal separating hyperplane for the finite data set

$X = X_+ \cup X_- = \{x_1, \ldots, x_N\}$ is determined by minimizing $\|z\|$ over the compact convex set

$C = Conv(X_+ - X_-) = Conv\{z_1, \ldots, z_M\} \subset \mathbb{R}^n$

The vectors z_i are simply obtained by taking all differences $x - y$ where $x \in X_+$ and $y \in X_-$. According to Lemma1.12, there exists a unique $z^* \in C$ with

$$\|z^*\| = \min_{z \in C} \|z\|$$

Definition1.14: Let $C \subset \mathbb{R}^n$ be a convex set. A point $z \in C$ is said to be an **extreme point** or **extremal point** of C if and only if $C - \{z\}$ is convex, i.e., z is not a (relatively) interior point of a segment $[z_1, z_2]$, where $z_1, z_2 \in C$.

The idea of the following line search algorithm is to consider, for a given point $z \in C$, the line segments connecting z with any of the extreme points $z_1, \ldots, z_M$ and to minimize $\|z\|$ along these. Iteratively, this produces a sequence of elements of C that will eventually converges to z^*,as shown below.

Step 0: Choose $\xi(0) = \xi(0,0) \in C$ and set $k := 0$

Step 1: For $j = 1, \ldots, M$: Minimize $\|t\, z_j + (1 - t)\xi(k, j-1)\|$ over $t \in [0,1]$. Let t^* be such that the minimum is achieved, set

$$\xi(k, j) := t^*\, z_j + (1 - t^*)\xi(k, j-1)$$

Step 2: Set $(k+1) := \xi(k+1, 0) := \xi(k, M)$. If $\|\xi(k+1) - \xi(k)\| < \delta$, then stop. Otherwise, augment k by one and go to step 1.

Lemma 1.14: Let $(\xi(k))_{k \in N}$ be the sequence generated according to the algorithm above. Then

$$\|\xi(k+1)\| \leq \|\xi(k)\|$$

and if equality holds , then $\xi(k) = z^*$

Proof: By construction, we have

$$\|\xi(k, j)\| = \min_{t \in [0,1]} \|t\, z_j + (1 - t)\xi(k, j-1)\| \leq \|\xi(k, j-1)\| \qquad (*)$$

Inductively, we get

$$\|\xi(k+1)\| = \|\xi(k, M)\| \leq \ldots \leq \|\xi(k, 0)\| = \|\xi(k)\| \text{ (by *)} \qquad (**)$$

Assume that $\|\xi(k+1)\| = \|\xi(k)\|$. Then we have equality in the relation above, in particular

$\|\xi(k,j)\| = \|\xi(k,j-1)\|$ (Since inequality in (**) is changed in to equality sign) i.e., the minimum is achieved at $t^* = 0$. This implies that $\xi(k,j) = \xi(k,j-1)$ because there is only one vector on a line close to zero and hence

$$\xi(k+1) = \xi(k,\mathrm{M}) = \cdots = \xi(k,0) = \xi(k).$$

Thus for all $= 1, \ldots, M$,

$$\|\xi(k)\| = \min_{\mathrm{t}\in[0,1]} \|t\,\mathrm{z_j} + (1-\mathrm{t})\xi(k)\|$$

that is, for all $\mathrm{t} \in [0,1]$,

$$\|\xi(k)\|^2 \le \|t\,\mathrm{z_j} + (1-\mathrm{t})\xi(k)\|^2 = t^2\|\mathrm{z_j} - \xi(k)\|^2 + 2t\langle \mathrm{z_j} - \xi(k), \xi(k)\rangle + \|\xi(k)\|^2$$

$$\Longrightarrow 0 \le t\|\mathrm{z_j} - \xi(k)\|^2 + 2\langle \mathrm{z_j} - \xi(k), \xi(k)\rangle$$

As $\mathrm{t} \longrightarrow 0$, so is $t\|\mathrm{z_j} - \xi(k)\|^2$. This implies that

$$0 \le \langle \mathrm{z_j} - \xi(k), \xi(k)\rangle$$

for all $j = 1, \ldots, M$ and thus

$$\langle \xi(k), \xi(k)\rangle \le \langle \mathrm{z}, \xi(k)\rangle$$

for all $\mathrm{z} \in \mathrm{C} = Conv\{\mathrm{z_1}, \ldots, \mathrm{z_M}\}$. Using the Cauchy – Schwarz inequality, we get

$$\|\xi(k)\| \le \|z\| \text{ for all } \mathrm{z} \in \mathrm{C}$$

showing that $\|\xi(k)\| = \min_{z\in C} \|z\| = \|z^*\|$. Lemma1.12 implies that $\xi(k) = z^*$.

Alternatively, the optimal separating hyperplane for $X = \mathrm{X_+} \cup \mathrm{X_-} = \{\mathrm{x_1}, \ldots, \mathrm{x_N}\}$ can be found by minimizing $\|w\|$, or equivalently, $\frac{1}{2}\|w\|^2$, subject to $\langle \mathrm{w}, \mathrm{z}\rangle \ge c$ for all $\mathrm{z} \in \mathrm{C} = Conv(\mathrm{X_+} - \mathrm{X_-})$. This is a consequence of Corollary 1.13. Similarly, as in the proof of Lemma 1.10, one shows that

$$\langle \mathrm{w}, \mathrm{z}\rangle \ge 2 \text{ for all } z \in C \Leftrightarrow \exists\theta: \begin{cases} \langle w, x\rangle - \theta \ge 1 & \text{for all } x \in X_+ \\ \langle w, y\rangle - \theta \ge -1 & \text{for all } y \in X_- \end{cases}$$

To see this, suppose $\langle \mathrm{w}, \mathrm{z}\rangle \ge 2$ for all $\mathrm{z} \in \mathrm{C}$. Let $x \in X_+$. Then, for $\theta = \theta_w$, we have

$$\langle w, z\rangle - \theta_w \ge \frac{1}{2}\min_{z\in C} \langle w, z\rangle \ge \frac{1}{2} \times 2 = 1$$

Let $y \in X_-$ .Then there exists $\theta = \theta_w$ such that

$\langle w, z\rangle - \theta_w \leq \frac{1}{2}\min_{z\in C} \langle w, z\rangle \leq -\frac{1}{2}\times 2 = -1$

Therefore, $\langle w, z\rangle \geq 2$ for all $z \in C$

$$\Leftrightarrow \exists\theta: \begin{cases} \langle w, x\rangle - \theta \geq 1 & \text{for all } x \in X_+ \\ \langle w, y\rangle - \theta \geq -1 & \text{for all } y \in X_- \end{cases} \quad (*)$$

Suppose there exist θ such that (*) holds. Let $z = \sum_{i=1}^{M} \lambda_i z_i$, $z_i \in X_+ - X_-$. Then

$\langle w, \sum_{i=1}^{M} \lambda_i z_i\rangle = \sum_{i=1}^{M} \lambda_i \langle w, z_i\rangle \geq 1\times 2 = 2.$ ■

Now for technical reason and instead of minimizing $\|w\|$ we minimize $\frac{1}{2}\|w\|^2$. The only difference from the first one is it is a quadratic optimization problem. Thus the optimal separating hyperplane can be found by solving the quadratic optimization problem

Minimize $\frac{1}{2}\|w\|^2$ subject to $y_i\langle w, x_i\rangle - \theta \geq 1 \quad for \quad 1 \leq i \leq N$ (1.16)

where $y_i = \pm 1$ for if $x_i \in X_\pm$.

For this problem there are standared methods of finding solution. The next theorem is the one. Before going to the Theorem consider the following definitions:

Definition1.15: Let $f : S \to \mathbb{R}$,where S is a non – empty convex set in $\mathbb{R}^n$. The function f is said to be **convex** on S if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

for each $x_1, x_2 \in S$ and each $\lambda \in [0,1]$ and f is concave if $-f$ is convex.

Example: Define $f: \mathbb{R}^n \longrightarrow R$ by $f(x) = x^2 - 2x$. Then f is convex over $\mathbb{R}^n$.

