On universality of fully-connected neural networks

Gabrijel Boduljak

Year 4 Project
School of Mathematics
University of Edinburgh
11 March 2022

Abstract

For a long time, it was impossible to imagine that a computer could accurately classify and segment images, summarise or generate text and play strategic computer games at a superhuman level. Recently, a family of machine learning algorithms involving artificial neural networks started to excel at those tasks, often outperforming humans and alternative methods. Artificial neural networks are a family of machine learning algorithms capable of approximating functions by extracting increasingly complex hierarchical representations from the data.
This thesis aims to present key results in the approximation theory of artificial neural networks assuming only undergraduate mathematics. This research field studies necessary and sufficient conditions under which neural networks can approximate an arbitrary function belonging to a particular family. The approximation is formalized within a function space. Theorems addressing those issues are known as the universal approximation theorems. We will state and prove various universal approximation theorems for continuous functions on compact sets. Those results will be generalized to spaces of Lebesgue integrable and square-integrable functions. We will also discuss the universal approximation of Borel measurable functions in a probabilistic sense. We will conclude with an experimental study of the relationship between established theoretical results and practical applications.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Gabrijel Boduljak*)

I dedicate this thesis to my family, for their endless support and encouragement.

Contents

Chapter 1

Introduction

1.1 Motivation

For a long time, it was difficult to imagine that a computer could accurately classify and segment images, summarise or generate text or play strategic computer games at a superhuman level. It is interesting to note that most of the progress in those problems is driven by deep learning. Deep learning refers to a family of machine learning algorithms related to artificial neural networks. Artificial neural networks are a family of machine learning algorithms capable of learning functions by extracting increasingly complex hierarchical representations from the data. The name comes from a biological inspiration for their structure. Mathematically, they are often nothing but a composition of nonlinear transformations of the input data. Those transformations are often initialized randomly and then learned from the data by some numerical optimization algorithm. Those transformations are often layered and parameterized. The computation of finding the optimal parameters for those transformations is called learning or training.

Although most of the computational problems mentioned above look seemingly unrelated, it turns out that they are all mathematically the same - they are all an instance of the problem of learning or approximating a (possibly) complex, an unknown function given data. Interestingly, the artificial neural networks excel at all mentioned tasks, often performing significantly better than different machine learning algorithms. Despite their impressive experimental performance, neural networks are often regarded as black-box models, due to the lack of theoretical guarantees and the difficulty of understanding their learning and decision-making process.

Given the recent success, it was natural to explore the mathematical properties of artificial neural networks and question their power of approximating functions. This thesis will focus on the simplest forms of neural networks - feed-forward, fully-connected neural networks. Despite their apparent simplicity, the rigorous analysis of the representation and approximation power of feed-forward neural networks turns out to be quite difficult, often involving various fields within mathematics, including general topology, measure theory, and functional analysis. Moreover, this is an open research problem, and papers addressing those issues are still published. You can read more about this in Literature review.

However, we have strong theoretical results regarding the approximation power of feed-forward neural networks. This project will tackle some of those in the increasing order of their complexity and generality. Although most of the results presented in this thesis are well-known in the machine learning community, they are often barely mentioned in machine learning textbooks and stated without proof or further rigorous discussion. A possible explanation of such a situation is the dependence on concepts and results from functional analysis, abstract measure theory, and general topology. Since the machine learning community is quite interdisciplinary, such demands on mathematical prerequisites are often out of the scope of those textbooks, aimed towards the more general audience. This thesis is an attempt to present the most fundamental results in approximation theory of neural networks assuming only undergraduate mathematics background. The necessary more advanced mathematical concepts are discussed in Appendix. In this thesis, the focus is on important theoretical results and proofs, presented in Universality of Neural Networks. The relationship between established theoretical guarantees and practical performance of neural networks is discussed in Experiments.

1.2 Machine Learning

In this section, we will introduce the standard machine learning concepts following the viewpoint of the statistical learning theory. Although this project is not concerned with issues studied in the statistical learning theory, definitions developed in that field are often useful to formalize common problems studied in machine learning, such as classification and regression. Definitions that will be presented are modifications of ones presented in [SB14].

1.2.1 Standard machine learning terminology

Let $\mathcal{X}$ be the set of inputs, also called observations. Let $\mathcal{Y}$ be the finite set of outputs, also called labels.

Definition 1 (a training set)**.** A training set is a finite subset of $\mathcal{X} \times \mathcal{Y}$, often denoted by S,

$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subseteq \mathcal{X} \times \mathcal{Y}.$$

Definition 2 (a learning algorithm)**.** A learning algorithm $\mathcal{A}$ is a map:

$$\mathcal{A} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Remark 1. The input of a learning algorithm is a training set S. The range of a learning algorithm is a set of functions which can be learned, denoted by $\mathcal{F}$.

Definition 3 (a hypothesis)**.** A function $h : \mathcal{X} \to \mathcal{Y}$ is called a **hypothesis**.

Remark 2. The output of a learning algorithm $\mathcal{A}$ given the training set S is the hypothesis $h_S = \mathcal{A}(S)$.

1.2.2 An important probabilistic assumption

Elements of a training set S, (x_i, y_i), are treated as outcomes of random variables (X_i, Y_i) which are independently and identically distributed according to an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. It is often assumed that the underlying σ-algebra is the product σ-algebra of Borel σ-algebras with respect to usual topologies. The joint distribution of a training set S is often denoted by $\mathcal{D}^n$.

1.2.3 Machine learning tasks

The main goal of a learning algorithm is to find the optimal hypothesis h with respect to the suitably chosen **loss function** $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The purpose of the loss function L is to measure the error between $h(x)$ and the expected true y corresponding to x. The quantity of the central interest is **generalization error**.

Definition 4 (generalization error)**. Generalization error** of the hypothesis h is given by

$$E(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[L(y, h(x))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, h(x))\, d\mathcal{D}(x, y).$$

Remark 3. In statistical learning literature, a generalization error is also known as a risk. In machine learning context, the term generalization error is more prevalent.

The central machine learning problem solved by learning algorithms is finding the hypothesis h minimising the generalization error. The main difficulty is the fact that the probability distribution D is unknown and the only information given to the learning algorithm is a training set S. In other words, the learning algorithm $\mathcal{A}$ is tasked to produce a hypothesis function h_S given only S, which can be a very small subset of $\mathcal{X} \times \mathcal{Y}$. Depending on whether $\mathcal{Y}$ is continuous or discrete, a learning algorithm is solving either a regression or a classification problem. Although this distinction seems unnecessary, regression and classification problems are fundamentally different from statistical perspective. This is often reflected in the choice of the loss function.

1.2.4 Model selection

Given the dataset, there are usually multiple candidate hypotheses. Candidate hypotheses often arise from different learning algorithms. Sometimes, hypotheses are also parameterized and the same learning algorithm can produce different hypotheses, each corresponding to a particular parameter set. Model selection is a problem of selecting the best hypothesis from the set of candidate hypotheses, given the data. Usually, the best model is defined as the model achieving the smallest generalization error. Since the generalization error is often computationally intractable, the generalization error of each candidate hypothesis is estimated from the data which was not used to train the model. This dataset is often known as a validation or a test set. However, there are many ways to estimate the generalization error. An alternative approach is k-fold cross-validation.

In Experiments, we will use a validation set to estimate generalization error.

1.3 Deep Learning

1.3.1 An artificial neuron

The fundamental building block of an artificial neural network is an artificial neuron.

Definition 5 (neuron)**.** A *d*-**dimensional neuron** is a function $f_{\phi,\mathbf{w},b} : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f_{\phi,\mathbf{w},b}(x_1, \ldots, x_d) = \phi\left(\sum_{k=1}^{d} w_k x_k + b\right)$$

where $\phi : \mathbb{R} \to \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. The vector $\mathbf{w}$ is known as the **weight vector** and its components w_i are known as **weights**. The constant b is known as a **bias** of the neuron. The function ϕ is known as an **activation function**.

Remark 4. The function $f_{\phi,\mathbf{w},b}$ is often expressed in the following matrix form

$$f_{\phi,\mathbf{w},b}(\mathbf{x}) = \phi(\langle \mathbf{w}, \mathbf{x} \rangle + b) = \phi(\mathbf{w}^\top \mathbf{x} + b).$$

Remark 5. We often want an activation function ϕ to be a function nonlinear in the input. The nonlinearity constraint plays an important role in the representation power of a single neuron and hence the neural network. Until recently, it was often imposed that ϕ is differentiable. Both of those constraints will be discussed in the upcoming sections.

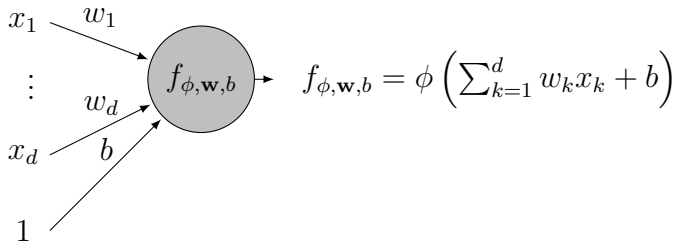


Figure 1.1: A neuron parameterized with a weight vector $\mathbf{w}$, a bias b and an activation function ϕ. Addition of a bias parameter b is often represented as a "virtual" input 1 connected to the neuron with a weight of value b.

The structure and computation of an artificial neuron $f_{\phi,\mathbf{w},b}$ is inspired by the structure of the biological neuron. Weights are inspired by synapses and the activation function ϕ is used to model the amount of information passed after the neuron processes the input. The model of an artificial neuron has no intention of emulating the much more complex biological counterpart. This analogy is visualized in Figure 1.1. However, there is a strong connection between linear regression, logistic regression, and the single neuron. For more about linear regression and logistic regression, see Chapter 10 and Chapter 11 in [Mur22].

1.3.2 Activation functions

Commonly used activation functions include the logistic sigmoid (σ), rectified linear unit (ReLU) and hyperbolic tangent (tanh). Less commonly used is the hard binary threshold, also known as the Heaviside step function. However, there are many other activation functions and some of them were invented quite recently. Those include PReLU[He+15b], SELU[Kla+17], ELU[CUH15], PELU[TGC18]. We will present the most commonly used activation functions.

Definition 6. The logistic sigmoid $\sigma : \mathbb{R} \to [0, 1]$ is given by $\sigma(x) = \frac{1}{1+\exp(-x)}$.

Definition 7. The hyperbolic tangent $\tanh : \mathbb{R} \to [-1, 1]$ is given by $\tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$.

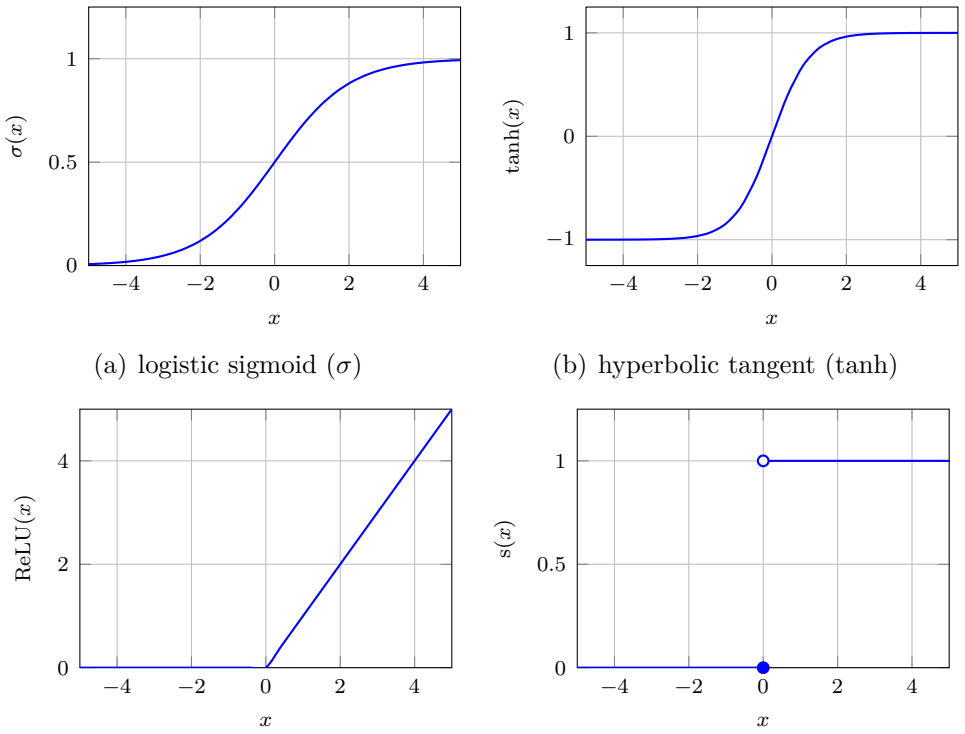
Definition 8. The rectified linear unit $\mathrm{ReLU} : \mathbb{R} \to [0, \infty)$ is given by $\mathrm{ReLU}(x) = \max(0, x)$.

Definition 9. The Heaviside step function $s : \mathbb{R} \to \{0, 1\}$ is given by

$$s(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}.$$

Lemma 1. *The logistic sigmoid is differentiable on $\mathbb{R}$. Moreover, its derivative satisfies $\frac{\partial\sigma}{\partial x}(x) = \sigma(x) \cdot (1 - \sigma(x))$.*

Proof. Let $x \in \mathbb{R}$. Then $\frac{\partial\sigma}{\partial x}(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} = \sigma(x) \cdot (1 - \sigma(x))$, as desired. ■



(a) logistic sigmoid (σ)

(b) hyperbolic tangent (tanh)

(c) rectified linear unit (ReLU))

(d) Heaviside step function (s)

Figure 1.2: Commonly used activation functions

1.3.3 A fully-connected layer

To perform more complex computations, artificial neurons are often organized to form layers. The following definition will introduce the simplest form of a layer - a **fully-connected layer**. Many modern neural network layers can be very complex.

Definition 10 (fully-connected layer)**.** A **fully-connected layer** is a function $f_{\phi,\mathbf{W},\mathbf{b}} : \mathbb{R}^n \to \mathbb{R}^m$ of the form

$$f_{\phi,\mathbf{W},\mathbf{b}}(x_1,\ldots,x_n) = \begin{bmatrix} f_{\phi,\mathbf{w_1},b_1}(x_1,\ldots,x_n) \\ f_{\phi,\mathbf{w_2},b_2}(x_1,\ldots,x_n) \\ \vdots \\ f_{\phi,\mathbf{w_m},b_m}(x_1,\ldots,x_n) \end{bmatrix},$$

where $\mathbf{W}$ is a matrix of weights corresponding to each neuron in the layer such that $\mathbf{W}_{i,j}$ is the weight connecting the ith input x_i to jth neuron in the layer. For $1 \leq k \leq m$, we denote the weight vector of kth neuron in the layer by $\mathbf{w}_k$ and we denote the bias of kth neuron in the layer by b_k. Hence, weight vectors $\mathbf{w_1},\ldots,\mathbf{w_m}$ are the columns of $\mathbf{W}$, and layer biases $b_1,\ldots,b_m$ are the components of the bias vector $\mathbf{b}$. The layer weight matrix $\mathbf{W}$ and the layer bias vector $\mathbf{b}$ are given by

$$\mathbf{W} = \begin{bmatrix} | & | & | & | \\ \mathbf{w_1} & \mathbf{w_2} & \ldots & \mathbf{w_m} \\ | & | & | & | \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}.$$

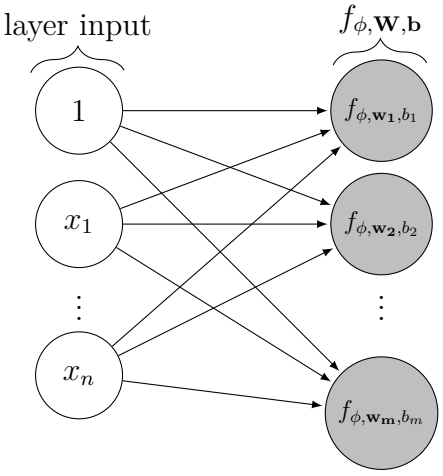


Figure 1.3: A fully-connected layer with n-dimensional input and m-dimensional output, parameterized with a weight matrix $\mathbf{W}$, bias vector $\mathbf{b}$ and an activation function ϕ. For the sake of clarity, weight and bias labels are omitted.

Remark 6. The layer $f_{\phi,\mathbf{W},\mathbf{b}}$ is often written in the more succint, matrix form

$$f_{\phi,\mathbf{W},\mathbf{b}}(\mathbf{x}) = \phi(\mathbf{W}^\top \mathbf{x} + \mathbf{b}).$$

The application of ϕ is understood component-wise.

Definition 11. We define the **width** of a layer to be the number of neurons in the layer. Equivalently, the width of a layer is the dimension of the layer output.

1.3.4 A fully-connected neural network

We are ready to introduce a fully-connected neural network, which is central to this thesis.

Definition 12 (fully-connected neural network)**.** A **fully-connected neural network** of depth L is a composition of L fully-connected layers. Suppose that the first layer is a function $f^{(1)}_{\phi_1,\mathbf{W_1},\mathbf{b_1}} : \mathbb{R}^{n_{(0)}} \to \mathbb{R}^{n_{(1)}}$ and the last layer is a function $f^{(L)}_{\phi_L,\mathbf{W_L},\mathbf{b_L}} : \mathbb{R}^{n_{(L-1)}} \to \mathbb{R}^{n_{(L)}}$. Suppose that for $1 \leq k \leq L$, the k-th layer is a function $f^{(k)}_{\phi_k,\mathbf{W_k},\mathbf{b_k}} : \mathbb{R}^{n_{(k-1)}} \to \mathbb{R}^{n_{(k)}}$. Now, a **fully-connected neural network** is a composite function parameterized by L layer weight matrices $\{\mathbf{W_k}\}_{k=1}^{L}$, L layer bias vectors $\{\mathbf{b_k}\}_{k=1}^{L}$ and L choices of layer activation functions $\{\phi_k\}_{k=1}^{L}$ of the form

$$f = f^{(L)}_{\phi_L,\mathbf{W_L},\mathbf{b_L}} \circ f^{(L-1)}_{\phi_{(L-1)},\mathbf{W_{(L-1)}},\mathbf{b_{(L-1)}}} \circ \ldots \circ f^{(1)}_{\phi_1,\mathbf{W_1},\mathbf{b_1}}.$$

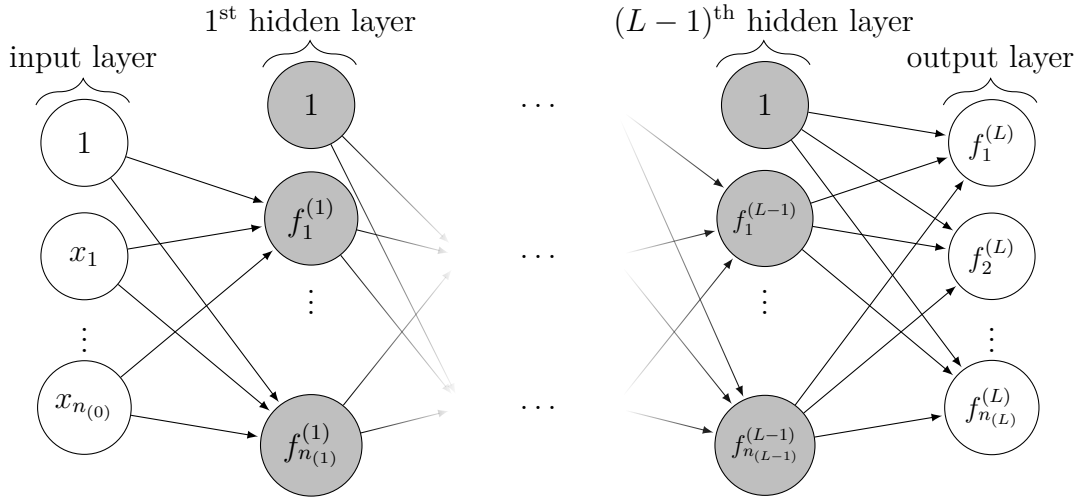


Figure 1.4: Network graph of a L-layer fully-connected neural network with $n_{(0)}$-dimensional input and $n_{(L)}$-dimensional output. This illustration corresponds to Definition 12. The k^{th} hidden layer contains $n_{(k)}$ neurons for $1 \leq k \leq L$. For the sake of clarity, parameters $\mathbf{W_k}, \mathbf{b_k}$ and activation functions ϕ_k are omitted from the graph. Hence, $f^{(k)}_j$ in the graph represents $(f_{\phi_k,\mathbf{W_k},\mathbf{b_k}})_j$, in the sense of Definition 10.

In literature, the layers $f_{\phi_k,\mathbf{W_k},\mathbf{b_k}} : \mathbb{R}^{n_{(k-1)}} \to \mathbb{R}^{n_k}$, for $1 \leq k \leq L$ are often called **hidden or latent layers**. To analyse interactions between layers, we will often decompose $f_{\phi^{(k)},\mathbf{W}^{(k)},\mathbf{b}^{(k)}}$ into the computation of the affine transformation and the application of the activation function,

$$f_{\phi^{(k)},\mathbf{W}^{(k)},\mathbf{b}^{(k)}}(\mathbf{x}) = \phi^{(k)}(\mathbf{a}^{(k)}(\mathbf{x})) \text{ where } \mathbf{a}^{(k)}(\mathbf{x}) = (\mathbf{W}^{(k)})^\top \mathbf{x} + \mathbf{b}^{(k)}.$$

1.4 Forward pass

Let $\mathbf{f} : \mathbb{R}^{n_{(0)}} \to \mathbb{R}^{n_{(L)}}$ be an L-layer fully-connected neural network parameterized by L weight matrices $\{\boldsymbol{W}^{(l)}\}_{l=1}^{L}$, bias vectors $\{\boldsymbol{b}^{(l)}\}_{l=1}^{L}$ and choices of activation functions $\{\sigma^{(l)}\}_{l=1}^{L}$. To evaluate predictions of the neural network $\mathbf{f}$ on input $\mathbf{x} \in \mathbb{R}^n$, we set $\mathbf{f}^{(0)} = \mathbf{x}$ and repeatedly evaluate following equations

$$\mathbf{a}^{(l)} = (\mathbf{W}^{(l)})^\top \mathbf{f}^{(l-1)} + \mathbf{b}^{(l)}, \text{ for } 1 \leq l \leq L, \tag{1.1a}$$

$$\mathbf{f}^{(l)} = \sigma^{(l)}(\mathbf{a}^{(l)}), \text{ for } 1 \leq l \leq L. \tag{1.1b}$$

Equations 1.1a and 1.1b are componentwise equivalent to

$$a_i^{(l)} = \sum_{k=1}^{n_{(l-1)}} w_{ki}^{(l)} f_k^{(l-1)} + b_i^{(l)}, \text{ for } 1 \leq i \leq n^{(l)}, \tag{1.2a}$$

$$f_i^{(l)} = \sigma^{(l)}(a_i^{(l)}), \text{ for } 1 \leq i \leq n^{(l)}. \tag{1.2b}$$

The process of evaluating $\mathbf{f}$ on input $\mathbf{x}$ is known as the forward pass.

1.5 Gradient descent and backpropagation algorithm

One of the most important ideas in machine learning is expressing the learning problem as an optimization problem. In this thesis, we will discuss only the supervised learning approach. In the supervised setting, we define a **loss function**, which is a function measuring the error between predictions of the model given the input and correct label corresponding to the input on which the model was evaluated. Assuming fixed neural network topology and a fixed choice of activation functions per layer, in the language of subsection 1.2.1, neural networks are a family of hypothesis functions, parameterized by weights and biases. The purpose of the backpropagation learning algorithm is to find the configuration of weights and biases corresponding to the minimal loss on the training set, using the method of gradient descent. Gradient descent is a first-order, iterative optimization algorithm based on the following fact from analysis in $\mathbb{R}^n$.

Claim. *If the multivariable differentiable function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}$ *is defined and differentiable in a neighbourhood of* $\mathbf{a} \in \mathbb{R}^n$*, then* $\mathbf{f}$ *decreases fastest at* $\mathbf{a}$ *in direction* $-\nabla f(\mathbf{a})$ *from* $\mathbf{a}$.

In gradient descent, the claim above is applied to the loss function. In this context, the loss function is a function of parameters of the neural network. The loss function is evaluated on the entire training set or its subset. Those parameters are L layer weight matrices $\{\mathbf{W}_l\}_{l=1}^{L}$ and L layer bias vectors $\{\mathbf{b}_l\}_{l=1}^{L}$. Gradient descent iteratively refines the current configuration of weights and biases by moving in the direction in the parameter space suggested by the claim above. This heuristic is expressed as equations 1.3a and 1.3b, discussed on the next page.

In our context, gradient descent is implemented by the following iterative process, repeated for a fixed number of steps or until convergence,

$$w_{ij}^{(l)}[t+1] = w_{ij}^{(l)}[t] - \eta \frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}[t]} \tag{1.3a}$$

$$b_{j}^{(l)}[t+1] = b_{j}^{(l)}[t] - \eta \frac{\partial \mathcal{L}}{\partial b_{j}^{(l)}[t]}, \tag{1.3b}$$

where $\eta > 0$ is a small constant known as **the learning rate**. The index $[t]$ indicates the current training step. Equations 1.3a and 1.3b can be expressed in the following vectorised form.

$$\mathbf{W}^{(l)}[t+1] = \mathbf{W}^{(l)}[t] - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}[t]}, \tag{1.4a}$$

$$\mathbf{b}^{(l)}[t+1] = \mathbf{b}^{(l)}[t] - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}[t]}. \tag{1.4b}$$

There are indeed many ways to initialize weights and biases. The initialization of weights and biases may significantly affect the gradient descent performance, especially in the case of sigmoidal activation functions. An example of a problem related to the initialization of weights and biases is the problem of *vanishing gradients*. We will briefly discuss initialization methods presented in [GB] and [He+15b]. However, many different methods exist and are used in practice. For sigmoidal and symmetric activation functions, a common initialization method is *Xavier/Glorot initialization* [GB],

$$w_{ij}^{(l)}[0] \sim \mathcal{U}\left(-\sqrt{\frac{6}{n_{(l)} + n_{(l-1)}}}, \sqrt{\frac{6}{n_{(l)} + n_{(l-1)}}}\right).$$

When it comes to rectified activation functions, more prevalent initialization method is *He initialization* [He+15b],

$$w_{ij}^{(l)}[0] \sim \mathcal{N}\left(0, \sqrt{\frac{2}{n_{(l-1)}}}\right).$$

Biases are often zero-initialized. Due to the structure of feed-forward neural networks and nonlinearity induced by activation functions, minimizing the training set loss is often not a convex optimization problem. Hence the gradient-based optimization presented above may not converge to a global minimum (if it exists) or even converge at all. However, this method works surprisingly well in practice and it recently became a topic of active research. For instance, [Li+18] has recently provided a method to visualize loss surfaces of modern neural networks. Using those visualizations, the paper provided a possible explanation of the trainability of modern deep networks. Until recently, it was thought that gradient-based optimization may struggle with local optima and saddle points. Although this is theoretically possible, it seems that optimizers work well in practice.

Many of the latest papers often use more sophisticated gradient-based optimizers. Those optimizers are designed to mitigate common problems related to gradient-based optimization, such as the performance on plateaus. Such optimizers are Adam, Nadam, Nesterov Accelerated Gradient, Adagrad, Adadelta, and RMSProp. For a comprehensive overview of those methods see [Rud17].

1.5.1 Stochastic and batch gradient descent

Since the training set can often be quite large, it quickly becomes computationally infeasible to compute gradients $\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}[t]}, \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}[t]}$. This problem arises if the loss function $\mathcal{L}$ is evaluated on the entire training set. A very common solution to this problem is to use stochastic gradient descent or batch gradient descent. Batch gradient descent procedure divides the training of the network into epochs. In each epoch, a small sample is sampled from the training set. This sampled subset of the training set is called a training **batch**. The loss function is evaluated with respect to the batch and equations 1.4a, 1.4b are also computed with respect to the batch. Stochastic gradient descent is a name for batch gradient descent when a training batch consists of precisely one training example.

1.5.2 Loss functions

To discuss loss functions, let $\mathbf{f} : \mathbb{R}^{n_{(0)}} \to \mathbb{R}^{n_{(L)}}$ be an L-layer fully-connected neural network parameterized by L weight matrices $\{\boldsymbol{W}^{(l)}\}_{l=1}^{L}$, bias vectors $\{\boldsymbol{b}^{(l)}\}_{l=1}^{L}$ and choices of activation functions $\{\sigma^{(l)}\}_{l=1}^{L}$. Suppose $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^{n}$ is the training set. The choice of the loss function depends on the type of a problem that the network $\mathbf{f}$ is designed to solve. The mean square error is widely-used for regression problems. In this setting, the output layer often has identity activation.

Definition 13 (mean square error loss)**.** The square loss corresponding to the single training item $(\mathbf{x}_k, \mathbf{y}_k)$ is given by

$$\mathcal{L}_k(\{\mathbf{W}^{(l)}\}_{l=1}^{L}, \{\mathbf{b}^{(l)}\}_{l=1}^{L}) = \frac{1}{2} \left\|\mathbf{f}(\mathbf{x}_k) - \mathbf{y_k}\right\|_2^2 = \frac{1}{2} \sum_{j=1}^{n_{(L)}} (f_j^{(L)}(\mathbf{x}_k) - y_{k_j})^2.$$

The mean square loss of the training set is given by

$$\begin{aligned}
\mathcal{L}(\{\mathbf{W}^{(l)}\}_{l=1}^{L}, \{\mathbf{b}^{(l)}\}_{l=1}^{L}) &= \frac{1}{n} \sum_{k=1}^{n} \mathcal{L}_k(\{\mathbf{W}^{(l)}\}_{l=1}^{L}, \{\mathbf{b}^{(l)}\}_{l=1}^{L}) \\
&= \frac{1}{2n} \sum_{k=1}^{n} \left\|\mathbf{f}(\mathbf{x}_k) - \mathbf{y_k}\right\|_2^2 \\
&= \frac{1}{2n} \sum_{k=1}^{n} \sum_{j=1}^{n_{(L)}} (f_j^{(L)}(\mathbf{x}_k) - y_{k_j})^2.
\end{aligned}$$

We will briefly discuss neural networks for multi-class classification. In this setting, we usually design a neural network such that its output layer width matches the number of different classes. Hence, each output layer neuron represents a class. Although it is possible to use the mean square error loss to train neural networks for classification, it is often a better idea to use the categorical cross-entropy loss. Categorical cross-entropy loss is a loss function designed for classification problems. In classification setting, this loss is theoretically superior to the mean squared error. For instance, it is tightly connected to Kullback-Leibler divergence between true class distribution and the class distribution the model is designed to learn. The cross-entropy loss is particularly effective when the last layer of the neural network implements *softmax*[Bri] activation function.

Definition 14 (categorical cross-entropy loss)**.** The cross entropy loss corresponding to $(\mathbf{x}_k, \mathbf{y}_k)$ is given by

$$\mathcal{L}_k(\{\mathbf{W}^{(l)}\}_{l=1}^L, \{\mathbf{b}^{(l)}\}_{l=1}^L) = -\sum_{j=1}^{n_{(L)}} y_{k_j} \ln f_j^{(L)}(\mathbf{x}_k).$$

The cross entropy loss of the training set is given by

$$\begin{aligned}\mathcal{L}(\{\mathbf{W}^{(l)}\}_{l=1}^L, \{\mathbf{b}^{(l)}\}_{l=1}^L) &= \sum_{k=1}^{n} \mathcal{L}_k(\{\mathbf{W}^{(l)}\}_{l=1}^L, \{\mathbf{b}^{(l)}\}_{l=1}^L) \\ &= \sum_{k=1}^{n} \sum_{j=1}^{n_{(L)}} y_{k_j} \ln f_j^{(L)}(\mathbf{x}_k).\end{aligned}$$

1.6 Backward pass

In this section, we will show how to systematically compute gradients 1.4a and 1.4b of any differentiable loss function for any fully-connected neural network. For the sake of brevity, we will denoted the lth layer of a neural network by $\mathbf{f}^{(l)}$.

Proposition 1 (Backpropagation equations)**.** *Let $\mathbf{f} : \mathbb{R}^{n_{(0)}} \to \mathbb{R}^{n_{(L)}}$ be a differentiable, fully-connected neural network of L layers, parameterized by L weight matrices $\{\boldsymbol{W}^{(l)}\}_{l=1}^L$, bias vectors $\{\boldsymbol{b}^{(l)}\}_{l=1}^L$ and choices of activation functions $\{\sigma^l\}_{l=1}^L$. Let $\mathcal{L}$ be a differentiable loss function for a single example $(\mathbf{x}, \mathbf{y})$, so $\mathbf{f}^{(0)} = \mathbf{x}$. Then*

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \delta_j^l \cdot f_i^{(l-1)}, \text{ for every } 1 \le i \le n_{(l-1)}, 1 \le j \le n_{(l)}, \tag{1.5a}$$

$$\frac{\partial \mathcal{L}}{\partial b_j^{(l)}} = \delta_j^l, \text{ for every } 1 \le j \le n_{(l)}, 1 \le l \le L, \tag{1.5b}$$

$$\delta_i^{(l-1)} = \left(\sum_{j=1}^{n_{(l)}} \delta_j^{(l)} w_{ij}^{(l)}\right) \cdot \left(\sigma^{(l-1)}\right)'(a_i^{(l-1)}), \text{ for } 1 \le i \le n_{(l-1)}, 1 \le l \le L, \tag{1.5c}$$

$$\delta_i^{(L)} = \frac{\partial \mathcal{L}}{\partial f_i^{(L)}} \cdot \left(\sigma^{(L)}\right)'(a_i^{(L)}), \text{ for } 1 \le i \le n_{(L)}. \tag{1.5d}$$

Proof.

Step 1 (Computation of $\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}}, \frac{\partial \mathcal{L}}{\partial b_i^{(l)}}$). Let $\mathcal{L}$ be a loss function, as in the statement. We are interested in quantitites $\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}}$ and $\frac{\partial \mathcal{L}}{\partial b_i^{(l)}}$. Consider single weight $w_{ij}^{(l)}$. By Definition 12, the weight $w_{ij}^{(l)}$ connects ith neuron in $(l-1)$th layer to jth neuron in (l)th layer. By 1.2a, the weight $w_{ij}^{(l)}$ contributes only to $a_j^{(l)}$ and no other component of $\mathbf{a}^{(l)}$. Now consider a single bias component $b_j^{(l)}$. By 1.2a, the bias component $b_j^{(l)}$ contributes only to $a_j^{(l)}$ and no other component of $\mathbf{a}^{(l)}$. Thus, by the Chain Rule,

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} \text{ and } \frac{\partial \mathcal{L}}{\partial b_j^{(l)}} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial b_j^{(l)}}. \tag{1.6}$$

Consider $\frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}}$ and $\frac{\partial a_j^{(l)}}{\partial b_j^{(l)}}$. By 1.2a,

$$\frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} = \frac{\partial}{\partial w_{ij}^{(l)}} \left(\sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} f_k^{(l-1)} + b_j^{(l)} \right) = f_i^{(l-1)}, \tag{1.7}$$

$$\frac{\partial a_j^{(l)}}{\partial b_j^{(l)}} = \frac{\partial}{\partial b_j^{(l)}} \left(\sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} f_k^{(l-1)} + b_j^{(l)} \right) = 1. \tag{1.8}$$

Substituting 1.7 into 1.6 gives

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}} \cdot f_i^{(l-1)}, \text{ and } \frac{\partial \mathcal{L}}{\partial b_j^{(l)}} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}}. \tag{1.9}$$

For the sake of simplicity, set $\delta_j^{(l)} = \frac{\partial \mathcal{L}}{\partial a_j^{(l)}}$. Then by 1.9

$$\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}} = \delta_j^l \cdot f_i^{(l-1)} \text{ and } \frac{\partial \mathcal{L}}{\partial b_j^{(l)}} = \delta_j^l. \tag{1.10}$$

By 1.9, to compute $\frac{\partial \mathcal{L}}{\partial w_{ij}^{(l)}}, \frac{\partial \mathcal{L}}{\partial b_j^{(l)}}$, it remains to compute δ_j.

Step 2 (Computation of δ_j). We begin with δ_L. By the Chain Rule,

$$\delta_i^{(L)} = \frac{\partial \mathcal{L}}{\partial f_i^{(L)}} \cdot \frac{\partial f_i^{(L)}}{\partial a_i^{(L)}}, \text{ for } 1 \leq i \leq n_{(L)}.$$

Now consider $\frac{\partial f_i^{(L)}}{\partial a_i^{(L)}}$. Differentiating 1.2b yields

$$\frac{\partial f_i^{(L)}}{\partial a_i^{(L)}} = \frac{\partial}{\partial a_i^{(L)}} \left(\sigma^{(L)}(a_i^{(L)}) \right) = (\sigma^{(L)})'(a_i^{(L)}). \tag{1.11}$$

We will express $\delta_i^{(l-1)}$ in terms of $\delta_j^{(l)}$. By 1.2a and 1.2b, for every $1 \leq i \leq n_{(l-1)}$, $a_i^{(l-1)}$ contributes to the value of every $a_j^{(l)}$ via $f_i^{(l-1)}$ and only via $f_i^{(l-1)}$. This relationship can be easily seen in Figure 1.5. By the Chain Rule, for every $1 \leq l \leq L$, for every $1 \leq i \leq n_{(l-1)}$,

$$\begin{aligned}\delta_i^{(l-1)} = \frac{\partial \mathcal{L}}{\partial a_i^{(l-1)}} &= \sum_{j=1}^{n_{(l)}} \frac{\partial \mathcal{L}}{\partial a_j^{(l)}} \cdot \frac{\partial a_j^{(l)}}{\partial a_i^{(l-1)}} \\ &= \sum_{j=1}^{n_{(l)}} \delta_j^{(l)} \cdot \frac{\partial a_j^{(l)}}{\partial a_i^{(l-1)}}. \end{aligned} \tag{1.12}$$

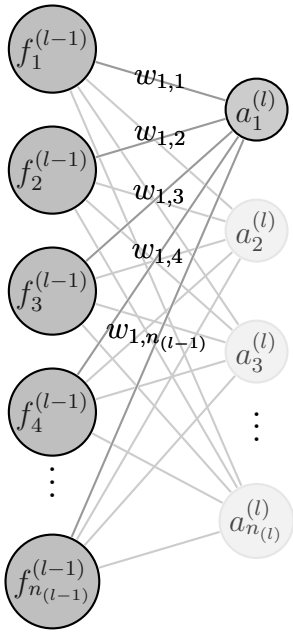


Figure 1.5: between two successive layers $f^{(l-1)}$ and $f^{(l)}$.

By 1.2a and 1.2b,

$$\begin{aligned} a_j^{(l)} &= \sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} f_k^{(l-1)} + b_j^{(l)} \\ &= \sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} \sigma^{(l-1)}(a_k^{(l-1)}) + b_j^{(l)}. \end{aligned} \tag{1.13}$$

Differentiating 1.13 with respect to $a_i^{(l)}$ gives

$$\begin{aligned} \frac{\partial a_j^{(l)}}{\partial a_i^{(l-1)}} &= \frac{\partial}{\partial a_i^{(l-1)}} \left(\sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} f_k^{(l-1)} + b_j^{(l)} \right) \\ &= \frac{\partial}{\partial a_i^{(l-1)}} \left(\sum_{k=1}^{n_{(l-1)}} w_{kj}^{(l)} \sigma^{(l-1)}(a_k^{(l-1)}) + b_j^{(l)} \right) \\ &= w_{ij}^{(l)} \cdot \left(\sigma^{(l-1)}\right)'(a_i^{(l-1)}). \end{aligned} \tag{1.14}$$

Substituting 1.14 into 1.12 gives

$$\delta_i^{(l-1)} = \sum_{j=1}^{n_{(l)}} \delta_j^{(l)} w_{ij}^{(l)} \left(\sigma^{(l-1)}\right)'(a_i^{(l-1)}) = \left(\sum_{j=1}^{n_{(l)}} \delta_j^{(l)} w_{ij}^{(l)} \right) \cdot \left(\sigma^{(l-1)}\right)'(a_i^{(l-1)}).$$

■

In practice, the standard way of implementing equations from Backpropagation equations is an implementation in the vectorized form. Two main advantages of vectorization are readability and performance. Recently, many deep learning programming frameworks started implementing a wide variety of optimized linear algebra routines [Li+20], often supporting efficient execution on GPUs and TPUs. Given the sheer size of modern datasets, vectorization is key to the computationally feasible implementation of both forward and backward passes.

Corollary 1 (Vectorised backpropagation equations)**.**

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(l)}} = \mathbf{f}^{(l-1)} \delta^{(l)\top} \tag{1.15a}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(l)}} = \delta^{(l)} \ \textit{where} \tag{1.15b}$$

$$\delta^{(l-1)} = \left(\mathbf{W}^{(l)} \delta^{(l)}\right) \odot \left(\sigma^{(l-1)}\right)'(\mathbf{a}_i^{(l-1)}) \ , \textit{for } 1 \leq l \leq L, \tag{1.15c}$$

$$\delta^{(L)} = \frac{\partial \mathcal{L}}{\partial \mathbf{f}^{(L)}} \odot \left(\sigma^{(L)}\right)'(\mathbf{a}^{(L)}). \tag{1.15d}$$

Remark 7. The symbol $\odot$ denotes the elementwise product, known as also Hadamard product.

Proof. The claimed equations follow directly from Backpropagation equations. Writing 1.5a in terms of matrix outer product yields 1.15a. Vectorizing 1.5b yields 1.15b. Recognising the inner product expansion $\sum_{j=1}^{n_{(l)}} \delta_j^{(l)} w_{ij}^{(l)}$ in 1.5c and applying the definition of matrix-vector multiplication $\mathbf{W}^{(l)} \delta^{(l)}$ yields 1.15c. Vectorizing 1.5d yields 1.15d. ■

Remark 8. It is worth discussing the computational complexity of Backpropagation equations. Recall that the forward pass equations are

$$\begin{aligned}
\mathbf{a}^{(l)} &= (\mathbf{W}^{(l)})^\top \mathbf{f}^{(l-1)} + \mathbf{b}^{(l)}, \ \text{for } 1 \leq l \leq L, \\
\mathbf{f}^{(l)} &= \sigma^{(l)}(\mathbf{a}^{(l)}), \ \text{for } 1 \leq l \leq L, \\
\mathbf{f}^{(0)} &= \mathbf{x}.
\end{aligned}$$

By comparing those equations to 1.15a, 1.15b, 1.15d, 1.15c, it is not difficult to observe that the time complexity of gradient computation is asymptotically equivalent to the time complexity of forward pass. This implies that backpropagation training is as computationally efficient as evaluating neural network predictions. The computational efficiency of backpropagation algorithm is a very important factor in widespread use of neural networks.

Remark 9. We will briefly discuss the importance of the differentiability of the loss function and the neural network. Loss functions are often differentiable, but many modern neural network architectures use rectified activation functions that are not differentiable. This implies that neural networks themselves are not differentiable so the Backpropagation equations simply do not hold. However, the gradient descent algorithm can be generalized to the (sub)gradient descent algorithm.

Remark 10. So far, we have discussed only the theoretical foundations for training fully-connected neural networks. In practice, there is a wide range of software frameworks designed to simplify the design of neural networks and to automate gradient computations we demonstrated in Backpropagation equations. Apart from just simplifying gradient computations, modern deep learning software frameworks perform very complex optimizations. The most popular frameworks are Tensorflow[Mar+15], PyTorch [Pas+19] and Jax[Bra+18]. We will use PyTorch in Experiments.

Chapter 2

Literature review

Since the invention of the backpropagation algorithm in "Learning representations by back-propagating errors" [RHW86], neural networks have been applied to a wide variety of problems in pattern recognition and machine learning.

> "The importance of neural networks in this context is that they offer a very powerful and very general framework for representing non-linear mappings from several input variables to several output variables, where the form of the mapping is governed by a number of adjustable parameters." (*Neural networks and machine learning* [Bis98], p.5)

Recently, very sophisticated neural networks began to significantly outperform alternative machine learning methods in a wide variety of tasks, including natural language processing ([Vas+17], [Bro+20]), computer vision ([Sze+14], [He+15a], [BWL20]), computational biology ([Sen+20]) and reinforcement learning ([Sil+17]). Interestingly, some of the most influential researchers in the field received a Turing award for their work in deep learning. However, most of the advancements in the field came from clever architectures, massive datasets, and experimental verification. Given such an experimental success, some research was devoted to demystification of mathematical capabilities of neural networks, even before the most of breakthroughs mentioned above. It turns out that the activation function plays an important role in the expressive power of neural networks. A simple observation is a fact if there was no activation function, neural networks would become nothing but a linear transformation of the input data. This follows from the fact that a composition of linear transformations is a linear transformation and each layer would be a linear transformation of the previous layer. Following [Bis98]'s terminology above, the key adjustable parameter governing the non-linearity is the choice of the activation function. When it comes to the approximation theory of neural networks, very important research problems are the following two questions.

- Under which necessary and sufficient conditions does the particular family of neural networks have the power to approximate any continuous function, possibly with a compact domain, given the desired approximation accuracy?
- Under which necessary and sufficient conditions does the particular family of neural networks have the power to approximate any Lebesgue p-integrable function, given the desired approximation accuracy?

The results addressing questions posed above are known as universal approximation theorems. This chapter contains a summary of conducted research in the last 35 years. Theorems are grouped based on the neural network topology and the relevant function space. There are two main groups of results based on the neural network topology.

Unbounded width, bounded depth This setup usually studies neural networks with a single hidden layer, imposing no bounds on the number of neurons in the hidden layer.

Bounded width, arbitrary depth This setup usually studies neural networks with several hidden layers, imposing bounds on the number of neurons in each hidden layer. The activation function usually remains the same for different layers.

Remark 11. In the previous chapter, we introduced the notation for neural network terminology. The introduced notation will be the main notation used in this thesis. However, the results cited from papers use the convention from the respective paper, so the notation in this chapter can be slightly different.

2.1 Unbounded width, bounded depth

Studies related to the approximation power of neural networks started with single-layer feed-forward neural networks. A possible explanation is the lack of computational power necessary to train deeper networks and the lack of massive datasets omnipresent today. This setting imposed no bounds on the hidden layer width. The classical result addressing that setup is Theorem 2 presented in [Cyb89], published by G.Cybenko in 1989.

Definition 15. Let $\sigma : \mathbb{R} \to \mathbb{R}$. We say that σ is sigmoidal if $\lim_{x\to\infty} \sigma(x) = 1$ and $\lim_{x\to-\infty} \sigma(x) = 0$.

Theorem 1 (Cybenko's Universal Approximation Theorem for Unbounded-Width Networks, Theorem 2 in [Cyb89])**.** *Let σ be any continuous sigmoidal function. Then the finite sums of the form $G(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle + \theta_j)$ are dense in $\mathcal{C}([0,1]^n)$.*

This result will be thoroughly discussed in Universal approximation of continuous functions via Cybenko's method. In 1991, K. Hornik proved that sigmoidal assumption can be dropped.

Theorem 2 (Hornik's Universal Approximation Theorem for Unbounded-Width Networks, Theorem 2 in [Hor91])**.** *If σ is continuous, bounded and nonconstant, then the finite sums of the form $G(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle + \theta_j)$ are dense in $\mathcal{C}(X)$ for all compact subsets X of $\mathbb{R}^n$.*

It turns out that both boundedness and continuity assumption can be dropped. Perhaps the most general result regarding single-layer networks and continuous functions is the following theorem, presented in [Les+93] from 1993. Before stating the theorem, we will introduce relevant notation used in the paper.

Definition 16 (the space $\mathcal{L}^\infty_{\text{loc}}(\Omega)$)**.** Let Ω be a domain in $\mathbb{R}^n$. A function $f : \Omega \to \mathbb{R}$, defined almost everywhere with respect to Lebesgue measure on Ω is locally essentially bounded on Ω if for every compact set $K \subset \Omega$, $f \in \mathcal{L}^\infty(\Omega)$. We denote the space of locally essentially bounded functions on Ω by $\mathcal{L}^\infty_{\text{loc}}(\Omega)$.

Definition 17 (the space $\mathcal{M}(\Omega)$)**.** Let Ω be a domain in $\mathbb{R}^n$. The space $\mathcal{M}(\Omega)$ is a subset of $\mathcal{L}^\infty_{\text{loc}}(\Omega)$ consisting of functions whose closure of the set of points of discontinuity is a set of Lebesgue measure zero.

Remark 12. We set $\mathcal{M} = \mathcal{M}(\mathbb{R})$.

Theorem 3 (Universal Approximation Theorem for Unbounded-Width Networks, Theorem 1 in [Les+93])**.** *Let $\sigma \in \mathcal{M}$. Set*

$$\Sigma_n = \text{span}\{\mathbf{x} \to \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + \theta) : \mathbf{w} \in \mathbb{R}^n, \theta \in \mathbb{R}\}.$$

Then Σ_n is dense in $\mathcal{C}(\mathbb{R}^n)$ if and only if σ is not an algebraic polynomial almost everywhere.

Interestingly, all proofs of theorems in this setup are non-constructive, offering no insight into the necessary number of neurons to achieve the desired approximation accuracy.

