



Treball Final de  
Grau en Matemàtiques

---

Machine learning: mathematical  
foundations

Alicia Chimeno Sarabia

---

Supervisor  
Roberto Rubio

Any  
2022/23

Convocatòria  
Juny

# Abstract

In today's world, many people employ machine learning models, yet only a few understand the underlying mathematics that support them. How can we find a predictive function from a given dataset and ascertain the existence of such a function? This research seeks to address these concerns by exploring the mathematical foundations of function approximation in machine learning. Especially focus on function approximation using neural networks. Our research presents a significant finding, demonstrating that a multilayer feedforward network equipped with a non-polynomial activation function can effectively approximate any continuous function. Through this study, we aim to bridge the gap between the practical application of machine learning and the mathematical principles that underpin its success.

# Resum

blslalblallb en català

# Preface

Inpirat per ?.

blablalbla amb glosary Universitat Autònoma de Barcelona (UAB) i ara curt  
UAB.

# Contents

Abstract	i
Glossary	ii
Preface	iii
Contents	iv
<b>1 Introduction</b>	<b>2</b>
<b>2 Machine Learning</b>	<b>3</b>
2.1 Machine Learning Basics . . . . .	3
2.1.1 Linear Regression . . . . .	3
2.1.2 Logistic Regression . . . . .	4
2.1.3 A learning problem . . . . .	4
2.2 Neural Networks . . . . .	6
2.3 Architecture of a Multilayer Feedforward Network . . . . .	7
<b>3 Function Approximation</b>	<b>8</b>
3.1 ??? . . . . .	8
3.1.1 Lebesgue measure . . . . .	9
3.1.2 Convolution . . . . .	10
3.1.3 Baire's thm . . . . .	10
<b>4 Theorem and proof</b>	<b>11</b>
4.1 Theorem . . . . .	11
4.1.1 Previous results . . . . .	11
4.2 Proof . . . . .	12
4.2.1 If $\sigma$ is not a polynomial then $\Sigma_n$ is dense in $\mathcal{C}(\mathbb{R}^n)$ . . . . .	12
4.3 Proof . . . . .	18
4.4 About the theorem . . . . .	18
4.4.1 Why does it not contradict the Weierstrass approximation theorem? . . . . .	18
<b>5 About the theorem</b>	<b>20</b>
5.1 Results . . . . .	20
<b>6 References</b>	<b>21</b>

**A Theory used****22**

# Chapter 1

## Introduction

Computers are like a bicycle for our minds.

— Steve Jobs, *Michael Lawrence Films*

Our brain is constantly classifying and recognizing. For instance, when we spot a dog on the street, one easy classification we can make is {dog, not dog}, which is probably too easy for our brain—it's almost instantaneous. However, things get a bit more complex when we read the teacher's whiteboard. What happens when we encounter a symbol that confuses us because it resembles another? We can interpret the mathematics behind this reasoning <sup>9</sup>as the brain seeking/creating a function that provides us with the certainty of recognizing that particular letter. Eventually, we reach a point where we feel confident enough to write it down in our notes.

Artificial intelligence aims to replicate the remarkable capabilities of our brains. It seeks to develop computational models and algorithms that can perform tasks such as classification, recognition, and decision-making with a level of accuracy and efficiency comparable to human intelligence. When AI first emerged, one of the initial challenges was hand-written digit recognition, exemplified by the MNIST digits dataset. This dataset comprises 60,000 examples of handwritten digits from 0 to 9. To enable a machine learning model to recognize these digits, it must effectively map each image to its corresponding number. This problem naturally aligns with a mathematician's perspective of function learning, where the goal is to approximate a function based on a given dataset consisting of points in space.

Neural Networks are a key approach used in artificial intelligence to tackle such problems. The theory of function approximation through neural networks has a long history dating back to the work by McCulloch and Pitts

This Bachelor's thesis aims to dig into the mathematical foundations of machine learning. Our main ... is to demonstrate that the "real-world" functions we seek to approximate can be effectively approximated by a specific type of functions.

# Chapter 2

## Machine Learning

### 2.1 Machine Learning Basics

*Machine Learning* focuses on the development of algorithms and models that enable computers to learn from data with the aim of making predictions without being explicitly programmed.

We can think about learning as the way we understand it as a human. We can classify a learning problem based on the degree of feedback. Machine learning models fall into three primary categories:

- Supervised learning, where we have immediate feedback.
- Reinforcement learning, where we have indirect feedback. For example when we are playing the game of chess.
- Unsupervised learning, where we have no feedback signal. For example, deducing which dog belongs to each owner.

Machine learning models simplify reality for the purposes of understanding or prediction. This prediction can be either a numerical prediction or a classification prediction. A number of machine learning algorithms are commonly used. These include:

#### 2.1.1 Linear Regression

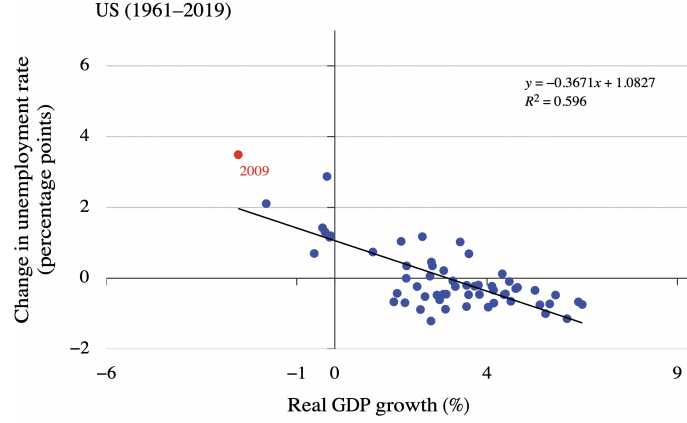
A linear regression algorithm is used to predict numerical values, based on a linear relationship between different values. A simple linear model has an outcome, denoted by  $y$ , also known as a response variable, and a predictor,  $x$ . It is defined by the following equation.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

where  $i = 1, \dots, n$  indexes the observations from 1 to  $n$  in the dataset.

We can add additional  $p$  predictors to a simple linear model, transforming it into a multivariate linear model, which we define as follows:





**Figure 2.1:**  $y$  - response variable: unemployment rate ,  $x$  predictor: GDP growth ,

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \epsilon_i$$

### 2.1.2 Logistic Regression

A logistic regression algorithm is used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. We need to define the concept of sigmoid function that will be important along te work.