Solution: Let $x, y \epsilon \mathbb{R}^n$ and let $\lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)y) = (\lambda x + (1-\lambda)y)^2 - 2(\lambda x + (1-\lambda)y)$$
$$= \lambda^2 x^2 + (1-\lambda)^2 y^2 + 2\lambda xy - 2\lambda^2 xy - 2\lambda x - 2(1-\lambda)y$$

Rewriting this, we get,

$$= (\lambda^2 - \lambda)(x-y)^2 + \lambda x^2 - 2\lambda x + (1-\lambda)y^2 - 2(1-\lambda)y$$

$$= (\lambda^2 - \lambda)(x-y)^2 + \lambda f(x) + (1-\lambda) f(y) \text{ (by definition)}$$

$$\leq \lambda f(x) + (1-\lambda) f(y) \text{ (since } (\lambda^2 - \lambda) \leq 0 \text{ for all } \lambda \epsilon [0,1]).$$

$\Longrightarrow f$ is convex on $\mathbb{R}^n$. ■

Definition1.16: Let S be a non – empty set in $\mathbb{R}^n$. Then a function $f: S \longrightarrow \mathbb{R}$ is said to be continuous at $\bar{x} \epsilon S$ if, for any given $\varepsilon > 0$, there is $\delta > 0$ such that

$x \epsilon S$ and $\|x - \bar{x}\| < \delta$ implies that

$$|f(x) - f(\bar{x})| < \varepsilon$$

Definition1.17: Let S be a non – empty set in $\mathbb{R}^n$, and let $f: S \longrightarrow \mathbb{R}$. Then f is said to be differentiable at $\bar{x} \epsilon intS$ if there exist a vector $\nabla f(\bar{x})$, called the gradient vector and a function $\alpha: \mathbb{R}^n \longrightarrow \mathbb{R}$ such that

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^{\mathrm{t}}(\mathrm{x} - \bar{\mathrm{x}}) + \|\mathrm{x} - \bar{\mathrm{x}}\|\alpha(\bar{x}; x - \bar{x})$$

for each $x \epsilon S$ where $\lim_{x \longrightarrow \bar{x}} \alpha(\bar{x}; x - \bar{x}) = 0$.

Note: If f is differentiable at $\bar{x}$, then there could only be one gradient vector , and this vector is given by

$$\nabla f(\bar{x}) = \left(\frac{\partial f(\bar{x})}{\partial x_1}, \dots, \quad \frac{\partial f(\bar{x})}{\partial x_n}\right)^t$$

$$\equiv (f_1(\bar{x}), \ , \ , \quad , f_n(\bar{x}))^t$$

where $\frac{\partial f(\bar{x})}{\partial x_i} = f_i(\bar{x})$ is the partial derivative of f with respect to x_i at $\bar{x}$

Theorem 1.15: (***Kuhn-Tucker***) Let $f: \mathbb{R}^m \to \mathbb{R}$ and $g_i: \mathbb{R}^m \to \mathbb{R}$, $1 \leq i \leq \mathrm{N}$ be continuously differentiable convex functions. Then the minimization problem

Minimize : $f(x)$

Subject to $g_1(x) \leq 0 \ , \dots, g_N(x) \leq 0$ is equivalent to the maximization problem

Maximize $F(x, \mu) = f(x) + \sum_{i=1}^{N} \mu_i g_i(x)$ with respect to μ

Subject to $G(\mathrm{x}, \mu) = \nabla f(x) + \sum_{i=1}^{N} \mu_i \nabla g_i(x) = 0$ and $\mu_1, \dots, \mu_N \geq 0$

The parameters $\mu = (\mu_1, \dots, \mu_N)$ can be interpreted as generalized Lagrange multipliers. In our particular situation, $m = n + 1$

$x = \begin{bmatrix} w \\ \theta \end{bmatrix}$ $f(w, \theta) = \frac{1}{2}\|w\|^2 = \frac{1}{2}\langle w, w \rangle = \frac{1}{2}{w_1}^2 + \cdots + \frac{1}{2}{w_n}^2$, and

$$g_i(w, \theta) = 1 - y_i(\langle w, x \rangle - \theta) = 1 - y_i({w_1}\ {x_i}^1 + \cdots + {w_n}\ {x_i}^n - \theta)$$

Then

$$\nabla f(w,\theta) = \begin{bmatrix} \frac{\partial f}{\partial w} \\ \frac{\partial f}{\partial \theta} \end{bmatrix} = \begin{bmatrix} w \\ 0 \end{bmatrix} \text{ and } \nabla g_i(w,\theta) = \begin{bmatrix} \frac{\partial g_i}{\partial w} \\ \frac{\partial g_i}{\partial \theta} \end{bmatrix} = \begin{bmatrix} -y_i\ x_i \\ y_i \end{bmatrix}$$

and thus $G(w,\theta,\mu) = 0$ reads

$$\nabla f(w,\theta) + \sum_{i=1}^{N} \mu_i \nabla g_i(w,\theta) = 0$$

$$\begin{bmatrix} w \\ 0 \end{bmatrix} + \sum_{i=1}^{N} \mu_i \begin{bmatrix} -y_i\ x_i \\ y_i \end{bmatrix} = \begin{bmatrix} w \\ 0 \end{bmatrix} + \mu_1 \begin{bmatrix} -y_1\ x_1 \\ y_1 \end{bmatrix} + \ldots + \mu_N \begin{bmatrix} -y_N\ x_N \\ y_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$\Rightarrow w = \sum_{i=1}^{N} \mu_i\ y_i\ x_i \text{ and } \sum_{i=1}^{N} \mu_i\ y_i = 0.$

Using this, we may write

$$\begin{aligned} F(w,\theta,\mu) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \mu_i\ g_i(w,\theta) \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} \mu_i\ y_i \langle w, x_i \rangle + \sum_{i=1}^{N} \mu_i\ y_i\ \theta \\ &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \mu_i - \langle w, \sum_{i=1}^{N} \mu_i\ y_i\ x_i \rangle \text{ (since } \sum_{i=1}^{N} \mu_i\ y_i = 0) \\ &= -\frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \mu_i \text{ (since } w = \sum_{i=1}^{N} \mu_i\ y_i\ x_i). \end{aligned}$$

$$\begin{aligned} \|w\|^2 &= \langle \sum_{i=1}^{N} \mu_i\ y_i\ x_i, \sum_{i=1}^{N} \mu_i\ y_i\ x_i \rangle \\ &= \sum_{i=1}^{N} \mu_i\ \mu_j\ y_i\ y_j \langle x_i, x_j \rangle \end{aligned}$$

and thus

$$F(w,\theta,\mu) = -\frac{1}{2}\sum_{i=1}^{N} \mu_i\ \mu_j\ y_i\ y_j \langle x_i, x_j \rangle + \sum_{i=1}^{N} \mu_i =: \widetilde{F}(\mu). \qquad \blacksquare$$

Proposition1.16: The optimal solution of (1.19) is given by

$$w^* = \sum_{i=1}^{N} \mu_i{}^*\ y_i\ x_i$$

where $\mu^* = (\mu_1{}^*, \ldots, \mu_N{}^*)$ is the solution to

$$\text{Maximize } \widetilde{F}(\mu) = -\frac{1}{2}\sum_{i=1}^{N} \mu_i\ \mu_j\ y_i\ y_j \langle x_i, x_j \rangle + \sum_{i=1}^{N} \mu_i \qquad (1.17)$$

$$\text{subject to } \sum_{i=1}^{N} \mu_i\ y_i = 0 \text{ and } \mu_1, \ldots, \mu_N \geq 0.$$

Note: 1. The vectors x_i with $\mu_i{}^* \neq 0$ are called **support vectors.**

2. Proposition 1.16 yields a reformulation of the original optimization problem over $\mathbb{R}^n$ as a n op timization pr oblem ove r N. T his w ill be a dvantageous f or **support vector learning.** The idea is t o pe rform n on – linear separation b y e mbedding the da ta i n a higher- dimensional Hilbert space in which they become linearly separable. Then n (the

Hilbert space dimension) is typically significantly larger than N (the number of data points).

3. The important features of (1.17) is the fact that the data vectors $x_1, \ldots, x_N$ enter only in the form $\langle x_i, x_j \rangle$.

1.6 Support Vector Learning

Definition1.18: A symmetric square matrix A is said to be positive semi – definite if $X^T AX \geq 0$ for all vector $X \, in \, \mathbb{R}^n$.

Definition1.19: A continuous symmetric map $k: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a positive kernel if for any finite data set $\{ x_1, \ldots, x_N \} \subset \mathbb{R}^n$, the matrix $K = (k(x_i, x_j))_{i,j=1,..,N}$ is positive semi – definite.

Definition1.20: An orthonormal set S in an inner product space is said to be **complete** if there exists no orthonormal set of which S is a proper subset.

Definition1.21: A complete inner product space **H** is called a **Hilbert space**.