2.2 Bounded width, arbitrary depth

Given the practical significance and convenient algebraic properties of ReLU, a lot of recent research focused on the multi-layer, width-bounded neural networks with ReLU activations. Contrary to the proofs of approximation theorems from the previous section, proofs related to multi-layer networks with ReLU activations are often constructive. Moreover, those proofs often provide explicit lower and upper bounds on layer widths which are necessary and sufficient for an approximation of the desired accuracy. However, the advantage of constructive proofs relies on very technical constructions, such as the Register Model (Proposition 4.2) from [KL20]. The first such a result is Theorem 1, published in [Lu+] from 2017.

Theorem 4 (Universal Approximation Theorem for Width-Bounded ReLU in L^p, Theorem 1 in [Lu+])**.** *For any Lebesgue-integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and any $\epsilon > 0$, there exists a fully-connected neural network with ReLU activation function, denoted by $\mathcal{A}$, with width $d_m \leq n + 4$, such that the function $F_\mathcal{A}$ represented by this network satisfies $\|f - F_\mathcal{A}\|_1 < \epsilon$.*

The following recent result from published in [Par+20] generalizes Theorem 1 in [Lu+] to a wider class of $\mathcal{L}^p$ spaces. Moreover, it characterizes the universal approximation in terms of the input dimension d_x and the output dimension d_y.

Theorem 5 (Universal Approximation Theorem for Width-Bounded ReLU in L^p, Theorem 1 in [Par+20])**.** *For any $p \in [1, \infty)$, ReLU networks of width w are dense in $L^p(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ if and only if $w \geq \max\{d_x + 1, d_y\}$.*

It turns out that a very similar result holds for $\mathcal{C}(K, \mathbb{R}^{d_y})$, for a compact set $K \subset \mathbb{R}^{d_x}$.

Theorem 6 (Universal Approximation Theorem for Width-Bounded ReLU + Step networks in L^p, Theorem 3 in [Par+20])**.** *ReLU + Step networks of width w are dense in $\mathcal{C}(K, \mathbb{R}^{d_y})$ if and only if $w \geq \max\{d_x + 1, d_y\}$, for every compact set $K \subset \mathbb{R}^{d_x}$.*

When it comes to continuous functions on compact sets, one of the most recent results is Theorem 3.2 in [KL20]. Before discussing this result, we will introduce some notation necessary for its statement.

Definition 18 ($\mathcal{NN}^p_{n,m,k}$, Definition 3.1 in [KL20])**.** Let $\rho : \mathbb{R} \to \mathbb{R}$ and $n, m, k \in \mathbb{N}$. Then let $\mathcal{NN}^p_{n,m,k}$ represent the class of functions $\mathbb{R}^n \to \mathbb{R}^m$ described by feedforward neural networks with n neurons in the input layer, m neurons in the output layer, and an arbitrary number of hidden layers, each with k neurons with activation function ρ. Every neuron in the output layer has the identity activation function.

Theorem 7 (Universal Approximation Theorem for Width-Bounded networks in $\mathcal{C}(X)$, Theorem 3.2 in [KL20])**.** *Let $\rho : \mathbb{R} \to \mathbb{R}$ be any nonaffine continuous function which is continuously differentiable at at least one point, with nonzero derivative at that point. Let $K \subseteq \mathbb{R}^n$ be compact. Then $\mathcal{NN}^\rho_{n,m,n+m+2}$ is dense in $\mathcal{C}(K; \mathbb{R}^m)$ with respect to the uniform norm.*

Proposition 2 (Proposition 4.11 in [KL20])**.** *Let $\rho : \mathbb{R} \to \mathbb{R}$ be any nonaffine polynomial. Let $K \subseteq \mathbb{R}^n$ be compact. Then $\mathcal{NN}^\rho_{n,m,n+m+2}$ is dense in $\mathcal{C}(K; \mathbb{R}^m)$ with respect to the uniform norm.*

In this thesis, the focus is on single-layer, feedforward neural networks. The purpose of this thesis is to present a variety of universal approximation theorems, based on Cybenko's argument published in [Cyb89].
We will begin with a simple proof of the universal approximation theorem for exp activation function and the space of continuous functions on a compact subset of $\mathbb{R}^n$, with respect to the uniform norm. The proof will be based on Stone-Weierstrass Theorem. Using Cybenko's method, we will generalize this result to a wider class of practically relevant activation functions, namely sigmoidal activation functions. The function space will be the space of continuous functions on the unit hypercube. The proof will be based on the relationship of (non)density in the normed linear space and its dual space. After this discussion, we will consider a different function space. More precisely, we will generalize previous results to the space of Lebesgue square-integrable and integrable functions on compact sets. We will conclude with a generalization to measurable functions, not necessarily of compact support, and we will study the universality in probabilistic sense. This setup will be thoroughly discussed, combining arguments from [Cyb89] and [HSW89].
Since the proofs of the results we are about to establish are non-constructive, we will conclude with an experimental study of the relationship between established theoretical results and practical applications. The practical application we will consider is the classification on *Fashion MNIST*.

Chapter 3

Universality of Neural Networks

3.1 Introduction

To discuss the approximation power of neural networks, it is necessary to set up a theoretical framework that enables the quantification of approximation accuracy. Since neural networks are a class of functions, a sensible and common approach is to consider various function spaces and discuss the approximation within the context of a given metric. We begin with a definition of the universal approximator family.

Definition 19. Let $(\mathcal{X}, \delta)$ be a metric space. Let $\mathcal{H} \subseteq \mathcal{X}$. We say that $\mathcal{H}$ is an universal approximator family or simply that $\mathcal{H}$ is universal for the space $(\mathcal{X}, \delta)$ if $\mathcal{H}$ is δ-dense in $\mathcal{X}$. Equivalently, in ϵ language, $\mathcal{H}$ is universal in $\mathcal{X}$ if

$$\forall x \in \mathcal{X}, \forall \epsilon > 0, \exists h \in \mathcal{H} \text{ such that } \delta(x, h) < \epsilon.$$

In this thesis, we will often set $\mathcal{X}$ to be some function space and δ to be some metric on the function space $\mathcal{X}$. We will also set $\mathcal{H}$ to be some family of neural networks, often parametrized by the activation function.

Example 1. We will denote the family of single-layer fully-connected neural networks with activation function σ by $\mathcal{H}_\sigma$, where

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Example 2. A classic example of a metric space with its universal approximator family is the space of real-valued continuous functions on $[0, 1]$, denoted by $\mathcal{C}([0, 1])$ and $\mathcal{H}$ as a family of Bernstein polynomials. This is discussed in detail in Bernstein Approximation Theorem. This family and space play an essential role in the proof of Stone-Weierstrass Theorem, which will be used to prove that $\mathcal{H}_{\exp}$ is dense in $\mathcal{C}(K)$, where K is a compact set in $\mathbb{R}^n$.

The key results in subsequent sections will be the universal approximation theorems - results about the universality of $\mathcal{H}_\sigma$ under various conditions in the following function spaces.

Example 3. Let K be a compact set in $\mathbb{R}^n$, where $\mathbb{R}^n$ is equipped with the standard topology. Let $\mathcal{X} = \mathcal{C}(K)$, equipped with the sup metric δ_∞

$$\delta_\infty(f, g) = \sup_{\mathbf{x}\in K}\{|f(\mathbf{x}) - g(\mathbf{x})|\}.$$

This metric arises from the sup norm $\|f\|_\infty = \sup_{\mathbf{x}\in K}\{|f(\mathbf{x})|\}$. The proof δ_∞ is indeed a metric and $\|.\|_\infty$ is a norm is an Example 10.6 in [Wad14].

Example 4. Let K be a compact set in $\mathbb{R}^n$, where $\mathbb{R}^n$ is equipped with the standard topology. Let $1 \leq p < \infty$. Let $\mathcal{X} = \mathcal{L}^p(K)$, where $\mathcal{L}^p(K)$ consists of equivalence classes of Borel measurable functions $f : K \to \mathbb{R}$ such that

$$\int_K |f(\mathbf{x})|^p \, d\mathbf{x} < \infty,$$

where two measurable functions are equivalent if they are equal almost everywhere, with respect to Lebesgue measure on $\mathbb{R}^n$. For the sake of simplicity, we will simply refer to representative functions of those equivalence classes instead of equivalence classes themselves. The space $\mathcal{L}^p(K)$ is equipped with a norm

$$\|f\|_p = \left(\int_K |f(\mathbf{x})|^p \, d\mathbf{x}\right)^{\frac{1}{p}}.$$

The fact $\|\cdot\|_p$ is indeed a norm is a consequence of Minkowski Inequality and Proposition 16.

Example 5. The last and the most general function space we will consider is the space $\mathcal{M}^n$, consisting of equivalence classes of Borel-measurable functions $f : \mathbb{R}^n \to \mathbb{R}$ where two measurable functions are equivalent if they are equal almost everywhere, with respect to a probability measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. For the sake of simplicity, we will simply refer to representative functions of those equivalence classes instead of equivalence classes themselves.

In this chapter, we will discuss the universal approximation capabilities of neural network family $\mathcal{H}_\sigma$ in various contexts. Those results are grouped in sections briefly discussed below.

Universal approximation of continuous functions via Stone-Weierstrass
In this section, we will focus on the normed linear space $\mathcal{C}(K)$, where K is a compact set in $\mathbb{R}^n$. We will show that the family $\mathcal{H}_{\text{exp}}$ is universal in $\mathcal{C}(K)$ using the Stone-Weierstrass Theorem.

Universal approximation of continuous functions via Cybenko's method
In this section, we will focus on the normed linear space $\mathcal{C}([0,1]^n)$, where $[0,1]^n$ denotes the unit hypercube in $\mathbb{R}^n$. This section is a generalization of the previous section to more practically relevant activation functions such as the logistic sigmoid. We will introduce an idea of discriminatory activation functions and use it to show that the family $\mathcal{H}_\sigma$ is universal in $\mathcal{C}([0,1]^n)$, where σ is a continuous sigmoidal activation function. Although

this result is a generalization of the result from the previous section, the argument will be significantly different and more sophisticated, based on paper [Cyb89].

Universal approximation of square-integrable functions In this section, we will focus on the normed linear space $\mathcal{L}^2([0,1]^n)$. We will introduce the idea of $\mathcal{L}^2([0,1]^n)$ discriminatory activation function and use it to show that $\mathcal{H}_\sigma$ is universal in $\mathcal{L}^2([0,1]^n)$, where σ is a continuous sigmoidal activation function. We will use an argument very similar to the Cybenko's method used to establish that $\mathcal{H}_\sigma$ is universal in $\mathcal{C}([0,1]^n)$. We will show that the most common $\mathcal{C}([0,1]^n)$-discriminatory activation functions remain $\mathcal{L}^2([0,1]^n)$-discriminatory.

Universal approximation of integrable functions In this section, we will focus on the normed linear space $\mathcal{L}^1([0,1]^n)$. We will introduce the idea of $\mathcal{L}^1([0,1]^n)$-discriminatory activation function and use it to show that $\mathcal{H}_\sigma$ is universal in $\mathcal{L}^2([0,1]^n)$, where σ is a $\mathcal{L}^1([0,1]^n)$-discriminatory activation function. We will use an argument very similar to one used to establish that $\mathcal{H}_\sigma$ is universal in $\mathcal{L}^2([0,1]^n)$.

Universal approximation of measurable functions on compact sets In this section, we will consider real-valued, Borel measurable functions with a compact domain in $\mathbb{R}^n$. We will show that σ remains universal in this setting, under slightly weakened conditions. The argument will be an application of the fact $\mathcal{H}_\sigma$ is universal in $\mathcal{C}([0,1]^n)$, where σ is a continuous sigmoidal activation function.

Universal approximation of measurable functions in probabilistic sense In this section, we will focus on the space $\mathcal{M}^n$. We will begin by introducing metrics on $\mathcal{M}^n$ which are equivalent to convergence in probability measure μ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. We will develop theoretical results which will enable us to apply the fact $\mathcal{H}_\sigma$ is universal in $\mathcal{C}([0,1]^n)$, where σ is a continuous sigmoidal activation function. The key result in this section will be a proof of The Probabilistic Universal Approximation Theorem.

3.2 Universal approximation of continuous functions via Stone-Weierstrass

In this section, we will focus on the space $\mathcal{C}(K)$, where $K \subset \mathbb{R}^n$ is a compact set. We aim to show that the family $\mathcal{H}_{\exp}$ is dense in $\mathcal{C}(K)$. When it comes to the approximation of continuous functions on compact sets, it is natural to consider one of the famous theorems addressing this issue - Stone-Weierstrass Theorem. Before discussing the proof of The Universal Approximation Theorem for exp activation, we will present the Stone-Weierstrass Theorem and concepts required for its application.

Theorem (Stone-Weierstrass Theorem)**.** *Suppose that X is a compact metric space. If $\mathcal{A}$ is an algebra in $\mathcal{C}(X)$ that separates points of X and contains constants then $\mathcal{A}$ is uniformly dense in $\mathcal{C}(X)$.*

Proof. See Stone-Weierstrass Theorem. ■

Remark 13. If $\mathcal{A}$ is uniformly dense in $\mathcal{C}(X)$, $\mathcal{A}$ is δ_∞-dense. In this context, uniform density is equivalent to density in δ_∞ metric.

Closely related to the Stone-Weierstrass Theorem is the idea of separating points of a topological (sub)space.

Definition (separation on $\mathcal{C}(X)$)**.** A subset $\mathcal{A}$ of $\mathcal{C}(X)$ separates points of X if and only if given $x, y \in X$ with $x \neq y$ there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

Informally, $\mathcal{A}$ is powerful enough to distinguish different inputs by mapping them to different outputs, while preserving continuity. In our context, $\mathcal{A}$ will be a family of fully-connected neural networks with a single hidden layer and exponential activation function, denoted by $\mathcal{H}_{\exp}$. The main reason why we consider exp activation will become evident soon. Using the Stone-Weierstrass Theorem, we can elegantly prove the first fundamental result about the approximation power of single-layer fully-connected neural networks.

Theorem 8 (The Universal Approximation Theorem for exp activation)**.** *Let $K \subset \mathbb{R}^n$ be a compact subset of $\mathbb{R}^n$, with respect to the standard topology on $\mathbb{R}^n$. Let $\mathcal{H}_{\exp}$ denote the family of single-layer fully-connected neural networks with* exp *activation function, given by*

$$\mathcal{H}_{\exp} = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \exp\left(\langle \mathbf{w_k}, \mathbf{x} \rangle\right) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

The family $\mathcal{H}_{\exp}$ is uniformly dense in $\mathcal{C}(K)$.

Proof Idea. By looking at the statement of the result we aim to prove and the conditions of the Stone-Weierstrass Theorem, it is natural to aim to apply the Stone-Weierstrass Theorem. We will demonstrate that $\mathcal{H}_{\exp}$ satisfies the necessary conditions.

Proof.

Step 1 ($\mathcal{H}_{\exp} \subset \mathcal{C}(K)$). This is obvious since a linear combination of continuous functions is still continuous.

Step 2 ($\mathcal{H}_{\exp}$ is an algebra). Let $f, g \in \mathcal{H}_{\exp}$ and suppose

$$f(\mathbf{x}) = \sum_{k=1}^{n} \alpha_k \exp\left(\langle \mathbf{w_k}, \mathbf{x} \rangle\right),$$
$$g(\mathbf{x}) = \sum_{l=1}^{m} \beta_l \exp\left(\langle \mathbf{v_l}, \mathbf{x} \rangle\right),$$

where $n, m \in \mathbb{N}, \alpha_1 \dots \alpha_n \in \mathbb{R}, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_1} \dots \mathbf{w_n} \in \mathbb{R}^n, \mathbf{v_1} \dots \mathbf{v_m} \in \mathbb{R}^n$. Clearly, $f + g \in \mathcal{H}_{\exp}$. Clearly $cf \in \mathcal{H}_{\exp}$, for every $c \in \mathbb{R}$. We will show that $fg \in \mathcal{H}_{\exp}$. We have

$$\begin{aligned} f(\mathbf{x})g(\mathbf{x}) &= \left(\sum_{k=1}^{n} \alpha_k \exp\left(\langle \mathbf{w_k}, \mathbf{x} \rangle\right)\right)\left(\sum_{l=1}^{m} \beta_l \exp\left(\langle \mathbf{v_l}, \mathbf{x} \rangle\right)\right) \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m} \alpha_k \beta_l \exp\left(\langle \mathbf{w_k}, \mathbf{x} \rangle\right) \exp\left(\langle \mathbf{v_l}, \mathbf{x} \rangle\right) \\ &= \sum_{k=1}^{n}\sum_{l=1}^{m} \alpha_k \beta_l \exp\left(\langle \mathbf{w_k} + \mathbf{v_l}, \mathbf{x} \rangle\right) \in \mathcal{H}_{\exp}. \end{aligned}$$

Hence $H_{\exp}$ is indeed an algebra on $\mathcal{C}(K)$.

Step 3 ($\mathcal{H}_{\exp}$ contains constants). This is evident from definition of $\mathcal{H}_{\exp}$. Let $\alpha \in \mathbb{R}$. Then $f(\mathbf{x}) = \alpha = \alpha \exp\left(\langle \mathbf{0}, \mathbf{x} \rangle\right) \in \mathcal{H}_{\exp}$.

Step 4 ($\mathcal{H}_{\exp}$ separates points of K). Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and suppose that $\mathbf{u} \neq \mathbf{v}$. Set $\mathbf{w} = \mathbf{u} - \mathbf{v}$. Define $f : K \to \mathbb{R}$ by $f(\mathbf{x}) = \exp\left(\langle \mathbf{w}, \mathbf{x} \rangle\right)$. Clearly, $f \in \mathcal{H}_{\exp}$. We claim that f separates $\mathbf{u}$ and $\mathbf{v}$. We have

$$\begin{aligned} \frac{f(\mathbf{u})}{f(\mathbf{v})} &= \frac{\exp\left(\langle \mathbf{w}, \mathbf{u} \rangle\right)}{\exp\left(\langle \mathbf{w}, \mathbf{v} \rangle\right)} = \exp\left(\langle \mathbf{w}, \mathbf{u} \rangle - \langle \mathbf{w}, \mathbf{v} \rangle\right) = \exp\left(\langle \mathbf{w}, \mathbf{u} - \mathbf{v} \rangle\right) \\ &= \exp\left(\|\mathbf{u} - \mathbf{v}\|^2\right). \end{aligned} \tag{3.1}$$

Since $\mathbf{u} \neq \mathbf{v}$, $\|\mathbf{u} - \mathbf{v}\| \neq 0$. By 3.1, $f(\mathbf{u}) \neq f(\mathbf{v})$. Since $\mathbf{u}, \mathbf{v}$ were arbitrary, $\mathcal{H}_{\exp}$ indeed separates points of K.

The result follows directly from Stone-Weierstrass Theorem. ■

Remark 14. To the author's knowledge, there has been no paper or textbook discussing this statement and proof addressing directly exp activation function. However, the statement and proof of the result above are a special case of the following theorem presented in [HSW89].

Theorem (Theorem 2.1 in [HSW89]). *Let $K \subset \mathbb{R}^n$ be a compact subset of $\mathbb{R}^n$, with respect to the standard topology on $\mathbb{R}^n$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be any continuous nonconstant activation function. Let $\Sigma\Pi$ denote the family of functions $\mathbb{R}^n \to \mathbb{R}$ given by*

$$\Sigma\Pi = \left\{ \mathbf{x} \to \sum_{i=1}^{m} \alpha_i \prod_{j=1}^{m_k} \sigma(\langle \mathbf{w}_{ji}, \mathbf{x} \rangle + \beta_{ji}) \right\},$$

where $m \in \mathbb{N}, m_k \in \mathbb{N}, \mathbf{w}_{ji} \in \mathbb{R}^n, \alpha_i \in \mathbb{R}, \beta_{ji} \in \mathbb{R}$. Then $\Sigma\Pi$ is dense in $\mathcal{C}(K)$.

It is worth discussing a few limitations of both statement and the proof of The Universal Approximation Theorem for exp activation. However elegant the argument is, we should be aware of its structural requirements from the family $\mathcal{H}_{\exp}$. Informally, any family satisfying the conditions of the Stone-Weierstrass Theorem resembles polynomials. More precisely, the algebraic closure under multiplication is a strong condition that many practically relevant activation functions do not guarantee. A clear counterexample is a neural network with a logistic sigmoid activation. This limitation will prevent us from easily generalizing the argument to more widely used and practically relevant activation functions. However, a generalization of such an argument was done in paper [HSW89].
In the next section, we will use a significantly different approach and argument style. Most of the work presented in the subsequent sections will be based on the relationship between the density of a family in a normed linear space and its dual space. This approach will enable us to support more widely used activation functions such as the logistic sigmoid.
Apart from the structural requirement on the approximation family, the proof gave us no insight into the underlying structure of the neural network accomplishing the desired error. For example, the argument told us nothing about the number of neurons required to achieve the desired error. Unfortunately, even the subsequent sections will give us no insight into this important question which is still an open research area.

3.3 Universal approximation of continuous functions via Cybenko's method

In this section, we will focus on the approximation in the space $\mathcal{C}([0,1]^n)$, where $[0,1]^n$ denotes the n-dimensional unit hypercube. It is worth noting that most of the results established in this section generalize to any compact subset of $\mathbb{R}^n$. However, we focus on $[0,1]^n$ for the sake of compatibility with [Cyb89]. We will discuss one of the most famous results regarding the approximation power of single-layer, fully-connected neural networks. The main result of this section is the following theorem, proved by G. Cybenko in the paper [Cyb89] from 1989.

Theorem (Cybenko, 1989)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with the logistic sigmoid activation function, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

The family $\mathcal{H}_\sigma$ is dense in $\mathcal{C}([0,1]^n)$.

Remark 15. We will prove a slightly more general version.

The proof of this theorem introduces a few novel concepts, such as the notion of discriminatory activation function and a generalization of the logistic sigmoid. Apart from those concepts, the proof relies on standard results from the functional analysis, Hahn-Banach Theorem and Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$. To establish the stated theorem, we will develop the necessary concepts in order very similar to [Cyb89]. The structure of the argument is outlined below.

Discriminatory activation functions In this subsection, we will introduce the concept of a discriminatory activation function. We will discuss a few examples of such functions and develop a lemma to identify discriminatory activation functions.

Sigmoidal activation functions In this subsection, we will discuss a generalization of the logistic sigmoid. Functions belonging to this generalized family are examples of discriminatory activation functions.

Density and the dual space In this subsection, we will explore the relationship between density in a normed linear space and its dual space. The analysis will use methods from functional analysis and measure theory.

The Universal Approximation Theorem for $\mathcal{C}([0,1]^n)$ We will state the main theorem and present an elegant proof based on results developed in the previous three subsections.

The Universal Approximation Theorem for $\mathcal{C}([0,1]^n, \mathbb{R}^m)$ We will generalize Cybenko's Universal Approximation Theorem, [Cyb89] to $\mathcal{C}([0,1]^n, \mathbb{R}^m)$.

3.3.1 Discriminatory activation functions

An essential part of Cybenko's argument is the notion of discriminatory activation function with respect to a given (signed) measure.

Definition 20 (discriminatory activation function with respect to a measure)**.** Let μ be a finite signed Borel measure on $[0,1]^n$. A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called discriminatory for μ if

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b)\, d\mu(\mathbf{x}) = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R} \implies \mu = 0.$$

Remark 16. In the definition above, σ is not necessarily the logistic sigmoid.

Definition 21 (discriminatory activation function)**.** A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called discriminatory if it is discriminatory for every finite signed Borel measure on $[0,1]^n$.

At first glance, the meaning of definition of a discriminatory activation function with respect to a measure is somewhat unclear. Thus, it is worth discussing the intuition behind this definition. By contrapositive, if σ is discriminatory for μ and μ is nonzero, then there exists at least one configuration of weights $\mathbf{w} \in \mathbb{R}^n$ and a bias $b \in \mathbb{R}$ such that $\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b)\, d\mu(\mathbf{x}) \neq 0$. Informally, if σ is discriminatory for μ, then σ is volumetrically non-destructive when it acts on the affine space $\{\langle \mathbf{w}, \mathbf{x} \rangle + b : \mathbf{x} \in [0,1]^n\}$. Recall that the affine space $\{\langle \mathbf{w}, \mathbf{x} \rangle + b : \mathbf{x} \in [0,1]^n\}$ is precisely the range of a single neuron parameterized by weights $\mathbf{w}$ and a bias $b \in \mathbb{R}$, immediately before the application of the activation function. This section aims to answer the following two questions.

- How to prove that an activation function is discriminatory for a given measure?
- Which practically useful activation functions are discriminatory?

To address the first question, we will develop a lemma which will help us prove that a given activation function is discriminatory for a given measure. To simplify claims of the following results, we shall introduce a bit of notation for hyperplanes and open half spaces of $[0,1]^n$. We define

$$\begin{aligned} \Pi_{\mathbf{w},b} &= \{\mathbf{x} : \mathbf{x} \in [0,1]^n, \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}, \\ \Pi^+_{\mathbf{w},b} &= \{\mathbf{x} : \mathbf{x} \in [0,1]^n, \langle \mathbf{w}, \mathbf{x} \rangle + b > 0\}, \\ \Pi^-_{\mathbf{w},b} &= \{\mathbf{x} : \mathbf{x} \in [0,1]^n, \langle \mathbf{w}, \mathbf{x} \rangle + b < 0\}. \end{aligned}$$

The following lemma will provide us with a method to identify discriminatory activation functions.

Lemma 2. *Let μ be a finite signed Borel measure on $[0,1]^n$. If μ vanishes on all hyperplanes and open half-spaces of $[0,1]^n$, then μ is identically zero. More formally, if for every configuration consisting of weights $\mathbf{w} \in \mathbb{R}^n$ and a bias $b \in \mathbb{R}$,*

$$\mu(\Pi_{\mathbf{w},b}) = 0 \text{ and } \mu(\Pi^+_{\mathbf{w},b}) = 0,$$

then $\mu = 0$.

Proof Idea. The idea is to apply "Lebesgue induction" to the cleverly constructed functional $F : \mathcal{L}^\infty(\mathbb{R}) \to \mathbb{R}$ given by

$$F_{\mathbf{w}}(h) = \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}).$$

We will "Lebesgue-inductively" show that $F_{\mathbf{w}} = 0$ on $\mathcal{L}^\infty(\mathbb{R})$. Surprisingly, the main difficulty will be proving that $F_{\mathbf{w}}(\chi_B) = 0$ for every Borel set $B \in \mathcal{B}(\mathbb{R})$. To prove this, we will use Dynkin's $\lambda - \pi$ theorem. After proving that the functional $F_{\mathbf{w}}$ vanishes on indicator functions of Borel sets, we will show that it vanishes on measurable simple functions. By appealing to Density of simple functions in $\mathcal{L}^p$, we will be able to generalize the result to $\mathcal{L}^\infty(\mathbb{R})$. Using $F_{\mathbf{w}}(\sin)$ and $F_{\mathbf{w}}(\cos)$ we will show that the Fourier transform of the finite signed Borel measure μ, denoted $\widehat{\mu}$, satisfies $\widehat{\mu} = 0$. But, this implies $\mu = 0$. The application of the Fourier transform of μ demystifies the definition and use of $F_{\mathbf{w}}$.

Proof.

Step 1 ($F_{\mathbf{w}}$ is a bounded linear functional.). Fix $\mathbf{w} \in \mathbb{R}^n$, and define the function $F_{\mathbf{w}} : \mathcal{L}^\infty(\mathbb{R}) \to \mathbb{R}$ by

$$F_{\mathbf{w}}(h) = \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}).$$

We claim that $F_{\mathbf{w}}$ is a bounded linear functional on $\mathcal{L}^\infty(\mathbb{R})$. Linearity follows from the linearity of an integral. To prove boundedness, suppose $h \in \mathcal{L}^\infty(\mathbb{R})$. By definition of $\mathcal{L}^\infty(\mathbb{R})$, without loss of generality, $h \leq \|h\|_\infty < \infty$. Then

$$|F_{\mathbf{w}}(h)| = \left| \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \right| \leq \int_{[0,1]^n} |h(\langle \mathbf{w}, \mathbf{x} \rangle)|\ d|\mu|(\mathbf{x}) \leq \|h\|_\infty\, |\mu|([0,1]^n). \tag{3.2}$$

Since μ is finite, by Hahn-Jordan decomposition, so is its total variation $|\mu|$. Hence $|\mu|([0,1]^n) < \infty$. By 3.2, $|F_{\mathbf{w}}(h)| < \infty$.

Step 2 ($F_{\mathbf{w}}$ vanishes on indicators of Borel sets in $\mathbb{R}$.). We begin by proving $F_{\mathbf{w}}$ vanishes on indicator functions of intervals. Consider the indicator function $\chi_{[b,\infty)}$, for $b \in \mathbb{R}$. We have

$$\begin{aligned} F_{\mathbf{w}}(\chi_{[b,\infty)}) &= \int_{[0,1]^n} \chi_{[b,\infty)}(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \\ &= \mu(\{\mathbf{x} \in [0,1]^n : b \leq \langle \mathbf{w}, \mathbf{x} \rangle < \infty\}) \\ &= \mu(\{\mathbf{x} \in [0,1]^n : 0 \leq \langle \mathbf{w}, \mathbf{x} \rangle - b\}) \\ &= \mu(\Pi_{\mathbf{w},-b}) + \mu(\Pi^+_{\mathbf{w},-b}) = 0. \end{aligned} \tag{3.3}$$

To establish 3.3, we applied the assumption that $\mu(\Pi_{\mathbf{w},-b}) = \mu(\Pi^+_{\mathbf{w},-b}) = 0$. Similarly,

$$F_{\mathbf{w}}(\chi_{(b,\infty)}) = \mu(\Pi^+_{\mathbf{w},-b}) = 0. \tag{3.4}$$

For every $a, b \in \mathbb{R}$ such that $a < b$, $\chi_{(a,b)} = \chi_{(a,\infty)} - \chi_{[b,\infty)}$. By linearity of $F_{\mathbf{w}}$ and by 3.3 and 3.4,

$$F_{\mathbf{w}}(\chi_{(a,b)}) = F_{\mathbf{w}}(\chi_{(a,\infty)} - \chi_{[b,\infty)}) = F_{\mathbf{w}}(\chi_{(a,\infty)}) - F_{\mathbf{w}}(\chi_{[b,\infty)}) = 0. \tag{3.5}$$

We claim that $F_{\mathbf{w}}(\chi_B) = 0$, for every Borel set $B \subseteq \mathbb{R}$. To show that $F_{\mathbf{w}}$ vanishes on the indicator function of every Borel set, we will appeal to Dynkin's $\lambda - \pi$ theorem. Define collections Π and Λ by

$$\Pi = \{(a, b) : -\infty \leq a \leq b \leq \infty\} \text{ and } \Lambda = \{A : A \in \mathcal{B}(\mathbb{R}) \text{ and } F_{\mathbf{w}}(\chi_A) = 0\}.$$

Since a finite intersection of open intervals is again an open, possibly degenerate interval, Π is a π-system. We will show that Λ is a λ-system. By 3.5, Λ contains Π. Clearly, $\mathbb{R} \in \Lambda$. Suppose that $A, B \in \Lambda$ where $B \subseteq A$. Then $\chi_{A\setminus B} = \chi_A - \chi_B$ so $F_{\mathbf{w}}(\chi_{A\setminus B}) = F_{\mathbf{w}}(\chi_A - \chi_B) = F_{\mathbf{w}}(\chi_A) - F_{\mathbf{w}}(\chi_B) = 0$, since $A, B \in \Lambda$. Thus $A \setminus B \in \Lambda$. Suppose that $\{B_n\}_{n=1}^{\infty}$ is a collection of disjoint sets in Λ. We will show that $B = \bigcup_{n=1}^{\infty} B_n \in \Lambda$. Clearly, $\chi_B = \sum_{k=1}^{\infty} \chi_{B_k}$ so $\sum_{k=1}^{m} \chi_{B_k} \uparrow \chi_B$, as $m \to \infty$. Then

$$\begin{aligned}
F_{\mathbf{w}}(\chi_B) &= \int_{[0,1]^n} \chi_B(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) = \int_{[0,1]^n} \sum_{k=1}^{\infty} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \\
&= \int_{[0,1]^n} \lim_{m\to\infty} \sum_{k=1}^{m} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \\
&= \lim_{m\to\infty} \int_{[0,1]^n} \sum_{k=1}^{m} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \text{ by Monotone Convergence Theorem} \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} \int_{[0,1]^n} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} F_{\mathbf{w}}(\chi_{B_k}) = 0.
\end{aligned}$$

It is worth justifying the application of Monotone Convergence Theorem. In the original form, Monotone Convergence Theorem applies only to measures. In the calculation above, we are integrating with respect to a signed measure. However, by Hahn-Jordan decomposition, μ admits decomposition $\mu = \mu^+ - \mu^-$, where μ^+ and μ^- are measures. We apply Monotone Convergence Theorem to integrals with respect to μ^+ and μ^-. By definition of the integral with respect to μ, we obtain the desired conclusion. Hence Λ is indeed a λ-system. By Dynkin's $\lambda - \pi$ theorem, $\mathcal{B}(\mathbb{R}) = \sigma(\Pi) = \lambda(\Pi) \subseteq \Lambda$. Since $\Lambda \subseteq \mathcal{B}(\mathbb{R})$, $\mathcal{B}(\mathbb{R}) = \Lambda$, as desired.

Step 3 ($F_{\mathbf{w}}$ vanishes on measurable simple functions). Suppose that φ is a $\mathcal{B}(\mathbb{R})$-measurable simple function. Without loss of generality, $\varphi = \sum_{k=1}^{m} \alpha_k \chi_{A_k}$, where A_k are disjoint $\mathcal{B}(\mathbb{R})$-measurable sets. By linearity of $F_{\mathbf{w}}$ and Step 2,

$$F_{\mathbf{w}}(\varphi) = F\left(\sum_{k=1}^{m} \alpha_k \chi_{A_k}\right) = \sum_{k=1}^{m} F_{\mathbf{w}}(\alpha_k \chi_{A_k}) = \sum_{k=1}^{m} \alpha_k F_{\mathbf{w}}(\chi_{A_k}) = 0. \tag{3.6}$$

Step 4 ($F_{\mathbf{w}}$ vanishes on $\mathcal{L}^\infty(\mathbb{R})$). Let $f \in \mathcal{L}^\infty(\mathbb{R})$. By Density of simple functions in $\mathcal{L}^p$, there exists a sequence of $\mathcal{B}(\mathbb{R})$-measurable simple functions $\{\varphi_n\}_{n=1}^\infty$ converging to f in $\|\cdot\|_\infty$. For every $m \in \mathbb{N}$, $f - \varphi_m \in \mathcal{L}^\infty(\mathbb{R})$. Without loss of generality, $|f - \varphi_m| \leq \|f - \varphi_m\|_\infty$. Then

$$\begin{aligned} |F_{\mathbf{w}}(f) - F_{\mathbf{w}}(\varphi_m)| &\leq \left| \int_{[0,1]^n} (f - \varphi_m)(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \right| \\ &\leq \int_{[0,1]^n} |(f - \varphi_m)(\langle \mathbf{w}, \mathbf{x} \rangle)|\, d|\mu|(\mathbf{x}) \\ &\leq \|f - \varphi_m\|_\infty\, |\mu|([0,1]^n). \end{aligned} \tag{3.7}$$

Since $\lim_{m\to\infty} \|f - \varphi_m\|_\infty = 0$ and $|\mu|([0,1]^n) < \infty$, by 3.7, $|F_{\mathbf{w}}(f) - F_{\mathbf{w}}(\varphi_m)| \to 0$, as $m \to \infty$. Combining $|F_{\mathbf{w}}(f) - F_{\mathbf{w}}(\varphi_m)| \to 0$ as $m \to \infty$ with 3.6 gives

$$F_{\mathbf{w}}(f) = \lim_{m\to\infty} F_{\mathbf{w}}(\varphi_m) = 0. \tag{3.8}$$

Since f was arbitrary, $F_{\mathbf{w}}$ vanishes on $\mathcal{L}^\infty(\mathbb{R})$.

Step 5 (μ is identically zero.). We will compute the Fourier transform of μ. Since cos and sin are bounded and measurable, $\cos, \sin \in \mathcal{L}^\infty(\mathbb{R})$. By 3.8, we have that $F_{\mathbf{w}}(\cos) = F_{\mathbf{w}}(\sin) = 0$. This implies

$$\begin{aligned} \widehat{\mu}(\mathbf{w}) &= \int_{[0,1]^n} e^{i\langle \mathbf{w}, \mathbf{x} \rangle}\, d\mu(\mathbf{x}) \\ &= \int_{[0,1]^n} \cos(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) + i \int_{[0,1]^n} \sin(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mu(\mathbf{x}) \\ &= F_{\mathbf{w}}(\cos) + iF_{\mathbf{w}}(\sin) \\ &= 0. \end{aligned}$$

By Corollary 10, $\mu = 0$.

■

We are ready to address the second question.

3.3.2 Sigmoidal activation functions

Definition 22 (sigmoidal activation function)**.** A function $\sigma : \mathbb{R} \to \mathbb{R}$ is called sigmoidal if

$$\lim_{x\to\infty} \sigma(x) = 1 \text{ and } \lim_{x\to-\infty} \sigma(x) = 0.$$

Proposition 3 (Bounded measurable sigmoidal functions are discriminatory)**.** *Let μ be a finite signed Borel measure on $[0,1]^n$. Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is bounded, Borel measurable and sigmoidal function. Then σ is discriminatory for μ.*

Proof Idea. We will reduce the problem to the application of Lemma 2.

Proof. Suppose that σ satisfies

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b)\, d\mu(\mathbf{x}) = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}. \tag{3.9}$$

We need to show that $\mu = 0$. We aim to apply Lemma 2. For fixed $\lambda, a \in \mathbb{R}$, define $\sigma_{\lambda,a} : [0,1]^n \to \mathbb{R}$ by

$$\sigma_{\lambda,a}(\mathbf{x}) = \sigma\left(\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b) + a\right) = \sigma\left(\langle \lambda \mathbf{w}, \mathbf{x} \rangle + b\lambda + a\right).$$

By 3.9,

$$\int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x})\, d\mu(\mathbf{x}) = 0. \tag{3.10}$$

Define $\gamma : [0,1]^n \to \mathbb{R}$ by $\gamma(\mathbf{x}) = \lim_{\lambda \to \infty} \sigma_{\lambda,a}(\mathbf{x})$. Observe that

$$\gamma(\mathbf{x}) = \begin{cases} 1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b > 0 \\ \sigma(a) & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \\ 0 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b < 0 \end{cases}. \tag{3.11}$$

By 3.11,

$$\begin{aligned} \int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) &= \int_{\Pi^+_{\mathbf{w},b}} \gamma(\mathbf{x}) d\mu(\mathbf{x}) + \int_{\Pi_{\mathbf{w},b}} \gamma(\mathbf{x}) d\mu(\mathbf{x}) + \int_{\Pi^-_{\mathbf{w},b}} \gamma(\mathbf{x}) d\mu(\mathbf{x}) \\ &= \int_{\Pi^+_{\mathbf{w},b}} 1 d\mu(\mathbf{x}) + \int_{\Pi_{\mathbf{w},b}} \sigma(a) d\mu(\mathbf{x}) + \int_{\Pi^-_{\mathbf{w},b}} 0 d\mu(\mathbf{x}) \\ &= \mu(\Pi^+_{\mathbf{w},b}) + \sigma(a) \cdot \mu(\Pi_{\mathbf{w},b}). \end{aligned}$$

Taking $\lim_{a \to \infty}$ and applying the fact σ is sigmoidal gives

$$\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) = \mu(\Pi^+_{\mathbf{w},b}) + \lim_{a \to \infty} \sigma(a) \cdot \mu(\Pi_{\mathbf{w},b}) = \mu(\Pi^+_{\mathbf{w},b}) + \mu(\Pi_{\mathbf{w},b}). \tag{3.12}$$

Taking $\lim_{a \to -\infty}$ and applying the fact σ is sigmoidal gives

$$\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) = \mu(\Pi^+_{\mathbf{w},b}) + \lim_{a \to -\infty} \sigma(a) \cdot \mu(\Pi_{\mathbf{w},b}) = \mu(\Pi^+_{\mathbf{w},b}). \tag{3.13}$$

Equating 3.12 and 3.13 gives

$$\mu(\Pi^+_{\mathbf{w},b}) + \mu(\Pi_{\mathbf{w},b}) = \mu(\Pi^+_{\mathbf{w},b}) \implies \mu(\Pi_{\mathbf{w},b}) = 0. \tag{3.14}$$

By 3.13, to prove $\mu(\Pi^+_{\mathbf{w},b}) = 0$, it is equivalent to prove $\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) = 0$. We will appeal to the Dominated Convergence Theorem. By Hahn-Jordan decomposition, μ can be decomposed as $\mu = \mu^+ - \mu^-$, where μ^+ and μ^- are measures and at least one of them is finite. Since μ is finite, so are both μ^+ and μ^-.

Since σ is bounded, for every $\lambda \in \mathbb{R}$, for every $a \in \mathbb{R}$, for every $\mathbf{x} \in [0,1]^n$,

$$|\sigma_{\lambda,a}(\mathbf{x})| \leq \|\sigma\|_\infty .$$

Since μ^+ and μ^- are finite, $\mathbf{x} \to \|\sigma\|_\infty$ is μ^+ and μ^- integrable. Now

$$\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) = \int_{[0,1]^n} \gamma(\mathbf{x}) d\mu^+(\mathbf{x}) - \int_{[0,1]^n} \gamma(\mathbf{x}) d\mu^-(\mathbf{x}). \tag{3.15}$$

By 3.11 and Dominated Convergence Theorem, we have

$$\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu^+(\mathbf{x}) = \lim_{\lambda \to \infty} \int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^+(\mathbf{x}). \tag{3.16}$$

Similarly, we have

$$\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu^-(\mathbf{x}) = \lim_{\lambda \to \infty} \int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^-(\mathbf{x}). \tag{3.17}$$

Applying 3.16 and 3.17 to 3.15 gives

$$\begin{aligned}
\int_{[0,1]^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) &= \lim_{\lambda \to \infty} \int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^+(\mathbf{x}) - \lim_{\lambda \to \infty} \int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^-(\mathbf{x}) \\
&= \lim_{\lambda \to \infty} \left(\int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^+(\mathbf{x}) - \int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu^-(\mathbf{x}) \right) \\
&= \lim_{\lambda \to \infty} \left(\int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu(\mathbf{x}) \right) \\
&= 0,
\end{aligned}$$

since by 3.10, $\int_{[0,1]^n} \sigma_{\lambda,a}(\mathbf{x}) d\mu(\mathbf{x}) = 0$. We have shown that μ vanishes on $\Pi^+_{\mathbf{w},b}$ and $\Pi_{\mathbf{w},b}$. By Lemma 2, μ is identically zero. We conclude σ is discriminatory for μ, as claimed. ■

Proposition 4 (Continuous sigmoidal functions are discriminatory)**.** *Let μ be a finite signed Borel measure on $[0,1]^n$. Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is a continuous sigmoidal function. Then σ is discriminatory for μ.*

Proof. Since σ is continuous, it is Borel measurable. Since it is continuous and sigmoidal, σ is bounded. The result follows directly from Proposition 3. ■

Corollary 2. *Consequently, any continuous sigmoidal function is discriminatory.*

Proof. Apply Proposition 4 to arbitrary μ. ■

Corollary 3. *Logistic sigmoid is discriminatory.*

Corollary 4. *Heaviside step function is discriminatory.*

It turns out that a wide variety of bounded functions are discriminatory.

Theorem 9 (Theorem 5 in [Hor91])**.** *Whenever σ is bounded and nonconstant, it is discriminatory.*

Proof. See Theorem 5 in [Hor91]. ■

3.3.3 Density and the dual space

In this section, we will focus on a real normed linear space $\mathcal{X}$ and its **non-dense** linear subspace $\mathcal{U}$. The density is understood with respect to the topology induced by the norm $\|\cdot\|$ on $\mathcal{X}$. It turns out that the fact $\mathcal{U}$ is not dense implies the existence of a non-trivial bounded linear functional which vanishes on $\mathcal{U}$.

Lemma 3 (Separation functional lemma)**.** *Let $\mathcal{U}$ be a **non-dense** linear subspace of a real normed linear space $\mathcal{X}$. Then there exists a bounded linear functional L on $\mathcal{X}$ such that $L \neq 0$ on $\mathcal{X}$ and $L_{|\mathcal{U}} = 0$.*

Proof Idea. To prove the lemma, we will give an explicit construction of the desired functional. We will begin with a subspace of $\mathcal{X}$, denoted by $\mathcal{T}$ where the construction of a candidate functional is more evident and the verification of required properties is relatively easy. The extension of the construction to the entire space $\mathcal{X}$ will follow from the Hahn-Banach Theorem.

Theorem (Hahn-Banach Theorem). *Let X be a real vector space, with a sublinear functional ρ defined on X. Suppose that W is a linear subspace of X and f_W a linear functional on W satisfying*

$$f_W(w) \leq \rho(w), w \in W. \tag{3.18}$$

Then f_W has an extension f on X such that

$$f(x) \leq \rho(x), x \in X. \tag{3.19}$$

Proof. See Hahn-Banach Theorem in Appendix. ■

Proof. Since $\mathcal{U}$ is not dense in $\mathcal{X}$, there exists $\mathbf{x_0} \in \mathcal{X}$ and there exists $\delta > 0$ such that

$$\|\mathbf{x_0} - \mathbf{u}\| \geq \delta, \text{ for every } \mathbf{u} \in \mathcal{U}. \tag{3.20}$$

Step 1 (Construction of a suitable linear subspace $\mathcal{T}$). To define the desired functional, we restrict our attention to a subset $\mathcal{T} \subseteq \mathcal{X}$ defined by

$$\mathcal{T} = \{\mathbf{u} + \lambda \mathbf{x_0} : \lambda \in \mathbb{R}, \mathbf{u} \in \mathcal{U}\}.$$

We claim that $\mathcal{T}$ is a linear subspace of $\mathcal{X}$. We begin by proving that every element in $\mathcal{T}$ has unique representation. To prove this, suppose that for $\mathbf{t} \in \mathcal{T}$, we have two representations,

$$\mathbf{t} = \mathbf{u} + \alpha \mathbf{x_0} = \mathbf{v} + \beta \mathbf{x_0}, \tag{3.21}$$

where $\mathbf{u}, \mathbf{v} \in \mathcal{U}$ and $\alpha, \beta \in \mathbb{R}$. Rearranging 3.21 gives

$$\mathbf{u} - \mathbf{v} = (\beta - \alpha)\mathbf{x_0}. \tag{3.22}$$

By 3.20, $\mathbf{x_0} \notin \mathcal{U}$. Since $\mathbf{u} - \mathbf{v} \in \mathcal{U}$, the only possible solution to 3.22 is $\mathbf{u} - \mathbf{v} = \mathbf{0} = (\beta - \alpha)\mathbf{x_0}$. This forces $\mathbf{u} = \mathbf{v}$ and $\alpha = \beta$. Thus, the representation of elements in $\mathcal{T}$ is indeed unique. Since $\mathcal{U}$ is linear, $\mathbf{0} \in \mathcal{U}$ and we may choose $\lambda = 0$ to deduce $\mathbf{0} \in \mathcal{T}$. Since $\mathcal{U}$ is closed under addition and scalar multiplication, so is $\mathcal{T}$.