A sigmoid function is a mathematical function that maps input values to a range between 0 and 1. We have the following sigmoid function, the logit function:

$$\text{logit} : \mathbb{R} \rightarrow (0, 1) \quad \text{and is expressed as:} \quad \text{logit}(x) = \log \left( \frac{x}{x - 1} \right)$$

The transformed result,  $\text{logit}(x)$ , is expressed in logarithms of probabilities. The probabilities of the result (also known as odds ratios) can be written as:

$$P(y_i = 1) = p_i \quad \text{and} \quad \text{logit}(p_i) = x_i \beta$$

The logistic model can be alternatively written using the inverse logit:

$$P(y_i = 1 | x_i) = \text{logit}^{-1}(x_i \beta)$$

where  $y_i$  is the binary response,  $\text{logit}^{-1}$  is the inverse logit function, and  $x_i \beta$  is the linear predictor.

### 2.1.3 A learning problem

Consider the problem of assessing the eligibility of a consumer for a credit. We are provided with the following set of data:

---

**Costumer application:**


---

Age	23 years
Gender	Male
Annual Salary	\$30,000
Years in Residence	1 year
Years in Job	1 year
Current Debt	\$15,000
...	...

---

- Input:  $x_c = (x_{c_1}, \dots, x_{c_d})$  "attributes of the costumer that we want to classify".
- Output:

$$y = \begin{cases} approve \\ deny \end{cases}$$

- Target funtion:  $f$  "ideal credit approval formula"
- Data: The set  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  corresponds of a historical records of credit customers where  $x_i$  is the attributes of the costumer and  $y_i$  classification awarded.

We are looking for the function  $f$  such that  $f(x_c) = y$ .

A fundamental problem of machine learning is the following. Given data of the form  $\{(x_i, y_i)\}_{i=1}^m \subset \mathbb{R}^n \times \mathbb{R}$ , drawn randomly from a probability distribution  $\mu$ , find a model  $P$  such that  $P(x_i) = y_i$ . An important aspect of machine learning is that *many supervised learning tasks are about function learning.*

**Example 1.** An example of a supervised learning task is digit recognition. The objective is to identify handwritten digits (0-9) based on input images. In this task, we aim to learn a probability distribution function denoted as  $f$ , which maps a set of pixel values ranging from 0 (black) to 255 (white), representing a 28x28 image, to a probability distribution over the digits 0 to 9.

$$f : \{0, \dots, 255\}^{28 \times 28} \longrightarrow \text{probability distribution on } \{0, 1, \dots, 9\}$$

**Example 2.** Example of a classification problem. We want to classify if an image is a dog or not a dog. We would like to produce a value which is correlated with the probability of this image being a dog or not a dog. We can approach the problem in the following way. We want to find a function that takes very high values when dog-image and very low val ues when non dog images and takes the value 0 when its uncertain.

$$d : \mathbb{R}^{\# \text{pixels in image}} \rightarrow \mathbb{R}$$

such that  $\mathbb{P}(d(\text{image})) = \text{probability that the image is a dog.}$

That is what we mean by many problems can be recast as function learning. Note that there is not a god-given reason why this function should exist. We know that certain points in space, and they have certain values associated to them, but we don't know that there is some big function.

*Important principle II:* Sometimes function learning can be recast as a classification problem.

Binary classification problem. Rather learning  $\mu : \mathbb{R}^{\text{\#bits}} \rightarrow \mathbb{R}$  where big values correspond to likely and small values to unlikely. It is better to learn  $\mu : \mathbb{R}^{\text{\#bits}} \rightarrow \text{probability distribution on } \{-1, 0, 1\}$ . In number theory the function  $\mu(n)$  is called Möbius function

$$\mu(n) = \begin{cases} 0 & \text{if } n \text{ has a repeated square factor,} \\ -1 & \text{if } n \text{ has an odd number of distinct prime factors,} \\ 1 & \text{if } n \text{ has an even number of distinct prime factors.} \end{cases}$$

## 2.2 Neural Networks

*Neural Networks* are the quintessential deep learning models, especially *multilayer feedforward networks*. They are widely used for nonlinear function approximation. The goal of a deep neural network is to approximate some function  $f^*$ . For example, for a classifier,  $y = f^*(x)$  maps an input  $x$  to a category  $y$ .

The term *neural* refers to the fact that this model was originally inspired by how biological neurons process information. These artificial neurons mimic the processing of information in biological neurons.

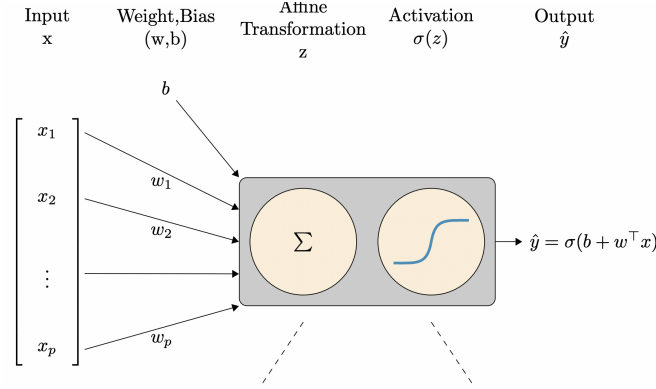
The term *feedforward* indicates the direction of information flow within the network, moving only forward in contraposition to backwards. Each layer processes the input data and passes its output to the next layer, creating a sequence of transformations until the final output is produced.  $f(f_1(f_2(f_3)))$

The term *network* refers to the interconnected structure of artificial neurons\*. An multilayer network consists of multiple layers, including an input layer, one or more hidden layers, and an output layer.

The architecture of the network entails determining its *depth*, *width*, and *activation functions* used. Depth is the number of hidden layers. Width is the number of units (nodes) on each hidden layer.

---

\*A single layer of a deep learning model is also called an *artificial neuron*.



## 2.3 Architecture of a Multilayer Feedforward Network

A single layer of a deep learning model, also called an *artificial neuron* is represented by

$$y = \sigma(w^T x + \theta)$$

Observe that the artificial neuron is composed of an affine transformation  $z = w^T x + \theta$  followed by a (generally) non-linear transformation  $\sigma(z)$ .