Proposition1.17 :(Multinomial Theorem) The binomial theorem can be extended to n elements in $\mathbb{R}^n$ and has a form

$$(\sum_{i=1}^{n} x_i)^d = \sum_{\alpha \in N^n, |\alpha|=d} \binom{d}{\alpha} x^\alpha \text{ , where } x^\alpha = x_1{}^{\alpha_1} \ldots x_n{}^{\alpha_n}, d \in N$$

$|\alpha| = \sum_{i=1}^{n} \alpha_i$, and $\binom{d}{\alpha} = \frac{d!}{\alpha_1 ! \ldots \alpha_n !}$

Theorem1.18: (**Aronszajn**) If k is a positive kernel, then there exists a Hilbert space H and a map

$\Phi: R^n \to$ H such that

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle \quad \text{for all } x, y \in \mathbb{R}^n.$$

The Hilbert space H is constructed in such a way that it contains all linear combination of functions

$k_x : \mathbb{R}^n \to \mathbb{R}$, y $\vdash$ $k_x(y) = k(x, y)$ and $\Phi(x) = k_x$. Thus

$\langle \Phi(x), \Phi(y) \rangle = \langle k_x, k_y \rangle = k(x, y)$

Example: Consider a map $k: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined by $k(x,y) = \langle x,y\rangle^2$. Let $x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix}$ and $y = \begin{bmatrix} y_{(1)} \\ y_{(2)} \end{bmatrix}$. k is positive kernel and

$$k(x,y) = (x^T y)^2 = (x_{(1)}\, y_{(1)} + x_{(2)}\, y_{(2)})^2$$

$$= x_{(1)}^2\, y_{(1)}^2 + 2\, x_{(1)}\, y_{(1)}\, x_{(2)}\, y_{(2)} + y_{(2)}^2\, x_{(2)}^2$$

$$= \begin{bmatrix} x_{(1)}^2 & x_{(2)}^2 & \sqrt{2}\, x_{(1)}\, x_{(2)} \end{bmatrix} \begin{bmatrix} y_{(1)}^2 \\ y_{(2)}^2 \\ \sqrt{2}\, y_{(1)}\, y_{(2)} \end{bmatrix}$$

$$\Rightarrow k_x = \Phi(x) = (x_{(1)}^2 \quad x_{(2)}^2 \quad \sqrt{2}\, x_{(1)}\, x_{(2)})$$

Examples:

1. $k(x,y) = \langle x,y\rangle^d$ for $d \in N$. k is a polynomial kernel. Since

$k(x,y) = (\sum_{i=1}^n x_i\ y_i)^d$. By multinomial theorem,

$k(x,y) = \sum_{\alpha \in N^n, |\alpha| = d} \binom{d}{\alpha} x^\alpha\, y^\alpha$, where x^α , $\binom{d}{\alpha}$ and $|\alpha|$as shown above.

We put

$$\Phi(x) = \left(\binom{\mathrm{d}}{\alpha}^{\frac{1}{2}} x^\alpha \right)_{\alpha \in N^n, |\alpha| = d}$$

Thus the components of $\Phi(x)$ are (up to the scaling factor) the monomials of total degree d. Since the number of such monomials is finite, $\mathrm{H} = \mathbb{R}^M$ for some $M \in N$ with the Euclidean scalar product.

2. $k(x,y) = (1 + \langle x,y\rangle)^d$ for $d \in N$. Then

$k(x,y) = (1 + \sum_{i=1}^n x_i\ y_i)^d = (x_1\ y_1 + \cdots + (x_n\ y_n + 1))^d = (\sum_{i=1}^{n-1} x_i\ y_i +$ *(xn yn +1))d*

$$= \sum_{\sum_{i=1}^{n-1} \alpha_i + \alpha = d} \binom{d}{\alpha_1, \ldots, \alpha_{n-1}, \alpha} x_1^{\alpha_1} \ldots x_{n-1}^{\alpha_{n-1}}\ y_1^{\alpha_1} \ldots y_{n-1}^{\alpha_{n-1}} \times (x_n\ y_n + 1)^{\alpha_n}$$

$$= \sum_{\sum_{i=1}^{n-1} \alpha_i + \alpha = d} \binom{d}{\alpha_1, \ldots, \alpha_{n-1}, \alpha} x_1^{\alpha_1} \ldots x_{n-1}^{\alpha_{n-1}}\ y_1^{\alpha_1} \ldots y_{n-1}^{\alpha_{n-1}} \times \sum_{\alpha_n + 1 = \alpha} \binom{\alpha}{1, \alpha_n} x_n^{\alpha_n}\ y_n^{\alpha_n}$$

$$= \sum_{\sum_{i=1}^{n} \alpha_i + 1 = d} \binom{d}{\alpha_1, \ldots, \alpha_{n-1}, \alpha} \binom{\alpha}{1, \alpha_n} x_1^{\alpha_1} \ldots x_n^{\alpha_n}\ y_1^{\alpha_1} \ldots y_n^{\alpha_n}$$

Then $|\alpha| + 1 = d \Rightarrow 1! = (d - |\alpha|)!$

$= \sum_{\alpha \in N^n, |\alpha| \leq d} \frac{1}{1!} \frac{d!}{\alpha_1 !... \alpha_{n!}} x^\alpha \, y^\alpha$,where x^α, $|\alpha|$ as in above.

$= \sum_{\alpha \in N^n, |\alpha| \leq d} \frac{1}{(\mathrm{d} - |\alpha|)!} \frac{d!}{\alpha_1 !... \alpha_{n!}} x^\alpha \, y^\alpha$

and we put

$$\Phi(x) = \left(\frac{1}{(\mathrm{d} - |\alpha|)!} \binom{\mathrm{d}}{\alpha}^{\frac{1}{2}} x^\alpha \right)_{\alpha \in N^n, |\alpha| \leq d}$$

and H is a gain f inite – dimensional. T he c omponents of Φ corresponding t o t he monomials of total degree at most d.

Definition1.22: Let $X_+, X_- \subset \mathbb{R}^n$ be a us ual no n – empty finite s ets s uch th at $X = X_+ \cup X_- = \{ x_1, \ldots, x_N \}$. L et $y_i = \pm 1$ if $x_i \in X_\pm$. We say t hat $\Phi : \mathbb{R}^n \to$ H separates X_+, X_- if $\Phi(X_+), \Phi(X_-)$ are affinely separable in H.

Note: If Φ and H are constructed from a p ositive k ernel k as in theorem1.18, this notion depends only on k, and therefore it is justified to say that k separates X_+, X_-.

Let k be a p ositive k ernel an d let Φ and H be as i n Theorem1.18. Suppose t hat k separates X_+, X_-. Then we may solve the optimal separation problem, that is ,

$$\text{Maximize } \widetilde{F}(\mu) = -\frac{1}{2} \sum_{i=1}^N \mu_i \, \mu_j \, y_i \, y_j \, \langle \Phi(x_i), \Phi(x_j) \rangle + \sum_{i=1}^N \mu_i$$

$$\text{subject to } \sum_{i=1}^N \mu_i \, y_i = 0 \text{ and } \mu_1, \ldots, \mu_N \geq 0.$$

Note that

$$\widetilde{F}(\mu) = -\frac{1}{2} \sum_{i=1}^N \mu_i \, \mu_j \, y_i \, y_j \, k(x_i, x_j) + \sum_{i=1}^N \mu_i = -\frac{1}{2} z^T K z + \sum_{i=1}^N \mu_i$$

where $K \in \mathbb{R}^{N \times N}$ and $z \in \mathbb{R}^N$ are defined by

$$K_{ij} = k(x_i, x_j) \text{ and } z_i = \mu_i \, y_i$$

Let $\mu^* = (\mu_1{}^*, \ldots, \mu_N{}^*)$ be a solution, then

$$w^* = \sum_{i=1}^N \mu^*{}_i \, y_i \, \Phi(x_i) \in H$$

and

$$\theta^* = \theta_{w^*} = \frac{1}{2} \left(\min_{x \in X_+} \sum_{i=1}^N \mu^*{}_i \, y_i \, k(x_i, x) + \min_{y \in X_-} \sum_{i=1}^N \mu^*{}_i \, y_i \, k(x_i, y) \right) \in \mathbb{R}$$

Define the separating hyperplane

$$H^* = \{ y \in H : \langle w^*, y \rangle = \theta^* \}$$

For $x \in X_+$,we have $\Phi(x) \in \Phi(X_+)$ and therefore

$\langle w^*, \Phi(x) \rangle = \sum_{i=1}^{N} \mu^*{}_i \; y_i \langle \Phi(x_i), \Phi(x) \rangle = \sum_{i=1}^{N} \mu^*{}_i \; y_i \, k(x_i, x) > \theta^*$

and similarly, for $y \in X_-$

$\sum_{i=1}^{N} \mu^*{}_i \; y_i \, k(x_i, y) < \theta^*$

In other words, we have realized the original function $f: X \to \{\pm 1\}$ with $f(X_\pm) = \pm 1$ by

$$f(x) = \text{Sign}(\sum_{i=1}^{N} \mu^*{}_i \; y_i \, k(x_i, x) - \theta^*)$$

that is , a formal neuron with activation function s(x)= $\sum_{i=1}^{N} \mu^*{}_i \; y_i \, k(x_i, x) - \theta^*$ and output function Sign.