Step 2 (Construction of a desired functional on $\mathcal{T}$). Now define $L : \mathcal{T} \to \mathbb{R}$ by

$$L(\mathbf{t}) = L(\mathbf{u} + \lambda \mathbf{x_0}) = \lambda\delta.$$

Since the representation of elements in $\mathcal{T}$ is unique, L is well defined. We claim that L is a bounded linear functional on $\mathcal{T}$. Let $\alpha \in \mathbb{R}$ and $\mathbf{t_1}, \mathbf{t_2} \in \mathcal{T}$ and suppose that $\mathbf{t_1} = \mathbf{u_1} + \lambda_1 \mathbf{x_0}, \mathbf{t_2} = \mathbf{u_2} + \lambda_2 \mathbf{x_0} \in \mathcal{T}$, for $\mathbf{u_1}, \mathbf{u_2} \in \mathcal{U}, \lambda_1, \lambda_2 \in \mathbb{R}$. Now

$$\begin{aligned} L(\mathbf{t_1} + \mathbf{t_2}) &= L(\mathbf{u_1} + \mathbf{u_2} + (\lambda_1 + \lambda_2)\mathbf{x_0}) \\ &= (\lambda_1 + \lambda_2)\delta = \lambda_1\delta + \lambda_2\delta \\ &= L(\mathbf{t_1}) + L(\mathbf{t_2}), \\ L(\alpha \mathbf{t_1}) &= L(\alpha \mathbf{u_1} + \alpha\lambda_1 \mathbf{x_0}) \\ &= \alpha\lambda_1\delta \\ &= \alpha L(\mathbf{t_1}). \end{aligned}$$

Hence, L is linear on $\mathcal{T}$. We will show that for every $\mathbf{t} \in \mathcal{T}$, $L(\mathbf{t}) \leq \|\mathbf{t}\|$. Write $\mathbf{t} = \mathbf{u} + \lambda \mathbf{x_0}$. If $\lambda = 0$ then $L(\mathbf{t}) = 0 \leq \|\mathbf{t}\|$. Now suppose $\lambda \neq 0$. Notice that if $\mathbf{u} \in \mathcal{U}$, then $\frac{\mathbf{u}}{\lambda} \in \mathcal{U}$. Since $\mathbf{x_0} + \frac{\mathbf{u}}{\lambda} \in \mathcal{T}$, by 3.20,

$$\left\| \mathbf{x_0} + \frac{\mathbf{u}}{\lambda} \right\| \geq \delta > 0. \tag{3.23}$$

By 3.23, $\frac{\delta}{\left\|\mathbf{x_0} + \frac{\mathbf{u}}{\lambda}\right\|} \leq 1$. Since $\frac{1}{\lambda}(\lambda \mathbf{x_0} + \mathbf{u}) = \mathbf{x_0} + \frac{1}{\lambda}\mathbf{u}$, we have

$$|\lambda|\delta \leq \|\mathbf{u} + \lambda \mathbf{x_0}\|. \tag{3.24}$$

By 3.24,

$$L(\mathbf{t}) = L(\mathbf{u} + \lambda \mathbf{x_0}) = \lambda\delta \leq |\lambda|\delta \leq \|\mathbf{u} + \lambda \mathbf{x_0}\| = \|\mathbf{t}\|.$$

Step 3 (Extension to $\mathcal{X}$). By Hahn-Banach Theorem applied with the norm $p(\mathbf{x}) = \|\mathbf{x}\|$, the constructed linear functional L can be extended to a linear functional $\widetilde{L} : \mathcal{X} \to \mathbb{R}$ such that $\widetilde{L}(\mathbf{x}) \leq \|\mathbf{x}\|$, for every $\mathbf{x} \in \mathcal{X}$. This implies $\left\|\widetilde{L}\right\| \leq 1$ so $\widetilde{L}$ is bounded. By definition of L and the fact $\widetilde{L}_{|\mathcal{T}} = L$,

$$\begin{aligned} \widetilde{L}(\mathbf{u}) &= L(\mathbf{u} + 0 \cdot \mathbf{x_0}) = 0 \cdot \delta = 0, \\ \widetilde{L}(\mathbf{x_0}) = L(\mathbf{x_0}) &= L(\mathbf{0} + 1 \cdot \mathbf{x_0}) = 1 \cdot \delta > 0. \end{aligned}$$

This implies $\widetilde{L}_{|\mathcal{U}} = 0$ and $\widetilde{L} \neq 0$ on $\mathcal{X}$.

■

The Separation functional lemma is a quite general result that guarantees the existence of a separation functional vanishing on a non-dense linear subspace of some normed linear space. Unfortunately, this result does not reveal any structure of the desired functional.

However, in the context of a normed linear space $\mathcal{C}([0,1]^n)$, there is a natural correspondence between bounded linear functionals on $\mathcal{C}([0,1]^n)$ and finite signed Borel measures on $[0,1]^n$. This is a consequence of a result known as the Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$, stated below.

Theorem (Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$)**.** *Let X be a compact metric space and I be a bounded linear functional on $\mathcal{C}(X)$. Then there exists the unique finite signed regular measure μ on $\mathcal{B}(X)$ such that*

$$I(f) = \int_X f \, d\mu, \text{ for every } f \in \mathcal{C}(X). \tag{3.25}$$

Proof. See Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$. ■

We will apply Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$ to reveal the structure of a functional given by Separation functional lemma.

Lemma 4. *Let $\mathcal{U}$ be a* ***non-dense*** *linear subspace of a normed linear space $\mathcal{C}([0,1]^n)$. Then there exists the unique finite signed regular measure μ on $\mathcal{B}([0,1]^n)$ such that*

$$\int_{[0,1]^n} h \, d\mu = 0, \text{ for every } h \in \mathcal{U},$$

but $\mu \neq 0$.

Proof. By Separation functional lemma applied to $\mathcal{X} = \mathcal{C}([0,1]^n)$ equipped with the sup-norm, there exists a bounded linear functional on $\mathcal{C}([0,1]^n)$, denoted by $L : \mathcal{C}([0,1]^n) \to \mathbb{R}$, such that $L \neq 0$ on $\mathcal{C}([0,1]^n)$ and $L_{|\mathcal{U}} = 0$. By Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$, there exists the unique finite signed regular measure μ on $\mathcal{B}([0,1]^n)$ such that

$$L(f) = \int_{[0,1]^n} f \, d\mu, \text{ for every } f \in \mathcal{C}([0,1]^n). \tag{3.26}$$

Combining 3.26 and the fact $L_{|\mathcal{U}} = 0$ gives

$$L(h) = 0 = \int_{[0,1]^n} h \, d\mu, \text{ for every } h \in \mathcal{U}.$$

Since $L \neq 0$ on $\mathcal{C}([0,1]^n)$, by 3.26, $\mu \neq 0$. ■

3.3.4 The Universal Approximation Theorem for $\mathcal{C}([0,1]^n)$

Theorem 10 (The Universal Approximation Theorem for continuous functions)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer, fully-connected neural networks with any continuous discriminatory activation function, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

The family $\mathcal{H}_\sigma$ is dense in $\mathcal{C}([0,1]^n)$.

Proof Idea. We will argue by contradiction and apply Lemma 4.

Proof. Since σ is continuous, the family $\mathcal{H}_\sigma$ is a linear subspace of $C([0,1]^n)$. To prove that the family $\mathcal{H}_\sigma$ is dense in $\mathcal{C}([0,1]^n)$, we will argue by contradiction. Suppose that $\mathcal{H}_\sigma$ is not dense. By Lemma 4, there exists the unique finite signed regular measure μ on $\mathcal{B}([0,1]^n)$ such that for every $h \in \mathcal{H}_\sigma$,

$$\int_{[0,1]^n} h \, d\mu = 0, \text{ but } \mu \neq 0. \tag{3.27}$$

By linearity of the integral and definition of the family $\mathcal{H}_\sigma$, 3.27 is equivalent to

$$\sum_{k=1}^{N} \alpha_k \int_{[0,1]^n} \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) \, d\mu(\mathbf{x}) = 0, \forall N \in \mathbb{N}, \mathbf{w_k} \in \mathbb{R}^n, \alpha_k, \beta_k \in \mathbb{R}. \tag{3.28}$$

Let $\mathbf{w} \in \mathbb{R}^n$, $b \in \mathbb{R}$ be arbitrary. By 3.28,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b) \, d\mu(\mathbf{x}) = 0. \tag{3.29}$$

Since σ is discriminatory for μ and 3.29 holds for arbitrary configuration of weights $\mathbf{w}$ and a bias b, μ is identically zero. However, by 3.27, $\mu \neq 0$. This is a contradiction. We conclude $\mathcal{H}_\sigma$ is indeed dense, as required. ■

As a corollary of Theorem 10, we present the original Cybenko's Universal Approximation Theorem, [Cyb89].

Theorem 11 (Cybenko's Universal Approximation Theorem, [Cyb89])**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer, fully-connected neural networks with any continuous sigmoidal activation function, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

The family $\mathcal{H}_\sigma$ is dense in $\mathcal{C}([0,1]^n)$.

Proof. By Proposition 4, any continuous sigmoidal function is discriminatory. The result follows from The Universal Approximation Theorem for continuous functions. ■

3.3.5 The Universal Approximation Theorem for $\mathcal{C}([0,1]^n, \mathbb{R}^m)$

In this subsection, we will focus on the metric space $\mathcal{C}([0,1]^n, \mathbb{R}^m)$ equipped with the sup-norm distance δ_∞. We begin with a generalization of single-layer, real-valued, fully-connected neural networks - $\mathcal{H}_\sigma$.

Definition 23. Let $\mathcal{H}_\sigma$ denote the family of single-layer, real-valued, fully-connected neural networks with any continuous sigmoidal activation function, given by

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{N} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : N \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

We define the family $\mathcal{H}_\sigma^m$ by

$$\mathcal{H}_\sigma^m = \left\{ \mathbf{x} \to \begin{bmatrix} h_\sigma^{(1)}(\mathbf{x}) \\ h_\sigma^{(2)}(\mathbf{x}) \\ \vdots \\ h_\sigma^{(m)}(\mathbf{x}) \end{bmatrix} : h_\sigma^{(k)} \in \mathcal{H}_\sigma, \text{ for every } 1 \le k \le m \right\}.$$

We are ready to state and prove the main theorem of this subsection.

Theorem 12 (The Universal Approximation Theorem for continuous, vector--valued functions on compact sets)**.** *$\mathcal{H}_\sigma^m$ is dense in $\mathcal{C}([0,1]^n, \mathbb{R}^m)$.*

Proof. Let $\mathbf{f} \in \mathcal{C}([0,1]^n, \mathbb{R}^m)$. Then $f(\mathbf{x}) = \begin{bmatrix} f^{(1)}(\mathbf{x}) & f^{(2)}(\mathbf{x}) & \dots & f^{(m)}(\mathbf{x}) \end{bmatrix}^\top$, where for every $k \in \{1, 2, \dots, m\}$, $f^{(k)} \in \mathcal{C}([0,1]^n)$. Let $\epsilon > 0$. By Cybenko's Universal Approximation Theorem, [Cyb89], there exists a finite family $\{h_\sigma^{(k)}\}_{k=1}^m$ such that $h_\sigma^{(k)} \in \mathcal{H}_\sigma$ and

$$\delta_\infty(f^{(k)}, h_\sigma^{(k)}) = \sup_{\mathbf{x} \in [0,1]^n} |f^{(k)}(\mathbf{x}) - h_\sigma^{(k)}(\mathbf{x})| < \frac{\epsilon}{\sqrt{m}}, \text{ for every } 1 \le k \le m. \quad (3.30)$$

Set $\mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_\sigma^{(1)}(\mathbf{x}) & h_\sigma^{(2)}(\mathbf{x}) & \dots & h_\sigma^{(m)}(\mathbf{x}) \end{bmatrix}^\top$. Clearly, $\mathbf{h} \in \mathcal{H}_\sigma^m$. We will show that $\delta_\infty(\mathbf{f}, \mathbf{h}) \le \epsilon$. For every $\mathbf{x} \in [0,1]^n$,

$$\begin{aligned} \|f(\mathbf{x}) - h(\mathbf{x})\|_2 &= \left(\sum_{k=1}^{m} |f_k(\mathbf{x}) - h_k(\mathbf{x})|^2 \right)^{\frac{1}{2}} \\ &\le \left(\sum_{k=1}^{m} \delta_\infty(f^{(k)}, h_\sigma^{(k)})^2 \right)^{\frac{1}{2}} \\ &< \left(\sum_{k=1}^{m} \left(\frac{\epsilon}{\sqrt{m}} \right)^2 \right)^{\frac{1}{2}} \qquad \text{by 3.30} \\ &\le \epsilon. \qquad (3.31) \end{aligned}$$

By 3.31, $\delta_\infty(\mathbf{f}, \mathbf{h}) \le \epsilon$. ∎

3.4 Universal approximation of square-integrable functions

Another important function space is the space of Lebesgue square-integrable functions. Those functions often arise in signal processing and physics. They are often interpreted as signals of finite energy. We will focus on functions in $\mathcal{L}^2([0,1]^n)$ because $\mathcal{L}^2([0,1]^n)$ is a Hilbert space. Due to this fact, it is sensible to discuss the notion of orthogonality. The orthogonality will help us gain more intuition about $\mathcal{L}^2([0,1]^n)$-discriminatory functions. To establish the The Universal Approximation Theorem for square-integrable functions, we will use Cybenko's method. The argument will be divided into two main sections outlined below.

$\mathcal{L}^2$-discriminatory activation functions In this subsection, we will introduce the concept of a $\mathcal{L}^2([0,1]^n)$-discriminatory activation function. We will discuss a few examples of such functions and develop a lemma to identify discriminatory activation functions. This section will be very similar to Discriminatory activation functions.

The Universal Approximation Theorem for $\mathcal{L}^2([0,1]^n)$ In this subsection, we will prove the The Universal Approximation Theorem for square-integrable functions. We will reuse a lot of theory developed in Density and the dual space.

Remark 17. It is worth noting that results from this section can be generalized to $\mathcal{L}^2(K)$, where K is any compact subset of $\mathbb{R}^n$. However, we consider $\mathcal{L}^2([0,1]^n)$ for the sake of simplicity and compatibility with [Cyb89].

3.4.1 $\mathcal{L}^2$-discriminatory activation functions

In the case of $\mathcal{L}^2([0,1]^n)$, it is possible to derive the notion of discriminatory activation function, using the orthogonality. Recall that the $\|\cdot\|_2$ on $\mathcal{L}^2([0,1]^n)$ arises from the inner product

$$\langle f, g\rangle = \int_{[0,1]^n} f(\mathbf{x})g(\mathbf{x})\,d\mathbf{x}, \text{ for every } f, g \in \mathcal{L}^2([0,1]^n).$$

Suppose that $\mathcal{H}$ is dense in $\mathcal{L}^2([0,1]^n)$ and consider $g \in \mathcal{L}^2([0,1]^n)$, satisfying $g \perp \mathcal{H}$. Since $\mathcal{H}$ is dense in $\mathcal{L}^2([0,1]^n)$, there exists a sequence of functions $\{g_m\}_{m=1}^{\infty}$ in $\mathcal{H}$ such that $g_m \to g$ in $\|\cdot\|_2$ as $m \to \infty$. Since $g \perp \mathcal{H}$, $\langle g_m, g\rangle = 0$. Now

$$\begin{aligned} |\langle g, g\rangle| &= |\langle g - g_m, g\rangle + \langle g_m, g\rangle| \\ &= |\langle g - g_m, g\rangle| \\ &\leq \|g - g_m\|_2 \cdot \|g\|_2 . \end{aligned} \tag{3.32}$$

To obtain 3.32, we applied the Cauchy–Schwarz inequality in the last step.

Applying $\lim_{m\to\infty}$ on 3.32 gives $|\langle g, g\rangle| = 0$. But then $\|g\|_2 = 0$. By Proposition 15, $\|g\|_2 = 0$ implies $g = 0$ almost everywhere. The condition $g \perp \mathcal{H}$ is equivalent to

$$\int_{[0,1]^n} h(\mathbf{x})g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } h \in \mathcal{H}. \tag{3.33}$$

Now recall the definition of the family of single-layer fully-connected neural networks with activation function σ,

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x}\rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m \in \mathbb{R}, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

If $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^2([0,1]^n)$ and 3.33 holds, then $g = 0$ almost everywhere. Thus, it is natural to define $\mathcal{L}^2$-discriminatory activation function in the following way.

Definition 24 ($\mathcal{L}^2$-discriminatory activation function)**.** Let $\sigma : \mathbb{R} \to [0,1]$. We say σ is $\mathcal{L}^2([0,1]^n)$-discriminatory if for every $g \in \mathcal{L}^2([0,1]^n)$,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R},$$

implies $g = 0$ almost everywhere.

We will develop an $\mathcal{L}^2([0,1]^n)$-analog of Lemma 2.

Lemma 5. *Suppose that $g \in \mathcal{L}^2([0,1]^n)$ satisfies*

$$\int_{\Pi^+_{\mathbf{w},b}} g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}. \tag{3.34}$$

Then $g = 0$ almost everywhere.

Proof Idea. We will mimick the proof of Lemma 2 and use the injectivity property of the Fourier Transform on $\mathcal{L}^1([0,1]^n)$. The first observation is that $g \in \mathcal{L}^1([0,1]^n)$ because $([0,1]^n, \mathcal{B}([0,1]^n), \lambda_{|[0,1]^n})$ is a finite measure space. Thus, if we can show that the Fourier transform of g is identically zero, we can appeal to the injectivity of the Fourier Transform on $\mathcal{L}^1([0,1]^n)$ to deduce $g = 0$ almost everywhere. The idea is to apply classical "Lebesgue induction" to the cleverly constructed functional $F : \mathcal{L}^\infty(\mathbb{R}) \to \mathbb{R}$ given by

$$F_{\mathbf{w}}(h) = \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x}\rangle)g(\mathbf{x})\,d\mathbf{x}.$$

We will use $F_{\mathbf{w}}$ to prove that $\widehat{g} = 0$. We will "Lebesgue-inductively" show that $F_{\mathbf{w}} = 0$ on $\mathcal{L}^\infty(\mathbb{R})$. As in the proof of Lemma 2, the main difficulty will be proving that $F_{\mathbf{w}}(\chi_B) = 0$ for every Borel set $B \in \mathcal{B}(\mathbb{R})$. To prove this, we will use Dynkin's $\lambda - \pi$ theorem.

Proof.

Step 1 ($F_{\mathbf{w}}$ is a bounded linear functional.). Fix $\mathbf{w} \in \mathbb{R}^n$, and define the function $F_{\mathbf{w}} : \mathcal{L}^\infty(\mathbb{R}) \to \mathbb{R}$ by

$$F_{\mathbf{w}}(h) = \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x} \rangle) g(\mathbf{x})\, d\mathbf{x}.$$

We claim that $F_{\mathbf{w}}$ is a bounded linear functional on $\mathcal{L}^\infty(\mathbb{R})$. Linearity follows from the linearity of an integral. To prove boundedness, suppose $h \in \mathcal{L}^\infty(\mathbb{R})$. By definition of $\mathcal{L}^\infty(\mathbb{R})$, without loss of generality, $h \leq \|h\|_\infty < \infty$. Then

$$\begin{aligned} |F_{\mathbf{w}}(h)| &= \left| \int_{[0,1]^n} h(\langle \mathbf{w}, \mathbf{x} \rangle) g(\mathbf{x})\, d\mathbf{x} \right| \leq \int_{[0,1]^n} |h(\langle \mathbf{w}, \mathbf{x} \rangle)|\, |g(\mathbf{x})|\, d\mathbf{x} \\ &\leq \|h\|_\infty \int_{[0,1]^n} |g(\mathbf{x})|\, d\mathbf{x} \leq \|h\|_\infty \left(\lambda([0,1]^n)\right)^{1/2} \|g\|_2 \text{ by Hölder inequality} \\ &\leq \|h\|_\infty \|g\|_2 . \end{aligned} \tag{3.35}$$

Since $g \in \mathcal{L}^2([0,1]^n)$ and h was arbitrary, $F_{\mathbf{w}}$ is indeed bounded by $\|g\|_2$.

Step 2 ($F_{\mathbf{w}}$ vanishes on indicators of Borel sets in $\mathbb{R}$.). We begin by proving $F_{\mathbf{w}}$ vanishes on indicator functions of intervals in $\mathbb{R}$. Consider the indicator function $\chi_{(b,\infty)}$, for $b \in \mathbb{R}$. Then

$$\begin{aligned} F_{\mathbf{w}}(\chi_{(b,\infty)}) &= \int_{[0,1]^n} \chi_{(b,\infty)}(\langle \mathbf{w}, \mathbf{x} \rangle) g(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\{\mathbf{x} \in [0,1]^n : b < \langle \mathbf{w}, \mathbf{x} \rangle < \infty\}} g(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\Pi^+_{\mathbf{w},-b}} g(\mathbf{x})\, d\mathbf{x} \\ &= 0. \qquad \text{by 3.34} \end{aligned} \tag{3.36}$$

Consider the indicator function $\chi_{[b,\infty)}$, for $b \in \mathbb{R}$. Then

$$\begin{aligned} F_{\mathbf{w}}(\chi_{[b,\infty)}) &= \int_{[0,1]^n} \chi_{[b,\infty)}(\langle \mathbf{w}, \mathbf{x} \rangle) g(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\{\mathbf{x} \in [0,1]^n : b \leq \langle \mathbf{w}, \mathbf{x} \rangle < \infty\}} g(\mathbf{x})\, d\mathbf{x} \\ &= \int_{\Pi_{\mathbf{w},-b}} g(\mathbf{x})\, d\mathbf{x} + \int_{\Pi^+_{\mathbf{w},-b}} g(\mathbf{x})\, d\mathbf{x} \\ &= 0. \qquad \text{by 3.34} \end{aligned} \tag{3.37}$$

To show $F_{\mathbf{w}}(\chi_{[b,\infty)}) = 0$, we also applied the fact that $\Pi_{\mathbf{w},-b}$ is a set of Lebesgue measure zero. Hence $\int_{\Pi_{\mathbf{w},-b}} g(\mathbf{x})\, d\mathbf{x} = 0$. Note that for every $a, b \in \mathbb{R}$,

$$\chi_{(a,b)} = \chi_{(a,\infty)} - \chi_{[b,\infty]}. \tag{3.38}$$

Applying linearity of $F_{\mathbf{w}}$ to 3.38 and together with 3.36 and 3.37 yields

$$F_{\mathbf{w}}(\chi_{(a,b)}) = F_{\mathbf{w}}(\chi_{(a,\infty)} - \chi_{[b,\infty)}) = F_{\mathbf{w}}(\chi_{(a,\infty)}) - F_{\mathbf{w}}(\chi_{[b,\infty)}) = 0. \tag{3.39}$$

We claim that $F_{\mathbf{w}}(\chi_B) = 0$, for every Borel set $B \subseteq \mathbb{R}$. To show that $F_{\mathbf{w}}$ vanishes on indicator functions of Borel sets, we will appeal to Dynkin's $\lambda - \pi$ theorem. Define the collections Π and Λ by

$$\Pi = \{(a,b) : -\infty \leq a \leq b \leq \infty\} \text{ and } \Lambda = \{A : A \in \mathcal{B}(\mathbb{R}) \text{ and } F_{\mathbf{w}}(\chi_A) = 0\}.$$

Since the finite intersection of open intervals is again an open, possibly degenerate interval, Π is a π-system. We will show that Λ is a λ-system. By 3.39, Λ contains Π. Clearly, $\mathbb{R} \in \Lambda$.

Suppose that $A, B \in \Lambda$ where $B \subseteq A$. Then $\chi_{A\setminus B} = \chi_A - \chi_B$ so $F_{\mathbf{w}}(\chi_{A\setminus B}) = F_{\mathbf{w}}(\chi_A - \chi_B) = F_{\mathbf{w}}(\chi_A) - F_{\mathbf{w}}(\chi_B) = 0$, since $A, B \in \Lambda$. Thus $A \setminus B \in \Lambda$. Suppose that $\{B_n\}_{n=1}^{\infty}$ is a collection of disjoint sets in Λ. We will show that $B = \bigcup_{n=1}^{\infty} B_n \in \Lambda$. Clearly, $\chi_B = \sum_{k=1}^{\infty} \chi_{B_k}$ so $\sum_{k=1}^{m} \chi_{B_k} \uparrow \chi_B$, as $m \to \infty$. Then

$$\begin{aligned}
F_{\mathbf{w}}(\chi_B) &= \int_{[0,1]^n} \chi_B(\langle \mathbf{w}, \mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} \\
&= \int_{[0,1]^n} \sum_{k=1}^{\infty} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} \\
&= \int_{[0,1]^n} \lim_{m\to\infty} \sum_{k=1}^{m} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} \\
&= \lim_{m\to\infty} \int_{[0,1]^n} \sum_{k=1}^{m} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} \text{ by Monotone Convergence Theorem} \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} \int_{[0,1]^n} \chi_{B_k}(\langle \mathbf{w}, \mathbf{x}\rangle) g(\mathbf{x})\, d\mathbf{x} \\
&= \lim_{m\to\infty} \sum_{k=1}^{m} F_{\mathbf{w}}(\chi_{B_k}) = 0.
\end{aligned}$$

Thus $B \in \Lambda$. Hence Λ is indeed a λ-system.

We will show that $\mathcal{B}(\mathbb{R}) = \Lambda$. By Dynkin's $\lambda - \pi$ theorem, $\mathcal{B}(\mathbb{R}) = \sigma(\Pi) = \lambda(\Pi)$. Recall that the λ-system generated by Π, denoted by $\lambda(\Pi)$, is the smallest λ-system on $\mathbb{R}$ containing Π. Since Λ is also a λ-system containing Π, we deduce $\lambda(\Pi) \subseteq \Lambda$. Since $\lambda(\Pi) = \mathcal{B}(\mathbb{R})$, we have that $\mathcal{B}(\mathbb{R}) \subseteq \Lambda$. By construction, $\Lambda \subseteq \mathcal{B}(\mathbb{R})$. Hence $\mathcal{B}(\mathbb{R}) = \Lambda$, as desired.

Step 3 ($F_{\mathbf{w}}$ vanishes on measurable simple functions). Suppose that φ is a $\mathcal{B}(\mathbb{R})$-measurable simple function. Without loss of generality, $\varphi = \sum_{k=1}^{m} \alpha_k \chi_{A_k}$, where A_k are disjoint $\mathcal{B}(\mathbb{R})$-measurable sets. By linearity of $F_{\mathbf{w}}$ and by Step 2,

$$F_{\mathbf{w}}(\varphi) = F\left(\sum_{k=1}^{m} \alpha_k \chi_{A_k}\right) = \sum_{k=1}^{m} F_{\mathbf{w}}(\alpha_k \chi_{A_k}) = \sum_{k=1}^{m} \alpha_k F_{\mathbf{w}}(\chi_{A_k}) = 0. \tag{3.40}$$

Step 4 ($F_{\mathbf{w}}$ vanishes on $\mathcal{L}^\infty(\mathbb{R})$). Let $f \in \mathcal{L}^\infty(\mathbb{R})$. By Density of simple functions in $\mathcal{L}^p$, there exists a sequence of $\mathcal{B}(\mathbb{R})$-measurable simple functions $\{\varphi_m\}_{m=1}^\infty$ converging to f in $\|\cdot\|_\infty$ as $m \to \infty$. For every $m \in \mathbb{N}$, $f - \varphi_m \in \mathcal{L}^\infty(\mathbb{R})$. Without loss of generality, $|f - \varphi_m| \leq \|f - \varphi_m\|_\infty$. Then

$$\begin{aligned}
|F_{\mathbf{w}}(f) - F_{\mathbf{w}}(\varphi_m)| &\leq \left| \int_{[0,1]^n} (f - \varphi_m)(\langle \mathbf{w}, \mathbf{x} \rangle) g(\mathbf{x})\, d\mathbf{x} \right| \\
&\leq \int_{[0,1]^n} |(f - \varphi_m)(\langle \mathbf{w}, \mathbf{x} \rangle)||g(\mathbf{x})|\, d\mathbf{x} \\
&\leq \|f - \varphi_m\|_\infty \int_{[0,1]^n} |g(\mathbf{x})|\, d\mathbf{x} \\
&\leq \|f - \varphi_m\|_\infty \left(\lambda([0,1]^n)\right)^{\frac{1}{2}} \|g\|_2 . \qquad \text{by Hölder inequality}
\end{aligned}$$

Since $\lim_{m\to\infty} \|f - \varphi_m\|_\infty = 0$ and $\|g\|_2 < \infty$, $|F_{\mathbf{w}}(f) - F_{\mathbf{w}}(\varphi_m)| \to 0$ as $m \to \infty$. Combining this result with 3.40 gives

$$F_{\mathbf{w}}(f) = \lim_{m\to\infty} F_{\mathbf{w}}(\varphi_m) = 0. \tag{3.41}$$

Since f was arbitrary, $F_{\mathbf{w}}$ vanishes on $\mathcal{L}^\infty(\mathbb{R})$.

Step 5 (g is zero almost everywhere.). We will compute the Fourier transform of g. Since cos and sin are bounded and measurable, $\cos, \sin \in \mathcal{L}^\infty(\mathbb{R})$. By 3.41, $F_{\mathbf{w}}(\cos) = F_{\mathbf{w}}(\sin) = 0$. However, this implies

$$\begin{aligned}
\widehat{g}(\mathbf{w}) &= \int_{[0,1]^n} e^{i\langle \mathbf{w}, \mathbf{x} \rangle} g(\mathbf{x})\, d\mathbf{x} \\
&= \int_{[0,1]^n} \cos(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mathbf{x} + i \int_{[0,1]^n} \sin(\langle \mathbf{w}, \mathbf{x} \rangle)\, d\mathbf{x} \\
&= F_{\mathbf{w}}(\cos) + i F_{\mathbf{w}}(\sin) \\
&= 0.
\end{aligned}$$

By Inclusion of $\mathcal{L}^p$ spaces of a finite measure, $g \in \mathcal{L}^1([0,1]^n)$. Since $\widehat{g} = 0$, by Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$, $g = 0$ almost everywhere.

■

As a corollary of Proposition Continuous sigmoidal functions are discriminatory, we have shown that the Heaviside step function and the logistic sigmoid were discriminatory in the sense of Definition 20. The natural question is whether those two activation functions remain $\mathcal{L}^2$-discriminatory. It turns out they do, and we will apply the Lemma 5 to justify that fact.

Lemma 6. *The Heaviside step function, denoted by s, is $\mathcal{L}^2$-discriminatory.*

Proof. Clearly, $0 \leq s \leq 1$. Assume that for $g \in \mathcal{L}^2([0,1]^n)$,

$$\int_{[0,1]^n} s(\langle \mathbf{w}, \mathbf{x} \rangle + b) g(\mathbf{x})\, d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}. \tag{3.42}$$

By definition of s and 3.42,

$$\int_{[0,1]^n} s(\langle \mathbf{w}, \mathbf{x} \rangle + b) g(\mathbf{x})\, d\mathbf{x} = \int_{\Pi^+_{\mathbf{w},-b}} g(\mathbf{x})\, d\mathbf{x} = 0.$$

By Lemma 5, $g = 0$ almost everywhere. Since g was arbitrary, s is $\mathcal{L}^2$-discriminatory. ■

Lemma 7. *The logistic sigmoid σ is $\mathcal{L}^2$-discriminatory.*

Proof Idea. We will present the argument very similar to the proof of Proposition 3. The main idea is to reduce the proof to an application of Lemma 5.

Proof. Clearly, $0 \leq \sigma \leq 1$. Assume that for $g \in \mathcal{L}^2([0,1]^n)$,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b) g(\mathbf{x})\, d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}. \tag{3.43}$$

Fix $\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$. For $\lambda \in \mathbb{R}$, define $\sigma_\lambda(\mathbf{x}) = \sigma(\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b))$. By 3.43,

$$\int_{[0,1]^n} \sigma_\lambda(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} = 0, \text{ for every } \lambda \in \mathbb{R}. \tag{3.44}$$

Define $\gamma : [0,1]^n \to \mathbb{R}$ by $\gamma(\mathbf{x}) = \lim_{\lambda \to \infty} \sigma_\lambda(\mathbf{x})$. Observe that

$$\gamma(\mathbf{x}) = \begin{cases} 1 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b > 0 \\ \frac{1}{2} & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \\ 0 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b < 0 \end{cases}. \tag{3.45}$$

Clearly, $\gamma \in \mathcal{L}^2([0,1]^n)$. We will show that $\sigma_\lambda \to \gamma$ in $\|\cdot\|_2$, as $\lambda \to \infty$. It is sufficient to show that $\|\sigma_\lambda - \gamma\|_2^2 \to 0$ as $\lambda \to \infty$. Note that

$$|\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2 = \begin{cases} \frac{1}{\left(1+e^{\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b)}\right)^2} & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b > 0 \\ \frac{1}{\left(1+e^{-\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b)}\right)^2} & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b < 0 \\ 0 & \text{if } \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \end{cases}.$$

Since $\Pi_{\mathbf{w},b}$ is a set of Lebesgue measure zero,

$$\begin{aligned} \int_{[0,1]^n} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} &= \int_{\Pi^+_{\mathbf{w},b}} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} + \int_{\Pi^-_{\mathbf{w},b}} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} \\ &= \int_{\Pi^+_{\mathbf{w},b}} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} + \int_{\Pi^+_{-\mathbf{w},-b}} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} \\ &= \int_{\Pi^+_{\mathbf{w},b}} \frac{1}{\left(1 + e^{\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b)}\right)^2}\, d\mathbf{x} + \int_{\Pi^+_{-\mathbf{w},-b}} \frac{1}{\left(1 + e^{-\lambda(\langle \mathbf{w}, \mathbf{x} \rangle + b)}\right)^2}\, d\mathbf{x}. \end{aligned} \tag{3.46}$$

For every λ, $\mathbf{x} \to \frac{1}{\left(1+e^{\lambda(\langle \mathbf{w},\mathbf{x}\rangle+b)}\right)^2}$ and $\mathbf{x} \to \frac{1}{\left(1+e^{\lambda(\langle -\mathbf{w},\mathbf{x}\rangle-b)}\right)^2}$ are bounded on $\Pi^+_{\mathbf{w},b}$, $\Pi^+_{-\mathbf{w},-b}$ respectively. Since $\lambda_{|[0,1]^n}$ is a finite measure, by Dominated Convergence Theorem,

$$\lim_{\lambda\to\infty} \int_{\Pi^+_{\mathbf{w},b}} \frac{1}{\left(1+e^{\lambda(\langle \mathbf{w},\mathbf{x}\rangle+b)}\right)^2}\, d\mathbf{x} = \int_{\Pi^+_{\mathbf{w},b}} \lim_{\lambda\to\infty} \frac{1}{\left(1+e^{\lambda(\langle \mathbf{w},\mathbf{x}\rangle+b)}\right)^2}\, d\mathbf{x} = 0, \tag{3.47}$$

and,

$$\lim_{\lambda\to\infty} \int_{\Pi^+_{-\mathbf{w},-b}} \frac{1}{\left(1+e^{-\lambda(\langle \mathbf{w},\mathbf{x}\rangle+b)}\right)^2}\, d\mathbf{x} = \int_{\Pi^+_{-\mathbf{w},-b}} \lim_{\lambda\to\infty} \frac{1}{\left(1+e^{-\lambda(\langle \mathbf{w},\mathbf{x}\rangle+b)}\right)^2}\, d\mathbf{x} = 0. \tag{3.48}$$

Taking $\lambda \to \infty$ on both sides of 3.46 and applying 3.47 and 3.48 gives

$$\lim_{\lambda\to\infty} \|\sigma_\lambda - \gamma\|_2^2 = \lim_{\lambda\to\infty} \int_{[0,1]^n} |\sigma_\lambda(\mathbf{x}) - \gamma(\mathbf{x})|^2\, d\mathbf{x} = 0. \tag{3.49}$$

Since $\Pi_{\mathbf{w},b}$ is a set of Lebesgue measure zero, by 3.45,

$$\int_{[0,1]^n} \gamma(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} = \int_{\Pi^+_{\mathbf{w},b}} g(\mathbf{x})\, d\mathbf{x}. \tag{3.50}$$

By Hölder inequality,

$$\begin{aligned}\left| \int_{[0,1]^n} \gamma(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} - \int_{[0,1]^n} \sigma_\lambda(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} \right| &\le \int_{[0,1]^n} |\gamma(\mathbf{x}) - \sigma_\lambda(\mathbf{x})|\, |g(\mathbf{x})|\, d\mathbf{x} \\ &\le \|\sigma_\lambda - \gamma\|_2 \cdot \|g\|_2 . \end{aligned} \tag{3.51}$$

Since $g \in \mathcal{L}^2([0,1]^n)$, taking $\lambda \to \infty$ on both sides of 3.51 and applying 3.49 and 3.44 gives

$$\int_{\Pi^+_{\mathbf{w},b}} g(\mathbf{x})\, d\mathbf{x} = \int_{[0,1]^n} \gamma(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} = \lim_{\lambda\to\infty} \int_{[0,1]^n} \sigma_\lambda(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} = 0. \tag{3.52}$$

By Lemma 5, $g = 0$ almost everywhere. ■

3.4.2 The Universal Approximation Theorem for $\mathcal{L}^2([0,1]^n)$

Theorem 13 (The Universal Approximation Theorem for square-integrable functions)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any $\mathcal{L}^2([0,1]^n)$-discriminatory activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x}\rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^2([0,1]^n)$.

Proof Idea. We will adapt the proof of The Universal Approximation Theorem for continuous functions. We will argue by contradiction, assuming $\mathcal{H}_\sigma$ is not dense in $\mathcal{L}^2([0,1]^n)$. By Separation functional lemma, this assumption implies existence of a non-trivial "separation" functional L, vanishing on $\mathcal{H}_\sigma$. Riesz Representation Theorem for the Dual of $\mathcal{L}^p$ will help us reveal the structure of the functional L. Using the fact σ is $\mathcal{L}^2([0,1]^n)$-discriminatory activation function, we will obtain the contradiction.

Proof. Firstly, we argue that $\mathcal{H}_\sigma$ is actually in $\mathcal{L}^2([0,1]^n)$. By definition of $\mathcal{L}^2([0,1]^n)$-discriminatory activation function, σ is bounded. Since $\lambda_{|[0,1]^n}$ is a finite measure, $\mathcal{H}_\sigma \subset \mathcal{L}^2([0,1]^n)$. Assume, for the sake of contradiction, that $\mathcal{H}_\sigma$ is not dense in $\mathcal{L}^2([0,1]^n)$. Since $\mathcal{L}^2([0,1]^n)$ is a normed linear space, $\mathcal{H}_\sigma$ is a vector subspace of $\mathcal{L}^2([0,1]^n)$. By Separation functional lemma, there exists a bounded linear functional L on $\mathcal{L}^2([0,1]^n)$ such that $L \neq 0$ on $\mathcal{L}^2([0,1]^n)$ and $L_{|\mathcal{H}_\sigma} = 0$. By Riesz Representation Theorem for the Dual of $\mathcal{L}^p$, there exists $g \in \mathcal{L}^2([0,1]^n)$ such that

$$L(f) = \int_{[0,1]^n} f(\mathbf{x})g(\mathbf{x})\,d\mathbf{x}, \text{ for every } f \in \mathcal{L}^2([0,1]^n). \tag{3.53}$$

Moreover, $\|L\| = \|g\|_2$. Since $L_{|\mathcal{H}_\sigma} = 0$, by 3.53,

$$\int_{[0,1]^n} h(\mathbf{x})g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } h \in \mathcal{H}_\sigma. \tag{3.54}$$

By 3.54,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n,\, b \in \mathbb{R}. \tag{3.55}$$

Since σ is $\mathcal{L}^2([0,1]^n)$-discriminatory activation function, 3.55 implies $g = 0$ almost everywhere. By Proposition 16, $\|g\|_2 = 0$. However, this implies that $\|L\| = 0$. We conclude L must be identically zero. But this is a contradiction since $L \neq 0$. Hence $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^2([0,1]^n)$, as desired. ■

As a corollary of Theorem 13, we present the following result addressing the logistic sigmoid.

Corollary 5. *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with the logistic sigmoid activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x}\rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^2([0,1]^n)$.

Proof. By Lemma 7, σ is $\mathcal{L}^2([0,1]^n)$-discriminatory activation function. The result follows directly from Theorem 13. ■

3.5 Universal approximation of integrable functions

In this section, we will outline the generalization of the universal approximation results from $\mathcal{L}^2([0,1]^n)$ to $\mathcal{L}^1([0,1]^n)$. Since the majority of results developed for $\mathcal{L}^2([0,1]^n)$ translate to $\mathcal{L}^1([0,1]^n)$ with almost no change, we will focus on the notion of $\mathcal{L}^1$-discriminatory activation function and The Universal Approximation Theorem for integrable functions.

3.5.1 $\mathcal{L}^1$-discriminatory activation functions

We will use a definition very similar to a definition of $\mathcal{L}^2([0,1]^n)$-discriminatory activation function.

Definition 25 ($\mathcal{L}^1$-discriminatory activation function)**.** Let $\sigma : \mathbb{R} \to [0,1]$. We say σ is $\mathcal{L}^1([0,1]^n)$-discriminatory if for every $g \in \mathcal{L}^\infty([0,1]^n)$,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b) g(\mathbf{x})\, d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}$$

implies $g = 0$ almost everywhere.

3.5.2 The Universal Approximation Theorem for $\mathcal{L}^1([0,1]^n)$

Theorem 14 (The Universal Approximation Theorem for integrable functions)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer, fully-connected neural networks with any $\mathcal{L}^1([0,1]^n)$-discriminatory activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^1([0,1]^n)$.

Proof Idea. We will adapt the proof of The Universal Approximation Theorem for square-integrable functions.

Proof. Firstly, we argue that $\mathcal{H}_\sigma$ is actually in $\mathcal{L}^1([0,1]^n)$. By definition of $\mathcal{L}^1([0,1]^n)$-discriminatory activation function, σ is bounded. Since $\lambda_{|[0,1]^n}$ is a finite measure, $\mathcal{H}_\sigma \subset \mathcal{L}^1([0,1]^n)$. Assume, for the sake of contradiction, that $\mathcal{H}_\sigma$ is not dense in $\mathcal{L}^1([0,1]^n)$. Since $\mathcal{L}^1([0,1]^n)$ is a normed linear space, $\mathcal{H}_\sigma$ is a vector subspace of $\mathcal{L}^1([0,1]^n)$. By Separation functional lemma, there exists a bounded linear functional L on $\mathcal{L}^1([0,1]^n)$ such that $L \neq 0$ on $\mathcal{L}^1([0,1]^n)$ and $L_{|\mathcal{H}_\sigma} = 0$. By Riesz Representation Theorem for the Dual of $\mathcal{L}^p$, there exists $g \in \mathcal{L}^\infty([0,1]^n)$ such that

$$L(f) = \int_{[0,1]^n} f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}, \text{ for every } f \in \mathcal{L}^1([0,1]^n) \text{ and } \|L\| = \|g\|_\infty. \quad (3.56)$$

Since $L_{|\mathcal{H}_\sigma} = 0$, by 3.56,

$$\int_{[0,1]^n} h(\mathbf{x})g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } h \in \mathcal{H}_\sigma. \tag{3.57}$$

By 3.57,

$$\int_{[0,1]^n} \sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)g(\mathbf{x})\,d\mathbf{x} = 0, \text{ for every } \mathbf{w} \in \mathbb{R}^n,\, b \in \mathbb{R}. \tag{3.58}$$

Since σ is $\mathcal{L}^1([0,1]^n)$-discriminatory activation function, 3.58 implies $g = 0$ almost everywhere. But then, $\|g\|_\infty = 0$. By 3.56, $\|L\| = 0$. We conclude L must be identically zero. But this is a contradiction since $L \neq 0$. Hence $\mathcal{H}_\sigma$ is dense in $\mathcal{L}^1([0,1]^n)$, as desired. ■

3.6 Universal approximation of measurable functions on compact sets

The focus in this section remains on the unit hypercube $[0,1]^n$ in $\mathbb{R}^n$. We will work with the measure space $([0,1]^n, \mathcal{B}([0,1]^n), \lambda_{|[0,1]^n})$, where $\lambda_{|[0,1]^n}$ is the restriction of Lebesgue measure on $\mathbb{R}^n$ to $[0,1]^n$. We begin the discussion about approximation of real-valued, measurable functions on $[0,1]^n$ with the following theorem.

Theorem 15 (Cybenko's Universal Approximation Theorem for measurable functions, [Cyb89])**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any continuous sigmoidal activation function, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x}\rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Suppose that $f : [0,1]^n \to \mathbb{R}$ is Borel measurable and let $\epsilon > 0$. There exists a network $h \in \mathcal{H}_\sigma$ and a set $K \subseteq [0,1]^n$ such that

$$\delta_\infty(f_{|K}, h) < \epsilon$$

where $\lambda(K) > 1 - \epsilon$.

Informally, for every Borel measurable function f, there exists a fully-connected neural network $h \in \mathcal{H}_\sigma$ approximating f to a desired error, except possibly on sets of arbitrarily small measure.

Proof Idea. We will apply Lusin's Theorem, Theorem 7.4.4 in [Coh13].

Theorem (Lusin's Theorem, Theorem 7.4.4 in [Coh13]). *Let X be a locally compact Hausdorff space and let $\mathcal{A}$ be a σ-algebra that includes $\mathcal{B}(X)$. Let μ be a regular measure on $(X, \mathcal{A})$ and suppose $f : X \to \mathbb{R}$ is measurable. If $A \in \mathcal{A}$ and satisfies $\mu(A) < \infty$ and if $\epsilon > 0$, then there is a compact set $K \subseteq A$ such that $\mu(A \setminus K) < \epsilon$ and $f_{|K}$ is continuous. Moreover, there is a function $g \in \mathcal{C}(X)$ that agrees with f on K.*

Lusin's Theorem will provide us with a compact set K and a continuous function $g : [0,1]^n \to \mathbb{R}$ which agrees with f on K. The result will follow from Cybenko's Universal Approximation Theorem, [Cyb89].

Proof. Since $[0,1]^n$ is a compact metric space, it is locally compact and Hausdorff. By Lusin's Theorem, Theorem 7.4.4 in [Coh13], there exists a compact set $K \subseteq [0,1]^n$ such that $\lambda([0,1]^n \setminus K) < \epsilon$ and a continuous function $g : [0,1]^n \to \mathbb{R}$ such that $f(\mathbf{x}) = g(\mathbf{x})$ for every $\mathbf{x} \in K$. By Cybenko's Universal Approximation Theorem, [Cyb89], there exists a neural network $h \in \mathcal{H}_\sigma$ such that $\delta_\infty(g,h) < \epsilon$. Since $f_{|K} = g$, we have $\delta_\infty(f_{|K}, h) < \epsilon$. Since $\lambda([0,1]^n \setminus K) < \epsilon$, $\lambda(K) > 1 - \epsilon$. ■

Remark 18. The consequences of this result are briefly discussed in subsection 3.7.5.

Remark 19. Cybenko's Universal Approximation Theorem for measurable functions, [Cyb89] also holds for any compact set in $\mathbb{R}^n$.

3.7 Universal approximation of measurable functions in probabilistic sense

3.7.1 Introduction

Until this section, we focused on the approximation power of neural networks on a compact domain. We focused on spaces of continuous functions and Lebesgue spaces. However, another practically important space is the space of measurable functions whose domain is often unrestricted. An example of such a space would be the space of Borel measurable functions from $\mathbb{R}^n$ to $\mathbb{R}$. Often, neural networks are used to directly or indirectly learn a probability distribution of some random variable. Since random variables are defined as measurable functions, it is sensible and important to investigate approximation properties that neural networks possess when the approximation space is the space of measurable functions.
In this section, we will focus on the space of all Borel measurable functions from $\mathbb{R}^n$ to $\overline{\mathbb{R}}$, denoted by $\mathcal{M}^n$. We will drop the assumption about the compactness of an approximation domain and consider the approximation on the entire $\mathbb{R}^n$, albeit in a probabilistic sense. Since measurable functions are not necessarily bounded, it is not immediately obvious how to equip $\mathcal{M}^n$ with a topology with respect to which we can discuss the approximation properties.
To address those issues, we will begin by introducing a few measure-theoretic assumptions that will enable us to construct a few useful metrics. In this section, we will focus on the approximation in the probabilistic sense, so let μ be a probability measure on the measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

μ-a.e. equivalence When discussing the approximation in a probabilistic sense, it is reasonable not to distinguish between measurable functions $f, g \in \mathcal{M}^n$ if they are μ-almost everywhere equivalent. Recall that two measurable functions $f, g \in \mathcal{M}^n$ are μ-almost everywhere equivalent if

$$\mu(\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \neq g(\mathbf{x})\}) = 0.$$

Furthermore, μ-almost everywhere equivalence is compatible with the fact that Lebesgue integral cannot distinguish between two almost everywhere equivalent functions. It is not difficult to show that μ-almost everywhere equivalence on $\mathcal{M}^n$ is in fact an equivalence relation. For the sake of simplicity, we will denote the space of equivalence classes under the μ-almost everywhere equivalence also by $\mathcal{M}^n$. This abuse of notation is analogous to one often used in literature when discussing $\mathcal{L}^p(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \lambda)$.

μ-a.e. finiteness Another important and mostly technical assumption is the assumption about μ-almost everywhere finiteness. Recall $f \in \mathcal{M}^n$ is μ-almost everywhere finite if

$$\mu(\{\mathbf{x} \in \mathbb{R}^n : |f(\mathbf{x})| = \infty\}) = 0.$$

In this section, we will assume that $\mathcal{M}^n$ consists of equivalence classes of μ-almost everywhere equivalent functions that are also μ-almost everywhere finite.

Under the assumptions above, we can equip $\mathcal{M}^n$ with a metric. A metric will help us formalize the notion of distance between measurable functions and the approximation error. There are many ways one can equip $\mathcal{M}^n$ with a metric. Since we want to discuss approximation in a probabilistic sense, we would like to link our metric to convergence in probability measure μ. An example of such a metric is δ_μ metric, where $\delta_\mu : \mathcal{M}^n \times \mathcal{M}^n \to \mathbb{R}$ is given by

$$\delta_\mu(f, g) = \inf\{\epsilon > 0 : \mu(\{\mathbf{x} \in \mathbb{R}^n : |f(\mathbf{x}) - g(\mathbf{x})| > \epsilon\}) < \epsilon\}.$$

By **μ-a.e. finiteness** assumption, $f - g$ can be infinite only on sets of probability measure zero. Another metric we will discuss and use in calculations is the ρ_μ metric, $\rho_\mu : \mathcal{M}^n \times \mathcal{M}^n \to \mathbb{R}$ given by

$$\rho_\mu(f, g) = \mathbb{E}[\min(|f - g|, 1)] = \int_{\mathbb{R}^n} \min(|f - g|, 1)\, d\mu.$$

Remark 20. Observe that since $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$ is a finite measure space, ρ_μ is finite for every $f, g \in \mathcal{M}^n$. In literature, δ_μ metric is also known as **Ky Fan metric**.