We now get into more details on the precise definition of a deep neural network, which is after all a purely mathematical object.

**Definition 1.** A *multilayer feedforward network* is the function

$$f(x) = \sum_{j=1}^k \beta_j \cdot \sigma(w_j \cdot x - \theta_j)$$

where  $x \in \mathbb{R}^n$  is the input vector,  $k \in \mathbb{N}$  is the number of processing units in the hidden layer,  $w_j \in \mathbb{R}^n$  is the weight vector that connects the input to processing unit  $j$  in the hidden layer,  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  is an activation function,  $\theta_j \in \mathbb{R}$  is the threshold (or bias) associated with processing unit  $j$  in the hidden layer, and  $\beta_j \in \mathbb{R}$  is the weight that connects processing unit  $j$  in the hidden layer to the output of the network.

Let  $N_w$  be the family of all functions implied by the network's architecture. If we can show that  $N_w$  is dense in  $C(\mathbb{R}^n)$ , we can conclude that for every continuous function  $g \in C(\mathbb{R}^n)$  and each compact set  $K \subset \mathbb{R}^n$ , there is a function  $f \in N_w$  such that  $f$  is a good approximation to  $g$  on  $K$ .

Under which necessary and sufficient conditions on  $\sigma$  will the family of networks  $N_w$  be capable of approximating to any desired accuracy any given continuous function?

# Chapter 3

## Function Approximation

Creating a machine learning model to predict/classify from a given data is a similar process than when we calculate a function from a given points in the space. This is called function approximation and among the most famous techniques of function approximation, we find interpolation: such as Taylor polynomial, Chebyshev polynomial, the method of least squares, or spline approximation. In this chapter we are going to talk about ...

### 3.1 ????

In this section we present some mathematical definitions and results of function approximation. If we want to approximate functions, we need to define the following notions: distance between functions, density,

**Definition 2.** A *metric* (or *distance*) on a set  $X$  is a function  $d : X \times X \rightarrow \mathbb{R}$  such that for all  $s, t, u \in X$  the following properties are satisfied:

1.  $d(s, t) \geq 0$  and  $d(s, t) = 0$  if and only if  $s = t$ .
2.  $d(s, t) = d(t, s)$ .
3.  $d(s, t) \leq d(s, u) + d(u, t)$  (*triangle inequality*).

A *metric space* is a pair  $(X, d)$ , where  $X$  is a set and  $d$  is a distance in  $X$ .

If we take  $X$  to be a set of functions, the metric  $d(f, g)$  will enable us to measure the distance between functions  $f, g \in X$ .

**Definition 3.** We denote by  $\mathcal{C}(\mathbb{R}^n)$  the set of continuous functions defined on  $\mathbb{R}^n$ .

**Definition 4.** We denote by  $\mathcal{C}_0^\infty$  the set of infinitely differentiable functions (also called smooth functions),  $\mathcal{C}^\infty$ , with compact support. Recall that the support of a function  $u$  is denoted by

$$\text{supp}(u) = \overline{\{x | u(x) \neq 0\}}$$

**Proposition 5.** Let  $\rho$  be a metric defined on the set  $\mathcal{C}_0^\infty[a, b]$  as follows:

$$\rho(\varphi_1, \varphi_2) = \sum_{n=0}^{\infty} 2^{-n} \frac{\|\varphi_1 - \varphi_2\|_n}{1 + \|\varphi_1 - \varphi_2\|_n}$$

where

$$\|\varphi\|_n = \sum_{j=0}^n \sup_{x \in [a, b]} |\varphi^{(j)}(x)|.$$

Then the metric space  $(\mathcal{C}_0^\infty[a, b], \rho)$  is complete, also known as a Fréchet space.

### 3.1.1 Lebesgue measure

**Definition 6.** A box in  $\mathbb{R}^d$  is a set of the form

$$Q = [a_1, b_1] \times \dots \times [a_d, b_d] = \prod_{i=1}^d [a_i, b_i]$$

The volume of the box is

$$\text{vol}(Q) = (b_1 - a_1) \dots (b_d - a_d) = \prod_{i=1}^d (b_i - a_i)$$

The *exterior measure* (or outer measure) of a set  $E \subseteq \mathbb{R}^d$  is

$$|E|^* = \inf \left\{ \sum_k \text{vol}(Q_k) \right\}$$

where the infimum is taken over all finite or countable collection of boxes  $\{Q_k\}$  such that  $E \subseteq \cup_k Q_k$

**Definition 7.** A set  $E \subseteq \mathbb{R}^n$  is *Lebesgue measurable* (or measurable) if  $\forall \epsilon > 0$ , there exist  $U$  open set such that  $E \subseteq U$  and  $|U \setminus E|^* < \epsilon$

**Definition 8.** We say that a property holds almost everywhere (a.e.) if the set of points that doesn't hold it is null.

**Definition 9.** A function  $u$  defined almost everywhere on a measurable set  $\Omega \in \mathbb{R}^n$  is said to be *essentially bounded* on  $\Omega$  if  $|u(x)|$  is bounded almost everywhere on  $\Omega$ . We denote  $u \in L^\infty(\Omega)$  with the norm

$$\|u\|_{L^\infty(\Omega)} = \inf \{ \lambda \mid \{x : |u(x)| \geq \lambda\} = \emptyset \} = \text{ess sup}_{x \in \Omega} |u(x)|$$

We have that  $L^\infty(\mathbb{R})$  is the space of essentially bounded functions.

**Definition 10.** A function  $u$  defined almost everywhere on a domain  $\Omega$  (a domain is an open set in  $\mathbb{R}^n$ ) is said to be *locally essentially bounded* on  $\Omega$  if for every compact set  $K \subset \Omega$ ,  $u \in L^\infty(K)$ . We denote  $u \in L_{loc}^\infty(K)$ .

**Definition 11.** Let  $\mathcal{M}$  denote the set of functions which are in  $L_{loc}^\infty(\mathbb{R})$  and have the following property. The closure of the set of points of discontinuity of any function in  $\mathcal{M}$  is of zero Lebesgue measure.

This implies that for any  $\sigma \in \mathcal{M}$ , interval  $[a, b]$ , and  $\delta > 0$ , there exists a finite number of open intervals, the union of which we denote by  $U$ , of measure  $\delta$ , such that  $\sigma$  is uniformly continuous on  $[a, b]/U$ .