Note that the maximization problem as well as the resulting separation of X_+, X_- depend only on *k*.

Example: Let $f: \{0,1\}^2 \to \{\pm 1\}$ be the modified XOR function, that is we have *N*= 4 and $f(x_i) = y_i$, where

$$\begin{array}{ll} x_1 = (0,0) & y_1 = -1 \\ x_2 = (0,1) & y_2 = \;\; 1 \\ x_3 = (1,0) & y_3 = \;\; 1 \\ x_4 = (1,1) & y_4 = -1 \end{array}$$

Let $k(x,y) = \langle x,y \rangle^2$ for $x, y \epsilon \mathbb{R}^2$. Then $k \in \mathbb{R}^{4\times 4}$

Since $K_{ij} = k(x_i, x_j) = \langle x_i, x_j \rangle^2$ for *i,j =1,2,3,4* and $x_i, x_j \in \mathbb{R}^2$

$K_{11} = \langle x_1, x_1 \rangle^2 = 0$ and $K_{24} = \langle x_2, x_4 \rangle^2 = 1$

Computing similarly the rest value of K_{ij} ,we get

$$K = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 4 \end{bmatrix} \quad \text{and} \quad z \in \mathbb{R}^4, z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} \mu_1 \, y_1 \\ \mu_2 \, y_2 \\ \mu_3 \, y_3 \\ \mu_4 \, y_4 \end{bmatrix} = \begin{bmatrix} -\mu_1 \\ \mu_2 \\ \mu_3 \\ -\mu_4 \end{bmatrix}$$

and we have to maximize

$$\widetilde{F}(\mu) = -\frac{1}{2} z^T K z + \sum_{i=1}^{4} \mu_i$$

$$= \begin{bmatrix} \frac{-\mu_1}{2} & \frac{\mu_2}{2} & \frac{\mu_3}{2} & \frac{-\mu_4}{2} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 4 \end{bmatrix} \begin{bmatrix} -\mu_1 \\ \mu_2 \\ \mu_3 \\ -\mu_4 \end{bmatrix} + \sum_{i=1}^{4} \mu_i$$

$= \frac{-1}{2}(\mu_2{}^2 + \mu_3{}^2) + \mu_2\mu_4 + \mu_3\mu_4 - 2\mu_4{}^2 + 2\mu_2 + 2\mu_3$ (since $\sum_{i=1}^{N} \mu_i\ y_i = 0$)

$= \widetilde{F}(\mu_2, \mu_3, \mu_4)$

subject to $\sum_{i=1}^{N} \mu_i\ y_i = 0$ and $\mu_1, \mu_2, \mu_3, \mu_4 \geq 0$

The gradient of this function is

$$\nabla\widetilde{F} = \begin{bmatrix} \partial\widetilde{F}/\partial\,\mu_2 \\ \partial\widetilde{F}/\partial\,\mu_3 \\ \partial\widetilde{F}/\partial\,\mu_4 \end{bmatrix} = \begin{bmatrix} -\mu_2 + \mu_3 + 2 \\ -\mu_3 + \mu_4 + 2 \\ \mu_2 + \mu_3 - 4\mu_4 \end{bmatrix}$$

Solving $\nabla\widetilde{F} = 0$ yields $\mu_2{}^* = \mu_3{}^* = 4$, $\mu_4{}^* = 2$. This implies that $\mu_1{}^* = 6$.

$\mu^* = (6,4,4,2)$ is a solution of $\widetilde{F}$

Recall that

$\Phi(x) = \Phi(x_{(1)}, x_{(2)}) = (x_{(1)}{}^2, x_{(2)}{}^2, \sqrt{2}\,x_{(1)}\,x_{(2)})$.

$\Phi(x_1) = \Phi(x_{1(1)}{}^2, x_{1(2)}{}^2, \sqrt{2}\,x_{1(1)}\,x_{1(2)}) = (0,0,0)$

Thus

$$\begin{aligned} w^* &= \sum_{i=1}^{N} \mu^*{}_i\ y_i\ \Phi(x_i) \\ &= \mu_1{}^*\,y_1\ \Phi(x_1) + \mu_2{}^*\,y_2\ \Phi(x_2) + \mu_3{}^*\,y_3\ \Phi(x_3) + \mu_4{}^*\,y_4\ \Phi(x_4) \\ &= -6\,\Phi(x_1) + 4\,\Phi(x_2) + 4\,\Phi(x_3) - 2\,\Phi(x_4) \\ &= -6(0,0,0) + 4(0,1,0) + 4(1,0,0) - 2\,(1,1,\sqrt{2}) \\ &= 2\begin{bmatrix} 1 \\ 1 \\ -\sqrt{2} \end{bmatrix} \end{aligned}$$

Note that $X_+ = \{x_2, x_3\}$, $X_- = \{x_1, x_4\}$

$$\theta^* = \frac{1}{2}\left(\min_{x \in X_+} \sum_{i=1}^{N} \mu^*{}_i\ y_i\,k(x_i, x) + \min_{y \in X_-} \sum_{i=1}^{N} \mu^*{}_i\ y_i\,k(x_i, y)\right)$$

$$= \frac{1}{2}\min\{\sum_{i=1}^{4} \mu^*{}_i\ y_i\,k(x_i, x_2), \sum_{i=1}^{4} \mu^*{}_i\ y_i\,k(x_i, x_3)\} +$$

$$\frac{1}{2}\min\{\sum_{i=1}^{4} \mu^*{}_i\ y_i\,k(x_i, x_1), \sum_{i=1}^{4} \mu^*{}_i\ y_i\,k(x_i, x_4)\}$$

$$= \frac{1}{2}\min\{2,2\} + \frac{1}{2}\min\{0,0\}$$

$$= 1$$

Thus the optimal separating hyperplane in $H = \mathbb{R}^3$ is

$$H^* = \{y \epsilon H : \langle w^*, y\rangle = 1\}$$

$$= \{y \epsilon H : \langle w^*, (y_{(1)}, y_{(2)}, y_{(3)})\rangle = 1\}$$

$$= \{y \epsilon H : 2\, y_{(1)} + 2\, y_{(2)} - 2\sqrt{2} y_{(3)} = 1\}$$

with $\Phi(x) = \left(y_{(1)},\, y_{(2)},\, y_{(3)}\right)$, this corresponds to

$$f(x) = \text{Sign}\left(\sum_{i=1}^{4} \mu^*{}_i\; y_i\, k(x_i, x) - \theta^*\right)$$

$$= \text{Sign}\left(\sum_{i=1}^{4} \mu^*{}_i\; y_i\, k(x_i, x) - 1\right)$$

$$= \text{Sign}\left(2\left(x_{(1)}{}^2,\, x_{(2)}{}^2, \sqrt{2}\, x_{(1)}\, x_{(2)}\right) - 1\right)$$

$$= \text{Sign}\left(\left(x_{(2)} - x_{(1)} - \frac{1}{\sqrt{2}}\right)\left(x_{(2)} - x_{(1)} + \frac{1}{\sqrt{2}}\right)\right).$$

Thus the realization which cannot be done by one hyperplane can be done by two planes.

CHAPTER TWO: FEED – FORWARD NETWORKS

The efficiency of perceptron is quite limited. The construction of multilayer networks, which have the capacity to realize all switching functions, remedies the matter. Similarly to the perceptron, linear optimization is used in forward directed networks to determine an explicit separation of arbitrary finite sets. An ANN which allow signals to travel one way only; from input to output is called feed-forward ANN. This chapter ends with an overview of the Back-propagation Algorithm.

2.1 Structure of Feed – Forward Networks

As was illustrated in the previous chapter, the representation of a Boolean function by means of a perceptron is of limited generally, as not all switching functions can be realized. It is often advantageous to link several perceptrons in order to increase the number of functions that can be represented. To this end, the following graph theory definitions are necessary.