Remark 21. Proofs that δ_μ and ρ_μ are indeed metrics are discussed in detail in the subsection Metrics and modes of convergence.

Remark 22. In the subsequent sections, we will use the following shortened notation

$$\{|f - g| > \epsilon\} = \{\mathbf{x} \in \mathbb{R}^n : |f(\mathbf{x}) - g(\mathbf{x})| > \epsilon\}.$$

Also, we will denote by $f \wedge g$ the function $\min(f, g)$.

We will briefly discuss the connection between metrics δ_μ and ρ_μ to the convergence in probability measure μ. Firstly, recall the definition of convergence in measure μ.

Definition 26 (convergence in measure)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $\{f_m\}_{m=1}^{\infty}$ be a sequence of $\mathcal{F}$-measurable, $\overline{\mathbb{R}}$-valued functions and suppose that f is $\mathcal{F}$-measurable. We say that $f_m \to f$ in measure μ if for every $\epsilon > 0$,

$$\mu(\{\omega \in \Omega : |f_m(\omega) - f(\omega)| > \epsilon\}) \to 0 \text{ as } m \to \infty.$$

Interestingly, δ_μ and ρ_μ are equivalent metrics. Consequently, they induce the same topology on $\mathcal{M}^n$. This fact is stated precisely as Characterization of convergence in probability, Lemma 2.1 in [HSW89], stated below.

Proposition (Characterisation of convergence in probability, Lemma 2.1 in [HSW89])**.** *Let $\{f_m\}_{m=1}^{\infty}$ be a sequence of functions in $\mathcal{M}^n$ and let $f \in \mathcal{M}^n$. Then the following statements are equivalent:*

(a) $\delta_\mu(f_m, f) \to 0$ *as* $m \to \infty$*;*
(b) $\mu(\{|f_m - f| > \epsilon\}) \to 0$ *as* $m \to \infty$*;*
(c) $\rho_\mu(f_m, f) \to 0$ *as* $m \to \infty$*.*

Remark 23. As a consequence of Characterization of convergence in probability, Lemma 2.1 in [HSW89], to prove convergence in probability measure μ, it is sufficient and in fact equivalent to prove convergence in either δ_μ or ρ_μ. Thanks to the existence of δ_μ and ρ_μ, the arguments addressing the convergence in measure μ will be much easier to read and understand. They will often reduce to exact computation or an estimation of the integral ρ_μ.

The main goal of this section is the statement and the proof of The Probabilistic Universal Approximation Theorem, stated below.

Theorem (The Probabilistic Universal Approximation Theorem)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any continuous sigmoidal activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is δ_μ-dense in $\mathcal{M}^n$.

To prove this theorem, we will need to develop a few auxiliary results, organised in the subsections described below.

Metrics and modes of convergence In this subsection, we will prove that δ_μ and ρ_μ are indeed metrics on $\mathcal{M}^n$. Apart from this, we will link convergence on compacta to convergence in measure μ via Proposition 8.

Towards the Probabilistic Universal Approximation Theorem In this subsection, we will build upon concepts developed in Metrics and modes of convergence. We will begin this subsection by discussing density of neural networks on compacta, in the sense of Theorem 16. To prove Theorem 16, we will apply Cybenko's Universal Approximation Theorem, [Cyb89]. The key result in this subsection is the Proposition 9, which guarantees that $\mathcal{C}(\mathbb{R}^n)$ is ρ_μ-dense in $\mathcal{M}^n$.

The Probabilistic Universal Approximation Theorem In this subsection, we will prove the The Probabilistic Universal Approximation Theorem. The key idea will be to combine density of neural networks on compacta, in the sense of Theorem 16 with the Proposition 9.

Relationship between measurable functions and classification In this subsection, we will discuss the application of Cybenko's Universal Approximation Theorem for measurable functions, [Cyb89] to classification problems.

3.7.2 Metrics and modes of convergence

We begin this subsection with a proof that δ_μ is indeed a metric on $\mathcal{M}^n$.

Proposition 5. *δ_μ is indeed a metric on $\mathcal{M}^n$.*

Proof Idea. The proof is essentially a verification of properties a metric must satisfy. The key observation is that the statement $\delta_\mu(f,g)=0$ is equivalent to $f=g$ μ-almost everywhere. The justification of triangle inequality is not difficult, but relatively technical.

Proof. Clearly, $\delta_\mu \geq 0$. Symmetry of δ_μ is obvious.

Step 1 ($\delta_\mu(f,g)=0 \iff f=g$ μ-almost everywhere). Suppose that $f=g$ μ-almost everywhere. We will show $\delta_\mu(f,g)=0$. Let $\epsilon>0$. Since $f=g$ μ-almost everywhere, $|f-g|=0$ μ-almost everywhere. Hence, there exists a set $M\in\mathcal{B}(\mathbb{R}^n)$ such that $\mu(M)=0$ and $|f-g|>0$ on M, while $|f-g|=0$ on $\mathbb{R}^n\setminus M$. Since $\epsilon>0$, by subadditivity of measure μ,

$$\mu(\{|f-g|>\epsilon\}) \leq \mu(\{|f-g|>0\}) \leq \mu(M)=0<\epsilon. \tag{3.59}$$

Since δ_μ is an infimum, by 3.59, $0\leq\delta_\mu(f,g)<\epsilon$. Since ϵ was arbitrary, we conclude $\delta_\mu(f,g)=0$.
Conversely, suppose that $\delta_\mu(f,g)=0$. We have

$$\{f\neq g\}=\{|f-g|>0\}\subseteq\bigcup_{n=1}^{\infty}\left\{|f-g|>\frac{1}{n}\right\}. \tag{3.60}$$

For every $n\in\mathbb{N}$, since $\delta_\mu(f,g)<\frac{1}{n}$ and $\delta_\mu(f,g)$ is an infimum, there exists $0<\epsilon_n<\frac{1}{n}$ such that $\mu(\{|f-g|>\epsilon_n\})<\epsilon_n$. Since $\epsilon_n<\frac{1}{n}$, $\left\{|f-g|>\frac{1}{n}\right\}\subseteq\{|f-g|>\epsilon_n\}$ and so

$$\mu\left(\left\{|f-g|>\frac{1}{n}\right\}\right)\leq\mu(\{|f-g|>\epsilon_n\})<\epsilon_n<\frac{1}{n}. \tag{3.61}$$

By 3.61, $\lim_{n\to\infty}\mu\left(\left\{|f-g|>\frac{1}{n}\right\}\right)=0$. Observe that

$$\left\{|f-g|>\frac{1}{n}\right\}\subseteq\left\{|f-g|>\frac{1}{n+1}\right\},\ \text{for every } n\in\mathbb{N}. \tag{3.62}$$

By 3.62,

$$\mu\left(\bigcup_{n=1}^{\infty}\left\{|f-g|>\frac{1}{n}\right\}\right)=\lim_{n\to\infty}\mu(\left\{|f-g|>\frac{1}{n}\right\})=0.$$

By 3.60, $\mu(\{f\neq g\})=0$.

Step 2 (The triangle inequality). To prove triangle inequality, let $f,g,h\in\mathcal{M}^n$. Let $\alpha>0$. Since $\delta_\mu(f,h)$ and $\delta_\mu(h,g)$ are infima, there exist $r_1,r_2>0$ such that

$$\delta_\mu(f,h)\leq r_1<\delta_\mu(f,h)+\frac{\alpha}{2}\text{ and }\delta_\mu(h,g)\leq r_2<\delta_\mu(h,g)+\frac{\alpha}{2},\tag{3.63}$$

satisfying

$$\mu(\{|f-h|>r_1\})<r_1\text{ and }\mu(\{|h-g|>r_2\})<r_2.\tag{3.64}$$

For every $\mathbf{x}\in\mathbb{R}^n$, if $|f(\mathbf{x})-h(\mathbf{x})|\leq r_1$ and $|h(\mathbf{x})-g(\mathbf{x})|\leq r_2$ then

$$|f(\mathbf{x})-g(\mathbf{x})|\leq|f(\mathbf{x})-h(\mathbf{x})|+|h(\mathbf{x})-g(\mathbf{x})|\leq r_1+r_2.$$

By contrapositive,

$$\{|f-g|>r_1+r_2\}\subseteq\{|f-h|>r_1\}\cup\{|h-g|>r_2\}.\tag{3.65}$$

Applying subadditivity of μ to 3.65 gives

$$\mu(\{|f-g|>r_1+r_2\})\leq\mu(\{|f-h|>r_1\})+\mu(\{|h-g|>r_2\}).\tag{3.66}$$

Applying 3.64 and 3.63 to 3.66 gives

$$\mu(\{|f-g|>r_1+r_2\})\leq r_1+r_2<\delta_\mu(f,h)+\delta_\mu(h,g)+\alpha.\tag{3.67}$$

Since $\delta_\mu(f,g)$ is an infimum, by 3.67, $\delta_\mu(f,g)\leq r_1+r_2<\delta_\mu(f,h)+\delta_\mu(h,g)+\alpha$. Since α was arbitrary, $\delta_\mu(f,g)\leq\delta_\mu(f,h)+\delta_\mu(h,g)$, as desired.

■

We continue with a proof that ρ_μ is indeed a metric on $\mathcal{M}^n$.

Proposition 6. *ρ_μ is indeed a metric on $\mathcal{M}^n$.*

Proof. The symmetry and non-negativity of ρ_μ are obvious.

Step 1. The fact $\rho_\mu(f,g)=0$ is equivalent to $f=g$ μ-almost everywhere follows directly from properties of Lebesgue integral of a non-negative function, namely Proposition 16.

Step 2 (The triangle inequality). Let $f,g,h\in\mathcal{M}^n$. By triangle inequality on $\mathbb{R}$,

$$|f(\mathbf{x})-g(\mathbf{x})|\leq|f(\mathbf{x})-h(\mathbf{x})|+|h(\mathbf{x})-g(\mathbf{x})|,\text{ for every }\mathbf{x}\in\mathbb{R}^n.\tag{3.68}$$

By 3.68,

$$\begin{aligned} |f(\mathbf{x}) - g(\mathbf{x})| \wedge 1 &\leq (|f(\mathbf{x}) - h(\mathbf{x})| + |h(\mathbf{x}) - g(\mathbf{x})|) \wedge 1, \text{ for every } \mathbf{x} \in \mathbb{R}^n, \\ &\leq |f(\mathbf{x}) - h(\mathbf{x})| \wedge 1 + |h(\mathbf{x}) - g(\mathbf{x})| \wedge 1, \text{ for every } \mathbf{x} \in \mathbb{R}^n. \end{aligned} \tag{3.69}$$

Integrating the inequality 3.69 gives

$$\begin{aligned} \rho_\mu(f, g) = \int_{\mathbb{R}^n} |f - g| \wedge 1 \, d\mu &\leq \int_{\mathbb{R}^n} |f - h| \wedge 1 \, d\mu + \int_{\mathbb{R}^n} |h - g| \wedge 1 \, d\mu \\ &\leq \rho_\mu(f, h) + \rho_\mu(h, g). \end{aligned}$$

■

The following proposition will provide us with a characterization of convergence in μ in terms of convergence in metrics δ_μ and ρ_μ.

Proposition 7 (Characterization of convergence in probability, Lemma 2.1 in [HSW89])**.** *Let $\{f_m\}_{m=1}^\infty$ be a sequence of functions in $\mathcal{M}^n$ and let $f \in \mathcal{M}^n$. Then the following statements are equivalent:*

(a) $\delta_\mu(f_m, f) \to 0$ as $m \to \infty$;
(b) For every $\epsilon > 0$, $\mu(\{|f_m - f| > \epsilon\}) \to 0$ as $m \to \infty$;
(c) $\rho_\mu(f_m, f) \to 0$ as $m \to \infty$.

Proof Idea. It is reasonable to attempt proving the implication chain $(a) \implies (b) \implies (c) \implies (a)$. However, it turns out that proving $(c) \implies (a)$ is relatively difficult, so we follow a slightly different method. The strategy is to prove that (a) and (b) are equivalent and separately prove that (b) and (c) are equivalent. The proof will be divided in four steps where each step will be a proof a single implication.

Proof.

Step 1 $(a \implies b)$. Suppose $\delta_\mu(f_m, f) \to 0$ as $m \to \infty$. Let $\epsilon > 0$. Then there exists $M \in \mathbb{N}$ such that

$$\delta_\mu(f_m, f) = \inf\{\alpha > 0 : \mu(\{|f_m - f| > \alpha\}) < \alpha\} < \epsilon, \text{ for } m \geq M. \tag{3.70}$$

By definition of δ_μ, there exists $0 < \alpha < \epsilon$ such that $\mu(\{|f_m - f| > \alpha\}) < \alpha < \epsilon$, for $m \geq M$. Since $0 < \alpha < \epsilon$, $\{|f_m - f| > \epsilon\} \subseteq \{|f_m - f| > \alpha\}$, so we have

$$\mu(\{|f_m - f| > \epsilon\}) \leq \mu(\{|f_m - f| > \alpha\}) < \alpha \leq \epsilon, \text{ for } m \geq M.$$

Step 2 $(b \implies a)$. Suppose that for every ϵ, $\mu(\{|f_m - f| \geq \epsilon\}) \to 0$ as $m \to \infty$. Then fix $\epsilon > 0$. Now there exists $M \in \mathbb{N}$ such that

$$\mu(\{|f_m - f| > \epsilon\}) < \epsilon, \text{ for } m \geq M. \tag{3.71}$$

Since δ_μ is an infimum, by 3.71, $\delta_\mu(f_m, f) \leq \epsilon$, for $m \geq M$.

Step 3 ($b \implies c$). Suppose that for every $\epsilon > 0, \mu(\{|f_m - f| > \epsilon\}) \to 0$ as $m \to \infty$. Then there exists $M \in \mathbb{N}$ such that

$$\mu(\left\{|f_m - f| > \frac{\epsilon}{2}\right\}) < \frac{\epsilon}{2}, \text{ for every } m \geq M. \tag{3.72}$$

We have

$$\begin{aligned}
\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu &= \int_{\left\{|f_m - f| \leq \frac{\epsilon}{2}\right\}} |f_m - f| \wedge 1 \, d\mu + \int_{\left\{|f_m - f| > \frac{\epsilon}{2}\right\}} |f_m - f| \wedge 1 \, d\mu \\
&\leq \int_{\mathbb{R}^n} \frac{\epsilon}{2} \, d\mu + \int_{\left\{|f_m - f| > \frac{\epsilon}{2}\right\}} 1 \, d\mu \\
&\leq \frac{\epsilon}{2}\mu(\mathbb{R}^n) + \mu(\left\{|f_m - f| > \frac{\epsilon}{2}\right\}) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \text{ for } m \geq M, \text{ by 3.72.}
\end{aligned}$$

Hence $\rho_\mu(f_m, f) < \epsilon$, for $m \geq M$.

Step 4 ($c \implies b$). Suppose $\rho_\mu(f_m, f) \to 0$ as $m \to \infty$, so $\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu \to 0$ as $m \to \infty$. Let $\epsilon > 0$. To prove that $\mu(\{|f_m - f| > \epsilon\}) \to 0$ as $m \to \infty$, it is sufficient to show that every subsequence $\mu(\{|f_{m_k} - f| > \epsilon\})$ has a subsubsequence $\mu(\{|f_{m_{k_l}} - f| > \epsilon\})$ such that $\mu(\{|f_{m_{k_l}} - f| > \epsilon\}) \to 0$ as $l \to \infty$. Thus, let $\{f_{m_k}\}_{k=1}^{\infty}$ be an arbitrary subsequence of $\{f_m\}_{m=1}^{\infty}$. By Markov's Inequality,

$$\mu(\{|f_{m_k} - f| \wedge 1 > \epsilon\}) \leq \frac{1}{\epsilon} \int_{\mathbb{R}^n} |f_{m_k} - f| \wedge 1 \, d\mu. \tag{3.73}$$

Since $\{f_{m_k}\}_{k=1}^{\infty}$ is a subsequence of $\{f_m\}_{m=1}^{\infty}$,

$$\lim_{k\to\infty} \int_{\mathbb{R}^n} |f_{m_k} - f| \wedge 1 \, d\mu = \lim_{m\to\infty} \int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu. \tag{3.74}$$

Taking $\lim_{k\to\infty}$ on both sides of 3.73 and applying 3.74 gives

$$\lim_{k\to\infty} \mu(\{|f_{m_k} - f| \wedge 1 > \epsilon\}) = 0. \tag{3.75}$$

Since $|f_{m_k} - f| \wedge 1 \to 0$ in μ by 3.75, by Proposition 3.1.3 in [Coh13], there exists a subsubsequence $|f_{m_{k_l}} - f| \wedge 1$ converging to 0 μ-almost everywhere. Thus, $|f_{m_{k_l}} - f| \to 0$ μ-almost everywhere. By Proposition 3.1.2 in [Coh13], $|f_{m_{k_l}} - f|$ converges to 0 in μ. Therefore, $\mu(\{|f_{m_{k_l}} - f| > \epsilon\} \to 0$ as $l \to \infty$.

■

Remark 24. Consequently, the convergence in probability measure is metrisable. By Proposition 7, metrics δ_μ and ρ_μ are equivalent. Moreover, they characterize the convergence in probability measure and induce the same topology on $\mathcal{M}^n$.

Before stating the next result, we will introduce the notation for closed balls in $\mathbb{R}^n$. Let $\mathbf{a} \in \mathbb{R}^n$ and $r > 0$. We denote the closed ball of radius r centred at $\mathbf{a}$ by $B(\mathbf{a}, r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{a}\| \leq r\}$. The following proposition provides a link between convergence on compacta and convergence in measure μ.

Proposition 8 (Uniform convergence on compacta implies convergence in μ)**.** *Let $\{f_m\}_{m=1}^{\infty}$ be a sequence of functions in $\mathcal{M}^n$ that converges uniformly to $f \in \mathcal{M}^n$ on compacta. In other words, for every compact set $K \subset \mathbb{R}^n$,*

$$\sup_{\mathbf{x}\in K}\{|f_m(\mathbf{x}) - f(\mathbf{x})|\} \to 0, \text{ as } m \to \infty.$$

Then $\delta_\mu(f_m, f) \to 0$ as $m \to \infty$.

Proof Idea. We will appeal to Characterization of convergence in probability, Lemma 2.1 in [HSW89]. Instead of directly proving $\delta_\mu(f_m, f) \to 0$ as $m \to \infty$, we will show $\rho_\mu(f_m, f) \to 0$ as $m \to \infty$. The main idea is to estimate the integral $\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu$ by integrating over a sufficiently large compact set K where we can control the absolute value of the integrand and the complement of such a set. We will find such a compact set by writing $\mathbb{R}^n$ as an increasing union of compact sets, namely closed balls. The desired compact set will arise from continuity of μ. After extracting the desired compact set, we will estimate above mentioned integrals and combine those estimates to complete the proof.

Proof. By Proposition 7, it is equivalent to show that $\rho_\mu(f_m, f) \to 0$ as $m \to \infty$. We need to show that

$$\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu \to 0 \text{ as } m \to \infty.$$

Let $\epsilon > 0$. We will estimate the integral $\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu$ by integrating over a sufficiently large compact set K and its complement. We begin by constructing such a compact set K. For $k > 0$, consider $B(\mathbf{0}, k)$. By Heine-Borel Theorem, $B(\mathbf{0}, k)$ is compact. Note that $\mathbb{R}^n = \bigcup_{k=1}^{\infty} B(\mathbf{0}, k)$. Since for every $k \in \mathbb{N}$, $B(\mathbf{0}, k) \subset B(\mathbf{0}, k+1)$, by continuity of the probability measure μ,

$$\mu(\mathbb{R}^n) = 1 = \lim_{k\to\infty} \mu(B(\mathbf{0}, k)). \tag{3.76}$$

By 3.76, there exists $k_0 \in \mathbb{N}$ such that

$$1 - \frac{\epsilon}{2} < \mu(B(\mathbf{0}, k)) \leq 1, \text{ for every } k \geq k_0.$$

Set $K = B(\mathbf{0}, k_0)$. Then $\mu(K) > 1 - \frac{\epsilon}{2}$. Hence

$$\mu(\mathbb{R}^n \setminus K) = \mu(\mathbb{R}^n) - \mu(K) < 1 - \left(1 - \frac{\epsilon}{2}\right) = \frac{\epsilon}{2}. \tag{3.77}$$

Write

$$\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu = \int_K |f_m - f| \wedge 1 \, d\mu + \int_{\mathbb{R}^n \setminus K} |f_m - f| \wedge 1 \, d\mu. \tag{3.78}$$

Consider $\int_K |f_m - f| \wedge 1 \, d\mu$.

Since $\sup_{\mathbf{x}\in K}\{|f_m(\mathbf{x}) - f(\mathbf{x})\} \to 0$ as $m \to \infty$, there exists $M \in \mathbb{N}$ such that

$$\sup_{\mathbf{x}\in K}\{|f_m(\mathbf{x}) - f(\mathbf{x})\} < \frac{\epsilon}{2}, \text{ for every } m \geq M. \tag{3.79}$$

Applying 3.79 gives

$$\begin{aligned}\int_K |f_m - f| \wedge 1 \, d\mu \leq \int_K |f_m - f| \, d\mu \leq \int_K \sup_{\mathbf{x}\in K}\{|f_m(\mathbf{x}) - f(\mathbf{x})|\} \, d\mu \\ \leq \int_{\mathbb{R}^n} \frac{\epsilon}{2} \, d\mu = \frac{\epsilon}{2} \cdot \mu(\mathbb{R}^n) = \frac{\epsilon}{2}, \text{ for every } m \geq M. \end{aligned} \tag{3.80}$$

Consider $\int_{\mathbb{R}^n \setminus K} |f_m - f| \wedge 1 \, d\mu$. Applying 3.77 gives

$$\int_{\mathbb{R}^n \setminus K} |f_m - f| \wedge 1 \, d\mu \leq \int_{\mathbb{R}^n \setminus K} 1 \, d\mu = \mu(\mathbb{R}^n \setminus K) < \frac{\epsilon}{2}. \tag{3.81}$$

By 3.78 and 3.80, 3.77, we have

$$\int_{\mathbb{R}^n} |f_m - f| \wedge 1 \, d\mu < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \text{ for every } m \geq M.$$

■

3.7.3 Towards the Probabilistic Universal Approximation Theorem

We will begin with a discussion of the uniform density of single-layer fully-connected neural networks on compacta in $\mathcal{C}(\mathbb{R}^n)$.

Theorem 16. *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any continuous sigmoidal activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m \in \mathbb{R}, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is uniformly dense on compacta in $\mathcal{C}(\mathbb{R}^n)$.

Proof Idea. By Cybenko's Universal Approximation Theorem, [Cyb89], for a fixed compact set $K \subset \mathbb{R}^n$, $\mathcal{H}_\sigma$ is dense on $\mathcal{C}(K)$. However, we need to show that for every $f \in \mathcal{C}(\mathbb{R}^n)$, there exists a sequence $\{h_m\}_{m=1}^{\infty}$, $h_m \in \mathcal{H}_\sigma$ such that $h_m \to f$ on every compact set $K \subset \mathbb{R}^n$. Informally, for every $f \in \mathcal{C}(\mathbb{R}^n)$, we need to exhibit a **single sequence** $\{h_m\}_{m=1}^{\infty}$ converging to f uniformly on every compact set $K \subset \mathbb{R}^n$. We will fix $f \in \mathcal{C}(\mathbb{R}^n)$. The key observation is that $\mathbb{R}^n$ can be covered with a countable union of nested compact sets. Using the Cybenko's Universal Approximation Theorem, [Cyb89], for every compact set, we will exhibit a sequence of neural networks from $\mathcal{H}_\sigma$ converging to f uniformly on the compact set. Using the fact our compact sets are nested, we will apply diagonalisation argument to produce a single sequence of neural networks from $\mathcal{H}_\sigma$ converging to f uniformly on the every compact set.

Proof. Fix $f \in \mathcal{C}(\mathbb{R}^n)$. By Heine-Borel Theorem, for every $r > 0$, $B(\mathbf{0}, r)$ is compact. Observe that $f_{|B(\mathbf{0},r)} \in \mathcal{C}(B(\mathbf{0}, r))$, for every $r > 0$. Hence by Cybenko's Universal Approximation Theorem, [Cyb89], for every $r > 0$, there exists a sequence in $\mathcal{H}_\sigma$ denoted by $\{h_m^{(r)}\}_{m=1}^{\infty}$ such that $h_m^{(r)} \to f_{|B(\mathbf{0},r)}$ uniformly as $m \to \infty$. Hence $h_m^{(r)} \to f$ uniformly on $B(\mathbf{0}, r)$, as $m \to \infty$. Consider following sequences.

$$
\begin{array}{ccccccc}
\mathbf{h}_1^{(1)} & h_2^{(1)} & h_3^{(1)} & h_4^{(1)} & \dots & \to f, \text{ uniformly on } B(\mathbf{0}, 1) \\
h_1^{(2)} & \mathbf{h}_2^{(2)} & h_3^{(2)} & h_4^{(2)} & \dots & \to f, \text{ uniformly on } B(\mathbf{0}, 2) \\
h_1^{(3)} & h_2^{(3)} & \mathbf{h}_3^{(3)} & h_4^{(3)} & \dots & \to f, \text{ uniformly on } B(\mathbf{0}, 3) \\
h_1^{(4)} & h_2^{(4)} & h_3^{(4)} & \mathbf{h}_4^{(4)} & \dots & \to f, \text{ uniformly on } B(\mathbf{0}, 4) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots
\end{array}
$$

Clearly, for every $r > 0$, $B(\mathbf{0}, r) \subset B(\mathbf{0}, r + 1)$. Thus, for every $r > 0$,

$$
h_m^{(r)} \to f \text{ uniformly on every } B(\mathbf{0}, r'), \text{ as } m \to \infty, \text{ for every } r' \leq r. \tag{3.82}
$$

Define $\{h_m\}_{m=1}^{\infty}$ by $h_m = h_m^{(m)}$ for every $m \in \mathbb{N}$. Note that $\mathbb{R}^n = \bigcup_{r=1}^{\infty} B(\mathbf{0}, r)$. Let $K \subset \mathbb{R}^n$ be a compact set. Since $B(\mathbf{0}, r) \uparrow \mathbb{R}^n$ as $r \to \infty$, there exists some $r_0 \in \mathbb{N}$ such that

$$
K \subseteq B(\mathbf{0}, r_0) \subset B(\mathbf{0}, r_0 + 1) \subset B(\mathbf{0}, r_0 + 2) \dots \tag{3.83}
$$

Assume, without loss of generality, r_0 is the smallest such a natural number. By 3.83 and 3.82, $\{h_m\}_{m \geq r_0} \to f$ uniformly on $B(\mathbf{0}, r_0)$ as $m \to \infty$. Since $K \subseteq B(\mathbf{0}, r_0)$ and K was arbitrary, the proof is complete. ■

The following two lemmas are auxiliary results that will be help us establish the Proposition 9 - the most important result in this subsection.

Lemma 8. *Let $A \in \mathcal{B}(\mathbb{R}^n)$. Suppose that μ is a probability measure on $\mathcal{B}(\mathbb{R}^n)$. Then for every $\epsilon > 0$, there exist a closed set $F \in \mathcal{B}(\mathbb{R}^n)$ and a continuous function $g \in \mathcal{C}(\mathbb{R}^n)$ such that g and χ_A agree on F and $\mu(\mathbb{R}^n \setminus F) < \epsilon$.*

Proof Idea. The proof relies on the regularity of the probability measure μ and Urysohn lemma. By appealing to the regularity of μ, we will extract the compact set approximating A to desired accuracy in measure μ. The existence of a desired function g will follow from Urysohn lemma, Theorem 33.1 [Mun14], stated below.

Lemma (Urysohn lemma, Theorem 33.1 [Mun14]). *Let X be a normal space. Let A, B be disjoint closed subsets of X. There exists a continuous map $f : X \to [a, b]$ such that $f(x) = a$, for every $a \in A$ and $f(x) = b$, for every $b \in B$.*

We will apply Urysohn lemma, Theorem 33.1 [Mun14] to the compact set approximating A in measure μ.

Proof. Since μ is a probability measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, by Proposition 1.5.6 in [Coh13], μ is regular. Hence, there exist a compact set K and an open set U such that $K \subset A \subset U$ satisfying

$$\mu(A \setminus K) < \frac{\epsilon}{2} \text{ and } \mu(U \setminus A) < \frac{\epsilon}{2}. \tag{3.84}$$

Since $\mathbb{R}^n$ with the standard topology is a metric space, $\mathbb{R}^n$ is Hausdorff and normal. Since K is compact and $\mathbb{R}^n$ Hausdorff, K is closed. Since U is open, $\mathbb{R}^n \setminus U$ is closed. Since $K \subset U$, $K \cap (\mathbb{R}^n \setminus U) = \emptyset$. By Urysohn lemma, Theorem 33.1 [Mun14], there exists $g \in \mathcal{C}(\mathbb{R}^n)$ such that $0 \leq g \leq 1$ and $g = 1$ on K while $g = 0$ on $\mathbb{R}^n \setminus U$. Set $F = K \cup (\mathbb{R}^n \setminus U)$. Since K and $(\mathbb{R}^n \setminus U)$ are closed, F is closed. We will show that χ_A and g agree on F. Suppose that $\mathbf{x} \in F$. Since $K \cap (\mathbb{R}^n \setminus U) = \emptyset$, either $\mathbf{x} \in K$ or $\mathbf{x} \in (\mathbb{R}^n \setminus U)$. If $\mathbf{x} \in K$, then $\chi_K(\mathbf{x}) = 1 = g(\mathbf{x})$. If $\mathbf{x} \in (\mathbb{R}^n \setminus U)$, then $\chi_K(\mathbf{x}) = 0$ since $K \subset U$. Hence $\chi_K(\mathbf{x}) = 0 = g(\mathbf{x})$. Therefore, $\chi_{A_{|F}} = g$. Hence

$$\begin{aligned}
\mu(\mathbb{R}^n \setminus F) &= \mu((\mathbb{R}^n \setminus K) \cap U) = \mu(U \setminus K) \\
&= \mu(A \setminus K) + \mu(U \setminus A) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \text{ by 3.84.}
\end{aligned}$$

■

The following lemma is a natural generalization of Lemma 8 to simple, $\mathcal{B}(\mathbb{R}^n)$-measurable functions.

Lemma 9. *Let φ be a simple, $\mathcal{B}(\mathbb{R}^n)$-measurable function. Suppose that μ is a probability measure on $\mathcal{B}(\mathbb{R}^n)$. Then for every $\epsilon > 0$, there exist a closed set $F \in \mathcal{B}(\mathbb{R}^n)$ and a continuous function $g \in \mathcal{C}(\mathbb{R}^n)$ such that φ and g agree on F and $\mu(\mathbb{R}^n \setminus F) < \epsilon$.*

Proof. Write $\varphi = \sum_{k=1}^m a_k \chi_{A_k}$ where for every $k \in \{1, 2, \ldots, m\}$, $a_k \in \mathbb{R}$ and $A_k \in \mathcal{B}(\mathbb{R}^n)$. Without loss of generality, for every $i, j \in \{1, 2, \ldots, m\}, i \neq j$, $A_i \cap A_j = \emptyset$. By Lemma 8, for every $k \in \{1, 2, \ldots, m\}$, there exist a closed set $F_k \in \mathcal{B}(\mathbb{R}^n)$ and a continuous function $g_k \in \mathcal{C}(\mathbb{R}^n)$ such that

$$\text{for every } \mathbf{x} \in F, \chi_{A_k}(\mathbf{x}) = g_k(\mathbf{x}) \text{ and } \mu(\mathbb{R}^n \setminus F_k) < \frac{\epsilon}{m}. \tag{3.85}$$

Set $F = \bigcap_{k=1}^m F_k$. Define $g : \mathbb{R}^n \to \mathbb{R}$ by $g = \sum_{k=1}^m a_k g_k$. Since g is a sum of continuous functions, $g \in \mathcal{C}(\mathbb{R}^n)$. We claim that φ and g agree on F. Observe

$$\begin{aligned}
\mathbf{x} \in F &\iff x \in F_k, \text{ for every } k \in \{1, 2, \ldots, m\} \\
&\implies \chi_{A_k}(\mathbf{x}) = g_k(\mathbf{x}), \text{ for every } \mathbf{x} \in F_k, \text{ for every } k \in \{1, 2, \ldots, m\} \\
&\implies a_k \chi_{A_k}(\mathbf{x}) = a_k g_k(\mathbf{x}), \text{ for every } \mathbf{x} \in F_k, \text{ for every } k \in \{1, 2, \ldots, m\}
\end{aligned} \tag{3.86}$$

Since for every $k \in \{1, 2, \ldots, m\}$, $F \subseteq F_k$, by 3.86,

$$a_k \chi_{A_k}(\mathbf{x}) = a_k g_k(\mathbf{x}), \text{ for every } \mathbf{x} \in F, \text{ for every } k \in \{1, 2, \ldots, m\}. \tag{3.87}$$

Summing over 3.87 gives

$$\varphi(\mathbf{x}) = \sum_{k=1}^{m} a_k \chi_{A_k}(\mathbf{x}) = \sum_{k=1}^{m} a_k g_k(\mathbf{x}) = g(\mathbf{x}), \text{ for every } \mathbf{x} \in F. \tag{3.88}$$

It remains to prove that $\mu(\mathbb{R}^n \setminus F) < \epsilon$. By definition of F and 3.85,

$$\mu(\mathbb{R}^n \setminus F) = \mu\left(\bigcup_{k=1}^{m} (\mathbb{R}^n \setminus F_k)\right) \leq \sum_{k=1}^{m} \mu(\mathbb{R}^n \setminus F_k) < \sum_{k=1}^{m} \frac{\epsilon}{m} = \epsilon. \tag{3.89}$$

■

The following proposition is the most important result in this section and it will play essential role in the proof of The Probabilistic Universal Approximation Theorem.

Proposition 9. *$\mathcal{C}(\mathbb{R}^n)$ is ρ_μ-dense in $\mathcal{M}^n$.*

Proof Idea. The argument will be divided in four steps. We begin by showing that f is bounded, except possibly on a set of arbitrarily small measure. Using this fact, we will construct a bounded function $h \in \mathcal{M}^n$ which is sufficiently close to f in ρ_μ metric. Since h is a bounded measurable function, there exists a sequence of measurable simple functions $\{\varphi_m\}_{m=1}^{\infty}$ such that $\varphi_m \to h$ pointwise. This will enable us to pick a simple function φ which is sufficiently close to h in ρ_μ metric. Using Lemma 9, we will extract a continuous function g which is sufficiently close to φ in ρ_μ metric. We will complete the proof by putting those estimates together.

Proof. Let $f \in \mathcal{M}^n$ and let $\epsilon > 0$. We will show that there exists $g \in \mathcal{C}(\mathbb{R}^n)$ such that $\rho_\mu(f, g) < \epsilon$.

Step 1 (f is almost bounded). Firstly, we will argue that f is bounded, except possibly on a set of arbitrarily small measure. Recall that $f \in \mathcal{M}^n$ implies that f is μ-almost everywhere finite. This implies

$$\mu(\{|f| = \infty\}) = 0. \tag{3.90}$$

Observe that $\{|f| = \infty\} = \bigcap_{m=1}^{\infty} \{|f| > m\}$. Since $\{|f| > m\} = f^{-1}(-\infty, -m) \cup f^{-1}(m, \infty)$ and f is $\mathcal{B}(\mathbb{R}^n)$-measurable, $\{|f| > m\} \in \mathcal{B}(\mathbb{R}^n)$. Clearly, $\{|f| > m+1\} \subseteq \{|f| > m\}$, for every $m \in \mathbb{N}$. By continuity of μ and 3.90,

$$\mu(\{|f| = \infty\}) = 0 = \mu\left(\bigcap_{m=1}^{\infty} \{|f| > m\}\right) = \lim_{m \to \infty} \mu(\{|f| > m\}). \tag{3.91}$$

By 3.91, there exists $N \in \mathbb{N}$ such that

$$\mu(\{|f| > m\}) < \frac{\epsilon}{2}, \text{ for every } m \geq N. \tag{3.92}$$

By 3.92, in particular, $\mu(\{|f| > N\}) < \frac{\epsilon}{2}$.

Step 2 (Approximate f with a bounded function h). Define $h : \mathbb{R}^n \to \mathbb{R}$ by $h = f\chi_{\{|f| \leq N\}}$. Since $f \in \mathcal{M}^n$ and $\{|f| \leq N\}$ is measurable, $h \in \mathcal{M}^n$. By definition of h, we have

$$|f(\mathbf{x}) - h(\mathbf{x})| = \begin{cases} 0 & \text{if } |f(\mathbf{x})| \leq N \\ |f(\mathbf{x})| & \text{otherwise} \end{cases}. \tag{3.93}$$

$$\begin{aligned} \rho_\delta(f,h) = \int_{\mathbb{R}^n} |f-h| \wedge 1 \, d\mu &= \int_{\{|f| \leq N\}} |f-h| \wedge 1 \, d\mu + \int_{\{|f| > N\}} |f-h| \wedge 1 \, d\mu \\ &= \int_{\{|f| > N\}} |f-h| \wedge 1 \, d\mu \text{ by 3.93} \\ &\leq \int_{\{|f| > N\}} 1 \, d\mu \\ &= \mu(\{|f| > N\}) < \frac{\epsilon}{2}. \text{ by 3.92} \end{aligned} \tag{3.94}$$

Step 3 (Approximate h with a simple function φ). Since $h \in \mathcal{M}^n$, by there exists a sequence of measurable simple functions $\{\varphi_m\}_{m=1}^\infty$ such that $\varphi_m \to h$ pointwise. Since $|h| \leq N$, without loss of generality, assume that $|\varphi_m| \leq N$, for every $m \in N$. We claim $\rho_\mu(\varphi_m, h) \to 0$ as $m \to \infty$. We aim to apply Dominated Convergence Theorem. Observe that for every $\mathbf{x} \in \mathbb{R}^n$,

$$|h(\mathbf{x}) - \varphi_m(\mathbf{x})| \wedge 1 \leq |h(\mathbf{x}) - \varphi_m(\mathbf{x})| \leq |h(\mathbf{x})| + |\varphi_m(\mathbf{x})| \leq N + N = 2N. \tag{3.95}$$

By 3.95, $\mathbf{x} \to 2N$ is μ-integrable and dominates $|h-\varphi_m| \wedge 1$. Hence, by Dominated Convergence Theorem,

$$\lim_{m\to\infty} \rho_\mu(h, \varphi_m) = \lim_{m\to\infty} \int_{\mathbb{R}^n} |h - \varphi_m| \wedge 1 \, d\mu = \int_{\mathbb{R}^n} \lim_{m\to\infty} |h - \varphi_m| \wedge 1 \, d\mu = 0. \tag{3.96}$$

By 3.96, there exists $M \in \mathbb{N}$ such that

$$\rho_\mu(h, \varphi_m) = \int_{\mathbb{R}^n} |h - \varphi_m| \wedge 1 \, d\mu < \frac{\epsilon}{4}, \text{ for every } m \geq M. \tag{3.97}$$

Set $\varphi = \varphi_M$. In particular, by 3.97,

$$\rho_\mu(h, \varphi) < \frac{\epsilon}{4}. \tag{3.98}$$

Step 4 (Approximate φ with a continuous function g). By Lemma 9, there exist a closed set $F \in \mathcal{B}(\mathbb{R}^n)$ and a continuous function $g \in \mathcal{C}(\mathbb{R}^n)$ such that

$$\varphi \text{ and } g \text{ agree on } F \text{ and } \mu(\mathbb{R}^n \setminus F) < \frac{\epsilon}{4}. \tag{3.99}$$

Then the following computation yields the desired estimate,

$$\begin{aligned}
\rho_\mu(\varphi, g) = \int_{\mathbb{R}^n} |\varphi - g| \wedge 1 \, d\mu &= \int_F |\varphi - g| \wedge 1 \, d\mu + \int_{\mathbb{R}^n \setminus F} |\varphi - g| \wedge 1 \, d\mu \\
&= \int_{\mathbb{R}^n \setminus F} |\varphi - g| \wedge 1 \, d\mu \text{ by 3.99} \\
&\leq \int_{\mathbb{R}^n \setminus F} 1 \, d\mu \\
&= \mu(\mathbb{R}^n \setminus F) < \frac{\epsilon}{4}. \text{ by 3.99}
\end{aligned} \tag{3.100}$$

Step 5 (Putting estimates regarding f, h, φ, g together.). We will show that g is a desired continuous function. We begin by proving that $\rho_\mu(h, g) < \frac{\epsilon}{2}$. Since ρ_μ is a metric, by 3.98 and 3.100,

$$\rho_\mu(h, g) \leq \rho_\mu(h, \varphi) + \rho_\mu(\varphi, g) < \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}. \tag{3.101}$$

Again, applying the triangle inequality of ρ_μ combined with estimates 3.94 and 3.101 yields

$$\rho_\mu(f, g) \leq \rho_\mu(f, h) + \rho_\mu(h, g) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

■

Corollary 6. *$\mathcal{C}(\mathbb{R}^n)$ is δ_μ-dense in $\mathcal{M}^n$.*

Proof. Follows directly from Proposition 9 and Characterization of convergence in probability, Lemma 2.1 in [HSW89]. ■

3.7.4 The Probabilistic Universal Approximation Theorem

We are ready to prove The Probabilistic Universal Approximation Theorem.

Theorem 17 (The Probabilistic Universal Approximation Theorem)**.** *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any continuous sigmoidal activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then $\mathcal{H}_\sigma$ is δ_μ-dense in $\mathcal{M}^n$.

Proof. Let $f \in \mathcal{M}^n$ and let $\epsilon > 0$. By Corollary 6, there exists $g \in \mathcal{C}(\mathbb{R}^n)$ such that $\delta_\mu(f, g) < \frac{\epsilon}{2}$. By Theorem 16, there exists a sequence of neural networks $\{h_m\}_{m=1}^{\infty}$ from $\mathcal{H}_\sigma$ such that $h_m \to g$ uniformly on every compact set $K \subset \mathbb{R}^n$. By Proposition 8, $h_m \to g$ in δ_μ. Hence there exists $M \in \mathbb{N}$ such that $\delta_\mu(h_m, g) < \frac{\epsilon}{2}$, for every $m \geq M$. Since δ_μ is a metric, applying estimates above yields

$$\delta_\mu(f, h_M) \leq \delta_\mu(f, g) + \delta_\mu(g, h_M) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

■

Remark 25. By Characterization of convergence in probability, Lemma 2.1 in [HSW89], density in δ_μ is equivalent to density in probability measure μ. Informally, a neural network from $\mathcal{H}_\sigma$ can approximate any Borel function on $\mathbb{R}^n$ up to desired accuracy, except possibly on set of arbitrarily small probability measure. This interpretation is stated formally as the following corollary of The Probabilistic Universal Approximation Theorem.

Corollary 7. *Let $\mathcal{H}_\sigma$ denote the family of single-layer fully-connected neural networks with any continuous sigmoidal activation function σ, given by*

$$\mathcal{H}_\sigma = \left\{ \mathbf{x} \to \sum_{k=1}^{m} \alpha_k \sigma(\langle \mathbf{w_k}, \mathbf{x} \rangle + \beta_k) : m \in \mathbb{N}, \alpha_1 \dots \alpha_m, \beta_1 \dots \beta_m \in \mathbb{R}, \mathbf{w_k} \in \mathbb{R}^n \right\}.$$

Then for every $f \in \mathcal{M}^n$, for every $\epsilon > 0$, there exists $h \in \mathcal{H}_\sigma$ such that

$$\mu(\{|f - h| > \epsilon\}) < \epsilon.$$

Proof. By The Probabilistic Universal Approximation Theorem, $\mathcal{H}_\sigma$ is δ_μ-dense in $\mathcal{M}^n$. By Characterization of convergence in probability, Lemma 2.1 in [HSW89], density in δ_μ is equivalent to density in probability measure μ. The result follows directly from definition of convergence in measure. ■

3.7.5 Relationship between measurable functions and classification

We will briefly discuss the relationship between measurable functions and classification, using the argument from the discussion following Theorem 2 in [Cyb89]. Let λ denote the restriction of Lebesgue measure to $[0, 1]^n$. Let $\{P_k\}_{k=1}^{m}$ be a partition of $[0, 1]^n$ into disjoint, λ-measurable subsets of $[0, 1]^n$. We can view the classification problem on $[0, 1]^n$ as a problem of approximating a classification function $f : [0, 1]^n \to \{1, 2, \dots, m\}$, given by

$$f(\mathbf{x}) = k \iff \mathbf{x} \in P_k.$$

Cybenko's Universal Approximation Theorem for measurable functions, [Cyb89] implies that neural networks with continuous sigmoidal activation functions can approximate classification functions to desired accuracy, except possibly on sets of arbitrarily small measure.

Chapter 4

Experiments

In this chapter, we will study the relationship between the performance of neural networks and various architectural decisions, including the choice of an activation function and the effect of neural network depth. The purpose of this chapter is to present a wide range of difficulties in the application of neural networks and illustrate the main differences between practical observations and theoretical guarantees. Studies in this chapter will be experimental and significantly less rigorous than counterparts in other chapters. We will also discuss the role of hyperparameters, including the optimizer and batch size.

4.1 Introduction

In the Literature review and Universality of Neural Networks, we have seen that the conditions on activation function play an important role in many proofs of the universal approximation. We have also seen that neural networks with only one "sufficiently" wide hidden layer can approximate continuous and measurable functions, in the appropriate sense.
In this chapter, we aim to experimentally examine the performance impact of the activation function and the (in)significance of neural network depth. We will discuss classification on *Fashion MNIST*. We want to provide insight into the following three questions of practical importance.

1. *Does the choice of activation function significantly affect the performance?*
2. *Does the neural network depth significantly affect the performance?*
3. *Does the training configuration significantly affect the performance?*

Training setup in Table 4.1 is used in every experiment unless stated otherwise.

epochs	10
batch size	32
optimizer	Adam
learning rate	PyTorch default
random seed	42
computer	Macbook Pro 2017 15"

Table 4.1: training configuration

4.2 Classification on Fashion MNIST

Fashion MNIST[XRV17] is a dataset consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example in the dataset is a 28x28 grayscale image belonging to precisely one of 10 classes, presented in Table 4.2.

label	description
0	T-shirt/top
1	trouser
2	pullover
3	dress
4	coat
5	sandal
6	shirt
7	sneaker
8	bag
9	ankle boot

Table 4.2: *Fashion MNIST* labels

The task for a machine learning algorithm is to classify a given 28x28 grayscale image into precisely one of 10 categories from Table 4.2. *Fashion MNIST* was invented in 2017 as a replacement for *MNIST*[Den12]. *MNIST*[Den12] has been used as a benchmark for novel and existing algorithms. However, it turns out to be quite easy for modern algorithms and it is not representative of modern computer vision tasks. *Fashion MNIST* is slightly more difficult, but still small enough for experimenting with limited computational resources. This makes it suitable for our purpose.

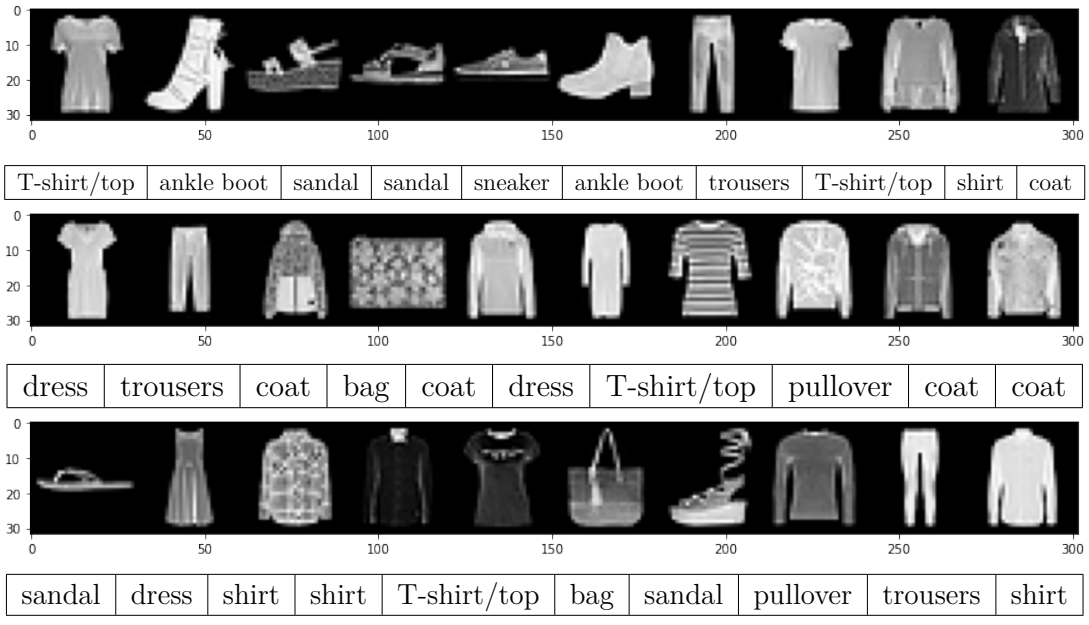


Figure 4.1: Some examples from *Fashion MNIST*

4.2.1 Model

To perform classification experiments, our task is to construct a single neural network that assigns to each given image precisely one of 10 labels from Table 4.2. We will begin the discussion by considering theoretical aspects of this problem. We will view each 28×28 image as a flattened vector in $\mathbb{R}^{784}$. A way to solve this problem is to interpret it as trying to approximate or learn a 'classification' function which maps an image viewed as a vector $\mathbb{R}^{784}$ to a vector in $\mathbb{R}^{10}$. In this context, each i-th component of an output vector in $\mathbb{R}^{10}$ corresponds to estimated probability that an image belongs to i-th class, for $0 \leq i \leq 9$. Hence to assign a label to an image, we assign a label that corresponds to the highest estimated probability. Formally, we want a neural network $\mathbf{f} : \mathbb{R}^{784} \to \mathbb{R}^{10}$. It turns out that our network can assign labels automatically by using the *softmax* activation function in the output layer. The *softmax* activation function is designed precisely for this use-case.