**Definition 12.** We say that a set of functions  $F \subset L_{loc}^\infty(\mathbb{R})$  is *dense* in  $C(\mathbb{R}^n)$  if for every function  $g \in C(\mathbb{R}^n)$  and for every compact  $K \subset \mathbb{R}^n$ , there exist a sequence of functions  $f_j \in F$  such that

$$\lim_{j \rightarrow \infty} \|g - f_j\|_{L^\infty(K)} = 0.$$

### 3.1.2 Convolution

**Definition 13.** Let  $f, g$  be real-valued functions with compact support. We define the *convolution* of  $f$  with  $g$  as

$$(f * g)(x) = \int f(x - t)g(t) dt$$

**Proposition 14.** If  $f$  is a smooth function that is compactly supported and  $g$  is a distribution, then  $f * g$  is a smooth function defined by

$$\int_{\mathbb{R}^d} f(y)g(x - y) dy = (f * g)(x) \in C^\infty(\mathbb{R}^d).$$

**Proposition 15.**

$$\frac{\partial}{\partial x_i}(f * g) = \frac{\partial f}{\partial x_i} * g = f * \frac{\partial g}{\partial x_i}.$$

### 3.1.3 Baire's thm

**Definition 16.** Let  $A$  be a subset of the metric space  $(X, d)$ .  $A$  is said to be *nowhere dense* if for every (nonempty) open subset  $U \subseteq X$ , the intersection  $U \cap \overline{A}$  is not dense in  $U$ , meaning that  $U$  contains a point that is not in the closure of  $A$ .

**Definition 17.** A set it is said to be category *I* if it can be written as a countable union of nowhere-dense sets. Otherwise it is said to be of *category II*

**Theorem 18.** (*Baire's Category Theorem*) Any complete metric space is of category *II*.

# Chapter 4

## Theorem and proof

The work revolves around the following theorem and its proof.

### 4.1 Theorem

**Theorem 19.** *Let  $\sigma \in \mathcal{M}$ . Set*

$$\Sigma_n = \text{span}\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}^n, \theta \in \mathbb{R}\}$$

*Then  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$  if and only if  $\sigma$  is not a polynomial.*

#### 4.1.1 Previous results

Previous research on the approximation capabilities of feedforward networks can be found. These studies show that if the network's activation functions obey an explicit set of assumptions then the network can indeed be shown to be a universal approximator. For example Hornik [1991] has proven two general results, as follows:

**Theorem 20.** *(Hornik Theorem 1). Multilayer feedforward networks with a bounded and nonconstant activation function can approximate any function in  $L^p(\mu)$  arbitrary well, given a sufficiently large number of hidden units.*

**Theorem 21.** *(Hornik Theorem 2) Multilayer feedforward networks with a continuous, bounded and nonconstant activation function can approximate any continuous function on  $X$  arbitrarily well (with respect to the uniform distance) given a sufficiently large number of hidden units.*

The theorem generalizes Hornik's Theorem 2 by establishing necessary and sufficient conditions for universal approximation. Note that the theorem merely requires "nonpolynomiality" in the activation function. Unlike Hornik's result, the activation functions do not need to be continuous or smooth. This has an important biological interpretation because the activation functions of real neurons may well be discontinuous or even non-elementary.



## 4.2 Proof

### 4.2.1 If $\sigma$ is not a polynomial then $\Sigma_n$ is dense in $\mathcal{C}(\mathbb{R}^n)$

Consider that  $\sigma$  is not an algebraic polynomial and we aim to show that  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ . In order to show that, we need to following Lemmas.

**Lemma 1.** If we have that  $\sigma * \varphi$  is a polynomial for all  $\varphi \in \mathcal{C}_0^\infty$ , then the degree of the polynomial  $\sigma * \varphi$  is finite, i.e. there exists an  $m \in \mathbb{N}$  such that  $\deg(\sigma * \varphi) \leq m$  for all  $\varphi \in \mathcal{C}_0^\infty$ .

*Proof.* We first prove the claim in the case of  $\varphi \in \mathcal{C}_0^\infty[a, b]$ , for some  $a < b$ .

By Proposition 5 we have that  $(\mathcal{C}_0^\infty[a, b], \rho)$  is a complete metric space.

Consider the following set,

$$V_k = \{\varphi \in \mathcal{C}_0^\infty[a, b] \mid \deg(\sigma * \varphi) \leq k\}.$$

We want to show that  $\mathcal{C}_0^\infty[a, b] \subseteq V_k$ .

The set  $V_k$  fulfills the following properties:  $V_k \subset V_{k+1}$ ,  $V_k$  is a closed subspace and  $\cup_{k=0}^\infty V_k = \mathcal{C}_0^\infty[a, b]$ .

As  $\mathcal{C}_0^\infty[a, b]$  is a complete metric space, by Baire's Category Theorem 18, this set is of category II, i.e.  $\mathcal{C}_0^\infty[a, b]$  cannot be written as a countable union of nowhere-dense sets. Recall that  $\mathcal{C}_0^\infty[a, b]$  can be written as a countable union of  $V_k$   $\mathcal{C}_0^\infty[a, b] = \cup_{k=0}^\infty V_k$ , therefore some  $V_m$  is not a nowhere-dense set, that is, there exists an open set  $U$  of  $\mathcal{C}_0^\infty[a, b]$  that is contained in the closure of  $V_m$ , but, as  $V_m$  is closed, we have that  $U$  is contained in  $V_m$ . Therefore, some  $V_m$  contains a nonempty open set. By results in topology, any open set of a vector space contains a basis of the vector space, therefore  $V_m$  is a vector space thus  $V_m = \mathcal{C}_0^\infty[a, b]$ .