In mathematics, a graph is abstract representation of a set of objects where some pairs of the objects are connected by links. The interconnected objects are represented by mathematical abstractions are called **vertices** or **nodes**, and the links that connect some pairs of vertices are called **edges**.

Definition2.1: **1.** Edges that have the same end vertices are ***parallel.***

2. An edge of the form (v, v) is a ***loop.***

3. A graph is ***simple*** if it has no parallel edges or loops.

4. An edge (i,j) $\in E$ is called **directed** , denoted $\boldsymbol{i \rightarrow j}$, (i,j) $\in E$ but (j,i) $\notin E$.

5. A graph G is **directed** or **digraph** if all edges (i,j) are directed.

Definition2.2: **a.** A (finite, simple, directed) graph $G = (V, E)$ is composed of a non-empty finite set of vertices or nodes V and a set of edges E $\subseteq V \times V$.

b. $P(i) = \{j \in V: \quad (j, i) \in E\}$ is the set of all direct predecessors of the node $i \in E$.

$S(i) = \{j \in V: \quad (i, j) \in E\}$ is the set of all direct successors of the node $i \in E$.

c. A vertex i with $P(i) = \emptyset$ is called a **source.**

A vertex i with $S(i) = \emptyset$ is said to be a **sink.**

d. For $i_0, \ldots, i_l \in V$, let $e_j := (i_{j-1}, i_j) \in E$ apply for all $j = 1, \ldots l$. The corresponding sequence of edges $(e_1, \ldots, e_l)$ is called a path from $\boldsymbol{i_0}$ to $\boldsymbol{i_l}$ and the number of edges $\boldsymbol{l}$ is its length.

f. A path whose first and last nodes coincide is called a **cycle**, and a graph that is devoid of cycles is called **acycle**

Example 1: Consider the graph shown below

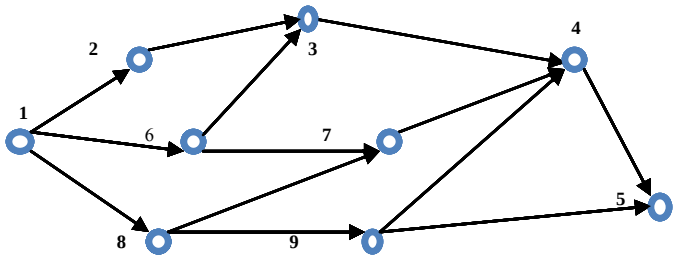

Fig 2.1: *A simple directed acyclic graph*

From this one can easily compute $P(4) = \{3, 7, 9\}$ and node 1 is a source since $P(1) = \emptyset$

Lemma 2.1: An acyclic graph contains a source and a sink.

Proof: Suppose that $G = (V, E)$ contains no sink. Then any vertex possesses a direct successor, which implies that there exists a path of arbitrary length. Consider a path $(e_1, \ldots, e_l)$ whose length exceeds the (finite) number of edges of the graph. The edges cannot be distinct, that is, we must have $e_i = e_j$ for some $1 \leq i < j \leq l$. But then $(e_i, \ldots, e_{j-1})$ has to be cycle. Dually we can show for other case. ■

Let $G = (V, E)$ be an acyclic graph. There exists an integer $k \geq -1$ and a natural partition of V into non-empty sets

$$V = V_0 \cup V_1 \cup \ldots \cup V_k \cup V_{k+1} \tag{2.1}$$

which is constructed as follows:

V_0 contains all sources. Remove from G all vertices in V_0 and all edges that start in one of them. The resulting graph is again acyclic. Let V_1 be the set of all sources in the new graph, etc. Thus,

$i \in V_j$ if and only if the longest path from a source to i has length j. Moreover,

$i \in V_j \Rightarrow P(i) \subseteq V_0 \cup V_1 \cup \ldots \cup V_{j-1}$ and $S(i) \subseteq V_{j+1} \cup \ldots \cup V_{k+1}$

The elements of V_{k+1} are sinks.

Consider an acyclic graph shown below

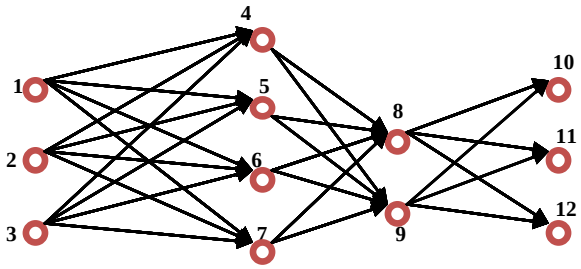

Fig 2.2

From the graph $V = \{1,2,3,4,5,6,7,8,9,10,11,12\}$ and $E \subseteq V \times V$

According to above construction V is partitioned into four non-empty sets, i.e.,

$V = V_0 \cup V_1 \cup V_2 \cup V_3$, where $\{1,2,3\} \subseteq V_0$, $\{4,5,6,7\} \subseteq V_1$, $\{8,9\} \subseteq V_2$, $\{10,11,12\} \subseteq V_3$. Now to illustrate the fact that $i \in V_j$ if and only if the longest path from a source to i has length, consider node **11**$\in V_3$ implies the longest path from a sources 1,2,and 3 to 11 is **3.**

Definition2.3: Let $G = (V, E)$ be a n a cyclic gr aph, a nd l et $V = V_0 \cup V_1 \cup \ldots \cup V_k \cup V_{k+1}$ be t he p artition according t o (2.1). A **Feed – Forward network** (FFN) $\mathcal{F}$ is composed of the graph G and a family of formal neurons $(\boldsymbol{X_i}, \boldsymbol{Y_i}, \boldsymbol{\sigma_i}, \boldsymbol{s_i})$, each associated to one of the non-source vertices $i \in V - V_0$. We have $X_i \subseteq \mathbb{R}^{n_i}$, where $n_i := |P(i)|$, $Y_i \subseteq \mathbb{R}$,

$s_i : X_i \to \mathbb{R}$ and $\sigma_i : \mathbb{R} \to Y_i$

The transfer functions of these neurons are $(f_i)_{i \in V - V_0}$ with $f_i : \sigma_i \, o \, s_i : X_i \to Y_i$

The compatibility requirement $\prod_{j\in P(i)} Y_j \subseteq X_i$ for all $i \in V - V_0$ will be satisfied if we set $X_i = Q^{n_i}$, $Y_i = Q$ for all i,for some $Q \subseteq \mathbb{R}$. From fig2.1a, for node 4, $P(4) = \{1,2,3\} \Rightarrow n_4 := |P(4)| = 3$.Thus $X_4 = Q^3$ and

$\prod_{j\in P(4)} Y_j \subseteq X_4 \Rightarrow Y_1 \times Y_2 \times Y_3 = Q \times Q \times Q = Q^3 \subseteq X_4$ for $4 \in V - V_0$.

For simplicity, we make this assumption in the following.

Then each node $i \in V$ has a state $q_i \in Q$.

Case1: For $i \in V_0$,the states will have to be imposed by some additional initial condition.

Case2: For $i \in V - V_0$, the states is determined by the states of the direct predecessors of i via

$$f_i\left((q_j)_{j\in P(i)}\right) = q_i$$

The nodes in V_0 , $V - (V_0 \cup V_{k+1})$, and V_{k+1} are called input, hidden, and output nodes, respectively. The transfer function of $\mathcal{F}$ is the mapping

f: $Q^{|V_0|} \to Q^{|V_{k+1}|}$ which maps the vector of input states to the output states.

Definition2.4: A FFN $\mathcal{F}$ with $V = V_0 \cup V_1 \cup \ldots \cup V_k \cup V_{k+1}$ according to (2.1) is called k −layer FFN if

$i \in V_j \Rightarrow P(i) \subseteq V_{j-1}$ and $S(i) \subseteq V_{j+1}$ for $j = 0, \ldots, k+1$ (putting $V_{-1} = V_{k+1} = \emptyset$)

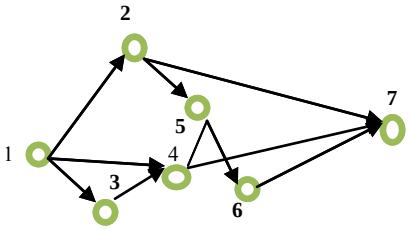

Fig2.3: *Graph of a FFN without a layer structure.*

Fig2.2: *Graph of a 2 – layer FFN.*

Note: We shall restrict to the case where the layers are fully interconnected, that is,

$i \in V_j \Rightarrow P(i) = V_{j-1}$ and $S(i) = V_{j+1}$ for $j = 0, \ldots, k+1$

Then V_0 is the set of all sources (by construction) and V_{k+1} is the set of all sinks. We write $V_0 = V_I$ and $V_{k+1} = V_O$ (the subscript refer to input(I) and output (O), respectively.