From a practical standpoint, we want to have a flexible interface to neural network architecture. This requirement is necessary to support a wide variety of experiments. We want to be able to quickly prototype a new neural network by adding layers or by changing activation function in each layer. Fortunately, PyTorch allows us to achieve both theoretical and practical requirements quite elegantly (see Figure 4.2).

```
from collections import namedtuple
import torch.nn as nn

image_width = 28
image_height = 28

LayerConfig = namedtuple("LayerConfig", ["width", "activation"])

class FashionNNMultiHiddens(nn.Module):

    def __init__(self, layers_config):
        super(FashionNNMultiHiddens, self).__init__()

        previous_width = image_width * image_height
        hidden_layers = []
        for config in layers_config:
            hidden_layers.append(
                nn.Linear(in_features=previous_width,
                          out_features=config.width))
            hidden_layers.append(config.activation)
            previous_width = config.width

        self.nn = nn.Sequential(*hidden_layers)

    def forward(self, x):
        return self.nn(x)
```

Figure 4.2: PyTorch implementation of model from Figure 4.3

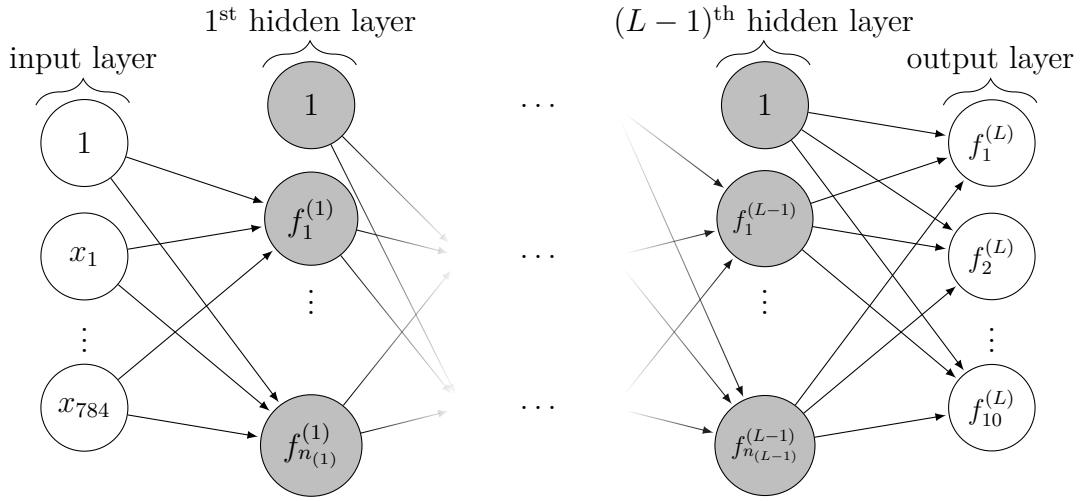


Figure 4.3: This figure illustrates a network graph for a *Fashion MNIST* classification neural network. Output layer is fixed and uses *softmax* activation function. Depending on experiment, other hidden layers vary in width and the activation function. More general implementation of this model is given in Figure 4.2.

4.2.2 Methodology

Each experiment consists of a neural network architecture configuration and a training configuration. Experiments are run independently. Each experiment run consists of instantiation of a neural network from the configuration and training from scratch using training configuration from Table 4.1. In each experiment, the neural network architecture is fixed and only weights and biases are changed by the optimizer. After every 100 batch iterations, training and validation set accuracy and loss are computed for the experiment neural network architecture. The model achieving the best validation accuracy is maintained and its weights are saved. Plots in the subsequent sections were created by manual inspection and analysis of the following statistics:

- training and validation loss collected during the training process,
- training and validation accuracy collected during the training process,
- validation set accuracy and loss corresponding to the best weights,
- class-specific validation set accuracy corresponding to the best weights,
- confusion matrix corresponding to the model with the best weights.

```
layer_width = 64
layers_number = 4
layers = [ LayerConfig(layer_width, nn.ReLU()) for _ in range(layers_number) ]
experiment = {
    'name' : 'nn-4x-64-relu-softmax',
    'epochs': 10,
    'batch_size' : 32,
    'layers_config': layers + [LayerConfig(10, nn.Softmax(dim=0))]
}
```

Figure 4.4: The configuration of *nn-4x-64-relu-softmax* experiment

4.2.3 The choice of an optimizer

When it comes to optimizers, in this thesis we only discussed stochastic gradient descent. We briefly mentioned other methods of practical importance. However, the convergence of stochastic gradient descent was too slow under configuration from Table 4.1 (see Table 4.3 and Table 4.4). Since the computational resources were limited, after experimenting with different optimizers, Adam produced the best results in the smallest number of training epochs. Adam is currently one of the most popular optimizers and it is a common default choice. The choice of an optimizer is a good example of a practical problem that is not discussed in the approximation theory of neural networks. Theoretical results are completely disconnected from choices regarding the optimization algorithm and its configuration. However, those may significantly affect the performance of the resulting neural network. Effects are especially noticeable when training for a fixed number of epochs, which is a very common practice. Given the fact an average training time of a network with a single hidden layer was about 30 minutes, the training time was indeed significantly affected. This observation was neatly summarised in *An overview of gradient descent optimization algorithms* [Rud17], quoted below.

> "Interestingly, many recent papers use vanilla SGD without momentum and a simple learning rate annealing schedule. As has been shown, SGD usually achieves to find a minimum, but it might take significantly longer than with some of the optimizers, is much more reliant on a robust initialization and annealing schedule, and may get stuck in saddle points rather than local minima. Consequently, if you care about fast convergence and train a deep or complex neural network, you should choose one of the adaptive learning rate methods." (*An overview of gradient descent optimization algorithms* [Rud17], p.10)

The following results were obtained by changing the optimizer in configuration from Table 4.1. All optimizers were initialized with a default configuration.

model	**SGD**	**Adam**
nn-16-sigmoid-softmax	70.9800%	79.2200%
nn-32-sigmoid-softmax	70.8300%	79.8700%
nn-64-sigmoid-softmax	70.1500%	79.0100%
nn-128-sigmoid-softmax	68.6500%	79.4500%

Table 4.3: validation set accuracy of sigmoid networks after 5 training epochs

model	**SGD**	**Adam**
nn-16-relu-softmax	72.5100%	76.7800%
nn-32-relu-softmax	72.5400%	76.6100%
nn-64-relu-softmax	72.8500%	77.2100%
nn-128-relu-softmax	72.8900%	77.4200%

Table 4.4: validation set accuracy of ReLU networks after 5 training epochs

4.2.4 Impact of batch size on validation accuracy

Although the batch size seems unimportant in comparison with neural network architecture and the choice of an optimizer, the experimental data from Figure 4.5 indicates that the batch size plays an important role when the network is trained for a fixed number of iterations. To examine the effects of various batch sizes, a reasonably large neural network architecture is chosen and such a network is trained using the configuration from Table 4.1. In this experiment, only the batch size and activation function are varied. It is very important to note that the number of epochs remained unchanged.

The results displayed in Figure 4.5 indicate that training using smaller batch sizes results in significantly greater validation accuracy. This can be attributed to the fact that a smaller batch size implies a greater number of parameter updates since the number of epochs is fixed. It is worth noting that such a training setup requires significantly more time. For instance, the average training time using a batch size of 32 examples was about 30 minutes, while the average training time using the batch size of 256 examples was about 5 minutes. The accuracy drop after the batch size of 128 is noticeable for every activation function. This could be expected since the number of parameter updates drops significantly. On the other hand, a larger batch size often results in a more accurate estimate of the gradient.

It seems that larger batch sizes generally require a larger number of training epochs, which is indeed sensible, as the optimization performs significantly less parameter updates. It is reasonable to conjecture that the sweet spot for the default training configuration is a batch size of 64 examples, as it seems that such a configuration performs slightly better than other batch sizes in 2 out of 3 activation functions.

To sum up, batch size likely plays an important role in training a neural network. The results discussed are relevant only for one neural network architecture and one default training setup. However, according to the Figure 1 from [HHS18] and Figure 2 from [Shi+17], the similar observations can be made on more complex datasets and neural networks. It is interesting to note that [HHS18] and [Shi+17] provide different explanations and analysis of this phenomenon.

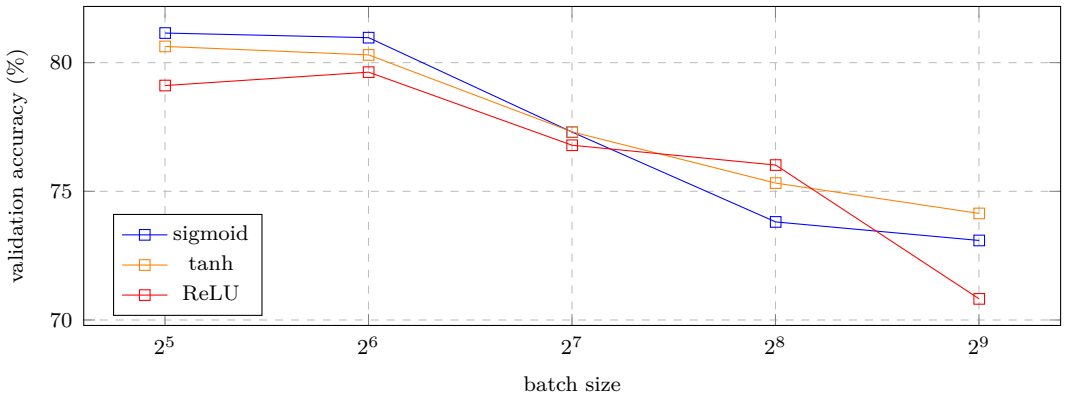


Figure 4.5: Effects of various batch sizes on training *nn-2x-128-sigmoid-softmax*

4.2.5 Impact of activation function on validation accuracy

Since the activation function gives rise to nonlinearity, it is reasonable to conjecture that the activation function affects performance. In Literature review and Universality of Neural Networks, we have also observed that the properties of activation function also play an essential role in many proofs of universality.
To begin an experimental study of the effects of the activation function, we will focus on neural networks with a single hidden layer. Each neural network is trained using the configuration from Table 4.1. In this experiment, only the hidden layer width and the activation function are varied. For each experiment configuration, the model achieving the best validation accuracy is saved and used in the following analysis.
According to the results displayed in Figure 4.6, it is likely that the activation function significantly affects validation set accuracy, given current training configuration. Thus, it is reasonable to conjecture that the activation function affects the generalization error. It is interesting to note that sigmoid and tanh significantly and consistently outperform ReLU regardless of hidden layer width. That is slightly surprising because many modern neural network architectures use ReLU and similar rectified activation functions instead of sigmoid and tanh. It can be shown that sigmoid and tanh are algebraically related. More precisely, for every $x \in \mathbb{R}, \tanh(x) = 2\sigma(2x) - 1$. Given the structure of a fully-connected neural network and this algebraic connection, it is reasonable to expect that sigmoid and tanh produce similar results. According to the results displayed in Figure 4.6, that is indeed the case. Another surprising observation is the significant validation accuracy drop for ReLU networks after hidden layer width of 128 neurons. Although ReLU networks perform worse as the hidden layer width increases, it seems that sigmoid and tanh networks perform better. The widest tanh and sigmoid achieve the best validation accuracy. We will briefly discuss the worst ReLU model. According to the Figure 4.7, validation and test accuracy are consistently very similar. Hence, there is no evidence of overfitting.

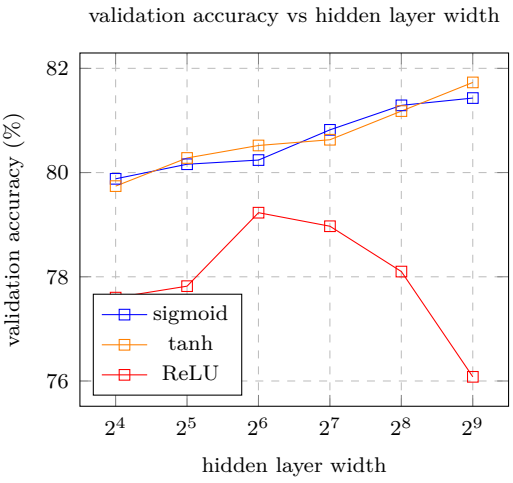


Figure 4.6: Validation accuracy of the best model with given hidden layer width and activation function

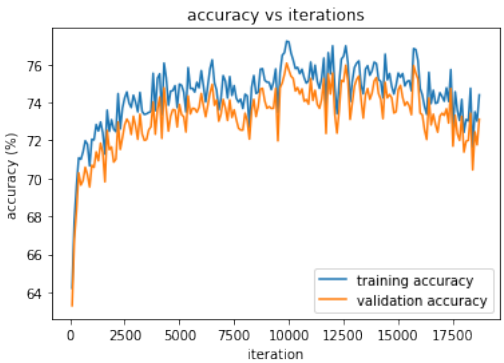


Figure 4.7: Training and validation accuracy of ReLU network with a single hidden layer of 512 neurons. This is the worst ReLU model in this experiment.

4.2.6 Adding a layer

In this subsection, we will focus on neural networks with two hidden layers of the same width. The experiment configuration is almost identical to one from the previous section. The only change is the addition of another identical hidden layer to each neural network. For each experiment configuration, the model achieving the best validation accuracy is saved and used in the following analysis.
From Figure 4.6, it is noticeable that the best validation accuracy increases consistently with the hidden layer width. This may suggest that sigmoid and tanh neural networks may benefit from the increased model complexity. However, it seems that this does not happen to the extent one may expect (see Figure 4.8).
In the previous section, we discussed the similarity between the sigmoid and tanh activation function. From the Figure 4.8, we can conclude that sigmoid and tanh networks remain achieving similar validation accuracy, what is compatible with results from the previous section.
From Figure 4.9, Figure 4.10 and Figure 4.11, it is evident that the layer addition results in slightly better validation accuracy for layer widths up to and including 128 neurons. However, the difference is mostly within 1.5% and this can be attributed to the noise. It is hard to say whether the validation accuracy boost is significant. According to the benchmark table in [Zal20], similar neural networks may achieve the validation accuracy of 88% (see the entry for *MLP 256-128-100*). The pattern is slightly unclear for layer widths larger than 128 neurons. For instance, the network with tanh activation and two hidden layers of 512 neurons achieved the validation accuracy of 82.33%, which is the best validation accuracy so far. However, such a configuration performed noticeably less well for other activation functions, especially ReLU.
Neural networks with tanh activation function benefited the most from the extra layer. According to Figure 4.10, tanh neural networks with two identical layers consistently performed at least as well as the corresponding single layer configuration. It is interesting to note that ReLU networks benefited the least from the extra layer. This observation is quite evident from the Figure 4.11. As in the case of neural networks with the single layer, ReLU networks seem to generalize slightly worse than tanh and sigmoid networks.

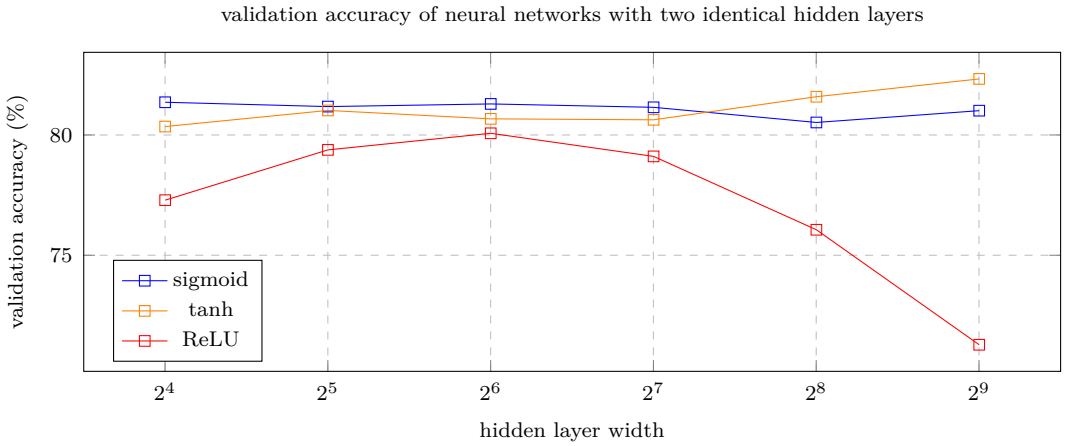


Figure 4.8: Validation accuracy for networks with two identical hidden layers

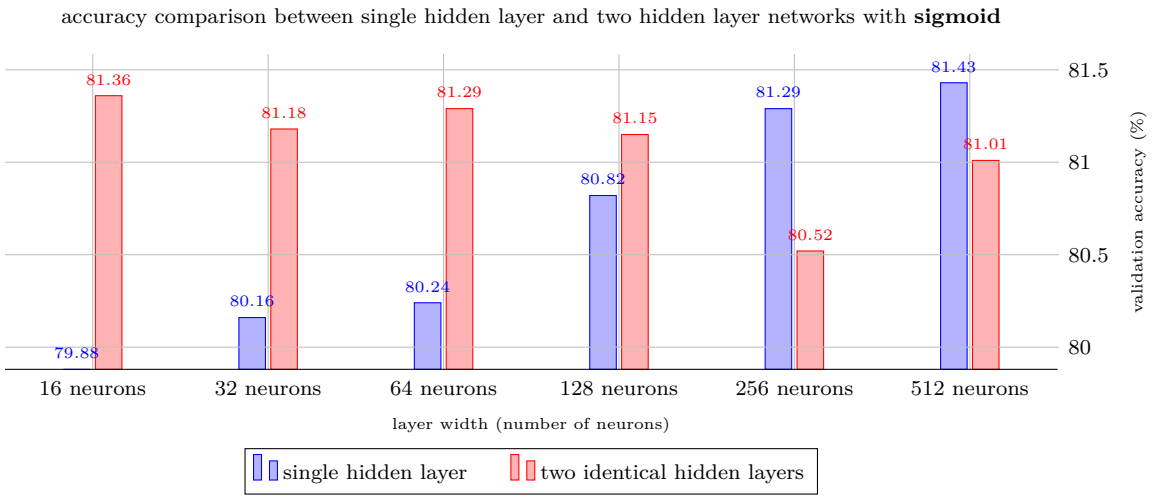


Figure 4.9: Validation accuracy for **sigmoid** grouped by number of layers

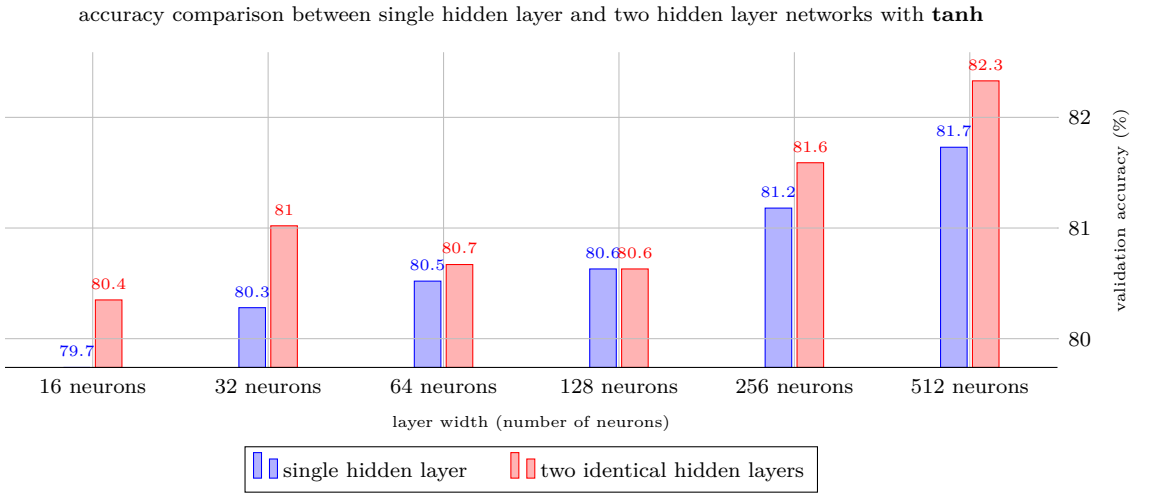


Figure 4.10: Validation accuracy for **tanh** grouped by the number of layers

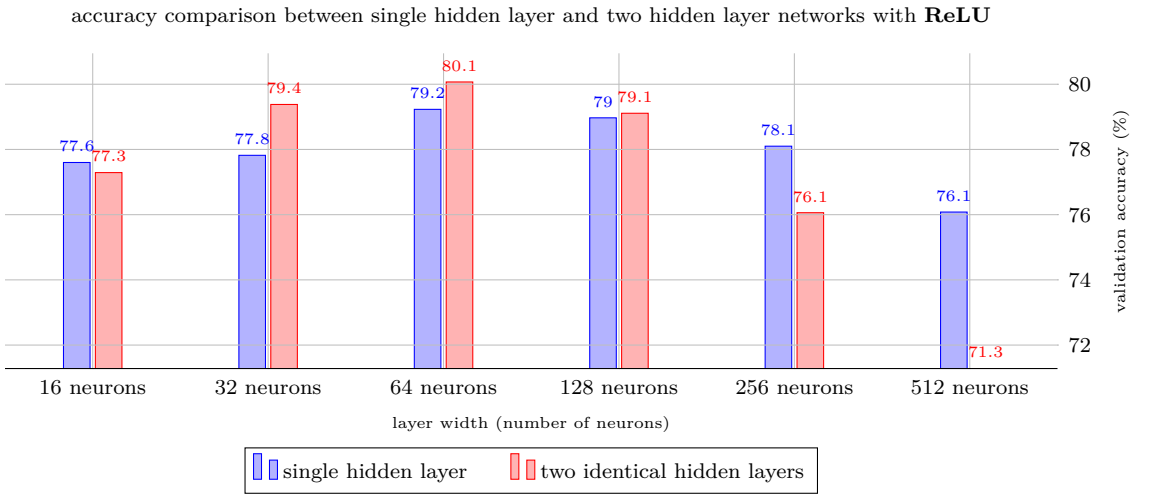


Figure 4.11: Validation accuracy for **ReLU** grouped by the number of layers

4.2.7 Impact of neural network depth on validation accuracy

In this subsection, we will discuss the relationship between the best validation accuracy and neural network depth. The experiment setup and methodology are almost identical to the previous subsection. The only change is the addition of multiple identical hidden layers instead of a single hidden layer. In the previous subsection, we observed that the addition of a single identical layer did not significantly improve the validation accuracy. According to the results presented in Figure 4.12 and in Figure 4.13, validation accuracy does not improve with addition of identically wide hidden layers. Hence, we may conjecture that on this dataset, deeper fully-connected networks do not generalize better than shallow fully-connected networks. It is important to note that results are consistent across all activation functions. Moreover, results in Figure 4.12 and Figure 4.13 are almost identical. This can be attributed to the apparent similarity in neural network architecture.

It is interesting to note that the validation accuracy remains close to 80% for networks with at most 4 layers. However, as the number of layers increases, the validation accuracy significantly decreases. This observation applies to all activation functions. However the validation accuracy drop is the largest for sigmoid and tanh. This can be attributed to a well-known *vanishing gradient* problem. We will briefly discuss that source of difficulty in training deep neural networks with sigmoid activation. We know that $\lim_{x\to\infty}\sigma(x) = 1$ and $\lim_{x\to-\infty}\sigma(x) = 0$. By Lemma 1, we conclude that $\lim_{x\to\infty}\sigma'(x) = 0$ and $\lim_{x\to-\infty}\sigma'(x) = 0$. Informally, as the inputs of a sigmoid layer become extremely small or extremely large, the gradient of the loss function with respect to parameters of a sigmoid layer vanishes. This causes serious problems in gradient-based optimization since the corresponding weights and biases cease to update. However, in case of ReLU, the drop in validation accuracy resulting from increased depth is noticeably smaller than in case of sigmoid and tanh. This is also a known result and one of key reasons why ReLU is preferred to sigmoid and tanh when training deep networks.

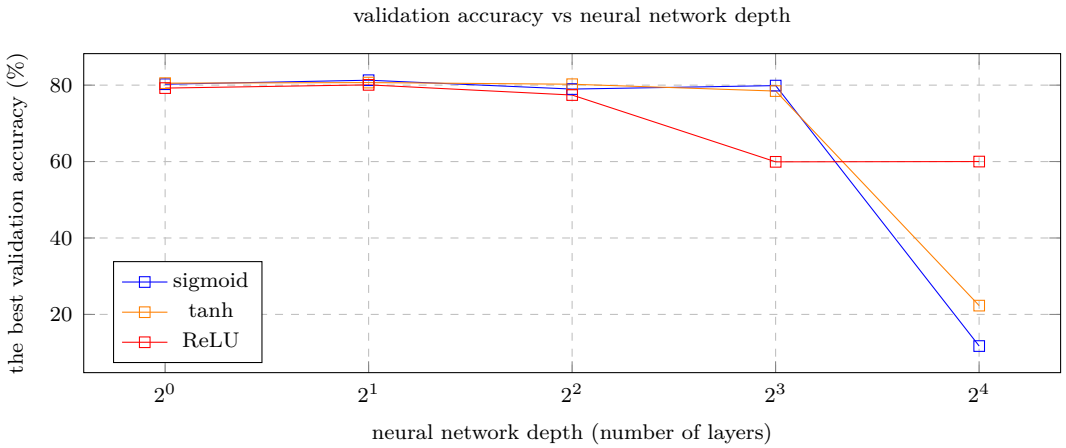


Figure 4.12: Effects of varying neural network depth while keeping layer width of **64 neurons**

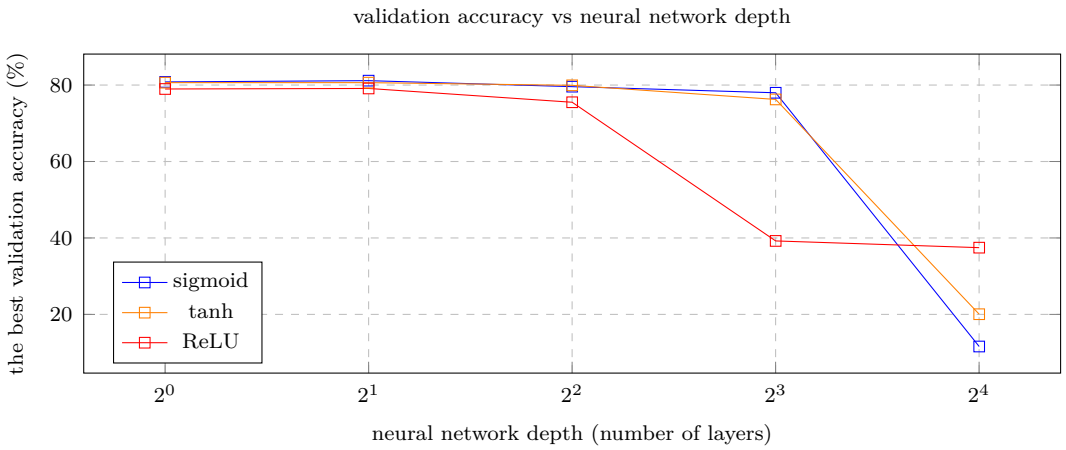


Figure 4.13: Effects of varying neural network depth while keeping layer width of **128 neurons**

4.2.8 Interesting observation

According to Figure 4.14, two models achieving the best validation accuracy, *nn-2x-512-tanh-softmax* and *nn-512-sigmoid-softmax*, demonstrate similar performance on every label. However, it is interesting to observe that both models struggle with *shirt*s. Although both models perform decently on *pullover* and *T-shirt*, they both perform significantly worse on *shirt*.
Figure 4.15 demonstrates that the training and accuracy gap of *nn-2x-512-tanh-softmax* remains quite small. However, in Figure 4.16, we can observe slightly larger gap between training and validation accuracy of *nn-512-sigmoid-softmax*. Since the training accuracy of both models peaks at about 82%, there is evidence to suggest both models struggle to fit the training set. Hence, there is no evidence that those networks overfit. In the previous section, we observed that the addition of more fully-connected layers does not significantly affect the generalization. This may indicate that fully-connected architecture is generally too simple for this dataset.

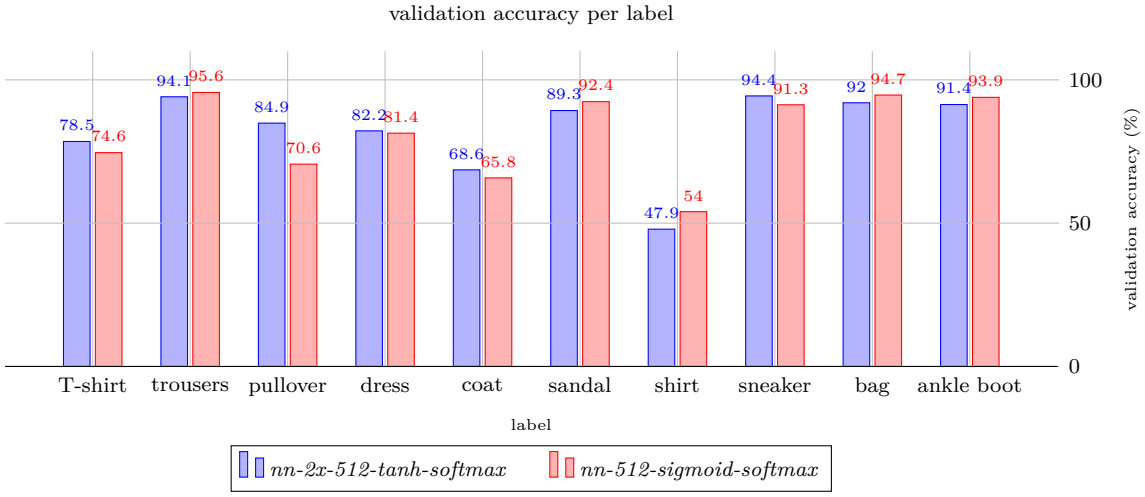


Figure 4.14: Validation accuracy per label for the best network configurations

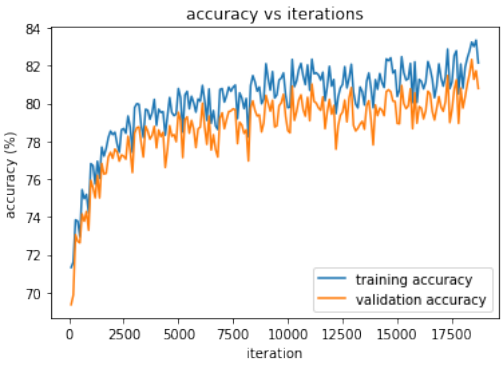


Figure 4.15: Accuracy pattern of *nn-2x-512-tanh-softmax*

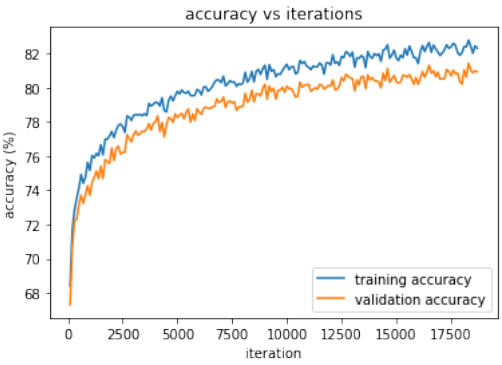


Figure 4.16: Accuracy pattern of *nn-512-sigmoid-softmax*

4.2.9 Conclusion

In this chapter, we discussed various practical problems related to the applications of neural networks. Despite impressive theoretical properties discussed in Literature review and Universality of Neural Networks, many practical applications of neural networks often suffer from issues not addressed in the universal approximation theory.

According to state of the art results from Literature review, ReLU, sigmoid and tanh have comparable theoretical properties. However, in Impact of activation function on validation accuracy, we observed that those activation functions result in noticeably different validation accuracy. This can be attributed to the little emphasis we put on the training configuration. This suggests that different activation functions may demand different training configurations. To address a mysterious performance gap between ReLU and alternatives, more research is necessary. For example, studying the effects of weight regularisation (weight decay) or employing a more flexible learning rate scheduler could provide more insight into this observation. According to the results reported on the Fashion MNIST benchmark dashboard, a neural network with a single layer of 100 neurons and ReLU activation function achieved the accuracy of 87.7%. Moreover, such a network outperformed tanh counterparts. This may suggest that ReLU networks perform better than observed in this thesis.

We have also observed that the best validation accuracy peaks at about 82%. According to the benchmark table in [Zal20], more sophisticated convolutional neural networks achieve the validation accuracy exceeding 90%. This may indicate that the fully-connected architecture is generally too simple to perform the classification on this dataset.

Seemingly unimportant hyperparameter choices discussed in Impact of batch size on validation accuracy and The choice of an optimizer are disconnected from the approximation theory of neural networks. However, we observed they may have significant practical consequences.

Chapter 5

Conclusion

In this thesis, we explored various important results in the approximation theory of single-layer, fully-connected neural networks.
In Introduction, we introduced fundamental machine learning terminology from statistical learning perspective. We formalized common machine learning tasks and presented main challenges. We also introduced fully-connected neural networks as a family of machine learning algorithms. We discussed the backpropagation training algorithm and we pointed out various subtleties related to training initialization and optimization.
In Literature review, we introduced the approximation theory of neural networks and discussed the main research questions. We briefly discussed the historical development of this field and we stated contemporary state-of-the-art results. Finally, we described the role of this thesis in the context of this research area.
In Universality of Neural Networks, we presented different universal approximation results, addressing different function spaces and activation functions. We discussed the universal approximation of continuous functions on compact sets from two different perspectives. We discussed the universal approximation of Lebesgue square-integrable and integrable functions. Finally, we discussed the universal approximation of Borel measurable functions in a probabilistic sense.
In Experiments, we presented and discussed various practical problems related to the applications of neural networks. Despite impressive theoretical properties discussed in Literature review and Universality of Neural Networks, many practical applications of neural networks often suffer from issues not addressed in the approximation theory. For instance, we discussed Impact of activation function on validation accuracy and conjectured that different activation functions may require different training configurations. We explored the impact of Adding a layer and expanded this study by considering Impact of neural network depth on validation accuracy. Finally, we observed that seemingly unimportant hyperparameter choices discussed in Impact of batch size on validation accuracy and The choice of an optimizer may have significant practical consequences.
In Universality of Neural Networks, we used a lot of standard results from measure theory and functional analysis, including Stone-Weierstrass Theorem, Riesz Representation Theorem for the Dual of $\mathcal{L}^p$, Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$ and Hahn-Banach Theorem. Those and many other used results are discussed, stated and proved in Appendix.

Chapter 6

Appendix

This chapter is devoted mainly to the mathematics backing the main results established in Universality of Neural Networks. Although many results discussed in this chapter are standard results in their respective fields, most of them are often not part of the undergraduate syllabus. The results that were part of the syllabus will be referenced. The most important results that were not part of the syllabus will be stated and proved. This chapter also contains the Code for Experiments and Statement of originality.

6.1 Set Theory

The only axiomatic set-theoretic result relevant to this work is Zorn lemma. To state the Zorn lemma, we recall various ordering concepts.

Definition 27 (partial order)**.** Suppose that Ω is a nonempty set and $\prec$ is a binary relation on Ω. We say $\prec$ is a partial order if:

P1. $x \prec x$ for every $x \in \Omega$;
P2. if $x \prec y$ and $y \prec x$ then $x = y$;
P3. if $x \prec y$ and $y \prec z$ then $x \prec z$.

If $\prec$ is a partial order, we say Ω is a partially ordered set.

Definition 28 (total order)**.** Suppose that Ω is a nonempty set and $\prec$ is a partial order on Ω. If for every $x, y \in \Omega$ either $x \prec y$ or $y \prec x$, we say that $\prec$ is a total order and we say Ω is a totally ordered set.

Definition 29 (maximal element)**.** Suppose that Ω is a nonempty set and $\prec$ is a partial order on Ω. Then $y \in \Omega$ is a maximal element of Ω if $y \prec x \implies y = x$.

Definition 30 (upper bound)**.** Suppose that Ω is a nonempty set and $\prec$ is a partial order on Ω. Let $\Theta \subseteq \Omega$. Then $y \in \Omega$ is an upper bound for Θ if for every $x \in \Theta$, $x \prec y$.

Lemma 10 (Zorn lemma)**.** *Let Ω be a nonempty, partially ordered set such that every totally ordered subset of Ω has an upper bound. Then there exists a maximal element in Ω.*

Remark 26. Zorn lemma is nontrivially equivalent to the **Axiom of Choice**.

6.2 Stone-Weierstrass Theorem

Weierstrass Approximation Theorem

We begin by tackling one of the oldest problems in analysis - whether it is possible to approximate continuous functions using polynomials. Weierstrass solved the problem with the approximation family of Bernstein polynomials.
To precisely define the notion of approximation, we begin by discussing the metric structure of a vector space of continuous functions - $\mathcal{C}(X)$. We equip $\mathcal{C}(X)$ with a metric δ_∞ based on the sup-norm $\|f\|_\infty = \sup_{x\in X} |f(x)|$. We define $\delta_\infty(f,g) = \|f-g\|_\infty$. The proof δ_∞ is a metric and $\|.\|_\infty$ is a norm is Example 10.6 in [Wad14].
Observe that the convergence in $\mathcal{C}(X)$ is equivalent to the uniform convergence. Before discussing the approximation, we define the uniform density of a family of functions.

Definition 31 (uniform density of a family)**.** A subset $\mathcal{A}$ of $\mathcal{C}(X)$ is uniformly dense in $\mathcal{C}(X)$ if and only if given $\epsilon > 0$, there exists $g \in \mathcal{A}$ such that $\delta_\infty(f,g) < \epsilon$.

We are ready to discuss the Bernstein polynomials.

Definition 32 (Bernstein polynomials)**.** Let f be a function defined on $[0,1]$. The n-th Bernstein polynomial of f, denoted by $B_n f$, is a polynomial given by

$$B_n f = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \cdot \binom{n}{k} x^k (1-x)^{n-k}.$$

Theorem 18 (Bernstein Approximation Theorem)**.** *Bernstein polynomials are uniformly dense in* $\mathcal{C}[0,1]$.

Before discussing the proof, we state a few technical results which will be used.

Lemma 11. *For every* $x \in \mathbb{R}$, $n \geq 0$, $\sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1$.

Proof. Follows from the Binomial Theorem applied to $a = x$ and $b = 1-x$. ■

Lemma 12. *For* $x \in \mathbb{R}$, $n \geq 0$, $\sum_{k=0}^{n} (nx-k)^2 \binom{n}{k} x^k (1-x)^{n-k} = nx(1-x) \leq \frac{n}{4}$.

Proof. This is a direct computation applying the combinatorial identites:

- for $k \geq 1$, $k\binom{n}{k} = n\binom{n-1}{k-1}$,
- for $k \geq 2$, $k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}$.

For detailed calculations, see the proof of Lemma 27.2 in [RL15]. ■

Proof of Bernstein Approximation Theorem. If f is identically zero the result follows immediatelly. Suppose f is not identically zero and let $\epsilon > 0$. Set $M = \sup\{|f(x)| : x \in [0,1]\}$. By compactness of $[0,1]$, M is finite and f is uniformly continuous, so there exists $\delta > 0$ such that

$$x, y \in [0,1] \text{ with } |x-y| < \delta \implies |f(x) - f(y)| < \frac{\epsilon}{2}. \tag{6.1}$$

Set $N = \frac{M}{\epsilon \cdot \delta^2}$. We will show $|B_n f(x) - f(x)| < \epsilon$, $\forall x \in [0,1]$, for every $n > N$. Let $x \in [0,1]$ and $n > N$. By Lemma 11, $f(x) = \sum_{k=0}^{n} f(x)\binom{n}{k}x^k(1-x)^{n-k}$, so

$$|B_n f(x) - f(x)| \leq \sum_{k=0}^{n} \left| f\left(\frac{k}{n}\right) - f(x) \right| \cdot \binom{n}{k} x^k (1-x)^{n-k}. \tag{6.2}$$

To estimate 6.2, partition $\{0, 1, \ldots, n\}$ into sets $\Delta_<$ and $\Delta_\geq$ such that $\Delta_< := \{k : \left|\frac{k}{n} - x\right| < \delta\}$ and $\Delta_\geq := \{k : \left|\frac{k}{n} - x\right| \geq \delta\}$.
Suppose $k \in \Delta_<$. By 6.1, $\left|f\left(\frac{k}{n}\right) - f(x)\right| < \frac{\epsilon}{2}$. By Lemma 11,

$$\sum_{k \in \Delta_<} \left| f\left(\frac{k}{n}\right) - f(x) \right| \cdot \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{\epsilon}{2} \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = \frac{\epsilon}{2}. \tag{6.3}$$

Suppose $k \in \Delta_\geq$. Then $\left|\frac{k}{n} - x\right| \geq \delta$ which is equivalent to $(k - nx)^2 \geq n^2\delta^2$, so

$$\begin{aligned} \sum_{k \in \Delta_\geq} \left| f\left(\frac{k}{n}\right) - f(x) \right| \cdot \binom{n}{k} x^k (1-x)^{n-k} &\leq 2M \sum_{k \in \Delta_\geq} \frac{n^2\delta^2}{n^2\delta^2} \binom{n}{k} x^k (1-x)^{n-k} \\ &\leq \frac{2M}{n^2\delta^2} \sum_{k=0}^{n} (nx - k)^2 \binom{n}{k} x^k (1-x)^{n-k}. \end{aligned}$$

Applying the Lemma 12,

$$\sum_{k=0}^{n} \left| f\left(\frac{k}{n}\right) - f(x) \right| \cdot \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{2M}{n^2\delta^2} \cdot \frac{n}{4} \leq \frac{M}{2n\delta^2} < \frac{M}{2N\delta^2} = \frac{\epsilon}{2}. \tag{6.4}$$

By 6.3 and 6.4, $|B_n f(x) - f(x)| < \epsilon$. Since x was arbitrary, the proof is complete. ■

Using Bernstein Approximation Theorem, we can easily prove the main result of this section - Weierstrass Approximation Theorem.

Theorem 19 (Weierstrass Approximation Theorem)**.** *Polynomials are uniformly dense in* $\mathcal{C}[a,b]$.

Proof. Consider the polynomial $\phi : [0,1] \to [a,b]$ given by $\phi(x) = (b-a)x + a$. It is easy to check ϕ is a continuous bijection, whose inverse $\phi^{-1} : [a,b] \to [0,1]$ given by $\phi^{-1}(y) = \frac{y-a}{b-a}$ is also a polynomial. By continuity of composition, $f \circ \phi : [0,1] \to \mathbb{R}$ is continuous. By Bernstein Approximation Theorem, there exists a sequence of polynomials φ_n such that $\lim_{n\to\infty} \sup_{y\in[0,1]} |(f \circ \phi^{-1})(y) - \varphi_n(y)| = 0$. We will show $\psi_n := \varphi_n \circ \phi^{-1}$ uniformly approximates f. Since

$$\begin{aligned} \sup_{x\in[a,b]} |f(x) - \psi_n(x)| &= \sup_{x\in[a,b]} |(f \circ \phi)(\phi^{-1}(x)) - \varphi_n(\phi^{-1}(x))| \\ &= \sup_{y\in[0,1]} |(f \circ \phi^{-1})(y) - \varphi_n(y)|, \end{aligned}$$

$\lim_{n\to\infty} \sup_{x\in[a,b]} |f(x) - \psi_n(x)| = \lim_{n\to\infty} \sup_{y\in[0,1]} |(f \circ \phi^{-1})(y) - \varphi_n(y)| = 0.$ ■

Stone-Weierstrass Theorem

In the following discussion, we assume that $\mathcal{C}(X)$ is a metric space equipped with d_∞ metric. We aim to generalize Weierstrass Approximation Theorem to $\mathcal{C}(X)$. To accomplish that, it is useful to generalize closure properties of polynomials by introducing algebras of continuous functions.

Definition 33 (algebra on $\mathcal{C}(X)$)**.** A nonempty subset $\mathcal{A} \subseteq \mathcal{C}(X)$ is said to be a real function algebra in $\mathcal{C}(X)$ if and only if

A1. If $f, g \in \mathcal{A}$, then $f + g \in \mathcal{A}$ and $fg \in \mathcal{A}$;
A2. If $c \in \mathbb{R}$ and $f \in \mathcal{A}$ then $cf \in \mathcal{A}$.

It turns out that algebras of functions are closed under pointwise minimums, maximums and the absolute value.

Lemma 13 (Closure under min and max)**.** *Suppose that X is a compact metric space and $\mathcal{A}$ a closed algebra containing constants in $\mathcal{C}(X)$. If $f, g \in \mathcal{A}$, then $|f|$,* $\min(f, g)$, $\max(f, g) \in \mathcal{A}$.

Proof. Recall that $\min(f, g) = \frac{1}{2}(f + g - |f - g|)$, $\max(f, g) = \frac{1}{2}(f + g + |f + g|)$. Since $\mathcal{A}$ is an algebra, to prove $\min(f, g), \max(f, g) \in \mathcal{A}$, it is sufficient to prove that $|f| \in \mathcal{A}$.
If f is identically zero, the result follows immediately. Thus, suppose f is not identically zero. Then set $M = \|f\|_\infty$ and observe that $M > 0$. Define the function $g : [-1, 1] \to [0, 1]$ by $g(t) = |t|$. By Weierstrass Approximation Theorem, there exists a sequence of polynomials $\{p_n\}_{n=1}^\infty$ defined on $[-1, 1]$ such that $p_n \to g$ in δ_∞.
For every $x \in X$, define $g_n : X \to \mathbb{R}$ by $g_n(x) := p_n(\frac{f(x)}{M})$. By definition of M, $t = \frac{f(x)}{M} \in [-1, 1]$. Thus, the composition g_n is indeed well-defined. Since composition preserves continuity, $g_n \in \mathcal{C}(X)$. We will show $g_n \in \mathcal{A}$. Since $\frac{1}{M} \in \mathcal{A}$ by A2, $\frac{1}{M} f \in \mathcal{A}$. Since p_n is a polynomial, by A1, $g_n \in \mathcal{A}$. By definition of g_n and the choice of p_n,

$$\delta_\infty \left(g_n, \frac{|f|}{M} \right) = \sup_{x \in X} \left| g_n(x) - \left| \frac{f(x)}{M} \right| \right| = \sup_{t \in [-1,1]} |p_n(t) - g(t)| = \delta_\infty (p_n, g) \quad (6.5)$$

Since $p_n \to g$ in δ_∞, $g_n \to \frac{|f|}{M}$ in δ_∞ by 6.5. Thus, $Mg_n \to |f|$ in δ_∞. By A2, $Mg_n \in \mathcal{A}$. Since $\mathcal{A}$ is closed, $|f| \in \mathcal{A}$.

■

It is interesting to note that a closure of an algebra remains an algebra.

Lemma 14 (The uniform closure lemma)**.** *If $\mathcal{A}$ is algebra on $\mathcal{C}(X)$ containing constants, so is $\overline{\mathcal{A}}$.*

Proof. Let $f, g \in \overline{\mathcal{A}}$. There exist $f_n, g_n \in \mathcal{A}$ such that $f_n \to f$, $g_n \to g$ in δ_∞. Then $f_n + g_n \to f + g$ in δ_∞ and $f_n g_n \to fg$ in δ_∞. Similarly, $\lambda f_n \to \lambda f$ in δ_∞. Since $\mathcal{A}$ is an algebra, $f_n + g_n, f_n g_n, \lambda f_n \in \mathcal{A}$. This establishes A1 and A2. ■

Finally, we introduce the notion of separation of points.

Definition 34 (separation on $\mathcal{C}(X)$)**.** A subset $\mathcal{A}$ of $\mathcal{C}(X)$ separates points of X if and only if given $x, y \in X$ with $x \neq y$ there exists $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

We are ready to state and prove the most important result in this section.

Theorem 20 (Stone-Weierstrass Theorem)**.** *Suppose that X is a compact metric space. If $\mathcal{A}$ is an algebra in $\mathcal{C}(X)$ that separates points of X and contains constants then $\mathcal{A}$ is uniformly dense in $\mathcal{C}(X)$.*

Proof. Since $\mathcal{A}$ is an algebra, so is $\overline{\mathcal{A}}$ by The uniform closure lemma. Let $f \in \mathcal{C}(X)$ and let $\epsilon > 0$.