For the general case where  $\varphi \in \mathcal{C}_0^\infty$ , we note that the number  $m$  does not depend on the interval  $[a, b]$ . This can be seen as follows. By translation  $m$  depends at most of the length of the interval. Let  $[A, B]$  be any interval. For  $\varphi \in \mathcal{C}_0^\infty[A, B]$  we can find  $\varphi_i \in \mathcal{C}_0^\infty[a_i, b_i]$  for  $i = 1, \dots, k$  such that  $[A, B] \subset \cup_{i=1}^k [a_i, b_i]$  where  $b_i - a_i = b - a$  and  $\varphi = \sum_{i=1}^k \varphi_i$  Thus

$$\sigma * \varphi = \sum_{i=1}^k \sigma * \varphi_i$$

and for every  $i = 1, \dots, k$  we have that  $\sigma * \varphi_i$  is a polynomial of degree less than or equal to  $m$ . Therefore  $\deg(\sigma * \varphi) \leq m$ .  $\square$

**Lemma 2.** If  $\sigma * \varphi$  is a polynomial such that  $\deg(\sigma * \varphi) \leq m$  for all  $\varphi \in \mathcal{C}_0^\infty$ , then  $\sigma$  is a polynomial of degree at most  $m$ .

*Proof.* If  $\sigma * \varphi$  is a polynomial of degree  $m$ . For all  $\varphi \in \mathcal{C}_0^\infty$ , using (15) we have that

$$(\sigma * \varphi)^{(m+1)}(x) = \int \sigma(x-y) \varphi^{(m+1)}(y) dy = 0$$

From standard results in Distribution Theory [aqui vam dir de ficar nomes la referencia](#)[e.g., Friedman(1963, 57-59)],  $\sigma$  is itself a polynomial of degree at most  $m$  (a.e.).  $\square$

**Proposition 1.** If  $\sigma$  is not a polinomial then  $\sigma * \varphi$  is not a polynomial for some  $\varphi \in \mathcal{C}_0^\infty$ .

*Proof.* We will show the contrapositive. Suppose that the convolution  $\sigma * \varphi$  is a polynomial for all  $\varphi \in \mathcal{C}_0^\infty$ , by Lemma 1 the degree of the convolution is finite. Now we have that  $\sigma * \varphi$  is a polynomial of finite degree, by Lemma 2 we have that  $\sigma$  is a polynomial.  $\square$

**Lemma 3.** For each  $\varphi \in \mathcal{C}_0^\infty$ ,  $\sigma * \varphi \in \overline{\Sigma_1}$ .

*Proof.* We recall that set

$$\Sigma_1 = \text{span}\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}, \theta \in \mathbb{R}\} \quad (1)$$

Consider

$$h_m = \sum_{i=1}^m \varphi(y_i) \Delta y_i \sigma(x - y_i)$$

The sequence  $(h_m)$  satisfies  $h_j \in \Sigma_1$  for  $j = 1, \dots, m$ . ( $w_i = 1, \theta_i = -y_i, \beta_i = \varphi(y_i) \Delta y_i$ ).

Where  $y_i = -\alpha + \frac{2i\alpha}{m}$ ,  $\Delta y_i = \frac{2\alpha}{m}$  for  $i = 1, \dots, m$ . Partition of the interval  $[-\alpha, \alpha]$

We want to show that  $h_m \rightrightarrows \sigma * \varphi$  in  $[-\alpha, \alpha]$ .

Given  $\epsilon > 0$ , we choose  $\delta > 0$  such that

$$10\delta \|\sigma\|_{L^\infty\{-2\alpha, 2\alpha\}} \|\varphi\|_{L^\infty} \leq \frac{\epsilon}{3}$$

Note that ...

We know that  $\sigma \in M$ . Hence, for this given  $\delta > 0$  and  $[-\alpha, \alpha]$  interval, there exists  $r(\delta)$  finite number of intervals the measure of whose union  $\mathcal{U}$  is  $\delta$  such that  $\sigma$  is uniformly continuous on  $[-2\alpha, 2\alpha]$ . We now choose  $m_i$  sufficiently large so that

1.  $m_1 \delta > \alpha r(\delta)$ . We can do this by Archimedes' principle.

2. From the uniform continuity of  $\varphi$ .

If  $|s - t| \leq \frac{2\alpha}{m_2}$  then

$$|\varphi(s) - \varphi(t)| \leq \frac{\epsilon}{2\alpha\|\sigma\|_{L^\infty[-2\alpha, 2\alpha]}}$$

3. From the previous,  $\sigma$  is uniformly continuous on  $[-2\alpha, 2\alpha]/U$ . We chose  $m$  such that, if  $s, t \in [-2\alpha, 2\alpha]/U$  and  $|s - t| \leq \frac{2\alpha}{m_3}$  then

$$|\sigma(s) - \sigma(t)| \leq \frac{\epsilon}{\|\varphi\|_L}$$

We choose  $m$  such that  $m = \max\{m_1, m_2, m_3\}$ .

Now, fix  $x \in [-\alpha, \alpha]$ . Set  $\Delta_i = [y_{i-1}, y_i]$  where  $y_0 = \alpha$ .

First, recall that,

$$\int \sigma(x - y)\varphi(y)dy = \sum_{i=1}^m \int_{\Delta_i} \sigma(x - y)\varphi(y)dy$$

Consider the following difference

$$\begin{aligned} & \left| \int \sigma(x - y)\varphi(y)dy - \sum_{i=1}^m \int_{\Delta_i} \sigma(x - y_i)\varphi(y)dy \right| \\ &= \left| \sum_{i=1}^m \int_{\Delta_i} \sigma(x - y)\varphi(y)dy - \sum_{i=1}^m \int_{\Delta_i} \sigma(x - y_i)\varphi(y)dy \right| \\ &= \left| \sum_{i=1}^m \int_{\Delta_i} \varphi(y) \left( \sigma(x - y) - \sigma(x - y_i) \right) dy \right| \\ &\leq \sum_{i=1}^m \int_{\Delta_i} |\varphi(y)| |\sigma(x - y) - \sigma(x - y_i)| dy \end{aligned}$$

If  $x - \Delta_i \cap U = \emptyset$ . Since  $x - y \notin U$ ,  $x - y_i \notin U$  and  $x - y_i \in [-2\alpha, 2\alpha]$ . For choice of  $m$  in property 2, we have

$$\begin{aligned} \sum_{i=1}^m \int_{\Delta_i} |\varphi(y)| |\sigma(x - y) - \sigma(x - y_i)| dy &\leq \frac{\epsilon}{\|\varphi\|_{L_1}} \sum_{i=1}^m \int_{\Delta_i} |\varphi(y)| \\ &= \frac{\epsilon}{3\|\varphi\|_{L_1}} \int |\varphi(y)| dy \\ &= \frac{\epsilon}{3\|\varphi\|_{L_1}} \|\varphi(y)\|_{L_1} \\ &= \frac{\epsilon}{3} \end{aligned}$$