Definition2.5: A k – layer FFN whose neurons are all ($\sigma -$) perceptrons, is called a k – layer $\sigma -$ perceptron.

The transfer function of a k – layer $\sigma -$ perceptron is therefore a composition

$f = f^{(k+1)} \circ \ldots \circ f^{(1)}: Q^{|V_I|} \to Q^{|V_O|}$ of $k+1$ mappings of the form

$f^{(i)}: Q^{|V_{i-1}|} \to Q^{|V_i|}$ defined by $q \vdash \underline{\sigma}(W^{(i)}q - \theta^{(i)})$ where $W^{(i)} \in \mathbb{R}^{|V_i| \times |V_{i-1}|}$ is a fixed weight matrix, $\theta^{(i)} \in \mathbb{R}^{|V_i|}$ is a threshold vector, and $\underline{\sigma}$ is a vector – valued sigmoid function defined by component – wise application of $\sigma : \mathbb{R} \to Q$, that is, for any integer $n > 0$

$$\underline{\sigma}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_n) \end{pmatrix}$$

For $i = 1, \ldots, k$, the mapping $f^{(i)}$ is called the transfer function of the $i - th$ hidden layer $f^{(k+1)}$ is referred to as the transfer function of the output layer

2.2: Realization By Multi – Layer Perceptrons

2.2.1: Realization of Boolean functions

As the XOR problem in Lemma 1.1 shows, the representation of arbitrary switching functions by a single perceptron is impossible. The situation is different when a multi – layer perceptron is used. Here, the set of possible states of neurons will be $Q = \{0, 1\}$ and the transfer functions considered will have the form $f: \{0,1\}^n \to \{0,1\}^p$, that is, we consider feed – forward networks with n input nodes and p output nodes. A simple observation is the fact that as long as the number of hidden neurons is unconstrained,

$f = (f_1, \dots, f_p)$ can be realized by such a FFN if and only if all its components $f_i : \{0,1\}^n \to \{0,1\}$ can be realized by a FFN (with n input nodes and one output neuron).

Theorem2.2: Any Boolean function $f: \{0,1\}^n \to \{0,1\}^p$ can be realized by a 1 – layer perceptron.

Proof: Let $p = 1$, without loss of generality. Let $X = \{0,1\}^n$, $X_- = f^{-1}(0)$, $X_+ = f^{-1}(1) = \{x_1, \dots, x_k\}$ and $X_i := \{x_i\}$ for $i = 1, \dots, k$ .Then $X_i \subseteq X_+ \subseteq X$ and $X_+ = \bigcup_{i=1}^{k} x_i \subseteq X$.

Note: a. If a set A is convex then, by definition of convex hull of set, $Conv(A) = A$.

b. A set containing a single element is convex.

Then $Conv(X_i) = X_i$ and $Conv(X_i) \cap Conv(X - X_i) = \emptyset$. This follows from the fact that x_i is an extreme point of the unit cube $Conv(X)$. For any set X and X_i, $Conv(X - X_i)$ the smallest convex set that contain $X - X_i$ but by the remark of convex hull of sets we have $Conv(X - X_i) = Conv(X) - Conv(X_i) = Conv(X) - X_i$ (since X_i convex set) , and therefore $Conv(X - X_i) \subseteq Conv(X) - X_i$. Thus X_i and $X - X_i$ are affinely separable. For $i = 1, \dots, k$. Let $\langle w^{(i)}, x \rangle = \theta^{(i)}$ be a separating hyperplane that separates X_i from $X - X_i$, that is, for $1 \le i \le k$,

$$X_i \subseteq H_i := \{x \in R^n : \langle w^{(i)}, x \rangle \ge \theta^{(i)}\} \text{ and } X_i = X \cap H_i$$

Define $g^{(1)}(x) = y = (y_1, \dots, y_k)^T$ by $y_i := \text{Sat}(\langle w^{(i)}, x \rangle - \theta^{(i)})$ for $i = 1, \dots, k$.

Then with

$g^{(2)}(y) = z = \text{Sat}(y_1 + \dots + y_k - 0.5)$ we have constructed a function $g = g^{(2)} \circ g^{(1)}$ that coincides with the given function f due to

$$g(x) = g^{(2)} \circ g^{(1)}(x) = g^{(2)}\left(g^{(1)}(x)\right) = g^{(2)}(y) = z$$

Now $z = g(x) = 1 \iff Sat(y_1 + \dots + y_k - 0.5) = 1$

$\iff \exists\, i \in \{1, \dots, k\}$ with $y_i = 1$

$\iff \exists\, i \in \{1, \dots, k\}$ with $\langle w^{(i)}, x \rangle \ge \theta^{(i)}$

$\iff \exists\, i \in \{1, \dots, k\}$ with $x \in X_i$

$\iff x \in X_+$

$\iff f(x) = 1.$ ■

Example: The XOR function shall be realizable as a perceptron with one hidden layer (see fig below)

Solution: For the XOR problem

$X_+ = \{(1,0);(0,1)\}$ and $X_- = \{(0,0);(1,1)\}$ with

$X_1 = \{(0,1)\}$ and $X_2 = \{(1,0)\}$ Hence, we have the half spaces

$H_1 = \{x \in R^2 : x_2 - x_1 - 0.5 \geq 0\}$ and

$H_2 = \{x \in R^2 : x_1 - x_2 - 0.5 \geq 0\}$ and thus

$$y_1 = Sat\,(x_2 - x_1 - 0.5)$$

$$y_2 = Sat\,(x_1 - x_2 - 0.5)$$

$$z = Sat\,(y_1 + y_2 - 0.5)$$

This can be summarized in the following table

x_1	x_2	y_1	y_2	z
0	0	0	0	0
0	1	1	0	1
1	0	0	1	1
1	1	0	0	0

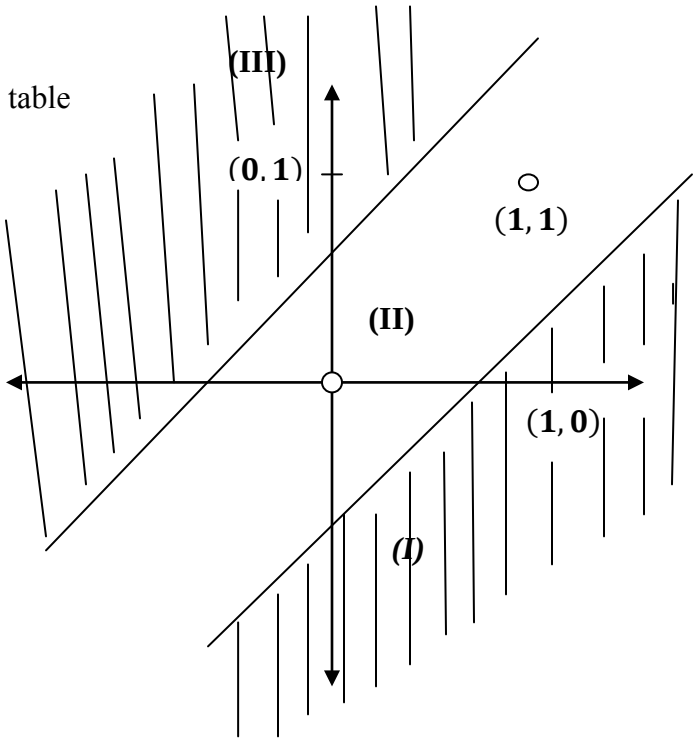

Fig2.4: *Realization of the XOR function by a perceptron with one hidden layer (in region* ***(I)*** *) neuron 1 fires ; in* ***(II)****no neuron fires; in* ***(III)*** *neuron 2 fires.)*

Remarks:

1. The number of neurons that lie in the hidden layer depends on the switching function,i.e.,if the number of hidden neurons is constrained, such a network cannot necessarily represent all switching functions.

2. As Q is dense in $\mathbb{R}$, the weight can always be chosen to be rational numbers. Since multiplication by positive numbers doesn't change the corresponding half – spaces, it even suffices to consider integer – valued weights.

2.2.2: Realization of Arbitrary Functions

Here $Q = \mathbb{R}$. Since we use the Heaviside function, the states of all neurons will nevertheless be in $\{0, 1\}$. Thus we consider transfer functions $f: \mathbb{R}^n \to \{0,1\}^p$, and the networks under consideration have n input nodes and p output neurons. Again it suffices to consider the case where $p = 1$.