Step 1. We begin by proving that for every $s \in X$, there exists $h_s \in \overline{\mathcal{A}}$ such that for every $x \in X, h_s(x) < f(x) + \frac{\epsilon}{2}$ and $h_s(s) = f(s)$. Fix $s \in X$ and consider $t \in X, s \neq t$. Since $\mathcal{A}$ separates points of X, there exists $h \in A$ such that $h(s) \neq h(t)$. Define

$$f_{s,t}(x) = s \cdot \frac{h(x) - h(t)}{h(s) - h(t)} + t \cdot \frac{h(x) - h(s)}{h(t) - h(s)}.$$

Observe that $f_{s,t}(s) = s$ and $f_{s,t}(t) = t$. Clearly, $f_{s,t} \in \mathcal{A}$ so $f_{s,t} \in \overline{\mathcal{A}}$. Define

$$V_{s,t} = \left\{x : x \in X, f_{s,t}(x) - f(x) < \frac{\epsilon}{2}\right\} = (f_{s,t} - f)^{-1}\left(-\infty, \frac{\epsilon}{2}\right).$$

Since $f_{s,t}(s) = s$ and $f_{s,t}(t) = t$, $s, t \in V_{s,t}$. Since $f_{s,t} \in \mathcal{C}(X)$, $(f_{s,t} - f) \in \mathcal{C}(X)$. By continuity of $f_{s,t}$, $V_{s,t}$ is open since it is the preimage of the open interval. Since $t \in V_{s,t}$, $\{V_{s,t}\}_{t \in X}$ is an open cover of X. By compactness of X, there exist $t_1, \ldots t_n$ such that $X = \bigcup_{k=1}^{n} V_{s,t_k}$. Now define $h_s = \min_{1 \leq k \leq n} f_{s,t_k}$. By Closure under min and max, $h_s \in \overline{\mathcal{A}}$. Since for every $f_{s,t}$, $f_{s,t}(s) = f(s)$, we have $h_s(s) = f(s)$. We will show $h_s(x) < f(x) + \frac{\epsilon}{2}$, for every $x \in X$. Let $x \in X$. Then $x \in V_{s,t_j}$, for at least one j such that $1 \leq j \leq n$. By definition of h_s and the fact $x \in V_{s,t_j}$ for at least one $j \in \{1, 2, \ldots, n\}$,

$$h_s(x) - f(x) \leq f_{s,t_j}(x) - f(x) < \frac{\epsilon}{2} \implies h_s(x) < f(x) + \frac{\epsilon}{2}.$$

Step 2. For $s \in X$, define

$$W_s = \left\{x \in X : h_s(x) > f(x) - \frac{\epsilon}{2}\right\} = (h_s - f)^{-1}\left(-\frac{\epsilon}{2}, \infty\right).$$

By continuity of $h_s - f$, W_s is open since it is the preimage of the open interval. Since $h_s(s) = f(s)$, $s \in W_s$. Therefore, the family $\{W_s\}_{s \in X}$ is an open cover of X. By compactness of X, there exist $s_1, \ldots s_m$ such that $X = \bigcup_{k=1}^{m} W_{s_k}$. Now define $g = \max_{1 \leq k \leq m} h_{s_k}$. By Closure under min and max, $g \in \overline{\mathcal{A}}$. Since for every $k \in \{1, 2, \ldots m\}$, for every $x \in X$, $h_{s_k}(x) < f(x) + \frac{\epsilon}{2}$, $g(x) < f(x) + \frac{\epsilon}{2}$. We will show that for every $x \in X$, $g(x) > f(x) - \frac{\epsilon}{2}$. Fix $x \in X$. Then $x \in W_{s_j}$, for some $j \in \{1, 2, \ldots m\}$. But then, $h_{s_j}(x) > f(x) - \frac{\epsilon}{2}$. Since for every $j \in \{1, 2, \ldots m\}$, for every $y \in X$, $h_{s_j}(y) \leq g(y)$, we have $f(x) - \frac{\epsilon}{2} < h_{s_j}(x) \leq g(x)$. Thus for every $x \in X, -\frac{\epsilon}{2} < f(x) - g(x) < \frac{\epsilon}{2}$. Hence $\delta_\infty(f, g) \leq \frac{\epsilon}{2} < \epsilon$.

■

6.3 Measure Theory and Integration

This section is a concise summary of elementary definitions and results from measure theory. The aim is to set up notation for later more advanced sections. The majority of the presented results have been discussed in *Essentials in Analysis and Probability*, although Carathéodory's theorem is given in a more general form suitable for applications in this thesis. The material about signed measures and decomposition theorems was not part of the syllabus and it is discussed in more detail.

6.3.1 Elementary definitions and notation

Definition 35 (σ-algebra)**.** Let Ω be a set. Let $\mathcal{F}$ be a family of subsets of Ω. We say that $\mathcal{F}$ is a σ-algebra on Ω if it satisfies all of the following

S1. $\Omega \in \mathcal{F}$;
S2. if $E \in \mathcal{F}$ then $\Omega \setminus E \in \mathcal{F}$;
S3. if $\{E_n\}_{n=1}^{\infty}$ where for every $n \in \mathbb{N}$, $E_n \in \mathcal{F}$ then $\bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$.

Definition 36 (measurable set)**.** Let Ω be a set and $\mathcal{F}$ be a σ-algebra on Ω. A subset $E \subseteq \Omega$ is said to be $\mathcal{F}$-measurable or simply measurable if $E \in \mathcal{F}$.

Definition 37 (measurable space)**.** Let Ω be a set and $\mathcal{F}$ be a σ-algebra on Ω. Then the ordered pair $(\Omega, \mathcal{F})$ is a measurable space.

Definition 38 (σ-algebra generated by the set)**.** Let Ω be a set and let $A \subseteq \mathcal{P}(\Omega)$. The σ-algebra generated by A is the smallest σ-algebra on Ω containing A.

Definition 39 (Borel σ-algebra)**.** Let (X, Γ) be a topological space. The Borel σ-algebra of X is the σ-algebra generated by open sets of X, denoted by $\mathcal{B}(X)$. Hence $\mathcal{B}(X) = \sigma(\Gamma)$.

Definition 40 (a measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. The function $\mu : \mathcal{F} \to [0, \infty]$ is a measure if it satisfies all of the following:

M1. $\mu(\emptyset) = 0$;
M2. if $\{E_n\}_{n=1}^{\infty}$ are pairwise disjoint $\mathcal{F}$-measurable sets, $\mu(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu(E_n)$.

A very useful property of a measure is the notion of continuity.

Lemma 15 (Continuity of a measure, Proposition 1.2.5 in [Coh13])**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let μ be a measure on $(\Omega, \mathcal{F})$.*

If $\{E_n\}_{n=1}^{\infty}$ is an increasing sequence of sets in $\mathcal{F}$, then $\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n)$.

If $\{E_n\}_{n=1}^{\infty}$ is a decreasing sequence of sets in $\mathcal{F}$ such that $\mu(E_n) < \infty$ for some $n \in \mathbb{N}$, then $\mu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \mu(E_n)$.

Definition 41 (finite measure)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We say that μ is a finite measure if for every $E \in \mathcal{F}$, $\mu(E) < \infty$.

Definition 42 (σ-finiteness)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We say that μ is a σ- finite measure if there exists a sequence $\{E_n\}_{n=1}^{\infty}$ of $\mathcal{F}$-measurable sets with $\mu(E_n) < \infty$ and $\Omega = \bigcup_{n=1}^{\infty} E_n$.

6.3.2 Construction of a measure and Carathéodory's theorem

Direct construction of a measure is often tedious and sometimes even impossible process. However, it is often possible to simplify the construction by starting with a somewhat weaker set function - outer measure.

Definition 43 (outer measure)**.** Let Ω be a set and let $\mathcal{P}(\Omega)$ be a powerset of Ω. A function $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ is an outer measure if it satisfies all of the following:

O1. $\mu^*(\emptyset) = 0$;
O2. if $A \subseteq B \subseteq \Omega$, then $\mu^*(A) \leq \mu^*(B)$;
O3. if $\{E_n\}_{n=1}^{\infty}$ are $\mathcal{F}$-measurable sets, $\mu^*(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu^*(E_n)$.

Using the concept of an outer measure, we can define outer measurable sets.

Definition 44 (outer measurable sets)**.** Let Ω be a set and let μ^* be an outer measure on $\mathcal{P}(\Omega)$. A subset $E \subseteq \Omega$ is said to be μ^*-measurable if for every $A \subseteq \Omega$, $\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap (\Omega \setminus E))$.

Carathéodory's theorem is a very important tool, guaranteeing existence of a measure induced by the outer measure.

Theorem 21 (Carathéodory's theorem)**.** *Let Ω be a set and let μ^* be an outer measure on $\mathcal{P}(\Omega)$. Suppose that $\mathcal{M}_{\mu^*}$ is a collection of all μ^*-measurable subsets of Ω. Then $\mathcal{M}_{\mu^*}$ is a σ-algebra and the restriction $\mu^*_{|\mathcal{M}_{\mu^*}}$ is a measure on $\mathcal{M}_{\mu^*}$.*

Proof. See Theorem 1.3.6 in [Coh13]. ■

For an application of Carathéodory's theorem and an example of a construction of a measure, see the proof of Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$.
Closely related to the construction of a measure is the concept of Dynkin classes, also known as λ-systems.

Definition 45 (Dynkin class (λ-system))**.** Let Ω be a set and let $\mathcal{P}(\Omega)$ be a powerset of Ω. The collection $\Lambda \subseteq \mathcal{P}(\Omega)$ is a Dynkin class or a λ-system if

L1. $\Omega \in \Lambda$;
L2. if $A \in \Lambda$ and $B \in \Lambda$ such that $B \subseteq A$, then $A \setminus B \in \Lambda$;
L3. if $\{A_n\}_{n=1}^{\infty}$ is a collection of disjoint sets in Λ, then $\bigcup_{n=1}^{\infty} A_n \in \Lambda$.

Definition 46 (π-system)**.** Let Ω be a set and let $\mathcal{P}(\Omega)$ be a powerset of Ω. The collection $\Pi \subseteq \mathcal{P}(\Omega)$ is a a π-system if it is closed under finite intersections.

Theorem 22 (Dynkin's $\lambda - \pi$ theorem)**.** *Let Ω be a set and let Π be a π-system on Ω. Then the σ-algebra generated by Π, denoted $\sigma(\Pi)$, coincides with the λ-system generated by Π.*

Proof. See Theorem 1.6.2 in [Coh13]. ■

Proposition 10 (Proposition 1.5.6 in [Coh13])**.** *Let μ be a finite measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^n))$. Then μ is regular. Moreover, each Borel subset A of $\mathbb{R}^d$ satisfies*

$$\mu(A) = \sup\{\mu(K) : K \subseteq A, \ K \text{ compact}\}.$$

6.3.3 Measurable functions and their properties

Definition 47 (a measurable function)**.** Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. A function $f : X \to Y$ is measurable if for every $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$.

In this thesis, we say that $f : X \to \mathbb{R}$ is measurable if it is measurable with respect to $\mathcal{B}(\mathbb{R})$. As one might expect, measurable functions are closed under various algebraic operations.

Proposition 11 (Proposition 5.7 in [Bas14])**.** *Let $c \in \mathbb{R}$. If f and g are measurable real-valued functions, so are $f + g$, cf, fg, $\max(f, g)$, $\min(f, g)$.*

The following decomposition of a measurable function and corresponding characterization of measurability are often applicable.

Proposition 12. *Let $(\Omega, \mathcal{F})$ be a measurable space. For $f : \Omega \to \mathbb{R}$, define $f^+ : \Omega \to \mathbb{R}$ by $f^+ = \max(f, 0)$ and define $f^- : \Omega \to \mathbb{R}$ by $f^- = \max(-f, 0)$. Then $f = f^+ - f^-$ and f is measurable if and only f^+ and f^- are measurable.*

Proof. $f = f^+ - f^-$ follows directly from definition of f^+ and f^- and the rest follows from Proposition 5.7 in [Bas14]. ■

One of the most useful facts about measurable functions is the closure under various limiting operations.

Proposition 13 (Proposition 2.1.5 in [Coh13])**.** *Let $(\Omega, \mathcal{F})$ be a measurable space, suppose $A \subseteq \Omega$ is $\mathcal{F}$-measurable. Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of $\overline{\mathbb{R}}$-measurable functions on A. Then $\sup_n f_n, \inf_n f_n, \limsup_n f_n, \liminf_n f_n$ are measurable.*

The simplest form of a measurable function is a characteristic function.

Definition 48 (characteristic function)**.** Let Ω be a set. If $E \subseteq \Omega$, we define the characteristic function of E, denoted χ_E, by

$$\chi_E(x) = \begin{cases} 1 & x \in E \\ 0 & x \notin E \end{cases}.$$

Linear combinations of characteristic functions are simple functions.

Definition 49 (simple function)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\varphi : \Omega \to \mathcal{F}$ is simple if it is of the form $\varphi = \sum_{k=1}^{n} a_k \chi_{A_k}$ where $\{A_k\}_{k=1}^{n}$ are $\mathcal{F}$-measurable sets and $\{a_k\}_{k=1}^{n}$ are real numbers.

It is interesting to note that measurable functions are pointwise limits of simple functions. This property is often used in proofs.

Proposition 14 (Proposition 5.14 in [Bas14])**.** *Suppose f is a non-negative and measurable function. Then there exists a sequence of non-negative measurable simple functions $\{\varphi_n\}_{n=1}^{\infty}$ increasing to f. Hence for every $n \in \mathbb{N}$, $0 \leq \varphi_n \leq \varphi_{n+1}$ and $\lim_{n\to\infty} \varphi_n = f$ pointwise.*

6.3.4 Lebesgue Integration

Development of the integration theory usually starts by defining the integral of some class of simple functions. In this case, we begin with non-negative simple functions.

Definition 50 (an integral of a non-negative, simple function)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that φ is a non-negative, simple function of the form $\varphi = \sum_{k=1}^{n} a_k \chi_{A_k}$. We define an integral of a φ by

$$\int_\Omega \varphi \, d\mu = \sum_{k=1}^{n} a_k \mu(A_k).$$

Remark 27. The proof that the integral of a non-negative, simple function is indeed well-defined is discussed on page 53 in [Coh13].

Now we generalize the integral to non-negative, measurable functions.

Definition 51. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that f is a non-negative, measurable function. The integral of f is defined by

$$\int_\Omega f \, d\mu = \sup \left\{ \int_\Omega g \, d\mu : 0 \leq g \leq f, \text{ g simple} \right\}.$$

Theorem 12 guarantees that every measurable function f can be expressed in the form $f = f^+ - f^-$, where f^+ and f^- are non-negative and measurable. We use this decomposition to define the integral of an arbitrary measurable function.

Definition 52. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that f is arbitrary, extended real-valued measurable function. If at least one of $\int_\Omega f^+ \, d\mu, \int_\Omega f^- \, d\mu$ is finite, we define integral of f by

$$\int_\Omega f \, d\mu = \int_\Omega f^+ \, d\mu - \int_\Omega f^- \, d\mu.$$

Definition 53. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that f is extended real-valued measurable function. We say that f is integrable if $\int_\Omega |f| \, d\mu < \infty$.

The following criteria for a function to be zero almost everywhere is very helpful.

Proposition 15 (Proposition 8.1 in [Bas14])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $f : \Omega \to [0, \infty]$ is $\mathcal{F}$-measurable, satisfying $\int_\Omega f \, d\mu = 0$. Then $f = 0$ μ-almost everywhere.*

Proposition 16. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $f : \Omega \to [0, \infty]$ is $\mathcal{F}$-measurable. Then $f = 0$ μ-almost everywhere if and only if $\int_\Omega f \, d\mu = 0$.*

Proof. By Proposition 15, it is sufficient to prove that $f = 0$ μ-almost everywhere implies $\int_\Omega f = 0$, $d\mu$. Suppose $f = 0$ μ-almost everywhere. Then there exists a set $E \subseteq \Omega$ such that $\mu(E) = 0$ and $f > 0$ on E while $f = 0$ on $\Omega \setminus E$. Then $\int_\Omega f \, d\mu = \int_E f \, d\mu + \int_{\Omega \setminus E} f \, d\mu = 0$, since f is identically zero on $\Omega \setminus E$ and E is a set of measure zero. ■

The following inequality is a well-known neat result.

Proposition 17 (Markov's Inequality)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $f : \Omega \to [0, \infty]$ is $\mathcal{F}$-measurable. If ϵ is a positive real number, then*

$$\mu(\{\omega \in \Omega : f(\omega) \geq \epsilon\}) \leq \frac{1}{\epsilon} \int_\Omega f \, d\mu.$$

Proof. See Proposition 2.3.10 in [Coh13]. ■

Proposition 18 and Proposition 18 guarantee that the integral satisfies expected properties.

Proposition 18 (Proposition 2.3.4 in [Coh13])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f, g : \Omega \to [0, \infty]$ be measurable and suppose that $\alpha \geq 0$. Then*

1. *$\int_\Omega \alpha f \, d\mu = \alpha \int_\Omega f \, d\mu$,*
2. *$\int_\Omega (f + g) \, d\mu = \int_\Omega f \, d\mu + \int_\Omega g \, d\mu$, and*
3. *if $f(\omega) \leq g(\omega)$ for every $\omega \in \Omega$, then $\int_\Omega f \, d\mu \leq \int_\Omega g \, d\mu$.*

Proposition 19 (Proposition 2.3.6 in [Coh13])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f, g : \Omega \to [-\infty, \infty]$ be measurable and suppose that $\alpha \in \mathbb{R}$. Then*

1. *αf is integrable and $\int_\Omega \alpha f \, d\mu = \alpha \int_\Omega f \, d\mu$,*
2. *$f + g$ is integrable and $\int_\Omega (f + g) \, d\mu = \int_\Omega f \, d\mu + \int_\Omega g \, d\mu$, and*
3. *if $f(\omega) \leq g(\omega)$ for every $\omega \in \Omega$, then $\int_\Omega f \, d\mu \leq \int_\Omega g \, d\mu$.*

The following three theorems are known as limiting or convergence theorems. Those theorems are fundamental results in Lebesgue integration theory and one of the main reasons for its success.

Theorem 23 (Monotone Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $\{f_n\}_{n=1}^\infty$ is a sequence of $[0, \infty]$-valued $\mathcal{F}$-measurable functions on Ω and $f : \Omega \to [0, \infty]$ such that*

$$f_n(\omega) \leq f_{n+1}(\omega) \text{ for every } n \in \mathbb{N} \quad \text{and } f(\omega) = \lim_{n \to \infty} f_n(\omega),$$

hold at μ-almost every $\omega \in \Omega$. Then

$$\int_\Omega f \, d\mu = \lim_{n \to \infty} \int_\Omega f_n \, d\mu.$$

Proof. See Theorem 2.4.1 and its proof in [Coh13]. ■

Fatou's Lemma is a very useful corollary of Monotone Convergence Theorem.

Theorem 24 (Fatou's Lemma)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $\{f_n\}_{n=1}^\infty$ is a sequence of $[0, \infty]$-valued $\mathcal{F}$-measurable functions on Ω. Then*

$$\int_\Omega \liminf_{n \to \infty} f_n \, d\mu \leq \liminf_{n \to \infty} \int_\Omega f_n \, d\mu.$$

Proof. See Theorem 2.4.4 and its proof in [Coh13]. ■

Remark 28. Usefulness of Fatou's Lemma follows from the fact it imposes almost no requirements on f_n. An example of such an use-case is in the proof of Riesz Representation Theorem for the Dual of $\mathcal{L}^p$. Using the Monotone Convergence Theorem and Fatou's Lemma, it is possible to prove the following very important theorem in Lebesgue Integration.

Theorem 25 (Dominated Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $g : \Omega \to [0, \infty]$ is an integrable function on Ω. Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of $[-\infty, \infty]$-valued $\mathcal{F}$-measurable functions on Ω such that*

$$|f_n(\omega)| \leq g(\omega) \text{ for every } n \in \mathbb{N}, \text{ and}$$
$$f(\omega) = \lim_{n\to\infty} f_n(\omega),$$

hold at μ-almost every $\omega \in \Omega$. Then f and f_n are integrable for each $n \in \mathbb{N}$. Moreover, $\int_\Omega f \, d\mu = \lim_{n\to\infty} \int_\Omega f_n \, d\mu$.

Proof. See Theorem 2.4.5 and its proof in [Coh13]. ■

The following theorem will help us evaluate Lebesgue integrals when Riemann integrals exist.

Theorem 26 (Equivalence Riemann - Lebesgue)**.** *A bounded Borel measurable real-valued function f on $[a, b]$ is Riemann integrable if and only if the set of points at which f is discontinuous has Lebesgue measure zero. In that case the Riemann integral of f is equal in value to the Lebesgue integral of f.*

Proof. See Theorem 9.1 and its proof in [Bas14]. ■

6.3.5 Modes of Convergence

The following three propositions are standard results addressing convergence in measure.

Proposition 20 (Proposition 3.1.2 in [Coh13])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that f and $\{f_n\}_{n=1}^{\infty}$ are real-valued, $\mathcal{F}$-measurable functions on Ω. If μ is finite and if $f_n \to f$ μ-almost everywhere, then $f_n \to f$ in μ.*

Proposition 21 (Proposition 3.1.3 in [Coh13])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that f and $\{f_n\}_{n=1}^{\infty}$ are real-valued, $\mathcal{F}$-measurable functions on Ω. If $f_n \to f$ in μ, then there exists a subsequence of $\{f_n\}_{n=1}^{\infty}$ converging to f μ-almost everywhere.*

Proposition 22 (Proposition 3.1.5 in [Coh13])**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that f and $\{f_n\}_{n=1}^{\infty}$ belong to $\mathcal{L}^1(\Omega, \mathcal{F}, \mu)$. If $f_n \to f$ in $\|\cdot\|_1$, then $f_n \to f$ in measure μ.*

Lusin's Theorem, Theorem 7.4.4 in [Coh13] relates continuous and measurable functions on locally compact Hausdorff space.

Theorem 27 (Lusin's Theorem, Theorem 7.4.4 in [Coh13])**.** *Let X be a locally compact Hausdorff space and let $\mathcal{A}$ be a σ-algebra that includes $\mathcal{B}(X)$. Let μ be a regular measure on $(X, \mathcal{A})$ and suppose $f : X \to \mathbb{R}$ is measurable. If $A \in \mathcal{A}$ and satisfies $\mu(A) < \infty$ and if $\epsilon > 0$, then there is a compact set $K \subseteq A$ such that $\mu(A \setminus K) < \epsilon$ and $f_{|K}$ is continuous. Moreover, there is a function $g \in \mathcal{C}(X)$ that agrees with f on K. If $A \neq \emptyset$ and f is bounded on A, g can be chosen to satisfy $\sup\{|g(x)| : x \in X\} \leq \sup\{|f(x)| : x \in A\}$.*

Proof. See Theorem 7.4.4 in [Coh13]. ■

6.3.6 Product Measure and Fubini's Theorem

Definition 54 (product σ-algebra)**.** Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be two measurable spaces. We define the product σ-algebra, written $\mathcal{A} \otimes \mathcal{B}$ by $\mathcal{A} \otimes \mathcal{B} = \sigma(\mathcal{A} \times \mathcal{B})$.

It is possible to construct the unique measure $\mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$ satisfying

$$(\mu \times \nu)(A \times B) = \mu(A) \cdot \nu(B), \text{ for every } A \in \mathcal{A},\ B \in \mathcal{B}.$$

That measure is called the product measure. The construction of product measure and verification of the claim above is thoroughly discussed in [Bas14]. The Fubini-Tonelli Theorem helps us evaluate integrals with respect to the product measure. We will use this result extensively in the section discussing Fourier Analysis.

Theorem 28 (Fubini-Tonelli Theorem)**.** *Suppose $f : \Omega \times \Gamma \to \mathbb{R}$ is measurable with respect to $\Omega \times \Gamma$. If either f is non-negative or $\int_{\Omega\times\Gamma} |f| \, d(\mu \times \nu) < \infty$ then*

$$\int_{\Omega\times\Gamma} f(x,y)\, d(\mu \times \nu) = \int_{\Omega} \left[\int_{\Gamma} f(x,y)\, d\nu \right] d\mu = \int_{\Gamma} \left[\int_{\Omega} f(x,y)\, d\mu \right] d\nu.$$

Proof. See Theorem 11.3 and its proof in [Bas14]. ■

Remark 29. Fubini-Tonelli Theorem also holds if we replace the assumption $\int_{\Omega\times\Gamma} |f| \, d(\mu\times\nu) < \infty$ with $\int_{\Omega} \left[\int_{\Gamma} |f(x,y)| \, d\nu \right] d\mu < \infty$ or $\int_{\Gamma} \left[\int_{\Omega} |f(x,y)| \, d\mu \right] d\nu < \infty$. For the justification, see the discussion following Theorem 11.3 in [Bas14].

6.3.7 Signed measures and their decompositions

This section is devoted to signed measures, which can be seen as a generalization of measures. We will also discuss the relationship between signed measures and measures. We begin with a definition.

Definition 55 (a signed measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. The function $\nu : \mathcal{F} \to \overline{\mathbb{R}}$ is a signed measure if it satisfies all of the following:

S1. $\nu(\emptyset) = 0$;
S2. ν does not attain both $+\infty$ and $-\infty$;
S3. if $\{E_n\}_{n=1}^{\infty}$ are pairwise disjoint $\mathcal{F}$-measurable sets, $\nu(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \nu(E_n)$.

As one might expect, the continuity results extend from measures.

Lemma 16 (Continuity of a signed measure)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let μ be a measure on $(\Omega, \mathcal{F})$.*

If $\{E_n\}_{n=1}^{\infty}$ is an increasing sequence of sets in $\mathcal{F}$, then $\nu\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \nu(E_n)$.

If $\{E_n\}_{n=1}^{\infty}$ is a decreasing sequence of sets in $\mathcal{F}$ such that $\nu(E_n)$ is finite for some $n \in \mathbb{N}$, then $\nu\left(\bigcap_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} \nu(E_n)$.

Proof. See Lemma 4.1.2 in [Coh13]. ■

To discuss the relationship between signed measures and measures, it is useful to introduce the idea of positive and negative sets.

Definition 56 (a positive set)**.** Let μ be a signed measure on the measurable space $(\Omega, \mathcal{F})$. We say that $P \subseteq \Omega$ is a positive set for μ if P is $\mathcal{F}$-measurable and each $\mathcal{F}$-measurable subset E of P satisfies $\mu(E) \geq 0$.

Definition 57 (a negative set)**.** Let μ be a signed measure on the measurable space $(\Omega, \mathcal{F})$. We say that $N \subseteq \Omega$ is a negative set for μ if N is $\mathcal{F}$-measurable and each $\mathcal{F}$-measurable subset F of N satisfies $\mu(F) \leq 0$.

Now we present the fundamental decomposition result.

Theorem 29 (Hahn Decomposition Theorem)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let μ be a signed measure on $(\Omega, \mathcal{F})$. Then there are disjoint sets $P \subseteq \Omega$, $N \subseteq \Omega$ such that P is positive for μ, N is negative for μ and $\Omega = P \cup N$.*

To simplify the proof of this theorem, we will extract the following lemma.

Lemma 17 (Negative set lemma)**.** *Let μ be a signed measure on the measurable space $(\Omega, \mathcal{F})$ and let $A \subseteq \Omega$ be $\mathcal{F}$-measurable, satisfying $-\infty < \mu(A) < 0$. Then there exists a negative set B included in A such that $\mu(B) \leq \mu(A)$.*

Proof. The idea is to remove a carefully constructed sequence of subsets from A and let B consist of the remaining elements of A. We begin by letting

$$\delta_1 = \sup\{\mu(E) : E \in \mathcal{F}, E \subseteq A\}.$$

Clearly, $\emptyset \subseteq A$ and $\emptyset \in \mathcal{F}$, so δ_1 exists, not necessarily finite. Moreover, since $\mu(\emptyset) = 0$, $\delta_1 \geq 0$. By definition of δ_1, we can choose $A_1 \in \mathcal{F}$ such that $A_1 \subseteq A$ and $\mu(A_1) \geq \min\{\frac{\delta_1}{2}, 1\}$. We proceed inductively, defining

$$\delta_n = \sup\left\{\mu(E) : E \in \mathcal{F}, E \subseteq A \setminus \left(\bigcup_{k=1}^{n-1} A_k\right)\right\}, n \in \mathbb{N}.$$

We choose a sequence $A_n \subseteq A \setminus \left(\bigcup_{k=1}^{n-1} A_k\right)$ satisfying $\mu(A_n) \geq \min\{\frac{\delta_n}{2}, 1\}$. By the same argument as for δ_1, we deduce $\delta_n \geq 0$. Observe that $\{A_n\}_{n=1}^{\infty}$ are disjoint by construction. Now define

$$A_\infty = \bigcup_{n=1}^{\infty} A_n \text{ and } B = A \setminus A_\infty.$$

Clearly, $B \in \mathcal{F}$ and $B \subseteq A$. We claim B is a desired set. Since for every $n \in \mathbb{N}, \mu(A_n) \geq 0$, by countable additivity, $\mu(A_\infty) \geq 0$. We also have $\mu(A) = \mu(A_\infty) + \mu(B)$. Since $\mu(A_\infty) \geq 0$, it follows that $\mu(B) \leq \mu(A)$, as desired. It remains to show B is a negative set. By assumption, $\mu(A)$ is finite. Since $A_\infty \subseteq A$, we have $\mu(A_\infty)$ is finite. By countable additivity of μ, $\mu(A_\infty) = \sum_{n=1}^{\infty} \mu(A_n)$. Since $\mu(A_\infty) < \infty$, $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. Hence $\lim_{n\to\infty} \mu(A_n) = 0$. Now consider the inequality

$$0 \leq \min\left\{\frac{\delta_n}{2}, 1\right\} \leq \mu(A_n), \forall n \in \mathbb{N}.$$

Since $\lim_{n\to\infty} \mu(A_n) = 0$, by Squeeze Theorem, $\lim_{n\to\infty} \delta_n = 0$. Suppose that $E \in \mathcal{F}$ is such that $E \subseteq B$. By definition of B, $E \subseteq A \setminus \left(\bigcup_{k=1}^{n-1} A_k\right)$ for every $n \in \mathbb{N}$. Therefore, $\mu(E) \leq \delta_n$, for every $n \in \mathbb{N}$. Hence $\mu(E) \leq \lim_{n\to\infty} \delta_n = 0$. Thus, B is indeed negative for μ. ■

Proof of the Hahn Decomposition Theorem. Since the signed measure μ cannot take both $+\infty$ and $-\infty$, without loss of generality, suppose that $\mu : \mathcal{F} \to (-\infty, \infty]$. The idea is to construct a negative set N. To perform the construction, consider

$$L := \inf\{\mu(A) : A \text{ is a negative set for } \mu\}.$$

Since $\emptyset$ is negative for μ, $L \leq 0$. By approximation property of sup, there exists a sequence of negative sets $\{A_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} \mu(A_n) = L$. Now define the family of sets $\{A'_n\}_{n=1}^{\infty}$ such that

$$\begin{aligned} A'_1 &= A_1 \\ A'_2 &= A_2 \setminus A_1 \\ &\vdots \\ A'_n &= A_n \setminus \bigcup_{k=1}^{n-1} A_k. \end{aligned}$$

Clearly, $A'_n \in \mathcal{F}$. By construction, $\{A'_n\}_{n=1}^{\infty}$ is a disjoint family such that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} A'_n$. Now define sets N, P as follows

$$N = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} A'_n \text{ and } P = \Omega \setminus N. \tag{6.6}$$

We claim N is a negative set. Let $E \in \mathcal{F}$ and suppose that $E \subseteq N$. By 6.6, $E = \bigcup_{n=1}^{\infty}(E \cap A'_n)$. Since each $(E \cap A'_n)$ is a subset of A_n which is negative, $\mu(E \cap A'_n) \leq 0$. By countable additivity, $\mu(E) = \sum_{n=1}^{\infty} \mu(E \cap A'_n) \leq 0$. Hence N is negative for μ, as claimed.
We claim $\mu(N) = L$. Since N is negative, $L \leq \mu(N)$. For every $n \in \mathbb{N}$, $N = A_n \cup (N \setminus A_n)$. Then $\mu(N) = \mu(A_n) + \mu(N \setminus A_n)$. Since N is negative for μ, $\mu(N \setminus A_n) \leq 0$. Thus for every $n \in \mathbb{N}$, $\mu(N) \leq \mu(A_n)$. Taking the limit as $n \to \infty$, $\mu(N) \leq \lim_{n\to\infty} \mu(A_n) = L$. Therefore, $\mu(N) = L$.

It remains to show P is positive for μ. We will argue by contradiction. Suppose P is not positive, so there exists $A \in \mathcal{F}$ such that $A \subseteq P$ and $\mu(A) < 0$. Since μ does not attain $-\infty$, $-\infty < \mu(A) < 0$. By Negative set lemma, there exists a negative set $B \subseteq A$ such that $\mu(B) \leq \mu(A) < 0$. Now consider the set $B \cup N$. Since B and N are negative sets, $B \cup N$ is negative. Since $B \subseteq A$, $B \subseteq P$. Since $P \cap N = \emptyset$, $B \cap N = \emptyset$. But then, $L \leq \mu(B \cup N) = \mu(B) + \mu(N) < \mu(N) = L$. This is a contradiction. Hence P is positive for μ, as desired. ■

Using the Hahn Decomposition Theorem, we can precisely describe the relationship between signed measures and measures.

Theorem 30 (Hahn-Jordan decomposition)**.** *Every signed measure is a difference of two measures, at least one of which is finite.*

Proof. Let μ be a signed measure on a the measurable space $(\Omega, \mathcal{F})$. Let (P, N) be the decomposition of Ω given by Hahn Decomposition Theorem. Now define functions μ^+, μ^- on $\mathcal{F}$ by

$$\mu^+(A) = \mu(A \cap P),$$
$$\mu^-(A) = -\mu(A \cap N).$$

Since P is a positive set for μ and N is a negative set for μ, μ^+ and μ^- are both measures, satisfying $\mu = \mu^+ - \mu^-$. Since μ cannot attain both $+\infty$ and $-\infty$, at least one of measures μ^+ and μ^- is finite. ■

Definition 58 (variation of a signed measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space and suppose that ν is a signed measure on $\mathcal{F}$. Write $\nu = \nu^+ - \nu^-$, as in Theorem 30. The variation of a signed measure ν is a measure $|\nu|$ defined by $|\nu| = \nu^+ + \nu^-$.

6.3.8 Absolute continuity and Radon-Nikodym Theorem

In this section, we introduce the idea of the absolute continuity of a measure and prove the key result about the absolutely continuous measures - Radon-Nikodym Theorem.

Definition 59 (absolute continuity of a measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space and let μ, ν be measures on $\mathcal{F}$. We say that ν is absolutely continuous with respect to μ, denoted by $\nu \ll \mu$, if

$$\forall A \in \mathcal{F}, \mu(A) = 0 \implies \nu(A) = 0.$$

Definition 60 (absolute continuity of a signed measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space. Suppose that ν is a signed measure on $\mathcal{F}$ and μ is a measure on $\mathcal{F}$. We say that ν is absolutely continuous with respect to μ if its variation $|\nu|$ is absolutely continuous with respect to μ, in the sense of Definition 59.

Example 6. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Suppose that $g : \Omega \to \mathbb{R}$ is measurable and nonnegative. Define $\nu : \mathcal{F} \to [0, \infty]$ by

$$\nu(A) = \int_A g \, d\mu, \text{ for every } A \in \mathcal{F}. \tag{6.7}$$

By appealing to additivity of the integral and Monotone Convergence Theorem, it is not difficult to show that ν is a measure on $\mathcal{F}$. Suppose that $A \in \mathcal{F}$ and $\mu(A) = 0$. Since the integral over a null set vanishes, $\nu(A) = \int_A g \, d\mu = 0$. Hence, ν is absolutely continuous with respect to μ.

The natural question is whether all absolutely continuous measures are in the form given by 6.7. Radon-Nikodym Theorem for measures gives conditions under which this is true and it can be seen as a partial converse to Example 6.

Theorem 31 (Radon-Nikodym Theorem for measures)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mu, \nu : \mathcal{F} \to [0, \infty]$ be σ-finite measures. If $\nu \ll \mu$, then there is a $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable function $g : \Omega \to [0, \infty)$ such that*

$$\nu(A) = \int_A g \, d\mu, \text{ for every } A \in \mathcal{F}.$$

The function g is unique up to μ-almost everywhere equality and $g \geq 0$.

Proof Idea. The proof of Radon-Nikodym Theorem for measures will be divided into three parts. In the first part, we will show that it is sufficient to prove the theorem for finite measures. In the second part, we demonstrate the existence of g for the case of finite measures. In the last part, we will prove the μ-almost everywhere uniqueness of g.

Proof.

Step 1 (Reduction to spaces of a finite measure). It is sufficient to prove this result for **finite** measures. In order to justify this, suppose that the theorem holds for **finite** measures and consider μ, ν as in the statement. By σ-finiteness of μ, there exists a sequence of $\mathcal{F}$-measurable sets $\{E_n\}_{n=1}^{\infty}$ such that each $\mu(E_n) < \infty$ and $\Omega = \bigcup_{n=1}^{\infty} E_n$. By σ-finiteness of ν, there exists a sequence of $\mathcal{F}$-measurable sets $\{F_m\}_{m=1}^{\infty}$ such that each $\nu(F_m) < \infty$ and $\Omega = \bigcup_{m=1}^{\infty} F_m$. Define $G_{n,m} = E_n \cap F_m$. Clearly, $G_{n,m} \in \mathcal{F}$ and $\mu(G_{n,m}) < \infty$ and $\nu(G_{n,m}) < \infty$. We claim $\bigcup_{n=1}^{\infty} \bigcup_{m=1}^{\infty} G_{n,m} = \Omega$. The inclusion $\bigcup_{n=1}^{\infty} \bigcup_{m=1}^{\infty} G_{n,m} \subseteq \Omega$ is trivial. Suppose that $\omega \in \Omega$. Since $\Omega = \bigcup_{n=1}^{\infty} E_n$, there exists $k \in \mathbb{N}$ such that $\omega \in E_k$. Similarly, there exists $l \in \mathbb{N}$ such that $\omega \in F_l$. Therefore, $\omega \in G_{k,l}$ so $\Omega \subseteq \bigcup_{n=1}^{\infty} \bigcup_{m=1}^{\infty} G_{n,m}$. Hence $\Omega = \bigcup_{n=1}^{\infty} \bigcup_{m=1}^{\infty} G_{n,m}$. It is clear that the family $\{G_{n,m}\}_{n,m \in \mathbb{N}}$ is countable so we may index it as $\{G_n\}_{n \in \mathbb{N}}^{\infty}$. Now define the family $\{H_n\}_{n \in \mathbb{N}}^{\infty}$ as follows

$$\begin{aligned} H_1 &= G_1 \\ H_2 &= G_2 \setminus G_1 \\ &\vdots \\ H_n &= G_n \setminus \bigcup_{k=1}^{n-1} G_k. \end{aligned} \tag{6.8}$$

By construction, $H_i \cap H_j = \emptyset$ for every $i, j \in \mathbb{N}, i \neq j$. Clearly, $\Omega = \bigcup_{n=1}^{\infty} H_n$ and $H_n \in \mathcal{F}$. Since $H_n \subseteq G_n$, by monotonicity of measure, $\mu(H_n) < \infty$ and $\nu(H_n) < \infty$. Now for each $n \in \mathbb{N}$, consider the measurable space $(H_n, \mathcal{F}_{|H_n})$.

Since $\nu \ll \mu$, $\nu_{|H_n} \ll \mu_{|H_n}$. By assumption that the theorem holds for finite measure spaces, there exists a $(\mathcal{F}_{|H_n}, \mathcal{B}(\mathbb{R}))$-measurable function $g_n : H_n \to [0, \infty)$ such that

$$\nu(A \cap H_n) = \int_{A \cap H_n} g_n \, d\mu, \text{ for every } A \in \mathcal{F}. \tag{6.9}$$

Moreover, g_n is unique up to μ-almost everywhere equality and $g_n \geq 0$. We will construct a $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable function $g : \Omega \to [0, \infty)$ such that

$$\nu(A) = \int_A g \, d\mu, \text{ for every } A \in \mathcal{F}.$$

We may extend g_n to $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable function $h_n : \Omega \to [0, \infty)$ by setting $h_n = g_n \chi_{H_n}$. Now define:

$$g = \sum_{n=1}^{\infty} h_n.$$

Since $h_n \geq 0$, $\sum_{k=1}^{n} h_k$ increases monotonically to g. Clearly, each $\sum_{k=1}^{n} h_k$ is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable. Since g is a pointwise limit of $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable, monotonically increasing nonnegative functions, g is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable and nonnegative. Since $\{H_n\}_{n=1}^{\infty}$ is a disjoint family, by definition of h_n and g,

$$\int_{A \cap H_n} g \, d\mu = \int_{A \cap H_n} h_n \, d\mu, \text{ for every } A \in \mathcal{F}. \tag{6.10}$$

Let $A \in \mathcal{F}$. Since $\{H_n\}_{n=1}^{\infty}$ partitions Ω, we have $A = \bigcup_{n=1}^{\infty}(A \cap H_n)$. Now,

$$\begin{aligned}
\nu(A) = \nu(\bigcup_{n=1}^{\infty}(A \cap H_n)) &= \sum_{n=1}^{\infty} \nu(A \cap H_n) \\
&= \sum_{n=1}^{\infty} \int_{A \cap H_n} g_n \, d\mu && \text{by 6.9} \\
&= \sum_{n=1}^{\infty} \int_{A \cap H_n} h_n \, d\mu && \text{since } h_n = g_n \chi_{H_n} \\
&= \sum_{n=1}^{\infty} \int_{A \cap H_n} g \, d\mu \\
&= \sum_{n=1}^{\infty} \int_{A} g \chi_{H_n} \, d\mu. && \text{by 6.10}
\end{aligned}$$

By additivity of the integral and the fact $\{H_n\}_{n=1}^{\infty}$ partitions Ω, we can write

$$\begin{aligned}
\sum_{n=1}^{\infty} \int_A g\chi_{H_n} \, d\mu = \lim_{N \to \infty} \sum_{n=1}^{N} \int_A g\chi_{H_n} \, d\mu &= \lim_{N \to \infty} \int_A \sum_{n=1}^{N} g\chi_{H_n} \, d\mu \\
&= \lim_{N \to \infty} \int_A g\chi_{\bigcup_{k=1}^{N} H_n} \, d\mu.
\end{aligned} \tag{6.11}$$

Now consider $\phi_N = g\chi_{\bigcup_{k=1}^N H_n}$. Clearly, ϕ_N is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable. By construction of $\{H_n\}_{n=1}^\infty$ (6.8), ϕ_N monotonically increases to g. Applying Monotone Convergence Theorem to 6.11 gives

$$\nu(A) = \lim_{N\to\infty} \int_A g\chi_{\bigcup_{k=1}^N H_n}\, d\mu = \int_A \lim_{N\to\infty} g\chi_{\bigcup_{k=1}^N H_n}\, d\mu = \int_A g\, d\mu.$$

Therefore, it is sufficient to prove the theorem assuming that μ, ν are finite.

Step 2 (Existence for finite measure spaces). Consider

$$\mathcal{H} = \left\{ f : \Omega \to [0,\infty] : f \text{ is } (\mathcal{F}, \mathcal{B}(\overline{\mathbb{R}})) \text{ -measurable}, \int_A f\, d\mu \le \nu(A), \forall A \in \mathcal{F} \right\}.$$

Clearly, $f = 0 \in \mathcal{H}$ so $\mathcal{H}$ is nonempty. We will show that there exists $g \in \mathcal{H}$ such that $\nu(A) = \int_A g\, d\mu$, for every $A \in \mathcal{F}$. We claim

$$f, g \in \mathcal{H} \implies \max(\mathrm{f}, \mathrm{g}) \in \mathcal{H}. \tag{6.12}$$

Let $f, g \in \mathcal{H}$ and let $A \in \mathcal{F}$. Set $F := \{\omega \in A : f(\omega) > g(\omega)\}$, $G := \{\omega \in A : f(\omega) \le g(\omega)\}$. Since F and G partition A, we have

$$\int_A \max(\mathrm{f}, \mathrm{g})\, d\mu = \int_F f\, d\mu + \int_G g\, d\mu \le \nu(F) + \nu(G) = \nu(A).$$

Hence $\max(\mathrm{f}, \mathrm{g}) \in \mathcal{H}$. Now set

$$\alpha = \sup \left\{ \int_\Omega f\, d\mu : f \in \mathcal{H} \right\}. \tag{6.13}$$

For every $f \in \mathcal{H}$, $\int_\Omega f\, d\mu \le \nu(\Omega)$ so $\alpha \le \nu(\Omega) < \infty$. Therefore, α is finite. By approximation property of sup, there exists a sequence $\{f_n\}_{n=1}^\infty$ in $\mathcal{H}$ such that

$$\lim_{n\to\infty} \int_\Omega f_n\, d\mu = \alpha. \tag{6.14}$$

Now set $g_n = \max_{1\le k\le n} f_k$. By 6.12, $g_n \in \mathcal{H}$. Clearly, $g_{n+1} \ge g_n \ge 0, \forall n \in \mathbb{N}$. But then, $\int_\Omega g_n\, d\mu \le \int_\Omega g_{n+1}\, d\mu$ so $\lim_{n\to\infty} \int_\Omega g_n\, d\mu$ exists. Since $g_n \in \mathcal{H}$, $\int_\Omega g_n\, d\mu \le \alpha$. Therefore, $\lim_{n\to\infty} \int_\Omega g_n\, d\mu \le \alpha$. By construction, $g_n \ge f_n \ge 0$. Hence, $\int_\Omega g_n\, d\mu \ge \int_\Omega f_n\, d\mu$. Now $\lim_{n\to\infty} \int_\Omega g_n\, d\mu \ge \lim_{n\to\infty} \int_\Omega f_n\, d\mu = \alpha$, by 6.14. We deduce

$$\lim_{n\to\infty} \int_\Omega g_n\, d\mu = \alpha. \tag{6.15}$$

Since $g_{n+1} \ge g_n \ge 0, \forall n \in \mathbb{N}$, $g = \lim_{n\to\infty} g_n$ is well-defined and $(\mathcal{F}, \mathcal{B}(\overline{\mathbb{R}}))$-measurable. Clearly, $g \ge 0$. By Monotone Convergence Theorem and 6.15,

$$\int_\Omega g\, d\mu = \lim_{n\to\infty} \int_\Omega g_n\, d\mu = \alpha.$$

Consider $A \in \mathcal{F}$. Since $g_n \in \mathcal{H}$, $\int_A g_n\, d\mu \le \nu(A)$ so $\lim_{n\to\infty} \int_A g_n\, d\mu \le \nu(A)$.

By Monotone Convergence Theorem,

$$\int_A g\,d\mu = \lim_{n\to\infty}\int_A g_n\,d\mu \le \nu(A).$$

We have shown $g \in \mathcal{H}$. It remains to show $\nu(A) = \int_A g\,d\mu, \forall A \in \mathcal{F}$. To prove this, define $\nu_0 : \mathcal{F} \to [0,\infty)$ by $\nu_0 := \nu(A) - \int_A g\,d\mu$. We will show $\nu_0 = 0$. Since $g \in \mathcal{H}$, ν_0 is nonnegative. Since ν is a measure and $\emptyset$ is a null set, $\nu_0(\emptyset) = 0$. Since ν is a measure and Monotone Convergence Theorem applies, ν_0 is countably additive and hence a measure. Since ν_0 is a measure, to prove $\nu_0 = 0$, it is sufficient to show $\nu_0(\Omega) = 0$. We will argue by contradiction. Suppose $\nu_0(\Omega) > 0$. Since $\mu(\Omega) < \infty$, there exists $\epsilon > 0$ such that

$$\nu_0(\Omega) > \epsilon\mu(\Omega). \tag{6.16}$$

Now consider the signed measure $\nu_0 - \epsilon\mu$. Let (P, N) be its Hahn-Jordan decomposition. Since P is a positive set for $\nu_0 - \epsilon\mu$, we have

$$(\nu_0 - \epsilon\mu)(A \cap P) \ge 0 \implies \nu_0(A \cap P) \ge \epsilon\mu(A \cap P), \forall A \in \mathcal{F}. \tag{6.17}$$

Consider $A \in \mathcal{F}$ and $(\mathcal{F}, \mathcal{B}(\overline{\mathbb{R}}))$-measurable function $g + \epsilon\chi_P$. We have

$$\begin{aligned}
\nu(A) &= \int_A g\,d\mu + \nu_0(A) \ge \int_A g\,d\mu + \nu_0(A \cap P) \\
&\ge \int_A g\,d\mu + \epsilon\mu(A \cap P) && \text{by 6.17} \\
&= \int_A (g + \epsilon\chi_P)\,d\mu.
\end{aligned}$$

Therefore, $g + \epsilon\chi_P \in \mathcal{H}$. We claim $\mu(P) > 0$. Suppose not. Then $\mu(P) = 0$. Since $\nu \ll \mu$, $\nu(P) = 0$. Then $\int_P g\,d\mu = 0$ so $\nu_0(P) = 0$. Since $\nu_0(P) = 0$, $(\nu_0 - \epsilon\mu)(P) = 0$. Since P and N partition Ω and N is a negative set for $\nu_0 - \epsilon\mu$

$$(\nu_0 - \epsilon\mu)(\Omega) = (\nu_0 - \epsilon\mu)(P) + (\nu_0 - \epsilon\mu)(N) \le 0.$$

This implies $\nu_0(\Omega) \le \epsilon\mu(\Omega)$ and this is a contradiction to 6.16. Hence, $\mu(P) > 0$. Since $g \in \mathcal{H}$, $\int_\Omega g\,d\mu \le \nu(\Omega) < \infty$. But then

$$\int_\Omega (g + \epsilon\chi_P)\,d\mu = \int_\Omega g\,d\mu + \epsilon\mu(P) > \int_\Omega g\,d\mu = \alpha. \tag{6.18}$$

Since $g + \epsilon\chi_P \in \mathcal{H}$, 6.18 contradicts 6.13. Therefore, $\nu_0 = 0$. This means $\nu(A) = \int_A g\,d\mu, \forall A \in \mathcal{F}$. Since $g \ge 0$ and $\int_\Omega g\,d\mu < \infty$, g is μ-almost everywhere finite, so g can be redefined to satisfy $g : \Omega \to [0,\infty)$. This proves existence of g.