If  $x - \Delta_i \cap U \neq \emptyset$

$$\sum_i |\widetilde{\Delta_i}| = \sum_i |(x - \Delta_i \cap U)| \leq |U| + 2|\Delta_i|r(\delta) \leq \delta + 2 \cdot \frac{2\alpha}{m}r(\delta) \leq \delta + 4\delta = 5\delta$$

True by our choice of  $m$ , satisfies  $m\delta > \alpha r(\delta) \iff \delta > \frac{\alpha \cdot r(\delta)}{m}$

$$\begin{aligned} \sum_{i=1}^m \int_{\widetilde{\Delta_i}} |\varphi(y)| |\sigma(x-y) - \sigma(x-y_i)| dy \\ \leq \sum_{i=1}^m \int_{\widetilde{\Delta_i}} \|\varphi\|_{L^\infty} 2\|\sigma\|_{L^\infty[-2\alpha, 2\alpha]} \\ = \|\varphi\|_{L^\infty} 2\|\sigma\|_{L^\infty[-2\alpha, 2\alpha]} \sum_i |\widetilde{\Delta_i}| \\ \leq \|\varphi\|_{L^\infty} 2\|\sigma\|_{L^\infty[-2\alpha, 2\alpha]} 5\delta \leq \epsilon/3 \end{aligned}$$

$$\begin{aligned} \left| \sum_{i=1}^m \int_{\Delta_i} \sigma(x-y_i)\varphi(y)dy - \sum_{i=1}^m \sigma(x-y_i)\varphi(y_i)\Delta y_i \right| \\ = \left| \sum_{i=1}^m \int_{\Delta_i} \sigma(x-y_i)[\varphi(y) - \varphi(y_i)]dy \right| \\ \leq \sum_{i=1}^m \int_{\Delta_i} |\sigma(x-y_i)| |\varphi(y) - \varphi(y_i)| dy \\ \leq \sum_{i=1}^m \int_{\Delta_i} |\sigma(x-y_i)| dy \left[ \frac{\epsilon/3}{2\alpha\|\sigma\|_{L^\infty[-2\alpha, 2\alpha]}} \right] \leq \frac{\epsilon}{3} \end{aligned}$$

Finally, we have the result  $h_m \rightrightarrows \sigma * \varphi$  because

$$\left| \int \sigma(x-y)\varphi(y)dy - \sum_{i=1}^m \sigma(x-y_i)\varphi(y_i)\Delta y_i \right| \leq \epsilon$$

for all  $x \in [-\alpha, \alpha]$  □

**Lemma 4.** If  $\sigma \in \mathcal{C}^\infty$ , then  $\Sigma_1$  is dense in  $\mathcal{C}(\mathbb{R})$ .

*Proof.* We suppose that  $\sigma \in \mathcal{C}^\infty$  and recall by the theorem hypothesis  $\sigma$  is not a polynomial. We can write any function  $f$  of the set  $\Sigma_1$  as

$$f = \sum_i \beta_i \sigma_i(w_i x + \theta_i) = \beta_1 \sigma_1(w_1 x + \theta_1) + \dots$$

We can see that the function

$$\frac{\sigma([w+h]x + \theta) - \sigma(wx + \theta)}{h} \in \Sigma_1$$

because is a linear combination, where  $\beta_1 = \frac{1}{h}, \beta_2 = \frac{-1}{h}$ .

By hypothesis,  $\sigma \in \mathcal{C}^\infty$ . By definition of derivative we have

$$\frac{d}{dw} \sigma(wx + \theta) = \lim_{h \rightarrow 0} \frac{\sigma([w+h]x + \theta) - \sigma(wx + \theta)}{h} \in \overline{\Sigma_1}$$

because the limit of a set belongs to the closure of the set.

By the same argument,

$$\frac{d^k}{dw^k} \sigma(wx + \theta) \in \overline{\Sigma_1}$$

for all  $k \in \mathbb{N}, w, \theta \in \mathbb{R}$ .

If we differentiate this expression  $k$  times, we obtain

$$\frac{d^k}{dw^k} \sigma(wx + \theta) = \sigma^{(k)}(wx + \theta) \cdot x^k$$

We are going to see that if  $\sigma$  is not a polynomial (by hypothesis) then there exists a  $\theta_k \in \mathbb{R}$  such that  $\sigma^{(k)}(\theta_k) \neq 0$ . To show that, let us assume that does not exist any  $\theta_k \in \mathbb{R}$  such that  $\sigma^{(k)}(\theta_k) \neq 0$ . This means that the  $k$ -th derivative at every point is 0, i.e.

$$\sigma^{(k)}(\theta) = 0 \quad \forall \theta \in \mathbb{R}$$

If we integrate this expression, we will have  $\int \sigma^{(k)} = \int 0$ . This implies that

$$\sigma^{(k-1)}(x) = C_1$$

for some constant  $C_1$ , as integrating zero results in a constant. If we integrate again, we have:

$$\sigma^{(k-2)}(x) = C_1 x + C_2$$

for some constants  $C_1$  and  $C_2$ .

Continuing this process, we arrive at

$$\sigma(x) = C_1 x^{k-1} + C_2 x^{k-2} + \dots + C_{k-1} x + C_k$$

for constants  $C_1, C_2, \dots, C_k$ . Hence,  $\sigma$  is a polynomial of degree  $k - 1$ , which contradicts our assumption that  $\sigma$  is not a polynomial. Therefore, there always exists a point where the derivative does not vanish.

Thus, we evaluate at the point where the derivative does not vanish, we call it  $\theta_k$ .

$$\sigma^{(k)}(\theta_k) \cdot x^k = \frac{d^k}{dw^k} \sigma(wx + \theta) \Big|_{w=0, \theta=\theta_k} \in \overline{\Sigma_1}$$

This implies that  $\overline{\Sigma_1}$  contains all polynomials, because the expression  $\sigma^{(k)}(\theta_k)x^k$  generates all polynomials. By the Weierstrass theorem, we know that the polynomials are dense in  $\mathcal{C}(\mathbb{R})$ . This concludes that the set  $\overline{\Sigma_1}$  contains a set which is dense in  $\mathcal{C}(\mathbb{R})$ , therefore  $\Sigma_1$  is dense in  $\mathcal{C}(\mathbb{R})$ . □

**Lemma 5.** If for some  $\varphi \in \mathcal{C}_0^\infty$  we have that  $\sigma * \varphi$  is not a polynomial, then  $\Sigma_1$  is dense in  $\mathcal{C}(\mathbb{R})$ .