Definition 2.6: A set $A \subset \mathbb{R}^n$ is called a **polyhedron** if it is the intersection of a finite number of half – spaces.

Note: A is convex set.

Theorem 2.3: **1.** Let $A \subset \mathbb{R}^n$ be a polyhedron. Then there exists a 1 – layer perceptron whose transfer function $f: \mathbb{R}^n \to \{0,1\}$ satisfies $f^{-1}(1) = A$.

2. Let $B \subset \mathbb{R}^n$ be a finite union of polyhedral. Then there exists a 2 – layer perceptron whose transfer function $f: \mathbb{R}^n \to \{0,1\}$ satisfies $f^{-1}(1) = B$.

Proof:1. Let $A = \{x \in \mathbb{R}^n : \langle x, w_1 \rangle \geq \theta_1, \dots, \langle x, w_k \rangle \geq \theta_k\}$. Let $\mathcal{F}$ be a 1 – layer perceptron with k hidden neurons with transfer functions $g_i : \mathbb{R}^n \to \{0,1\}$, $x \vdash \mathrm{Sat}(\langle w, x_i \rangle - \theta_i)$ and let the transfer function of the output neuron be $h: \{0,1\}^k \to \{0,1\}$, $y \vdash \mathrm{Sat}(y_1 + \cdots + y_k - (k - 0.5))$

Then $g = (g_1, \dots, g_k) \Rightarrow g(x) = (g_1(x), \dots, g_k(x)) \in \{0,1\}^k$

$$\Rightarrow g: R^n \to \{0,1\}^k \ \therefore f: \mathbb{R}^n \xrightarrow{\boldsymbol{g}} \{0,1\}^k \xrightarrow{\boldsymbol{h}} \{0,1\} \Rightarrow f = h \circ g$$

Then the transfer function of $\mathcal{F}$ *is* $f = h \circ g$

Now we have

$x \in A \iff \langle w, x_i \rangle - \theta_i$ for all $i = 1, \dots, k$

$\iff g_i(x) = 1$ for all $i = 1, \dots, k$

$\iff g(x) = (1, \dots, 1)$

$\Leftrightarrow h\left(g(x)\right) = h(1,\dots,1) = Sat\left(1+\cdots+1-(k-0.5)\right) = \text{Sat}(0.5) = 1$

$\Leftrightarrow h\left(g(x)\right) = 1$

$\Leftrightarrow f(x) = 1.$

$\Leftrightarrow x \in f^{-1}(1).$ Therefore, $f^{-1}(1) = A$ ■

2. If $B = \bigcup_{i=1}^{l} A_i$ with A_i the intersection of $k(i)$ closed halfspaces, we first build, as above, a 1 – layer network with n inputs, $\sum_{i=1}^{l} k(i)$ hidden neurons and l output neurons such that the transfer function $g: R^n \to \{0,1\}^l$ of this network satisfies $g(x)_i = 1$ if and only if $x \in A_i$. Then we add a new output layer with one output neuron and set $h: \{0,1\}^l \to \{0,1\}$, $z \vdash Sat(z_1 + \cdots + z_l - 0.5)$

Let $f = h \circ g: \mathbb{R}^n \to \{0,1\}$ be the transfer function of the whole network. We have

$f(x) = 1 \Leftrightarrow h\,(gx)) = 1.$

Put $g(x) = (z_1, \dots, z_l)$ then

$h\,(gx)) = 1 = \text{Sat}(z_1 + \cdots + z_l - 0.5)$

$\Leftrightarrow \exists\, i \in \{1,\dots,l\}$ with $z_i = 1$

$\Leftrightarrow (z_1, \dots, z_l) \neq 0$

$\Leftrightarrow g(x) \neq 0.$

$\Longrightarrow f(x) = 1 \Leftrightarrow g(x) \neq 0$

$\Leftrightarrow \exists\, i : g(x)_i = 1 \Leftrightarrow x \in A_i \Leftrightarrow x \in B$

Therefore, $f^{-1}(1) = B.$ ■

Theorem 2.4: Let $A \subset \mathbb{R}^n$ be a closed and let $B \subset \mathbb{R}^n$ be compact, with $A \cap B = \emptyset$. Then there exists a 2 – layer perceptron whose transfer function $f: \mathbb{R}^n \to \{0,1\}$ satisfies

$$B \subseteq f^{-1}(1) \;\&\; A \subseteq f^{-1}(0).$$

Proof: Since B is compact, there exists a set $B_1 \supseteq B$ which is a finite union of polyhedra. Since A is closed, and $A \cap B = \emptyset$, this set can be chosen such that $A \cap B_1 = \emptyset$. According to Theorem 2.3, there exists a 2 – layer perceptron whose transfer function f satisfies $f^{-1}(1) = B_1 \supseteq B$.

Then $f^{-1}(0) = (\mathbb{R}^n - B_1) \supseteq A$. ■

2.3. Back - Propagation Algorithm (BPA)

According to Theorem 2.4, disjoint finite sets can be separated by 2 – layer perceptron if the number of required neurons in the hidden layers is unconstrained. At this point, we will present another algorithm where the number of neurons is given priori.

The classical perceptron uses the Heaviside function or the sign function. Here, the functions involved have to be differentiable. Therefore, the standard function

$\sigma_t : \mathbb{R} \to [0, 1]$ with

$$\sigma_t(x) = \frac{1}{1+e^{-tx}}$$

will be used as the activation function. The parameter t determines its steepness. An advantage of this sigmoid function is the simple calculation of its derivative:

$$\sigma_t'(x) = \frac{te^{-tx}}{(1+e^{-tx})^2} = \frac{1}{1+e^{-tx}}\left(t - \frac{t}{1+e^{-tx}}\right) = t\sigma_t(x)(1-\sigma_t(x))$$

For simplicity, only the case $t = 1$ will be considered. We write $\sigma_1(x) = \sigma(x)$. Thus

$$\sigma'(x) = \sigma(x)(1-\sigma(x)) \qquad (*)$$

In the following, let a one – layer σ – perceptron be given with n input nodes, l hidden neurons and p output neurons.Suppose that the task of the network is to assign, to a given set of input vectors $x^{(i)} \in \mathbb{R}^n$, $i = 1, \dots, N$, certain prescribed output vectors $\xi^{(i)} \in (0,1)^p$,as precisely as possible. Typically, the network should learn to classify pattern $x \in \mathbb{R}^n$, based on the information contained in the set of given input – output – pairs $(x^{(i)}, \xi^{(i)})$. These correspond to known correct classification (training data). Instead of the discrete value set {0,1},this classification admits a continuous range of output values between 0 and 1.

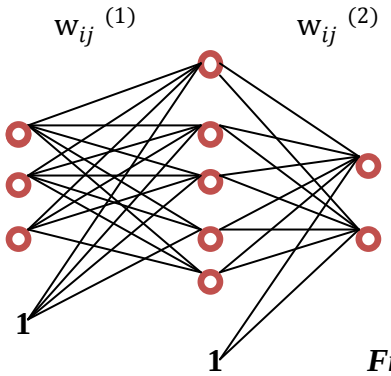

Fig 2.5: *Structure of a network for the BPA*

Let $w_{mk}^{(1)}$ denote the weight between input node m and hidden neuron k, and $w_{kj}^{(2)}$ the weight between hidden neuron k and output neuron j. Hence, the output value at neuron j is

$$z_j^{(i)} = \left(\sum_{k=1}^{l} w_{kj}^{(2)} \sigma\left(\sum_{k=1}^{l} w_{mk}^{(1)} x_m^{(i)} - \theta_k^{(1)}\right) - \theta_j^{(2)}\right)$$

If $x^{(i)}$ is applied as an input. Just like with perceptron, the threshold vectors can be eliminated by introducing auxiliary neurons. In order to simplify the notation, we assume this has already done,i.e., the input layer and the hidden layer have each been extended by one neuron whose state is always 1 (see fig 2.4). The output values of the layer are the input values of the next layer. Thus we write

$$y_j^{(i)} = \sigma\left(\sum_{k=1}^{n+1} w_{mj}^{(1)} x_m^{(i)}\right) \quad \text{and} \quad z_j^{(i)} = \sigma\left(\sum_{k=1}^{l+1} w_{kj}^{(2)} y_k^{(i)}\right)$$