Step 3 (Uniqueness). Suppose that g, h both satisfy the conclusion of the theorem. Then for every $A \in \mathcal{F}$, $\nu(A) = \int_A g\,d\mu = \int_A h\,d\mu$. Now define $G = \{\omega \in \Omega : g(\omega) > h(\omega)\}, H = \{\omega \in \Omega : g(\omega) < h(\omega)\}$. Clearly, $G, H \in \mathcal{F}$. Observe that $(g-h)^+ = (g-h)\chi_G$ and $(g-h)^- = (h-g)\chi_H$ and

$$\int_\Omega (g-h)^+\,d\mu = \int_\Omega (g-h)\chi_G\,d\mu = \int_G g\,d\mu - \int_G h\,d\mu = 0,$$
$$\int_\Omega (g-h)^-\,d\mu = \int_\Omega (h-g)\chi_H\,d\mu = \int_H h\,d\mu - \int_H g\,d\mu = 0.$$

Since $(g-h)^+ \geq 0, (g-h)^- \geq 0$, by Proposition 15, $(g-h)^+$ and $(g-h)^-$ vanish μ-almost everywhere. But then, $(g-h)$ vanishes μ-almost everywhere. We conclude $g = h$ μ-almost everywhere, as desired.

■

The Radon-Nikodym Theorem for measures is a very powerful theorem. There are numerous applications of this theorem in probability theory. For instance, using Radon-Nikodym Theorem for measures it is possible to neatly justify the existence of conditional expectation (See Exercise 13.15 in [Bas14]).
It is possible generalize Radon-Nikodym Theorem for measures to finite signed measures.

Theorem 32 (Radon-Nikodym Theorem for signed measures)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let $\mu : \mathcal{F} \to [0, \infty]$ be σ-finite measures. Suppose that ν is a finite signed measure. If $\nu \ll \mu$, then there is a function $g \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$ such that*

$$\nu(A) = \int_A g\,d\mu, \text{ for every } A \in \mathcal{F}.$$

The function g is unique up to μ-almost everywhere equality.

Proof Idea. The proof of Radon-Nikodym Theorem for signed measures is divided into two parts. In the first part, we will discuss the existence of the desired function. The existence will follow directly from the Hahn-Jordan decomposition and Radon-Nikodym Theorem for measures. The second part will be about uniqueness.

Proof.
Step 1 (Existence). Let $\nu = \nu^+ - \nu^-$ be the Hahn-Jordan decomposition of ν. By definition of $\nu \ll \mu$, we have $|\nu| \ll \mu$. Since $\nu^+ \leq |\nu|$ and $\nu^- \leq |\nu|$, we have $\nu^+ \ll \mu$ and $\nu^- \ll \mu$. By Radon-Nikodym Theorem for measures, there exist measurable functions $g^+ : \Omega \to [0, \infty)$ and $g^- : \Omega \to [0, \infty)$ such that for every $A \in \mathcal{F}$,

$$\nu^+(A) = \int_A g^+\,d\mu \text{ and } \nu^-(A) = \int_A g^-\,d\mu. \tag{6.19}$$

Since ν is finite, ν^+, ν^- are finite measures. Since $\nu^+(\Omega)$ and $\nu^-(\Omega) < \infty$, by 6.19, $g^+, g^- \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$. Define $g : \mathcal{F} \to \mathbb{R}$ by $g = g^+ - g^-$. We claim that g is the desired function. Clearly, $g \in \mathcal{L}^1(\Omega, \mathcal{F}, \mu)$.

By linearity of the integral and 6.19, for every $A \in \mathcal{F}$,

$$\nu(A) = \nu^+(A) - \nu^-(A) = \int_A g^+ \, d\mu - \int_A g^- \, d\mu = \int_A g \, d\mu.$$

This proves the existence of the desired function g.

Step 2 (Uniqueness). Suppose that g, h both satisfy the conclusion of the theorem. Then for every $A \in \mathcal{F}$, $\nu(A) = \int_A g \, d\mu = \int_A h \, d\mu$. Now define $G = \{\omega \in \Omega : g(\omega) > h(\omega)\}, H = \{\omega \in \Omega : g(\omega) < h(\omega)\}$. Clearly, $G, H \in \mathcal{F}$. Observe that $(g-h)^+ = (g-h)\chi_G$ and $(g-h)^- = (h-g)\chi_H$ and

$$\int_\Omega (g-h)^+ \, d\mu = \int_\Omega (g-h)\chi_G \, d\mu = \int_G g \, d\mu - \int_G h \, d\mu = 0,$$
$$\int_\Omega (g-h)^- \, d\mu = \int_\Omega (h-g)\chi_H \, d\mu = \int_H h \, d\mu - \int_H g \, d\mu = 0.$$

Since $(g-h)^+ \geq 0, (g-h)^- \geq 0$, by Proposition 15, $(g-h)^+$ and $(g-h)^-$ vanish μ-almost everywhere. But then, $(g-h)$ vanishes μ-almost everywhere. We conclude $g = h$ μ-almost everywhere, as desired.

■

For an application of this result, see the proof of Riesz Representation Theorem for the Dual of $\mathcal{L}^p$.

6.4 Functional Analysis

In this section, we will recall key definitions from *Linear Analysis*. We aim to prove Hahn-Banach Theorem and discuss analog of Hahn-Jordan decomposition for bounded linear functionals. Although many definitions in this section can be expressed in a more general form, we focus on real normed linear spaces, since we were studying neural networks with real, possibly vector-valued output.

Definition 61 (bounded linear operator)**.** Let X, Y be normed linear spaces and let $T : X \to Y$ be a linear operator. We say T is **bounded** if there exists a positive real number M such that $\|T(x)\|_Y \leq M \|x\|_X$, for every $x \in X$.

Lemma 18 (Lemma 4.1 in [RY08])**.** *Let X, Y be normed linear spaces and let $T : X \to Y$ be a linear operator. T is bounded if and only if T is uniformly continuous with respect to topologies induced by norms on X and Y.*

Definition 62 (norm of a bounded linear operator)**.** Let X, Y be normed linear spaces and let $T : X \to Y$ be a bounded linear operator. We define the norm of T, denoted $\|T\|$, by

$$\|T\| = \sup\{\|T(x)\|_Y : x \in X, \|x\|_X \leq 1\}.$$

Remark 30. Since T is bounded, the norm $\|T\|$ is finite. The proof it is indeed a norm is Lemma 4.15 in [RY08].

Definition 63 (linear functional)**.** Let X be a real vector space. A **linear functional** on X is a linear operator $l : X \to \mathbb{R}$.

Definition 64 (algebraic dual space)**.** Let X be a real vector space. The algebraic dual space of X, denoted by $X^\star$ is a vector space of all linear functionals on X.

Definition 65 (topological dual space)**.** Let X be a real normed linear space. The topological dual space of X, denoted by X' is a vector space of all continuous linear functionals on X.

Definition 66 (sublinear functional)**.** Let X be a real vector space. A **sublinear functional** on X is a function $\rho : X \to \mathbb{R}$ such that

SF1. $\rho(x + y) \leq \rho(x) + \rho(y)$ for every $x, y \in X$;
SF2. $\rho(\alpha x) = \alpha\rho(x)$ for every $x \in X, \alpha \in \mathbb{R}, \alpha \geq 0$.

6.4.1 Hahn-Banach Theorem

In this subsection, we will state and prove Hahn-Banach Theorem.

Theorem 33 (Hahn-Banach Theorem)**.** *Let X be a real vector space, with a sublinear functional ρ defined on X. Suppose that W is a linear subspace of X and f_W a linear functional on W satisfying*

$$f_W(w) \leq \rho(w), w \in W. \tag{6.20}$$

Then f_W has an extension f on X such that

$$f(x) \leq \rho(x), x \in X. \tag{6.21}$$

Proof Idea. The proof of this theorem relies on the Zorn lemma. It is divided in two parts. The first part is to show that it is possible to perform an extension along a 'single dimension'. The second part is a careful construction of the set of all possible extensions and the application of the Zorn lemma to produce the desired extension.

Lemma 19 (Single dimension extension lemma)**.** *Let X be a real vector space and $W \subset X$ its proper linear subspace. Suppose that f_W is a linear functional on W and let ρ be a sublinear functional on X. Furthermore, suppose that*

$$f_W(w) \leq p(w), \text{ for every } w \in W. \tag{6.22}$$

Suppose that $z \in X \setminus W$. Define W_z by

$$W_z = \{w + \alpha z : \alpha \in \mathbb{R}, w \in W\}.$$

W_z is a vector subspace of X. There exists a linear functional $f_{W_z} : W_z \to \mathbb{R}$ such that

$$f_{W_z}(w) = f_W(w), \text{ for every } w \in W.$$

Furthermore, for every $w \in W_z, f_{W_z}(w) \leq p(w)$.

Proof. We begin by showing that W_z is indeed a vector subspace of X. Suppose $\omega_1, \omega_2 \in W_z$, $\lambda \in \mathbb{R}$. Then $\omega_1 = w_1 + \alpha_1 z, \omega_2 = w_2 + \alpha_2 z$ for $\alpha_1, \alpha_2 \in \mathbb{R}, w_1, w_2 \in W$. Since W is linear, $0 \in W$ so $0 \in W_z$. Since W is linear, $\lambda w_1 \in W$ so $\lambda\omega_1 \in W_z$. Since W is linear, $w_1 + w_2 \in W$ so $\omega_1 + \omega_2 = (w_1 + w_2) + (\alpha_1 + \alpha_2)z \in W_z$.

Moreover, we claim that the representation of each element in W_z is unique. Suppose that $\omega = w_1 + \alpha_1 z = w_2 + \alpha_2 z$ for $\alpha_1, \alpha_2 \in \mathbb{R}, w_1, w_2 \in W$. This implies $(w_1 - w_2) = (\alpha_2 - \alpha_1)z$. Since $z \notin W$ and $w_1 - w_2 \in W$, the equality holds if and only if $\alpha_2 = \alpha_1$. But then $w_1 = w_2$.

By linearity of f_W on W, sublinearity of ρ on X and 6.22, for every $w_1, w_2 \in W$,

$$\begin{aligned} f_W(w_1) + f_W(w_2) = f_W(w_1 + w_2) &\leq \rho(w_1 + w_2) = \rho(w_1 - z + z + w_2) \\ &\leq \rho(w_1 - z) + \rho(w_2 + z), \end{aligned}$$

which implies

$$f_W(w_1) - \rho(w_1 - z) \leq \rho(w_2 + z) - f_W(w_2), \text{ for every } w_1, w_2 \in W. \tag{6.23}$$

By 6.23, $\sup\{f_W(w) - \rho(w - z) : w \in W\} \leq \inf\{\rho(w + z) - f_W(w) : w \in W\}$.

Now let ξ be any real number satisfying

$$\sup\{f_W(w) - \rho(w - z) : w \in W\} \leq \xi \leq \inf\{\rho(w + z) - f_W(w) : w \in W\}. \tag{6.24}$$

Define $f_{W_z} : W_z \to \mathbb{R}$ by $f(w+\alpha z) = f(w)+\alpha\xi$. By uniqueness of representation of elements in W_z, the map f_{W_z} is well-defined. We claim that f is a desired extension. It is clear that f_{W_z} agrees with f_W on W.

We begin by discussing linearity. Let $\omega_1, \omega_2 \in W_z$, $\lambda \in \mathbb{R}$. Then write $\omega_1 = w_1 + \alpha_1 z, \omega_2 = w_2 + \alpha_2 z$ for $\alpha_1, \alpha_2 \in \mathbb{R}, w_1, w_2 \in W$. By linearity of f_W,

$$\begin{aligned} f_{W_z}(\omega_1 + \omega_2) &= f_W(w_1 + w_2) + (\alpha_1 + \alpha_2)\xi = f_{W_z}(\omega_1) + f_{W_z}(\omega_2), \\ f_{W_z}(\lambda\omega_1) &= f_{W_z}(\lambda w_1 + \lambda\alpha_1 z) = f(\lambda w_1) + \lambda\alpha\xi = \lambda f(w_1) + \lambda\alpha\xi \\ &= \lambda f_{W_z}(\omega_1). \end{aligned}$$

Since f_{W_z} agrees with f_W on W, for $w \in W$, $f_{W_z}(w) \leq \rho(w)$. It remains to verify that $f_{W_z}(\omega) \leq \rho(\omega)$ for $\omega \in W_z$. Let $\omega \in W_z$, so $\omega = w + \alpha z$, $\alpha \in \mathbb{R}$.

If $\alpha = 0, \omega \in W$ so the claim holds by 6.22.

If $\alpha > 0$,

$$\begin{aligned} f_{W_z}(w + \alpha z) &= \alpha \left(f_W(\frac{1}{\alpha} w) + \xi \right) && \text{by linearity of } f_W \\ &\leq \alpha \left(f_W(\frac{1}{\alpha} w) + \rho(\frac{1}{\alpha} w + z) - f_W(\frac{1}{\alpha} w) \right) && \text{by 6.24} \\ &\leq \rho(w + \alpha z). && \text{by sublinearity of } \rho \end{aligned}$$

If $\alpha < 0$, $\alpha = -\beta$, $\beta > 0$ and

$$\begin{aligned} f_{W_z}(w + \alpha z) &= \beta \left(f_W(\frac{1}{\beta} w) - \xi \right) && \text{by linearity of } f_W \\ &\leq \beta \left(f_W(\frac{1}{\beta} w) - (f_W(\frac{1}{\beta} w) - \rho(\frac{1}{\beta} w - z)) \right) && \text{by 6.24} \\ &\leq \rho(w - \beta z) = \rho(w + \alpha z). && \text{by sublinearity of } \rho \end{aligned}$$

■

Proof of the Hahn-Banach Theorem. Let Ω be the set of all pairs (W_α, f_α) where:

(i) $W_\alpha \subseteq X$ is a vector subspace of X, $f_\alpha : W_\alpha \to \mathbb{R}$ a linear functional on W_α;
(ii) $f_\alpha(w) = f_W(w), w \in W$;
(iii) $f(w) \leq \rho(w), w \in W_\alpha$.

Since $(W, f_W) \in \Omega$, Ω is nonempty and we define a relation $\prec$ on Ω by

$$(W_\alpha, f_\alpha) \prec (W_\beta, f_\beta) \iff W_\alpha \subseteq W_\beta \text{ and } \forall x \in W_\alpha,\ f_\alpha(x) = f_\beta(x). \tag{6.25}$$

We claim that $\prec$ is a partial order on Ω. Clearly, $(W_\alpha, f_\alpha) \prec (W_\alpha, f_\alpha)$ so P1 holds. Suppose that $(W_\alpha, f_\alpha) \prec (W_\beta, f_\beta)$ and $(W_\beta, f_\beta) \prec (W_\alpha, f_\alpha)$. By 6.25, $W_\alpha \subseteq W_\beta$ and $W_\beta \subseteq W_\alpha$. Therefore $W_\alpha = W_\beta$ and $f_\alpha = f_\beta$. Hence $(W_\alpha, f_\alpha) = (W_\beta, f_\beta)$ so P2 holds. Suppose that $(W_\alpha, f_\alpha) \prec (W_\beta, f_\beta)$ and $(W_\beta, f_\beta) \prec (W_\gamma, f_\gamma)$. By 6.25, $W_\alpha \subseteq W_\beta$, $W_\beta \subseteq W_\gamma$ so $W_\alpha \subseteq W_\gamma$. By 6.25, f_α, f_β agree on W_α and f_β, f_γ agree on W_β and since $W_\alpha \subseteq W_\beta$, f_α and f_γ agree on W_α. Hence $(W_\alpha, f_\alpha) \prec (W_\gamma, f_\gamma)$. This establishes P3 and $\prec$ is indeed a partial order on Ω.
Suppose that $\Omega' \subseteq \Omega$ is totally ordered. Then Ω' is of the form

$$\Omega' = \{(W_\alpha, f_\alpha) : (W_\alpha, f_\alpha) \in \Omega, \alpha \in A\},$$

for some nonempty index set A. We will construct an upper bound for Ω' in Ω. Define

$$U = \bigcup_{\alpha \in A} W_\alpha. \tag{6.26}$$

We will show U is a vector subspace of X. Since A is nonempty, let $\alpha \in A$. Since W_α is a vector subspace of X, $0 \in W_\alpha$ so $0 \in U$. Let $w_1, w_2 \in U, \lambda \in \mathbb{R}$. Since $w_1 \in U$, $w_1 \in W_\alpha$ for $\alpha \in A$. Since $w_2 \in U$, $w_2 \in W_\beta$ for $\beta \in A$. Since W_α is a vector subspace of X, $\lambda w_1 \in W_\alpha$ so $\lambda w_1 \in U$. By the total ordering of Ω', either $W_\alpha \prec W_\beta$ or $W_\beta \prec W_\alpha$. Without loss of generality, $W_\alpha \prec W_\beta$. Since W_α is a vector subspace of X, $w_1 + w_2 \in W_\alpha$ so $w_1 + w_2 \in U$.
We define $f : U \to \mathbb{R}$ as follows. Let $w \in U$. By 6.26, $w \in W_\alpha$ for some $\alpha \in A$. Then set $f(w) = f_\alpha(w)$. We will show that f is well-defined. Suppose that $w \in W_\alpha \cap W_\beta$, $\alpha, \beta \in A$. By the total ordering of Ω', either $W_\alpha \prec W_\beta$ or $W_\beta \prec W_\alpha$. Without loss of generality, $W_\alpha \prec W_\beta$. By 6.25, $W_\alpha \subseteq W_\beta$ and $f_\alpha(w) = f_\beta(w)$ since $w \in W_\alpha$. Thus, f is well-defined.

We will show f is linear on U. Take $w_1, w_2 \in U$ and $\lambda \in \mathbb{R}$. Since $w_1 \in U$, $w_1 \in W_\alpha$ for $\alpha \in A$. Since $w_2 \in U$, $w_2 \in W_\beta$ for $\beta \in A$. Since f_α is linear on W_α, $f(\lambda w_1) = f_\alpha(\lambda w_1) = \lambda f_\alpha(w_1) = \lambda f(w_1)$. By the total ordering of Ω', either $W_\alpha \prec W_\beta$ or $W_\beta \prec W_\alpha$. Without loss of generality, $W_\alpha \prec W_\beta$. By 6.25, $W_\alpha \subseteq W_\beta$. Then $w_1, w_2 \in W_\beta$. Since f_β is linear on W_β, $f(w_1 + w_2) = f_\beta(w_1 + w_2) = f_\beta(w_1) + f_\beta(w_2) = f(w_1) + f(w_2)$.
We will prove that f is an extension of f_W in the sense of (ii). Let $w \in W$. Since for each $\alpha \in A$, (W_α, f_α) in Ω, by (ii), $f_\alpha(w) = f_W(w)$. By definition of f, $f(w) = f_\alpha(w) = f_W(w)$.
Now we prove that f is dominated by ρ in the sense of (iii). Consider $w \in U$. By 6.26, $w \in W_\alpha$ for some $\alpha \in A$. By (iii), $f_\alpha(w) \leq \rho(w)$. By definition of f, $f(w) = f_\alpha(w) \leq \rho(w)$.
Hence $(U, f) \in \Omega$. We claim (U, f) is an upper bound of Ω'. Let $(W_\alpha, f_\alpha) \in \Omega'$. Clearly, $W_\alpha \subseteq U$. Consider $w \in W_\alpha$. By definition of f, $f(w) = f_\alpha(w)$. So f and f_α agree on W_α. Thus, $(W_\alpha, f_\alpha) \prec (U, f)$. By Zorn lemma, Ω contains a maximal element, say $(\widetilde{U}, \widetilde{f})$. To prove $\widetilde{f}$ is a desired extension of f_W, we will prove $\widetilde{U} = X$. Suppose not. Then $\widetilde{U} \subset X$ and there exists $z \in X \setminus \widetilde{U}$. By Single dimension extension lemma, $\widetilde{f}$ can be extended to $f_{\widetilde{U}_z}$ on $\widetilde{U}_z$ such that $(\widetilde{U}_z, f_{\widetilde{U}_z}) \in \Omega$ with $(\widetilde{U}, \widetilde{f}) \prec (\widetilde{U}_z, f_{\widetilde{U}_z})$, contradicting the maximality of $(\widetilde{U}, \widetilde{f})$. Hence $\widetilde{U} = X$. ■

6.4.2 Hahn-Jordan decomposition for bounded linear functionals on $\mathcal{C}(X)$

In this section, we will discuss an interesting result about the decomposition of linear functionals. It turns out that linear functionals admit similar decomposition as measures.

Theorem 34 (Hahn-Jordan decomposition for bounded linear functionals on $\mathcal{C}(X)$)**.** *Let X be a topological space. If I is a bounded linear functional on $\mathcal{C}(X)$, then there exist two positive linear functionals I^+, I^- on $\mathcal{C}(X)$ such that $I = I^+ - I^-$.*

Proof.

Step 1 (Construction for non-negative functions). Let $\mathcal{H}_X^+$ be the set of all non-negative functions in $\mathcal{C}(X)$. For $f \in \mathcal{H}_X^+$, define the set $\Phi(f)$ by

$$\Phi(f) = \{I(g) : g \in \mathcal{C}(X), 0 \leq g \leq f\}.$$

Now define the function $\widetilde{I} : \mathcal{H}_X^+ \to \mathbb{R}$ by $\widetilde{I}(f) = \sup \Phi(f)$. Firstly, we observe that $0 \in \Phi(f)$. Since I is linear, $I(0) = 0$ so $\widetilde{I}(f) \geq 0$.
Consider $g \in \Phi(f)$. Since $0 \leq g \leq f$ and I is a bounded linear functional, we have $|I(g)| \leq \|I\| \, \|g\|_\infty \leq \|I\| \, \|f\|_\infty$. Consequently,

$$0 \leq \widetilde{I}(f) \leq \|I\| \, \|f\|_\infty . \tag{6.27}$$

We claim that $\widetilde{I}$ is linear on $\mathcal{H}_X^+$. We will show that for $\lambda \geq 0$, $\widetilde{I}(\lambda f) = \lambda \widetilde{I}(f)$.

Consider $\lambda = 0$. Since $\widetilde{I}(0) = 0$, the claim holds. Now suppose $\lambda > 0$. Then

$$\begin{aligned}
\widetilde{I}(\lambda f) &= \sup\{I(g) : g \in \mathcal{C}(X), 0 \leq g \leq \lambda f\} \\
&= \sup\{I(g) : g \in \mathcal{C}(X), 0 \leq \frac{1}{\lambda} g \leq f\} \\
&= \sup\{I(\lambda h) : h \in \mathcal{C}(X), 0 \leq h \leq f\} && \text{by substitution } h = \frac{1}{\lambda} g \\
&= \sup\{\lambda I(h) : h \in \mathcal{C}(X), 0 \leq h \leq f\} && \text{by linearity of } I \\
&= \lambda \sup\{I(h) : h \in \mathcal{C}(X), 0 \leq h \leq f\} \\
&= \lambda \widetilde{I}(f).
\end{aligned}$$

Now consider $f, h \in \mathcal{H}_X^+$. Let $g_1 \in \Phi(f)$, $g_2 \in \Phi(h)$. Then $0 \leq g_1 \leq f$, $0 \leq g_2 \leq h$. Hence $0 \leq g_1 + g_2 \leq f + h$ so $I(g_1 + g_2) \in \Phi(f + h)$. This implies $I(g_1 + g_2) \leq \widetilde{I}(f + h)$. By linearity of I, we have

$$I(g_1 + g_2) = I(g_1) + I(g_2) \leq \widetilde{I}(f + h).$$

Since g_1 and g_2 were arbitrary, $\widetilde{I}(f) + \widetilde{I}(h) \leq \widetilde{I}(f + h)$. Conversely, suppose $\epsilon > 0$. By definition of $\widetilde{I}$ and 6.27, we can choose $g \in C(X)$ such that $0 \leq g \leq f + h$ and

$$\widetilde{I}(f + h) < I(g) + \epsilon. \tag{6.28}$$

Set $g_1 = \min(f, g)$ and observe that $0 \leq g_1 \leq f$, $0 \leq g - g_1$. Since $0 \leq g_1 \leq f$, $I(g_1) \in \Phi(f)$. We claim $0 \leq g - g_1 \leq h$. Suppose not. Then there exists $x \in X$ such that

$$g(x) - g_1(x) > h(x). \tag{6.29}$$

By definition of g_1, $g_1(x) = g(x)$ or $g_1(x) = f(x)$.
Suppose $g_1(x) = g(x)$. By 6.29, $h(x) < 0$. This is a contradiction to $h \in \mathcal{H}_X^+$. Hence $g_1(x) = f(x)$. By 6.29, $f(x) + h(x) < g(x)$. This contradicts the choice $I(g) \in \Phi(f + h)$ which implied $0 \leq g \leq f + h$. Hence $0 \leq g - g_1 \leq h$ so $I(g - g_1) \in \Phi(h)$. By linearity of I, $I(g) = I(g - g_1 + g_1) = I(g - g_1) + I(g_1)$. By 6.28,

$$\begin{aligned}
\widetilde{I}(f + h) &< I(g) + \epsilon \\
&= I(g - g_1) + I(g_1) + \epsilon \\
&\leq \widetilde{I}(h) + I(g_1) + \epsilon && \text{since } I(g - g_1) \in \Phi(h) \\
&\leq \widetilde{I}(h) + \widetilde{I}(f) + \epsilon. && \text{since } I(g_1) \in \Phi(f)
\end{aligned}$$

Since ϵ was arbitrary, $\widetilde{I}(f + h) \leq \widetilde{I}(f) + \widetilde{I}(h)$. Hence $\widetilde{I}(f + h) = \widetilde{I}(f) + \widetilde{I}(h)$.

Step 2 (Generalization to $\mathcal{C}(X)$). Using $\widetilde{I}$, we construct the desired decomposition of I on $\mathcal{C}(X)$. To perform the decomposition, we define $\phi : \mathcal{C}(X) \to \mathbb{R}$ by

$$\phi(f) = \widetilde{I}(f^+) - \widetilde{I}(f^-).$$

We claim that ϕ is a positive linear functional.

Let $f \in \mathcal{C}(X)$ and assume $f \geq 0$. Then $f^+ = f$ and $f^- = 0$. We have $\widetilde{I}(f^-) = \widetilde{I}(0) = 0$. Now $\phi(f) = \widetilde{I}(f^+) - \widetilde{I}(f^-) = \widetilde{I}(f) \geq 0$, by 6.27. Hence ϕ is positive. Now consider the linearity of ϕ. Let $f, g \in \mathcal{C}(X)$. Write

$$f = f^+ - f^- \text{ where } f^+, f^- \in \mathcal{H}_X^+,$$
$$g = g^+ - g^- \text{ where } g^+, g^- \in \mathcal{H}_X^+.$$

By decompositions above, $f + g = (f^+ + g^+) - (f^- + g^-)$. We also have the decomposition $f + g = (f + g)^+ - (f + g)^-$. Combining those gives

$$(f^+ + g^+) + (f + g)^- = (f^- + g^-) + (f + g)^+. \tag{6.30}$$

Applying $\widetilde{I}$ to 6.30 gives

$$\widetilde{I}(f^+) + \widetilde{I}(g^+) + \widetilde{I}((f + g)^-) = \widetilde{I}(f^-) + \widetilde{I}(g^-) + \widetilde{I}((f + g)^+).$$

Rearranging gives

$$\widetilde{I}((f + g)^+) - \widetilde{I}((f + g)^-) = \widetilde{I}(f^+) - \widetilde{I}(f^-) + \widetilde{I}(g^+) - \widetilde{I}(g^-). \tag{6.31}$$

By definition of ϕ and 6.31, $\phi(f + g) = \phi(f) + \phi(g)$.
It remains to prove that for every $\lambda \in \mathbb{R}$ and for every $f \in \mathcal{C}(X)$, $\phi(\lambda f) = \lambda\phi(f)$. Consider $\lambda \geq 0$. By linearity of $\widetilde{I}$ on $\mathcal{H}_X^+$, we have

$$\begin{aligned}
\phi(\lambda f) &= \widetilde{I}((\lambda f)^+) - \widetilde{I}((\lambda f)^-) \\
&= \widetilde{I}(\lambda f^+) - \widetilde{I}(\lambda f^-) \\
&= \lambda\widetilde{I}(f^+) - \lambda\widetilde{I}(f^-)) \\
&= \lambda\left(\widetilde{I}(f^+) - \widetilde{I}(f^-)\right) \\
&= \lambda\phi(f).
\end{aligned}$$

By definition, $(-\lambda f)^+ = \lambda f^-$ and $(-\lambda f)^- = \lambda f^+$. Now

$$\begin{aligned}
\phi(-\lambda f) &= \widetilde{I}((-\lambda f)^+) - \widetilde{I}((-\lambda f)^-) \\
&= \widetilde{I}(\lambda f^-) - \widetilde{I}(\lambda f^+) \\
&= \lambda\widetilde{I}(f^-) - \lambda\widetilde{I}(f^+)) \\
&= \lambda\left(\widetilde{I}(f^-) - \widetilde{I}(f^+)\right) \\
&= -\lambda\phi(f).
\end{aligned}$$

We conclude ϕ is linear on $\mathcal{C}(X)$. Finally, we claim that the desired decomposition is $I = \phi - (\phi - I)$. We have shown that ϕ is a positive linear functional. We claim that $(\phi - I)$ is also a positive linear functional. Linearity of $(\phi - I)$ follows from the linearity of ϕ and I. We will show that $(\phi - I)$ is positive. Suppose $f \in \mathcal{H}_X^+$. Clearly, $I(f) \in \Phi(f)$. Then $\phi(f) = \widetilde{I}(f) \geq I(f)$. Hence $\phi(f) - I(f) \geq 0$.

■

6.5 $\mathcal{L}^p$ spaces

In this section, we will review key concepts in $\mathcal{L}^p$ spaces. We will discuss essential inequalities and density results. We will conclude with a discussion about Riesz Representation Theorem for the Dual of $\mathcal{L}^p$.

Definition 67. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose $1 \leq p < \infty$. The space $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ consists of equivalence classes of measurable functions $f : \Omega \to \mathbb{R}$ such that

$$\int_\Omega |f|^p \, d\mu < \infty.$$

We say that two measurable functions are equivalent if they are equal μ-almost everywhere. For the sake of simplicity, we will simply refer to representative functions of those equivalence classes instead of equivalence classes themselves. We will also write $\mathcal{L}^p(\Omega)$ instead of $\mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ when the measure space is unambiguous.

The $\mathcal{L}^p$ norm for $f \in \mathcal{L}^p(\Omega, \mathcal{F}, \mu)$ is defined by

$$\|f\|_p = \left(\int_\Omega |f|^p \, d\mu \right)^{\frac{1}{p}}.$$

Definition 68 (essential supremum)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that $f : \Omega \to \mathbb{R}$ is measurable. The essential supremum of f on Ω is defined by

$$\operatorname*{ess\,sup}_\Omega f = \inf \left\{ a \in \mathbb{R} : \mu(\{\omega \in \Omega : f(\omega) > a\}) \right\}.$$

Definition 69. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The space $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu)$ consists of pointwise μ-almost everywhere equivalence classes of essentially bounded measurable functions $f : \Omega \to \mathbb{R}$ with the norm

$$\|f\|_\infty = \operatorname*{ess\,sup}_\Omega |f|.$$

Again, for the sake of simplicity, we will simply refer to representative functions of those equivalence classes instead of equivalence classes themselves. We will also write $\mathcal{L}^\infty(\Omega)$ instead of $\mathcal{L}^\infty(\Omega, \mathcal{F}, \mu)$ when the measure space is unambiguous.

Remark 31. Observe that $|f| \leq \operatorname{ess\,sup}_\Omega |f|$ μ-a.e. Hence $|f| \leq \|f\|_\infty$ μ-a.e.

Remark 32. The fact $\|\cdot\|_p$ and $\|\cdot\|_\infty$ are norms on $\mathcal{L}^p$ and $\mathcal{L}^\infty$ is a consequence of Proposition 16 and Minkowski Inequality.

6.5.1 Essential inequalities

In this section, we will discuss two essential and very useful inequalities - Hölder inequality and Minkowski Inequality. We will apply them to prove Inclusion of $\mathcal{L}^p$ spaces of a finite measure.

Definition 70 (Hölder conjugate)**.** Let $1 \leq p < \infty$. We define the Hölder conjugate q of p by:

$$\frac{1}{p} + \frac{1}{q} = 1, \text{ if } 1 < p < \infty.$$

When $p = 1$ we define $q = \infty$.

Theorem 35 (Hölder inequality)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that $1 \leq p < \infty$. Let q be the Hölder conjugate of p. If $f \in \mathcal{L}^p(\Omega)$ and $g \in \mathcal{L}^q(\Omega)$ then $fg \in \mathcal{L}^1(\Omega)$ with*

$$\int_\Omega |fg| \, d\mu \leq \|f\|_p \|g\|_q .$$

Proof. See Proposition 3.3.2 in [Coh13]. ■

The following result is an important consequence of Hölder inequality.

Theorem 36 (Inclusion of $\mathcal{L}^p$ spaces of a finite measure)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a **finite** measure space and suppose that $1 \leq p_1 < p_2 \leq \infty$. Then $\mathcal{L}^{p_2}(\Omega) \subseteq \mathcal{L}^{p_1}(\Omega)$.*

Proof. Let $f \in \mathcal{L}^{p_2}(\Omega)$. Set $g = 1$. Since $\mu(\Omega) < \infty$, for $1 \leq p \leq \infty$, $g \in \mathcal{L}^p(\Omega)$. We need to show $\|f\|_{p_1} < \infty$. We split in two cases - $p_2 < \infty$ and $p_2 = \infty$.
Suppose $p_2 < \infty$. Set $p = \frac{p_2}{p_1}, q = 1 - p$. Then p, q are Hölder conjugates. Observe that $(|f|^{p_1})^p = |f|^{p_2}$. Hence $|f|^{p_1} \in \mathcal{L}^p(\Omega)$. By Hölder inequality and $f \in \mathcal{L}^{p_2}(\Omega)$,

$$\|f\|_{p_1}^{p_1} = \int_\Omega |f|^{p_1} \cdot 1 \, d\mu \leq \left(\int_\Omega (|f|^{p_1})^p \, d\mu \right)^{\frac{1}{p}} \left(\int_\Omega 1 \, d\mu \right)^{\frac{1}{q}} = \|f\|_{p_2}^{p_1} \mu(\Omega)^{\frac{1}{q}} < \infty.$$

Suppose $p_2 = \infty$. By Remark 31, $0 \leq |f|^{p_1} \leq \|f\|_\infty^{p_1}$ μ-a.e. Integrating gives

$$\|f\|_{p_1}^{p_1} \leq \int_\Omega |f|^{p_1} \, d\mu \leq \int_\Omega \|f\|_\infty^{p_1} \, d\mu = \|f\|_\infty^{p_1} \mu(\Omega) < \infty.$$

■

Theorem 37 (Minkowski Inequality)**.** *If $f, g \in \mathcal{L}^p(\Omega)$, where $1 \leq p \leq \infty$, then $f + g \in \mathcal{L}^p(\Omega)$ and*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p .$$

Proof. See Proposition 3.3.3 in [Coh13]. ■

6.5.2 Density in $\mathcal{L}^p$

We begin this section by proving a small lemma.

Lemma 20 ($\mathcal{L}^p$ convergence lemma)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that $1 \leq p < \infty$. If $f_n \to f$ pointwise, $|f_n| \leq g$ and $g \in \mathcal{L}^p(\Omega)$, then $f \in \mathcal{L}^p(\Omega)$ and $f_n \to f$ in $\mathcal{L}^p(\Omega)$.*

Proof. Consider $|f - f_n|^p$. By continuity, $|f - f_n|^p \to 0$ pointwise. Since $|f_n| \leq g$, $|f| = \lim_{n\to\infty} |f_n| \leq g$. Since $g \in \mathcal{L}^p(\Omega)$, $f \in \mathcal{L}^p(\Omega)$ by monotonicity of the integral. We have $|f - f_n|^p \leq (|f| + |f_n|)^p \leq (2g)^p$. Since $g \in \mathcal{L}^p(\Omega)$, $2g \in \mathcal{L}^p(\Omega)$. Hence $(2g)^p \in \mathcal{L}^1(\Omega)$. By Dominated Convergence Theorem,

$$\lim_{n\to\infty} \|f - f_n\|_p = \lim_{n\to\infty} \int_\Omega |f - f_n|^p \, d\mu = \int_\Omega \lim_{n\to\infty} |f - f_n|^p \, d\mu = 0.$$

■

Theorem 38 (Density of simple functions in $\mathcal{L}^p$)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that $1 \leq p \leq \infty$. The simple functions belonging to $\mathcal{L}^p(\Omega)$ are dense in $\mathcal{L}^p$.*

Proof. Suppose $p < \infty$ and $f \in \mathcal{L}^p(\Omega)$ and $f \geq 0$. Since f is measurable, there exists a monotonically nondecreasing sequence of nonnegative simple functions $\{\phi\}_{n=1}^\infty$ such that $\phi_n \uparrow f$. We will show $\lim_{n\to\infty} \|f - \phi_n\|_p = 0$.
To begin, we claim that $\phi_n \in \mathcal{L}^p(\Omega)$ for every $n \in \mathbb{N}$. Since $0 \leq \phi_n \leq \phi_{n+k}$ for every $n, k \in \mathbb{N}$ and $\lim_{k\to\infty} \phi_{n+k} = \lim_{n\to\infty} \phi_n = f$, we have

$$0 \leq \phi_n \leq \lim_{k\to\infty} \phi_{n+k} = f, \text{ for every } n \in \mathbb{N}. \tag{6.32}$$

By monotonicity of the integral, $0 \leq \|\phi_n\|_p \leq \|f\|_p < \infty$, for every $n \in \mathbb{N}$. Hence $\phi_n \in \mathcal{L}^p(\Omega)$ for every $n \in \mathbb{N}$. Since $\phi_n \uparrow f$ and 6.32, by $\mathcal{L}^p$ convergence lemma, $\lim_{n\to\infty} \|f - \phi_n\|_p = 0$, as desired.
Now consider $f \in \mathcal{L}^p(\Omega)$. Write $f = f^+ - f^-$. By the previous case, there exist sequences of simple functions $\{\phi\}_{n=1}^\infty, \{\psi\}_{n=1}^\infty$ such that $\phi_n \in \mathcal{L}^p(\Omega), \psi_n \in \mathcal{L}^p(\Omega)$ and $\phi_n \to f^+$ in $\mathcal{L}^p(\Omega)$ and $\psi_n \to f^-$ in $\mathcal{L}^p(\Omega)$. Clearly, $\phi_n - \psi_n \in \mathcal{L}^p(\Omega)$ and $\phi_n - \psi_n$ is simple. We have

$$\begin{aligned} 0 \leq \|f - (\phi_n - \psi_n)\|_p &= \left\|f^+ - f^- - (\phi_n - \psi_n)\right\|_p \\ &\leq \left\|f^+ - \phi_n\right\|_p + \left\|f^- - \psi_n\right\|_p . \text{ by Minkowski Inequality} \end{aligned}$$

Since $\phi_n \to f^+$ in $\mathcal{L}^p(\Omega)$ and $\psi_n \to f^-$ in $\mathcal{L}^p(\Omega)$, $\lim_{n\to\infty} \|f - (\phi_n - \psi_n)\|_p = 0$.
If $f \in \mathcal{L}^\infty(\Omega)$, f is equivalent to a bounded measurable function. Then there exists a sequence of simple functions converging to f uniformly and hence in $\|\cdot\|_\infty$ norm. ■

We discuss another density result.

Theorem 39 (Density of compactly supported functions in $\mathcal{L}^p$)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and suppose that $1 \leq p \leq \infty$. Suppose that Ω is also a locally compact Hausdorff space. Then continuous compactly supported functions on Ω are dense in $\mathcal{L}^p$.*

Proof. By Density of simple functions in $\mathcal{L}^p$, it is sufficient to prove that a continuous, compactly supported function can approximate a simple function in $\mathcal{L}^p$. Let φ be a simple function in $\mathcal{L}^p(\Omega)$ and $\epsilon > 0$. By Lusin's Theorem, Theorem

7.4.4 in [Coh13], there exists a continuous function $g : \Omega \to \mathbb{R}$ such that $\operatorname{supp} g$ is compact, $\mu(\{\omega : g(\omega) \neq \varphi(\omega)\}) < \epsilon$ and $|g| \leq \|\varphi\|_\infty$. Now,

$$\begin{aligned}\|g - \varphi\|_p &= \left(\int_\Omega |g-\varphi|^p \, d\mu\right)^{\frac{1}{p}} = \left(\int_{g\neq\varphi} |g-\varphi|^p \, d\mu + \int_{g=\varphi} |g-\varphi|^p \, d\mu\right)^{\frac{1}{p}} \\ &\leq \left(\int_{g\neq\varphi} |g-\varphi|^p \, d\mu\right)^{\frac{1}{p}} \leq \left(\int_{g\neq\varphi} 2^p \|\varphi\|_\infty^p \, d\mu\right)^{\frac{1}{p}} \leq 2\|\varphi\|_\infty \epsilon^{\frac{1}{p}}.\end{aligned}$$

Since ϵ was arbitrary, the proof is complete. ■

6.5.3 Duality and Riesz Representation Theorem for $\mathcal{L}^p$

The purpose of this section is to establish Riesz Representation Theorem for the Dual of $\mathcal{L}^p$. We begin by stating and proving an useful lemma.

Lemma 21. *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-**finite** measure space and suppose that $1 \leq p < \infty$. Let q be the Hölder conjugate of p. If $g \in \mathcal{L}^q$, then the map $F : \mathcal{L}^p(\Omega) \to \mathbb{R}$*

$$F(f) = \int_\Omega fg \, d\mu$$

is a bounded linear functional on $\mathcal{L}^p(\Omega)$. Moreover, $\|F\| = \|g\|_q$.

Proof Idea. The fact F is a bounded linear functional on $\mathcal{L}^p(\Omega)$ is reasonably obvious. However, the main difficulty is showing that $\|F\| = \|g\|_q$. By Hölder inequality, it is sufficient to show that $\|F\| \geq \|g\|_q$. Since the case for $p = 1$ is slightly trickier, we will handle the cases $1 < p < \infty$ and $p = 1$ separately. The main idea will be to construct a function f in $\mathcal{L}^p(\Omega)$ satisfying $\|f\|_p = 1$ and $|F(f)| \geq \|g\|_q$.

Proof. The fact F is a linear functional follows from the linearity of the integral. Now we demonstrate it is a bounded linear functional. By Hölder inequality, we have

$$0 \leq |F(f)| \leq \|f\|_p \|g\|_q .$$

This implies $\|F\| \leq \|g\|_q$. It suffices to prove the reverse inequality $\|F\| \geq \|g\|_q$. We may assume that $g \neq 0$. Otherwise, the result is trivial.

Step 1. We begin by considering $1 < p < \infty$. Set

$$f = (\operatorname{sgn} g)\left(\frac{|g|}{\|g\|_q}\right)^{\frac{q}{p}}. \tag{6.33}$$

Since $g \neq 0$, $|\operatorname{sgn} g| = 1$ so $|f|^p = \left|(\operatorname{sgn} g)\left(\frac{|g|}{\|g\|_q}\right)^{\frac{q}{p}}\right| = \frac{|g|^q}{\|g\|_q^q}$. Now

$$\|f\|_p^p = \int_\Omega |f|^p \, d\mu = \int_\Omega \frac{|g|^q}{\|g\|_q^q} \, d\mu = \frac{\|g\|_q^q}{\|g\|_q^q} = 1.$$

Therefore, $f \in \mathcal{L}^p(\Omega)$. Since p, q are Hölder conjugates, $\frac{q}{p} = q - 1$, we have

$$\begin{aligned} F(f) &= \int_\Omega (\operatorname{sgn} g) g \left(\frac{|g|}{\|g\|_q} \right)^{q-1} d\mu \\ &= \int_\Omega |g| \left(\frac{|g|}{\|g\|_q} \right)^{q-1} d\mu \\ &= \int_\Omega \frac{|g|^q}{\|g\|_q^{q-1}} d\mu \\ &= \|g\|_q . \end{aligned}$$

Since $\|f\|_p = 1$, it follows that $\|F\| \geq \|g\|_q$.

Step 2. Suppose $p = 1$. For $\epsilon > 0$, define

$$A_\epsilon = \{\omega \in \Omega : |g(\omega)| > \|g\|_\infty - \epsilon\} .$$

By definition of essential supremum, $\mu(A_\epsilon) > 0$. Since $(\Omega, \mathcal{F}, \mu)$ is a σ-finite measure space, there exists a family $\{A_n\}_{n=1}^\infty$ of $\mathcal{F}$-measurable sets of finite measure such that $\Omega = \bigcup_{n=1}^\infty A_n$. Without loss of generality, we may assume $A_n \uparrow \Omega$. Hence $A_\epsilon = \bigcup_{n=1}^\infty (A_\epsilon \cap A_n)$. This implies there exists an $\mathcal{F}$-measurable subset $B \subseteq A_\epsilon$ such that $0 < \mu(B) < \infty$. Define $f : \Omega \to \mathbb{R}$ by

$$f = (\operatorname{sgn} g) \frac{\chi_B}{\mu(B)}.$$

Clearly, f is measurable. Since $|\operatorname{sgn} g| = 1$, we have

$$\|f\|_1 = \int_\Omega \frac{\chi_B}{\mu(B)} d\mu = 1.$$

Hence $f \in \mathcal{L}^1(\Omega)$ satisfying $\|f\|_1 = 1$. Now

$$\begin{aligned} F(f) &= \int_\Omega (\operatorname{sgn} g) g \frac{\chi_B}{\mu(B)} d\mu \\ &= \int_B \frac{|g|}{\mu(B)} d\mu \geq \frac{1}{\mu(B)} \int_B (\|g\|_\infty - \epsilon) d\mu \\ &\geq \|g\|_\infty - \epsilon. \end{aligned}$$

Hence $\|F\| \geq \|g\|_\infty - \epsilon$. Since ϵ was arbitrary, $\|F\| \geq \|g\|_\infty$.

■

Remark 33. Observe that σ-finiteness assumption was not used in Step 1 of the proof, corresponding to $1 < p < \infty$. Thus, the assumption regarding σ-finiteness can be dropped assuming $1 < p < \infty$.

We are ready to state and prove the most important result in this section - Riesz Representation Theorem for the Dual of $\mathcal{L}^p$.

Theorem 40 (Riesz Representation Theorem for the Dual of $\mathcal{L}^p$)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space and suppose that $1 \leq p < \infty$. Let q be the Hölder conjugate of p. Let $F : \mathcal{L}^p(\Omega) \to \mathbb{R}$ be a bounded linear functional on $\mathcal{L}^p(\Omega)$. Then there exists a function $g \in \mathcal{L}^q(\Omega)$ such that*

$$F(f) = \int_\Omega f g \, d\mu. \tag{6.34}$$

Moreover, $\|F\| = \|g\|_q$.

Proof.