*Proof.* From Lemma 3,  $\sigma * \varphi \in \overline{\Sigma_1}$  for each  $\varphi \in \mathcal{C}_0^\infty$ . It thus follows that,  $\sigma * \varphi(w x + \theta) \in \overline{\Sigma_1}$ , for each  $w, \theta \in \mathbb{R}$ . Now for  $\sigma$  and  $\varphi \in \mathcal{C}_0^\infty$ , we have that  $\sigma * \varphi \in \mathcal{C}^\infty$ , see results in distributions 14. From Lemma 4 applied in  $\sigma = \sigma * \varphi$ , if  $\sigma * \varphi \in \mathcal{C}^\infty$ , then  $\Sigma_1$  dense in  $\mathcal{C}(\mathbb{R})$ .  $\square$

We will prove that approximating a  $\mathcal{C}(\mathbb{R})$  function with one from the set  $\Sigma_1$  implies approximating a function  $\mathcal{C}(\mathbb{R}^n)$  from the set  $\Sigma_n$ . Therefore, it is only necessary to approximate a continuous function. We can see this from the density characterization:

**Lemma 6.** If  $\Sigma_1$  is dense in  $\mathcal{C}(\mathbb{R})$ , then  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ .

*Proof.* Let

$$V := \text{span}\{f(ax) : a \in \mathbb{R}^n, f \in \mathcal{C}(\mathbb{R})\}$$

We shall see that  $V$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ . If we show that  $V$  contains the polynomials (which are dense in  $\mathcal{C}(\mathbb{R}^n)$  for Weierstrass Theorem) that would be enough.

**!!mirar** Let  $L(a)$  denote the span of the  $n$  rows of  $a$  for each  $a \in \mathbb{R}^n$ . Set  $L(\mathbb{R}^n) = \cup L(a)$ . Let

$$H_k^n = \left\{ \sum c_m s^m \right\}$$

denote the set of homogeneous polynomials of  $n$  variables of total degree  $k$ , and

$$H^n = \cup_{k=0}^{\infty} H_k^n$$

the set of all homogeneous polynomials of  $n$  variables.

Assume that for a given  $k \in \mathbb{N}$  no non-trivial  $p \in H_k^n \subseteq V$  for all  $k \in \mathbb{Z}$ , then  $V$  contains all polynomials. For that we have  $V$  dense in  $\mathcal{C}(\mathbb{R}^n)$ . Now, we only need to show that  $H_k^n \subseteq V$ . SOS

Let  $g \in \mathcal{C}(\mathbb{R})$ , for any compact subset  $K \subset \mathbb{R}^n$ ,  $V$  dense in  $\mathcal{C}(K)$ . That is, given  $\epsilon > 0$ , there exist  $f_i \in \mathcal{C}(\mathbb{R})$  and  $a_i \in \mathbb{R}^n$   $i = 1, \dots, k$  such that

$$\left| g(x) - \sum_{i=1}^k f_i(a^i \cdot x) \right| < \frac{\epsilon}{2}$$

for all  $x \in K$ . We now consider the set of all the points in the compact  $K$  multiplied by the vector  $a^i$ . That is  $\{a^i \cdot x | x \in K\} \subseteq [\alpha_i, \beta_i]$  for some finite interval  $[\alpha_i, \beta_i]$ ,  $i = 1, \dots, k$ . By hypothesis  $\Sigma_1$  dense in  $\mathcal{C}(\mathbb{R})$ , specifically  $\Sigma_1$  is dense in  $[\alpha_i, \beta_i]$   $i = 1, \dots, k$ . Hence there exist constants  $c_{ij}, w_{ij}$  and  $\theta_{ij}$ ,  $j = 1, \dots, m_i$ ,  $i = 1, \dots, k$  such that

$$\left| f_i(y) - \sum_{j=1}^{m_i} c_{ij} \sigma(w_{ij} y + \theta_{ij}) \right| < \frac{\epsilon}{2k}$$

for all  $x \in K$ .

Therefore,

$$\left| g(x) - \sum_{i=1}^k \sum_{j=1}^{m_i} c_{ij} \sigma(w_{ij}(a^i \cdot x) + \theta_{ij}) \right| < \epsilon$$

□

We showed that to approximate a  $\mathcal{C}(\mathbb{R}^n)$  function we only need to approximate a  $\mathcal{C}(\mathbb{R})$  function with the set  $\Sigma_1$ .

## 4.3 Proof

*Proof.*

$\Rightarrow$  To prove this implication statement, we aim to show that if  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ , then  $\sigma$  is not a polynomial. We will proceed to prove the contrapositive statement, assuming that  $\sigma$  is indeed a polynomial, and demonstrate that in this case,  $\Sigma_n$  cannot be dense in  $\mathcal{C}(\mathbb{R}^n)$ .

Let  $\sigma$  be a polynomial of degree  $k$ , then  $\sigma(wx + \theta)$  is a polynomial of degree  $k$  for every  $w, \theta$ . Recall that

$$\Sigma_n = \text{span}\{\sigma(w \cdot x + \theta) : w \in \mathbb{R}^n, \theta \in \mathbb{R}\}$$

that is the set of algebraic polynomials of degree at most  $k$ . To show that  $\Sigma_n$  is not dense in  $\mathcal{C}(\mathbb{R}^n)$ , for the definition of density, we need to find a function  $f(x) \in \mathcal{C}(\mathbb{R}^n)$ ,  $\epsilon > 0$  and  $K$  such that  $\|p - f\| > \epsilon$  for all  $p$  polynomial of degree  $k$ . For example, let  $f(x) = \cos(x)$ , and let  $p(x) = \sigma(wx + \theta)$  that has degree at most  $k$ . This implies  $p$  has maximum  $k$  roots. We can find an interval where  $\cos(x)$  has  $k+1$  roots. Therefore,  $\Sigma_n$  is not dense in  $\mathcal{C}(\mathbb{R}^n)$ .

$\Leftarrow$  In order to prove this implication, we need to show that if  $\sigma$  is not an algebraic polynomial, then  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ .

By hypothesis  $\sigma$  is not a polynomial, by Proposition 1 this implies that  $\sigma * \varphi$  is not a polynomial. For Lemma 3 we have that  $\sigma * \varphi \in \overline{\Sigma_1}$ . We showed in Lemma 5 that  $\sigma * \varphi \in \mathcal{C}^\infty$  and for Lemma 4 and Lemma 5 we have that  $\Sigma_1$  is dense in  $\mathcal{C}(\mathbb{R})$ . Finally, for Lemma 6  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$ .

□

## 4.4 About the theorem

### 4.4.1 Why does it not contradict the Weierstrass approximation theorem?

**Theorem 22.** (*Weierstrass approximation theorem*). Let  $f : [a, b] \rightarrow \mathbb{R}$  be a continuous function. Then, there exists polynomials  $p_n \in \mathcal{R}[x]$  such that the sequence  $(p_n)$  converge uniformly to  $f$  on  $[a, b]$ .

**Corollary 23.** The set of polynomial functions  $\mathcal{R}^n[x]$  is dense in the space of continuous functions on a compact set  $K \subset \mathbb{R}^n$ ,  $\mathcal{C}(K)$ . So any continuous function on a compact set can be approximated arbitrarily well by a polynomial.

The theorem states that: if  $\Sigma_n$  is dense in  $\mathcal{C}(\mathbb{R}^n)$  then  $\sigma$  is not an algebraic polynomial. But why this statement does not contradict the Weierstrass approximation theorem? This does not work because  $\sigma$  has degree fixed  $k$ , then any element in the set  $\Sigma_n$  has degree at most  $k$ . Hence, the set  $\Sigma_n$  is a finite vector space and can not be dense in  $\mathcal{C}(\mathbb{R}^n)$ . Not all continuous functions can be approximated with a polynomial of degree fixed, for example: (comment per afegir : per exemple una funció que sigui contínua que no es pugui aproximar per un polinomi de grau  $k$ , una  $k$  tingui grau més gran que  $k$  polinomi de  $k+1$  ?? )

Leshno et al. [1993]



# Chapter 5

## About the theorem

### 5.1 Results

**Definition 24.** The set  $L^p(\mu)$  contains all measurable functions  $f$  such that:

$$\|f\|_{L^p(\mu)} = \left( \int_{\mathbb{R}^n} |f(x)|^p d\mu(x) \right)^{1/p} < \infty$$

**Proposition 25.** Let  $\mu$  be a non-negative finite measure on  $\mathbb{R}$  with compact support, absolutely continuous with respect to Lebesgue measure. Then  $\Sigma_n$  is dense in  $L^p(\mu)$ ,  $1 \leq p < \infty$ , if and only if,  $\sigma$  is not a polynomial.

**Proposition 26.** If  $\sigma \in M$  is not a polynomial (a.e) then,

$$\Sigma_n(\mathcal{A}) = \text{span}\{\sigma(\lambda w \cdot x + \theta) : \lambda, \theta \in \mathbb{R}, w \in \mathcal{A}\}$$

is dense in  $\mathcal{C}(\mathbb{R}^n)$  for some  $\mathcal{A} \subset \mathbb{R}^n$  if and only if there does not exist a nontrivial polynomial vanishing on  $\mathcal{A}$ .

**Remark 1.** The theorem only requires for the activation function to be nonpolynomial, we don't need continuity on sigma. For example, let  $\sigma$  be continuous with a jump discontinuity at 0 such that:

$$\lim_{x \rightarrow 0^-} \sigma(x) = 0 \quad \lim_{x \rightarrow 0^+} \sigma(x) = 1$$

Given  $f \in \mathcal{C}(\mathbb{R})$  and  $K \subset \mathbb{R}$  compact, letting  $w \rightarrow 0$  in  $\sigma(wx)$  the function

$$h(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x > 0 \end{cases}$$

$h \in \overline{\Sigma_1}$ .

Linear combinations of  $h$  and its translates can uniformly approximate any continuous function on any finite interval (and thus any compact subset of  $\mathbb{R}$ ).

# Chapter 6

## References

- K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080. doi: [https://doi.org/10.1016/0893-6080\(91\)90009-T](https://doi.org/10.1016/0893-6080(91)90009-T). URL <https://www.sciencedirect.com/science/article/pii/089360809190009T>.
- M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

# Appendix A

## Theory used

**Definition 27.** Riemann integral reminder. The Riemann integral is a method for calculating the volume under a curve of a continuous function on a closed, bounded domain in  $\mathbb{R}^n$ . The method involves dividing the domain into smaller subregions and approximating the volume of each subregion with a rectangular solid whose height is the function value at a specific point in the subregion. The Riemann sum is the sum of the volumes of all the rectangular solids, and as the size of the subregions approaches zero, the Riemann sum converges to the Riemann integral.

**Definition 28.** Let  $\Sigma$  be a  $\sigma$ -algebra over a set  $\Omega$ . A *measure* over  $\Omega$  is any function

$$\mu : \Sigma \longrightarrow [0, \infty]$$

satisfying the following properties:

1.  $\mu(\emptyset) = 0$ .
2.  $\sigma$ -*additivity*: If  $(A_n) \in \Sigma$  are pairwise disjoint, then:

$$\mu \left( \bigsqcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} \mu(A_n)$$

**Definition 29.** The closure of a set  $A$  of a metric space  $(X, d)$  is defined as follows:

$$\text{closure}(A) = \overline{A} = \{t \mid \forall \epsilon > 0, \exists a \in A, d(a, t) < \epsilon\}.$$

**Proposition 30.** Let  $(X, \tau)$  be a topological space and  $A \subseteq X$  be a subset. Then,  $A$  is dense in  $(X, \tau)$  if and only if  $\overline{A} = X$ .

**Definition 31.** A metric space  $(X, d)$  is said to be *complete* if every Cauchy sequence in  $X$  converges to a point in  $X$ .

**Definition 32.** We say that a property holds almost everywhere (a.e.) if the set of points that doesn't hold it is null.

**Definition 33.**  $\varphi : I \rightarrow \mathbb{R}$  is uniformly continuous on  $I$  if  $\forall \epsilon > 0 \exists \delta > 0$  such that  $|\varphi(x) - \varphi(y)| < \epsilon$  whenever  $|x - y| < \delta$