In order to "train" a network with the BPA, the quality of the generated network has to be measurable. Therefore, we introduce the performance function

$$E : \mathbb{R}^{(n+1)l+p(l+1)} \to \mathbb{R}$$

$$\left(w_{11}^{(1)}, \dots, w_{n+1,l}^{(1)}, w_{11}^{(2)}, \dots, w_{l+1,p}^{(1)}\right) \vdash \frac{1}{2}\sum_{i=1}^{N} \left\| z^{(i)} - \xi^{(i)} \right\|_2^2$$

It assigns the mean squared error between the true and the desired network outputs $z^{(i)}$ and $\xi^{(i)}$, respectively, to the vector of weights describing the network. The weights of the network are to be chosen such that E becomes minimal. The partial derivatives of E are computed as follows

$$\frac{\partial E}{\partial w_{kj}^{(2)}} = \frac{\partial E}{\partial z_j^{(i)}} \times \frac{\partial z_j^{(i)}}{\partial w_{kj}^{(2)}} \text{ (Chain rule)}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial}{\partial z_j^{(i)}}\left(\left\| z^{(i)} - \xi^{(i)} \right\|_2^2\right)\right) \times \frac{\partial}{\partial w_{kj}^{(2)}}\left(\sigma\left(\sum_{k=1}^{l+1} w_{kj}^{(2)} y_k^{(i)}\right)\right)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial}{\partial z_j^{(i)}}\left(\left| z_1^{(i)} - \xi_1^{(i)} \right|^2 + \cdots + \left| z_j^{(i)} - \xi_j^{(i)} \right|^2 + \cdots + \left| z_p^{(i)} - \xi_p^{(i)} \right|^2\right)\right) \times$$

$$z_j^{(i)}\left(1 - z_j^{(i)}\right) \frac{\partial}{\partial w_{kj}^{(2)}}\left(\sum_{k=1}^{l+1} w_{kj}^{(2)} y_k^{(i)}\right) \text{ (since } \|x\|_2^2 = \sum_j |x_j|^2 \text{ and by (*))}$$

$$= \sum_{i=1}^{N}\left(z_j^{(i)} - \xi_j^{(i)}\right) z_j^{(i)}\left(1 - z_j^{(i)}\right) y_k^{(i)}$$

And

$$\frac{\partial E}{\partial w_{kj}^{(1)}} = \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial}{\partial w_{kj}^{(1)}}\left(\left| z_1^{(i)} - \xi_1^{(i)} \right|^2 + \cdots + \frac{\partial}{\partial w_{kj}^{(1)}} \left| z_p^{(i)} - \xi_p^{(i)} \right|^2\right)\right)$$

$$= \sum_{i=1}^{N}\left(\left(z_1^{(i)} - \xi_1^{(i)}\right)\frac{\partial z_1^{(i)}}{\partial w_{kj}^{(1)}} + \ldots + \left(z_p^{(i)} - \xi_p^{(i)}\right)\frac{\partial z_p^{(i)}}{\partial w_{kj}^{(1)}}\right.$$

$$= \sum_{i=1}^{N}\left(\left(z_1^{(i)} - \xi_1^{(i)}\right) z_1^{(i)}\left(1 - z_1^{(i)}\right)\frac{\partial}{\partial w_{kj}^{(1)}}\left(\sum_{k=1}^{l+1} w_{k1}^{(2)} y_k^{(i)}\right)\right.$$

$$\left. + \cdots + \left(z_p^{(i)} - \xi_p^{(i)}\right) z_p^{(i)}\left(1 - z_p^{(i)}\right)\frac{\partial z_p^{(i)}}{\partial w_{kj}^{(1)}}\sum_{k=1}^{l+1} w_{kp}^{(2)} y_k^{(i)}\right) \qquad \text{(by (*))}$$

$$= \sum_{i=1}^{N}\left(\sum_{q=1}^{p}\left(z_q^{(i)} - \xi_q^{(i)}\right) z_q^{(i)}\left(1 - z_q^{(i)}\right)\frac{\partial}{\partial w_{kj}^{(1)}}\left(\sum_{k=1}^{l+1} w_{kq}^{(2)} y_k^{(i)}\right)\right)$$

$$= \sum_{i=1}^{N}\left(\sum_{q=1}^{p}\left(z_q^{(i)} - \xi_q^{(i)}\right) z_q^{(i)}\left(1 - z_q^{(i)}\right)\frac{\partial}{\partial y_j^{(i)}}\left(\sum_{k=1}^{l+1} w_{kq}^{(2)} y_k^{(i)}\right)\frac{\partial y_j^{(i)}}{\partial w_{kj}^{(1)}}\right)$$

(by Chain rule)

$$= \sum_{i=1}^{N}\left(\sum_{q=1}^{p}\left(z_q^{(i)} - \xi_q^{(i)}\right) z_q^{(i)}\left(1 - z_q^{(i)}\right) w_{jq}^{(2)} y_j^{(i)}\left(1 - y_j^{(i)}\right) x_k^{(i)}\right)$$

Hence,

$$\frac{\partial E}{\partial w_{kj}^{(2)}} = \sum_{i=1}^{N} \delta_j^{(i)} y_k^{(i)} \quad \text{and} \quad \frac{\partial E}{\partial w_{kj}^{(1)}} = \sum_{i=1}^{N} \Delta_j^{(i)} x_k^{(i)}$$

where

$$\delta_j^{(i)} = \sum_{i=1}^{N}\left(z_j^{(i)} - \xi_j^{(i)}\right) z_j^{(i)}\left(1 - z_j^{(i)}\right) \tag{2.2}$$

$$\Delta_j^{(i)} = y_j^{(i)}\left(1 - y_j^{(i)}\right)\sum_q \delta_q^{(i)} y_{jq}^{(i)} \tag{2.3}$$

A well known method of minimizing E is the gradient descent method, where

$$w_{kj}^{(i)}(t+1) = w_{kj}^{(i)}(t) - \varphi_t \frac{\partial E}{\partial w_{kj}^{(i)}}$$

The BPA is based on such a method

BPA for one-Layer-σ − perceptrons: Initialize the weights $w_{kj}^{(i)}$ randomly.

Set: $i = 1, t = 0$

and choose $\varphi_t > 0, \epsilon > 0$.

Step 1: Feed – forward Calculation. The input vector $x^{(i)}$ is fed in to the network. Compute $y_k^{(i)}$ & $z_j^{(i)}$

Step 2: Back – Propagation to the Hidden Layer. Calculate $\delta_j^{(i)}$ according to (2.2)

Step 3: Back – Propagation to the Input Layer. Calculate $\Delta_j^{(i)}$ according to (2.3). Augment i by one. If $i > N$, go to Step 4, otherwise go to step 1.

Step 4: Evaluation a nd U pdate of Weights. I f $\frac{1}{2}\sum_{i=1}^{N}\left\| z^{(i)} - \xi^{(i)}\right\|_2^2 < \epsilon,$ then s top. Otherwise, define corrected weights by

$$w_{kj}{}^{(2)}(t+1) = w_{kj}{}^{(2)}(t) - \varphi_t \frac{\partial E}{\partial w_{kj}{}^{(2)}}$$

$$w_{kj}{}^{(1)}(t+1) = w_{kj}{}^{(1)}(t) - \varphi_t \frac{\partial E}{\partial w_{kj}{}^{(1)}}$$

Set $i = 1$ and choose $\varphi_{t+1} > 0$. Augment t by one and go to step 1.

It s hould b e c lear th at i n e ach i teration $t = 0,1,2,\ldots$ the c urrent w eights $w_{kj}{}^{(i)}(t)$ are used in step 1&3.

The algorithm is constructed in such a way that each node only needs local information for the calculation of the gradient.

Bibliography

[1]. **Eva Zerz, et al.** Mathematical Theory of Neural Networks, Fachbereich Mathematic Universtat Kaiserslautern.

[2]. **Mokhtar S. Bazaraa, Hanif D. Sherali and C.M. Shetty**, Non – linear programming Theory and Algorithms, Second Edition, John Wiley and Sons, Inc, 1993.

[3]. **R.Deumlich**, Optimization Theory I, Addis Ababa University, 1996.

[4]. **Jan Larsen**, Introduction to Artificial Neural Networks, First Edition, November 1999.

[5]. **Dr.M.Turhan (Tury) Taner**, Neural Networks and Their Supervised Training, Rock Solid Images, 1995.

[6]. **Reinard Diestel**, Graph Theory, Third edition, Springer Berlin Heidelberg New York, 1965.

[7]. **Sebastian M. Cioabä, Maruti Ram Murty**, A First Course in Graph Theory and Combinatorics, Volume55, Hindustan Book Agency, 2009.