Step 1 ($\mu(\Omega) < \infty$). We will show such a function g exists, assuming $\mu(\Omega) < \infty$. Define the map $\nu : \mathcal{F} \to \mathbb{R}$ by

$$\nu(E) = F(\chi_E).$$

Since $\mu(\Omega) < \infty$, $\chi_E \in \mathcal{L}^p(\Omega)$ so the map ν is indeed well-defined on $\mathcal{F}$. We will show that ν is a signed measure, $\nu \ll \mu$ and that the Radon-Nikodym derivative $\frac{\partial \nu}{\partial \mu}$ is a desired function. By linearity of F, $\nu(\emptyset) = F(\chi_\emptyset) = F(0) = 0$. Observe that for every $E \in \mathcal{F}$,

$$|\nu(E)| = |F(\chi_E)| \leq \|F\| \, \|\chi_E\|_p = \|F\| \, (\mu(E))^{\frac{1}{p}}. \tag{6.35}$$

Firstly, we will prove that ν is finitely additive. Let $A, B \in \mathcal{F}$ be disjoint. By linearity of F,

$$\nu(A \cup B) = F(\chi_{A \cup B}) = F(\chi_A + \chi_B) = F(\chi_A) + F(\chi_B) = \nu(A) + \nu(B).$$

Now let $\{E_n\}_{n=1}^\infty$ be a sequence of $\mathcal{F}$-measurable, pairwise disjoint sets. Set $E = \bigcup_{n=1}^\infty E_n$. For $m \in \mathcal{N}$, define E_m^* by $E_m^* = \bigcup_{k=m+1}^\infty E_k$. Clearly, $E_m^* \in \mathcal{F}$. For every $m \in N$, we have

$$E = \left(\bigcup_{k=1}^m E_k \right) \cup \left(\bigcup_{k=m+1}^\infty E_k \right) = \left(\bigcup_{k=1}^m E_k \right) \cup E_m^*. \tag{6.36}$$

By 6.36 and finite additivity of ν, for every $m \in \mathbb{N}$,

$$\nu(E) = \nu \left(\bigcup_{k=1}^m E_k \right) + \nu(E_m^*) = \sum_{k=1}^m \nu(E_k) + \nu(E_m^*). \tag{6.37}$$

We will show that $\lim_{m \to \infty} \nu(E_m^*) = 0$. By definition of E_m^*, $E_m^* \downarrow \emptyset$ as $m \to \infty$. By continuity of μ and $\mu(\Omega) < \infty$, $\mu(E_m^*) \downarrow \mu(\emptyset) = 0$, as $m \to \infty$. Then, by 6.35, $|\nu(E_m^*)| \to 0$ as $m \to \infty$. Thus $\nu(E_m^*) \to 0$ as $m \to \infty$. Taking $\lim_{m \to \infty}$ on both sides of 6.37 gives

$$\nu(E) = \lim_{m \to \infty} \sum_{k=1}^m \nu(E_k) + \lim_{m \to \infty} \nu(E_m^*) = \sum_{n=1}^\infty \nu(E_n).$$

Since F is bounded, by 6.35, $|\nu| \ll \mu$.

By Radon-Nikodym Theorem for signed measures, there exists a μ-almost everywhere unique map $g \in \mathcal{L}^1(\Omega)$ such that for every $E \in \mathcal{F}$,

$$\nu(E) = F(\chi_E) = \int_\Omega \chi_E g \, d\mu. \tag{6.38}$$

Now consider a simple function $\varphi \in \mathcal{L}^p(\Omega)$. Without loss of generality, there exist disjoint $\mathcal{F}$-measurable sets $\{A_k\}_{k=1}^n$ and constants $a_k \in \mathbb{R}, 1 \le k \le n$ such that $\varphi = \sum_{k=1}^n a_k \chi_{A_k}$. By linearity of F and 6.38,

$$\begin{aligned} F(\varphi) &= F\left(\sum_{k=1}^n a_k \chi_{A_k}\right) = \sum_{k=1}^n a_k F(\chi_{A_k}) \\ &= \sum_{k=1}^n a_k \int_\Omega \chi_{A_k} g \, d\mu = \sum_{k=1}^n \int_\Omega a_k \chi_{A_k} g \, d\mu \\ &= \int_\Omega \left(\sum_{k=1}^n a_k \chi_{A_k}\right) g \, d\mu \\ &= \int_\Omega \varphi g \, d\mu. \end{aligned} \tag{6.39}$$

Since F is bounded, for every simple function $\varphi \in \mathcal{L}^p(\Omega)$,

$$|F(\varphi)| = \left|\int_\Omega \varphi g \, d\mu\right| \le \|F\| \, \|\varphi\|_p . \tag{6.40}$$

We will show that $g \in \mathcal{L}^q(\Omega)$. If g is μ-almost everywhere equivalent to 0, the claim is trivial. Therefore, assume $g \neq 0$. Since g is measurable, there exists a sequence of measurable simple functions $\{\varphi_n\}_{n=1}^\infty$ such that $\varphi_n \to g$ pointwise μ-almost everywhere and $|\varphi_n| \le |g|$. Since g is not equivalent to 0, we have $\|g\|_q > 0$. Eventually, $\|\varphi_n\|_q > 0$. For sufficiently large $n \in \mathbb{N}$, define

$$f_n = (\operatorname{sgn} g) \left(\frac{|\varphi_n|}{\|\varphi_n\|_q}\right)^{\frac{q}{p}}.$$

Since $g \neq 0$, $|\operatorname{sgn} g| = 1$. Observe that

$$\|f_n\|_p^p = \int_\Omega |f_n|^p \, d\mu = \int_\Omega \frac{|\varphi_n|^q}{\|\varphi_n\|_q^q} \, d\mu = \frac{\|\varphi_n\|_q^q}{\|\varphi_n\|_q^q} = 1. \tag{6.41}$$

Hence $\|f_n\|_p = 1$. Since φ_n and $\operatorname{sgn} g$ are simple, f_n is simple. Since f_n is simple, $f_n \varphi_n$ is simple. Since $\mu(\Omega) < \infty$, $f_n \in \mathcal{L}^p(\Omega)$ and hence $f_n \varphi_n \in \mathcal{L}^p(\Omega)$. Since p, q are Hölder conjugates, $\frac{q}{p} = q - 1$,

$$\int_\Omega |f_n \varphi_n| \, d\mu = \int_\Omega \frac{|\varphi_n|^{q-1}}{\|\varphi_n\|_q^{q-1}} \cdot |\varphi_n| \, d\mu = \frac{\|\varphi_n\|_q^q}{\|\varphi_n\|_q^{q-1}} = \|\varphi_n\|_q . \tag{6.42}$$

Observe that $|f_n g| = f_n g$. To show $g \in \mathcal{L}^q(\Omega)$, we estimate $\|g\|_q$. Since $\varphi_n \to g$ pointwise μ-almost everywhere, $|\varphi_n|^q \to |g|^q$ pointwise μ-almost everywhere. By Fatou's Lemma,

$$\begin{aligned}
\|g\|_q = \left(\int_\Omega |g|^q \, d\mu\right)^{\frac{1}{q}} &\le \liminf_{n\to\infty} \left(\int_\Omega |\varphi_n|^q \, d\mu\right)^{\frac{1}{q}} = \liminf_{n\to\infty} \|\varphi_n\|_q \\
&= \liminf_{n\to\infty} \int_\Omega |f_n \varphi_n| \, d\mu && \text{by 6.42} \\
&\le \liminf_{n\to\infty} \int_\Omega |f_n g| \, d\mu && \text{since } |\varphi_n| \le |g| \\
&\le \|F\| < \infty. && \text{by 6.41 and 6.40}
\end{aligned}$$

Hence $g \in \mathcal{L}^q(\Omega)$.

Now we show that 6.34 holds on $\mathcal{L}^p(\Omega)$. Let $f \in \mathcal{L}^p(\Omega)$. By Density of simple functions in $\mathcal{L}^p$, there exists a sequence of simple functions $\{\varphi_n\}_{n=1}^\infty$ such that $\varphi_n \to f$ in $\mathcal{L}^p(\Omega)$. By 6.39, there exists $g \in \mathcal{L}^q(\Omega)$ such that for every $n \in \mathbb{N}$,

$$F(\varphi_n) = \int_\Omega \varphi_n g \, d\mu.$$

By Hölder inequality, $\|fg\|_1 \le \|f\|_p \cdot \|g\|_q < \infty$, so $fg \in \mathcal{L}^1(\Omega)$. Similarly, by Hölder inequality, $\|\varphi_n g\|_1 \le \|\varphi_n\|_p \cdot \|g\|_q < \infty$, so $\varphi_n g \in \mathcal{L}^1(\Omega)$. We have

$$\begin{aligned}
\left|\int_\Omega fg \, d\mu - F(\varphi_n)\right| = \left|\int_\Omega fg \, d\mu - \int_\Omega \varphi_n g \, d\mu\right| &= \left|\int_\Omega (f - \varphi_n) g \, d\mu\right| \\
&\le \int_\Omega |(f - \varphi_n)||g| \, d\mu \text{ by Hölder inequality} \\
&\le \|f - \varphi_n\|_p \cdot \|g\|_q .
\end{aligned}$$

Since $\|g\|_q < \infty$ and $\lim_{n\to\infty} \|f - \varphi_n\|_p = 0$, $\lim_{n\to\infty} \left|\int_\Omega fg \, d\mu - F(\varphi_n)\right| = 0$. Hence,

$$\lim_{n\to\infty} F(\varphi_n) = \int_\Omega fg \, d\mu. \tag{6.43}$$

Since F is bounded on $\mathcal{L}^p(\Omega)$, we have

$$|F(f) - F(\varphi_n)| = |F(f - \varphi_n)| \le \|F\| \, \|f - \varphi_n\|_p . \tag{6.44}$$

Since $\lim_{n\to\infty} \|f - \varphi_n\|_p = 0$, by 6.44,

$$F(f) = \lim_{n\to\infty} F(\varphi_n). \tag{6.45}$$

Applying 6.43 to 6.45 yields

$$F(f) = \int_\Omega fg \, d\mu.$$

By Lemma 21, $\|F\| = \|g\|_q$. By 6.38, g is unique up to a μ-null set. This completes the proof for a finite measure space.

Step 2. Now we extend the result to a σ-finite measure space. Let $\{A_n\}_{n=1}^{\infty}$ be the sequence of $\mathcal{F}$-measurable sets such that $\Omega = \bigcup_{n=1}^{\infty} A_n$. Without loss of generality, assume $A_n \subseteq A_{n+1}$. Now consider the measure space $(A_n, \mathcal{F}_{|A_n}, \mu_{|A_n})$. By the previous case, there exists a $g_n \in \mathcal{L}^q(A_n, \mathcal{F}_{|A_n}, \mu_{|A_n})$ such that for all $f \in \mathcal{L}^p(\Omega)$,

$$F(f\chi_{A_n}) = \int_{\Omega} (f\chi_{A_n}) g_n \, d\mu.$$

Furthermore $\left\|F_{|A_n}\right\| = \|g_n\|_q$ and g_n is unique up to a μ-null set. Suppose $m \geq n$. Since $A_n \subseteq A_m$ and g_m is μ-a.e unique, $g_m = g_n$ μ-a.e. Hence $g = \lim_{n\to\infty} g_n \chi_{A_n}$ is well-defined μ-a.e. By redefining g at a null-set, we may assume that g is defined on Ω. We claim that g is a desired function. By Fatou's Lemma,

$$\|g\|_q^q \leq \liminf_{n\to\infty} \int_{\Omega} |g_n \chi_{A_n}|^q \, d\mu \leq \liminf_{n\to\infty} \|g_n\|_q^q \leq \liminf_{n\to\infty} \left\|F_{|A_n}\right\|^q \leq \|F\|^q < \infty.$$

Hence $g \in \mathcal{L}^q(\Omega)$. Consider $f \in \mathcal{L}^p(\Omega)$. Since $\Omega = \bigcup_{n=1}^{\infty} A_n$, $f\chi_{A_n} \to f$ pointwise. Since $|f| \in \mathcal{L}^p(\Omega)$ and $|f\chi_{A_n}| \leq |f|$, by $\mathcal{L}^p$ convergence lemma, $f\chi_{A_n} \to f$ in $\mathcal{L}^p(\Omega)$. Since $f \in \mathcal{L}^p(\Omega), f\chi_{A_n} \in \mathcal{L}^p(\Omega)$. Since F is bounded,

$$|F(f) - F(f\chi_{A_n})| = |F(f - f\chi_{A_n})| \leq \|F\| \, \|f - f\chi_{A_n}\|_p . \tag{6.46}$$

Since $f\chi_{A_n} \to f$ in $\mathcal{L}^p(\Omega)$, by 6.46,

$$F(f) = \lim_{n\to\infty} F(f\chi_{A_n}). \tag{6.47}$$

By Hölder inequality, $\|fg\|_1 \leq \|f\|_p \cdot \|g\|_q < \infty$, so $fg \in \mathcal{L}^1(\Omega)$. We have

$$\begin{aligned}
\left| \int_{\Omega} fg \, d\mu - F(f\chi_{A_n}) \right| &= \left| \int_{\Omega} fg \, d\mu - \int_{\Omega} f\chi_{A_n} g \, d\mu \right| = \left| \int_{\Omega} (f - f\chi_{A_n}) g \, d\mu \right| \\
&\leq \int_{\Omega} |(f - f\chi_{A_n})||g| \, d\mu \text{ by Hölder inequality} \\
&\leq \|f - f\chi_{A_n}\|_p \cdot \|g\|_q .
\end{aligned}$$

Since $\|g\|_q < \infty$ and $\lim_{n\to\infty} \|f - f\chi_{A_n}\|_p = 0$,

$$\lim_{n\to\infty} F(f\chi_{A_n}) = \int_{\Omega} fg \, d\mu. \tag{6.48}$$

Applying 6.48 to 6.47 yields

$$F(f) = \int_{\Omega} fg \, d\mu.$$

By Lemma 21, $\|F\| = \|g\|_q$.

■

6.6 Linear functionals on $\mathcal{C}(X)$

In this section, we will discuss duality theory on $\mathcal{C}(X)$.

6.6.1 Construction of partitions of unity

In this subsection, we will prepare foundations for the proof of the Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$. We begin by recalling the definition of the support of a function.

Definition 71 (support of a function)**.** Let X be a topological space and let $f : X \to \mathbb{R}$. The support of f, denoted by $\operatorname{supp} f$ is the set

$$\operatorname{supp} f = \overline{\{x : f(x) \neq 0\}}.$$

If G is an open subset of X, we define $\mathcal{F}_G$ by

$$\mathcal{F}_G = \{f : f \in \mathcal{C}(X), 0 \leq f \leq 1, \operatorname{supp} f \subset G\}.$$

Lemma 22. *Let X be a compact metric space such that $K \subset G \subset X$, K compact, G open. Then there exists a function $h \in \mathcal{F}_G$ such that $h = 1$ on K.*

To prove this lemma, we will recall the Urysohn lemma.

Lemma 23 (Urysohn lemma, Theorem 33.1 [Mun14])**.** *Let X be a normal space. Let A, B be disjoint closed subsets of X. There exists a continuous map $f : X \to [a, b]$ such that $f(x) = a$, for every $a \in A$ and $f(x) = b$, for every $b \in B$.*

Proof of Lemma 22. Since K is compact and X is Hausdorff, K is closed. Since G is open, $X \setminus G$ is closed and $K \cap (X \setminus G) = \emptyset$. By Urysohn Lemma, there exists a continuous function $h : X \to [0, 1]$ separating K and $X \setminus G$. We may choose h satisfying $h = 1$ on K and $h = 0$ on $X \setminus G$. Since $h = 0$ on $X \setminus G$, $\operatorname{supp} h \subseteq G$. Since $\operatorname{supp} h$ is closed and G is open, $\operatorname{supp} h \subset G$, as claimed. ∎

Remark 34. Since we are working in a metric space X, it was not necessary to invoke Urysohn Lemma, due to the structure of X induced by the metric d. For example, a desired function is $f : X \to [0, 1]$ given by

$$f(x) = \frac{d(x, X \setminus G)}{d(x, K) + d(x, X \setminus G)}.$$

where $d(x, A) = \inf_{y \in A}\{d(x, y)\}$. The constructive proof is omitted for the sake of brevity.

Definition 72 (partition of unity, p.225 [Mun14])**.** Let $\{U_i\}_{i=1}^n$ be a finite indexed open covering of the topological space X. An indexed family of continuous functions $\{\varphi_i\}_{i=1}^n$ is said to be a **partition of unity** dominated by $\{U_i\}_{i=1}^n$ if for each i, $\operatorname{supp} \varphi_i \subset U_i$ and for each $x \in X$, $\sum_{k=1}^n \varphi_i(x) = 1$.

Using the Lemma 22, we can demonstrate the existence of partition of unity under the conditions relevant to the Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$.

Lemma 24 (Existence of partition of unity of a compact set)**.** *Let X be a compact metric space and suppose that $K \subset X$ is compact and $K \subset \bigcup_{k=1}^{n} G_k$ where G_k are open in X. There exists $g_i \in \mathcal{F}_{G_i}$, for $1 \leq i \leq n$ such that $\sum_{k=1}^{n} g_k(x) = 1$ whenever $x \in K$.*

Proof. Let $x \in K$. Then $x \in G_i$, for some $i \in \{1, 2, \ldots, n\}$. Since $\{x\}$ is compact, by Lemma 22, there exists $h_x \in \mathcal{F}_{G_i}$ such that $h_x = 1$ on K.
Set $N_x = \{y : h_x(y) > 0\}$. Since h_x is continuous, N_x is open. Since $h_x(x) = 1$, $x \in N_x$. Clearly, $N_x \subseteq \operatorname{supp} h_x \subset G_i$ so $\overline{N_x} \subseteq \sup h_x \subset G_i$. Since $x \in N_x$, $\{N_x\}_{x \in X}$ is an open cover for X. Since X is compact, there exists $x_1, x_2, \ldots, x_m$ such that

$$X = \bigcup_{k=1}^{m} N_{x_k}. \tag{6.49}$$

Now for each $i \in \{1, 2, \ldots, n\}$, define $F_i = \bigcup\{\overline{N_{x_j}} : \overline{N_{x_j}} \subset G_i\}$. Clearly, $F_i \subset G_i$ and F_i is closed. Since F_i is closed and X is compact, F_i is compact. By Lemma 22, choose a function $f_i \in \mathcal{F}_{G_i}$ such that $f_i = 1$ on F_i. Define $g_i : X \to [0, 1]$ by

$$\begin{aligned} g_1 &= f_1 \\ g_2 &= (1 - f_1) f_2 \\ &\vdots \\ g_n &= \prod_{k=1}^{n-1} (1 - f_k) f_n \end{aligned}$$

Since $0 \leq f_k \leq 1$, by definition of g_i, $0 \leq g_i \leq 1$. Clearly, $\operatorname{supp} g_i \subseteq \operatorname{supp} f_i \subset G_i$. Hence $g_i \in \mathcal{F}_{g_i}$. It remains to prove that for $x \in K$, $\sum_{k=1}^{n} g_k(x) = 1$. We claim that

$$\sum_{k=1}^{l} g_k = 1 - \prod_{k=1}^{l} (1 - f_k), \text{ for every } 1 \leq l \leq n. \tag{6.50}$$

We will proceed by induction. When $l = 1$, $g_1 = f_1 = 1 - (1 - f_1)$ so the base case holds. Suppose that $\sum_{k=1}^{l-1} g_k = 1 - \prod_{k=1}^{l-1} (1 - f_k)$. Now

$$\begin{aligned} \sum_{k=1}^{l} g_k &= \sum_{k=1}^{l-1} g_k + g_l \\ &= 1 - \prod_{k=1}^{l-1} (1 - f_k) + \prod_{k=1}^{l-1} (1 - f_k) f_l && \text{by induction, definition of } g_l \\ &= 1 - \prod_{k=1}^{l-1} (1 - f_k)(1 - f_l) = 1 - \prod_{k=1}^{l} (1 - f_k). \end{aligned}$$

This proves the induction step and hence the claim holds. Now let $x \in K$. By 6.49, $x \in N_{x_j}$, for at least one $j \in \{1, 2, \ldots, m\}$. This implies $x \in F_i$, for some $i \in \{1, 2, \ldots, n\}$. By construction of F_i, $f_i(x) = 1$. By 6.50, $\sum_{k=1}^{n} g_k(x) = 1 - \prod_{k=1}^{n} (1 - f_k(x)) = 1$, as desired. ∎

6.6.2 Measures on compact metric spaces

Definition 73 (regular measure)**.** Let X be a topological space and let $\mathcal{F}$ be a σ-algebra. Suppose $\mu : \mathcal{F} \to [0, \infty]$ is a measure. We say μ is **regular** if for every $E \in \mathcal{F}$,

$$\mu(E) = \inf\{\mu(G) : \text{G open}, E \subseteq G\}, \text{ and}$$
$$\mu(E) = \sup\{\mu(K) : \text{K compact}, K \subseteq E\}.$$

Example 7. Lebesgue measure on $\mathbb{R}^n$ is regular.

Lemma 25 (Approximation Lemma, Proposition 17.6 in [Bas14])**.** *Let X be a compact metric space, $\mathcal{B}(X)$ a Borel σ-algebra and suppose that μ is a **finite** measure on the measurable space $(X, \mathcal{B}(X))$. If $E \in \mathcal{B}(X)$, for every $\epsilon > 0$, there exist sets K and U such that $K \subset E \subset U$ where K is compact, U is open and*

$$\mu(U \setminus E) < \epsilon \text{ and } \mu(E \setminus K) < \epsilon.$$

Proof Idea. We will apply a common method in measure theory where we prove the property on the whole σ-algebra $\mathcal{B}(X)$ by defining a set $\mathcal{H} \subseteq \mathcal{B}(X)$ where the property holds and then proving that $\mathcal{H}$ is actually $\mathcal{B}(X)$.

Proof. Say that $E \in \mathcal{B}(X)$ is **approximable** if for every $\epsilon > 0$, there exist sets K and U such that $K \subset E \subset G$ where K is compact, U is open and $\mu(U \setminus E) < \epsilon$ and $\mu(E \setminus K) < \epsilon$. Define $\mathcal{H}$ by

$$\mathcal{H} = \{E : E \in \mathcal{B}(X) \text{ such that } E \text{ is } \textbf{approximable}\}.$$

Suppose that K is compact. For $n \in \mathbb{N}$, define $U_n = \{x : x \in X, d(x, K) < \frac{1}{n}\}$ Since $d(\cdot, K)$ is continuous, U_n is open. Note that, $x \in K \iff d(x, K) = 0 \iff \forall n \in \mathbb{N}, d(x, K) < \frac{1}{n} \iff \forall n \in \mathbb{N}, x \in U_n$. Hence $K = \bigcap_{n=1}^{\infty} U_n$. Since μ is continuous and **finite**, $\mu(K) = \lim_{n\to\infty} \mu(U_n)$. But then, for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $n \geq N \implies \mu(U_n) - \mu(K) < \epsilon$. In particular $\mu(U_N) - \mu(K) < \epsilon$. Since μ is finite, $\mu(U_N \setminus K) = \mu(U_N) - \mu(K) < \epsilon$. Hence $K \in \mathcal{H}$. Since X is compact, $X \in \mathcal{H}$.
If $E \in \mathcal{H}$ and $\epsilon > 0$, choose $K \subset E \subset U$ with K compact, U open such that $\mu(E \setminus K) < \epsilon$ and $\mu(U \setminus E) < \epsilon$. Then $X \setminus U \subset X \setminus E \subset X \setminus K$. Since U is open, $X \setminus U$ is closed. Since $X \setminus U$ is closed and X is compact, $X \setminus U$ is also compact. Since K is compact and X Hausdorff, K is also closed. But then, $X \setminus K$ is open. Observe that

$$\mu((X \setminus E) \setminus (X \setminus U)) = \mu(U \setminus E) < \epsilon,$$
$$\mu((X \setminus K) \setminus (X \setminus E)) = \mu(E \setminus K) < \epsilon.$$

Hence $X \setminus E \in \mathcal{H}$.
Let $\{E_i\}_{i=1}^{\infty}$ be a family of pairwise disjoint sets in $\mathcal{H}$. We will show that $\bigcup_{i=1}^{\infty} E_i \in \mathcal{H}$. Let $\epsilon > 0$. For each $i \in \mathbb{N}$, choose K_i, K_i compact, U_i open such that $K_i \subset E_i \subset U_i$ and $\mu(U_i \setminus E_i) < \frac{\epsilon}{2^i}$, $\mu(E_i \setminus K_i) < \frac{\epsilon}{2^{i+1}}$.

Then $\bigcup_{i=1}^{\infty} U_i$ is open and $\bigcup_{i=1}^{\infty} E_i \subseteq \bigcup_{i=1}^{\infty} U_i$. Now

$$\mu\left(\bigcup_{i=1}^{\infty} U_i \setminus \bigcup_{j=1}^{\infty} E_j\right) \leq \mu\left(\bigcup_{i=1}^{\infty}(U_i \setminus E_i)\right) \leq \sum_{i=1}^{\infty} \mu(U_i \setminus E_i) < \epsilon.$$

We have $\bigcup_{i=1}^{\infty} K_i \subseteq \bigcup_{i=1}^{\infty} E_i$ and

$$\mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{j=1}^{\infty} K_j\right) \leq \sum_{i=1}^{\infty} \mu\left(E_i \setminus K_i\right) < \frac{\epsilon}{2}. \tag{6.51}$$

Since $\bigcup_{i=1}^{n} K_i \uparrow \bigcup_{i=1}^{\infty} K_i$, by continuity of μ, $\lim_{n\to\infty} \mu(\bigcup_{i=1}^{n} K_i) = \mu(\bigcup_{i=1}^{\infty} K_i)$. Hence we can choose $n \in N$ such that

$$\mu\left(\bigcup_{i=n+1}^{\infty} K_i\right) = \mu\left(\bigcup_{i=1}^{\infty} K_i\right) - \mu\left(\bigcup_{j=1}^{n} K_j\right) < \frac{\epsilon}{2}. \tag{6.52}$$

Since each K_j is compact, so is $\bigcup_{j=1}^{n} K_j$. Moreover, $\bigcup_{j=1}^{n} K_j \subseteq \bigcup_{i=1}^{\infty} E_i$ and

$$\begin{aligned}\mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{j=1}^{n} K_j\right) &= \mu\left(\bigcup_{i=1}^{\infty} E_i \setminus \bigcup_{j=1}^{\infty} K_j\right) + \mu\left(\bigcup_{i=1}^{\infty} K_i \setminus \bigcup_{j=1}^{n} K_j\right) \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \text{ by 6.51 and 6.52}\end{aligned}$$

Hence $\bigcup_{i=1}^{\infty} E_i \in \mathcal{H}$.
It remains to prove $\mathcal{H} = \mathcal{B}(X)$. Clearly, $\mathcal{H} \subseteq \mathcal{B}(X)$. Since X is compact, every closed subset of X is compact. Since $\mathcal{H}$ contains all compact sets, $\mathcal{H}$ contains all closed subsets of X. Since $\mathcal{H}$ is a σ-algebra containing all closed sets, $\mathcal{H}$ contains all open sets. Since $\mathcal{B}(X)$ is the smallest σ-algebra containing all open sets, $\mathcal{B}(X) \subseteq \mathcal{H}$. Hence $\mathcal{H} = \mathcal{B}(X)$, as desired. ■

Using Approximation Lemma, Proposition 17.6 in [Bas14], we can prove the main result of this section - a generalization of Proposition 1.5.6 in [Coh13].

Corollary 8 (Regularity of finite measures on compact metric spaces)**.** *Let X be a compact metric space, let $\mathcal{B}(X)$ be the Borel σ-algebra on X and suppose that μ is a **finite** measure on the measurable space $(X, \mathcal{B}(X))$. Then μ is **regular**.*

Proof. Let $\epsilon > 0$. By Approximation Lemma, Proposition 17.6 in [Bas14], there exist sets K and U such that $K \subseteq E \subseteq U$ where K is compact, U is open and

$$\mu(U \setminus E) < \epsilon \text{ and } \mu(E \setminus K) < \epsilon. \tag{6.53}$$

Since μ is finite, $\mu(U \setminus E) = \mu(U) - \mu(E)$ and $\mu(E \setminus K) = \mu(E) - \mu(K)$. By 6.53,

$$\mu(U) - \epsilon < \mu(E) < \mu(K) + \epsilon.$$

Hence $\mu(E) = \inf\{\mu(U) : \text{U open}, E \subseteq U\} = \sup\{\mu(K) : \text{K compact}, K \subseteq E\}$, as desired. ■

6.6.3 Riesz Representation Theorem for the dual of $\mathcal{C}(X)$

We are ready to state and prove the main result of this section - Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$.

Theorem 41 (Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$)**.** *Let X be a compact metric space and I be a positive linear functional on $\mathcal{C}(X)$. Then there exists the unique regular finite measure μ on $\mathcal{B}(X)$ such that*

$$I(f) = \int_X f \, d\mu, \text{ for every } f \in \mathcal{C}(X). \tag{6.54}$$

Proof Idea. Although the statement of Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$ resembles the statement of Riesz Representation Theorem for the Dual of $\mathcal{L}^p$, the proof will be significantly different. Unlike the proof for the case of the dual of $\mathcal{L}^p$, this proof will be constructive. In other words, we will construct the measure μ satisfying the claimed conditions. Thanks to the Radon-Nikodym Theorem for signed measures, such an important and relatively difficult construction could be abstracted. The argument is divided into nine steps. The first four steps are a construction of desired measure μ based on Carathéodory's method. The first part is a definition of the outer measure μ^*. In Step 2, we will show that μ^* is indeed an outer measure. In Step 3, we argue that every open set is μ^*-measurable. In Step 4, we will apply the Carathéodory's theorem to extend μ^* to a measure. In Step 5, we will discuss some regularity conditions. In Step 6, we will prove that the resulting measure is finite. In Step 7, we will establish 6.54. In Step 8, we will discuss the regularity of the resulting measure. In the last step, we will discuss uniqueness.

Proof.

Step 1 (Definition of μ^*). For an open set U, define

$$\nu(U) = \sup\{I(f) : f \in \mathcal{F}_U\}.$$

For an arbitrary set A, set

$$\mu^*(A) = \inf\{\nu(U) : A \subseteq U, U \text{ open}\}.$$

We claim that for an open set U, $\mu^*(U) = \nu(U)$. Since $U \subseteq U$ and U is open, $\mu^*(U) \leq \nu(U)$. Now let V be an open set containing U. If $f \in \mathcal{F}_U$, then $0 \leq f \leq 1$ and $\operatorname{supp} f \subset U \subseteq V$. Hence $f \in \mathcal{F}_V$. Hence $\nu(U) \leq \nu(V)$. Since V was arbitrary, $\nu(U) \leq \mu^*(U)$. Therefore, $\mu^*(U) = \nu(U)$, as desired.

Step 2 (μ^* is an outer measure). Since I is positive, $\nu \geq 0$ so $\mu^* \geq 0$.
We will show $\mu^*(\emptyset) = 0$. Consider $f \in \mathcal{F}_\emptyset$. Since $\operatorname{supp} f \subset \emptyset$, $\operatorname{supp} f = \emptyset$ so $f = 0$. By linearity of I, $I(0) = 0$. Since f was arbitrary, $\nu(\emptyset) = 0$. Since $\emptyset$ is open, $\mu^*(\emptyset) = \nu(\emptyset) = 0$.
We will show μ^* is monotonic. Suppose $A \subseteq B$. Suppose that V is an open set such that $B \subseteq V$. Then $A \subseteq B \subseteq V$. Hence $\{\nu(U) : B \subseteq U, U \text{ open}\} \subseteq \{\nu(U) : A \subseteq U, U \text{ open}\}$. Thus, $\mu^*(A) \leq \mu^*(B)$.

We will prove countable subadditivity on open sets. Let $\{U_n\}_{n=1}^{\infty}$ be a family of open sets. Set $U = \bigcup_{n=1}^{\infty} U_n$. We want to show

$$\mu^*(U) = \mu^*\left(\bigcup_{n=1}^{\infty} U_n\right) \leq \sum_{n=1}^{\infty} \mu^*(U_n). \tag{6.55}$$

Let $f \in \mathcal{F}_U$. Then $0 \leq f \leq 1$ and $\operatorname{supp} f \subset U$. Since $\operatorname{supp} f$ is closed and X is compact, $\operatorname{supp} f$ is compact. Observe that $\{U_n\}_{n=1}^{\infty}$ is an open cover for $\operatorname{supp} f$. By compactness of $\operatorname{supp} f$, without loss of generality, $\operatorname{supp} f \subset \bigcup_{k=1}^{m} U_k$. By Existence of partition of unity of a compact set, there exists a family of functions $\{\varphi_k\}_{k=1}^{m}$ such that $\varphi_k \in \mathcal{F}_{U_k}$ and for every $x \in \operatorname{supp} f$, $\sum_{k=1}^{m} \varphi_k(x) = 1$. Since for every $x \in \operatorname{supp} f$, $\sum_{k=1}^{m} \varphi_k(x) = 1$,

$$f = \sum_{k=1}^{m} f\varphi_k. \tag{6.56}$$

Consider φ_k for fixed $k \in \{1, 2, \ldots, m\}$. Since $0 \leq f \leq 1$ and $0 \leq \varphi_k \leq 1$, we have $0 \leq f\varphi_k \leq 1$. Clearly, $\operatorname{supp}(f\varphi_k) \subseteq \operatorname{supp} \varphi_k \subset U_k$. Hence $f\varphi_k \in \mathcal{F}_{U_k}$. This implies $\mu^*(U_k) = \nu(U_k) \geq I(f\varphi_k)$. Summing over all $k \in \{1, 2, \ldots, m\}$ and applying the linearity of I along with 6.56 gives

$$\sum_{k=1}^{m} \mu^*(U_k) \geq \sum_{k=1}^{m} I(f\varphi_k) = I\left(\sum_{k=1}^{m} f\varphi_k\right) = I(f).$$

Since $\mu^* \geq 0$, $I(f) \leq \sum_{k=1}^{m} \mu^*(U_k) \leq \sum_{k=1}^{\infty} \mu^*(U_k)$. Since f was arbitrary in $\mathcal{F}_U$ and U is open, $\mu^*(U) = \nu(U) \leq \sum_{k=1}^{\infty} \mu^*(U_k)$. This establishes 6.55.

Using 6.55, we can prove countable subadditivity for arbitrary sets. Let $\{A_n\}_{n=1}^{\infty}$ be a family of sets. We want to show

$$\mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mu^*(A_n). \tag{6.57}$$

If $\sum_{n=1}^{\infty} \mu^*(A_n) = \infty$, 6.57 holds. Assume $\sum_{n=1}^{\infty} \mu^*(A_n) < \infty$. Hence, for every $n \in \mathbb{N}$, $\mu^*(A_n) < \infty$. Let $\epsilon > 0$. For every $n \in \mathbb{N}$, choose an open set U_n containing A_n such that $\mu^*(U_n) < \mu^*(A_n) + \frac{\epsilon}{2^n}$. Since $A_n \subseteq U_n$ for every $n \in N$, $\bigcup_{n=1}^{\infty} A_n \subseteq \bigcup_{n=1}^{\infty} U_n$. By monotonicity of μ^*,

$$\begin{aligned}
\mu^*\left(\bigcup_{n=1}^{\infty} A_n\right) &\leq \mu^*\left(\bigcup_{n=1}^{\infty} U_n\right) \\
&\leq \sum_{n=1}^{\infty} \mu^*(U_n) && \text{by 6.55} \\
&\leq \sum_{n=1}^{\infty} \mu^*(A_n) + \sum_{n=1}^{\infty} \frac{\epsilon}{2^n} \leq \sum_{n=1}^{\infty} \mu^*(A_n) + \epsilon.
\end{aligned}$$

Since ϵ was arbitrary, 6.57 holds.

Step 3 (Open subsets of X are μ^*-measurable). We will show that every open set in X is μ^*-measurable. Let $U \subseteq X$ be open. We begin by showing that for every open V,

$$\mu^*(V) = \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U)). \tag{6.58}$$

By 6.57, it suffices to show $\mu^*(V) \geq \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U))$. If $\mu^*(V) = \infty$, the claim holds. Suppose $\mu^*(V) < \infty$. By monotonicity of μ^*, $\mu^*(V \cap U), \mu^*(V \cap (X \setminus U)) < \infty$. Let $\epsilon > 0$ and choose $f \in \mathcal{F}_{V \cap U}$ such that

$$\mu^*(V \cap U) - \frac{\epsilon}{2} < I(f). \tag{6.59}$$

Since $f \in \mathcal{F}_{V \cap U}$, $0 \leq f \leq 1$, $\operatorname{supp} f \subset V \cap U$. Since $\operatorname{supp} f$ is closed, $X \setminus \operatorname{supp} f$ is open. This implies $V \cap (X \setminus \operatorname{supp} f)$ is open. Since $\operatorname{supp} f \subset V \cap U$, $(X \setminus V) \cup (X \setminus U) \subset X \setminus \operatorname{supp} f$. Thus,

$$V \cap (X \setminus U) \subset V \cap (X \setminus \operatorname{supp} f). \tag{6.60}$$

Since $V \cap (X \setminus \operatorname{supp} f) \subseteq V$ and $\mu^*(V) < \infty$, by monotonicity of μ^*, $\mu^*(V \cap (X \setminus \operatorname{supp} f)) < \infty$. Hence, we may choose $g \in \mathcal{F}_{V \cap (X \setminus \operatorname{supp} f)}$ such that

$$\mu^*(V \cap (X \setminus \operatorname{supp} f)) - \frac{\epsilon}{2} < I(g). \tag{6.61}$$

Since $g \in \mathcal{F}_{(V \cap X \setminus \operatorname{supp} f)}$, $0 \leq g \leq 1$, $\operatorname{supp} g \subset V \cap (X \setminus \operatorname{supp} f)$. Since $\operatorname{supp} g \subset V$ and $\operatorname{supp} f \subset V$, $\operatorname{supp}(f+g) \subset V$. Since $\operatorname{supp} g \subset (X \setminus \operatorname{supp} f)$, $0 \leq f + g \leq 1$. Hence $f + g \in \mathcal{F}_V$. This implies

$$\begin{aligned}
\mu^*(V) &\geq I(f+g) = I(f) + I(g) && \text{by linearity of } I \\
&> \mu^*(V \cap U) - \frac{\epsilon}{2} + \mu^*((V \cap X \setminus \operatorname{supp} f)) - \frac{\epsilon}{2} && \text{by 6.59 and 6.61} \\
&\geq \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U)) - \epsilon. && \text{by 6.60}
\end{aligned}$$

Since ϵ was arbitrary, 6.58 holds.

Now suppose that V is not necessarily open. We will show that

$$\mu^*(V) = \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U)), \text{ for every open set } U. \tag{6.62}$$

By 6.57, it suffices to prove

$$\mu^*(V) \geq \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U)). \tag{6.63}$$

If $\mu^*(V) = \infty$, 6.63 holds. Suppose $\mu^*(V) < \infty$ and let $\epsilon > 0$. Then choose an open set E containing V such that $\mu^*(V) \leq \mu^*(E) < \mu^*(V) + \epsilon$. Then

$$\begin{aligned}
\mu^*(V) + \epsilon &> \mu^*(E) \\
&= \mu^*(E \cap U) + \mu^*(E \cap (X \setminus U)) && \text{by 6.58} \\
&\geq \mu^*(V \cap U) + \mu^*(V \cap (X \setminus U)). && \text{since } V \subseteq E
\end{aligned}$$

Since ϵ was arbitrary, 6.63 holds.

Step 4 ($\mu^*_{|\mathcal{B}(X)}$ is a measure on $\mathcal{B}(X)$). By Carathéodory's theorem, $\mu^*_{|\mathcal{M}(X)}$ is a measure on the σ-algebra of μ^*-measurable sets, denoted by $\mathcal{M}(X)$. By Step 3, $\mathcal{M}(X)$ contains all open sets. Since $\mathcal{B}(X)$ is generated by the collection of all open sets of X, $\mathcal{B}(X) \subseteq \mathcal{M}(X)$. Hence, $\mu^*_{|\mathcal{B}(X)}$ is a measure on $\mathcal{B}(X)$.

Step 5 (Regularity lemma).

Lemma 26. *Suppose that $f \in \mathcal{C}(X)$ and $K \subseteq X$ is compact. If $0 \leq \chi_K \leq f$ then $\mu(K) \leq I(f)$.*

Proof. For $\epsilon \in (0,1)$, define

$$U_\epsilon = \{x \in X : f(x) > 1 - \epsilon\} = f^{-1}(1-\epsilon, \infty).$$

Since f is continuous, U_ϵ is open. Since $0 \leq \chi_K \leq f$ and $\chi_K = 1$ on K, $K \subseteq U_\epsilon$. By monotonicity of μ, $\mu(K) \leq \mu(U_\epsilon)$. Hence, it is sufficient to prove $\mu(U_\epsilon) \leq I(f)$. Since U_ϵ is open, $\mu(U_\epsilon) = \nu(U_\epsilon) = \sup\{I(g) : g \in \mathcal{F}_{U_\epsilon}\}$. Observe that on U_ϵ, $\frac{1}{1-\epsilon} f > 1$. Consider $g \in \mathcal{F}_{U_\epsilon}$. Since $0 \leq g \leq 1$ and $\operatorname{supp} g \subset U_\epsilon$, it follows that $0 \leq g \leq 1 < \frac{1}{1-\epsilon} f$. Thus, $I(g) \leq I(\frac{1}{1-\epsilon} f) = \frac{1}{1-\epsilon} I(f)$. Since g was arbitrary, $\mu(U_\epsilon) \leq \frac{1}{1-\epsilon} I(f)$. Since ϵ was arbitrary, $\mu(U_\epsilon) \leq I(f)$, as desired. ∎

Step 6 (μ is finite). We will show μ is finite. Since μ is a measure, by subadditivity, it suffices to prove that $\mu(X) < \infty$. Since X is open, $\mu(X) = \nu(X) = I(1) < \infty$.

Step 7 (Verification of 6.54). Let $f \in \mathcal{C}(X)$ and suppose $f \geq 0$. Since f is continuous and X is compact, f is bounded. By linearity of integral and the functional I, we may assume that $0 \leq f \leq 1$.

To apply Lemma 26, we partition the range of f, in segments of equal length. Fix $n \in \mathbb{N}$. Define $K_k = \{x \in X : f(x) \geq \frac{k}{n}\}$, for $k \in \{0, 1, \dots n\}$. Since f is continuous, K_k is a closed set. Since X is compact, K_k is compact. Observe that $K_0 = X$ and for every k, $K_{k+1} \subseteq K_k$. Define

$$f_k(x) = \begin{cases} 0 & x \in (X \setminus K_{k-1}) \\ f(x) - \frac{k-1}{n} & x \in (K_{k-1} \setminus K_k) \\ \frac{1}{n} & x \in K_k \end{cases}.$$

By construction, $f = \sum_{k=1}^n f_k$ and $\chi_{K_k} \leq n f_k \leq \chi_{K_{k-1}}$. Integration gives

$$\frac{\mu(K_k)}{n} \leq \int_X f_k \, d\mu \leq \frac{\mu(K_{k-1})}{n}. \tag{6.64}$$

Summing 6.64 over all $k \in \{0, 1, \dots n\}$ and applying $f = \sum_{k=1}^n f_k$ gives

$$\frac{1}{n} \sum_{k=1}^{n} \mu(K_k) \leq \int_X f \, d\mu \leq \frac{1}{n} \sum_{k=0}^{n-1} \mu(K_k). \tag{6.65}$$

Let $\epsilon > 0$. Since μ is finite, let G be an open set containing K_{k-1} such that $\mu(G) < \mu(K_{k-1}) + \epsilon$. By definition of f_k, $\operatorname{supp} n f_k \subseteq K_{k-1} \subset G$. Since $\chi_{K_k} \leq n f_k \leq \chi_{K_{k-1}}$, $n f_k \in \mathcal{F}_G$. Then $I(n f_k) \leq \mu(G) < \mu(K_{k-1}) + \epsilon$.

By linearity of I, $I(f_k) < \frac{\mu(K_{k-1})}{n} + \frac{\epsilon}{n}$. Since ϵ was arbitrary, $I(f_k) \leq \frac{\mu(K_{k-1})}{n}$. By Lemma 26, $\mu(K_k) \leq I(nf_k)$ which implies $\frac{\mu(K_k)}{n} \leq I(f_k)$. We have

$$\frac{\mu(K_k)}{n} \leq I(f_k) \leq \frac{\mu(K_{k-1})}{n}. \tag{6.66}$$

Summing 6.66 over all $k \in \{0, 1, \dots n\}$ and applying $f = \sum_{k=1}^{n} f_k$ gives

$$\frac{1}{n}\sum_{k=1}^{n} \mu(K_k) \leq I(f) \leq \frac{1}{n}\sum_{k=0}^{n-1} \mu(K_k). \tag{6.67}$$

Combining 6.65 and 6.67 gives

$$\left| I(f) - \int_X f \, d\mu \right| \leq \frac{\mu(K_0) - \mu(K_n)}{n} \leq \frac{\mu(X)}{n}. \tag{6.68}$$

Letting $n \to \infty$ in 6.68 gives $I(f) = \int_X f \, d\mu$, as desired.
We will now generalize the result to $\mathcal{C}(X)$. Consider an arbitrary $f \in \mathcal{C}(X)$. Write $f = f^+ - f^-$, where $f^+, f^- \in \mathcal{C}(X)$ and $f^+, f^- \geq 0$. There exists a regular finite measure μ on $\mathcal{B}(X)$ such that $I(f^+) = \int_X f^+ \, d\mu$ and $I(f^-) = \int_X f^- \, d\mu$. By linearity of the functional I and the integral, we have

$$I(f) = I(f^+ - f^-) = I(f^+) - I(f^-) = \int_X f^+ \, d\mu - \int_X f^- \, d\mu = \int_X f \, d\mu,$$

as desired.

Step 8 (Regularity). Regularity of μ follows directly from Corollary 8.

Step 9 (Uniqueness). Let η be another regular finite measure on $\mathcal{B}(X)$ satisfying

$$I(f) = \int_X f \, d\mu = \int_X f \, d\eta, \text{ for every } f \in \mathcal{C}(X).$$

Regularity of μ and η implies that μ and η are completely determined by their values on compact sets in $\mathcal{B}(X)$. Thus, it suffices to prove that μ and η agree on compact sets in $\mathcal{B}(X)$. Let $K \in \mathcal{B}(X)$ be compact and let $\epsilon > 0$. Since μ is finite and regular, we may choose an open set U such that $K \subset U$ and $\mu(U) < \mu(K) + \epsilon$. By Lemma 22, there exists $h \in \mathcal{C}(X)$ with $\operatorname{supp} h \subset U$ such that $h = 1$ on K and $0 \leq h \leq 1$. Then

$$\begin{aligned} \eta(K) = \int_X \chi_K \, d\eta &\leq \int_X h \, d\eta \\ &\leq \int_X h \, d\mu \leq \int_X \chi_U \, d\mu = \mu(U) < \mu(K) + \epsilon. \end{aligned}$$

Hence $\eta(K) < \mu(K) + \epsilon$. Since ϵ was arbitrary, $\eta(K) \leq \mu(K)$. Reversing roles of μ and η and applying the previous argument gives $\mu(K) \leq \eta(K)$. Hence $\mu(K) = \eta(K)$, as desired.

■

Using the Riesz Representation Theorem for positive linear functionals, we can prove the similar result for bounded linear functionals.

Theorem 42 (Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$)**.** *Let X be a compact metric space and I be a bounded linear functional on $\mathcal{C}(X)$. Then there exists the unique finite signed regular measure μ on $\mathcal{B}(X)$ such that*

$$I(f) = \int_X f \, d\mu, \text{ for every } f \in \mathcal{C}(X). \tag{6.69}$$

Proof Idea. The proof consists of two parts. In the first part, we will prove the existence of such a signed measure by reduction to Theorem 41. We will discuss uniqueness in the second part.

Proof.

Step 1 (Existence). By Hahn-Jordan decomposition for bounded linear functionals on $\mathcal{C}(X)$, I can be expressed as $I = I^+ - I^-$, where I^+ and I^- are positive linear functionals on $\mathcal{C}(X)$. By Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$, there exist two finite regular Borel measures μ^+ and μ^- such that for every $f \in \mathcal{C}(X)$,

$$I^+(f) = \int_X f \, d\mu^+ \text{ and } I^-(f) = \int_X f \, d\mu^-. \tag{6.70}$$

Define $\mu : \mathcal{B}(X) \to \mathbb{R}$ by $\mu = \mu^+ - \mu^-$. Then μ is a signed measure on $\mathcal{B}(X)$. By linearity of the integral and 6.70, for every $f \in \mathcal{C}(X)$,

$$I(f) = I^+(f) - I^-(f) = \int_X f \, d\mu^+ - \int_X f \, d\mu^- = \int_X f \, d\mu.$$

Step 2 (Regularity). Regularity of μ follows from the regularity of μ^+ and μ^-.

Step 3 (Uniqueness). Let μ and ν be finite signed measures on $\mathcal{B}(X)$ such that for every $f \in \mathcal{C}(X)$,

$$I(f) = \int_X f \, d\mu = \int_X f \, d\nu. \tag{6.71}$$

Define $\eta : \mathcal{B}(X) \to \mathbb{R}$ by $\eta = \mu - \nu$. Clearly, η is a signed measure on $\mathcal{B}(X)$. By 6.71, for every $f \in \mathcal{C}(X)$, $\int_X f \, d\eta = \int_X f \, d\mu - \int_X f \, d\nu = 0$. By Hahn-Jordan decomposition, η admits decomposition $\eta = \eta^+ - \eta^-$, where η^+ and η^- are measures on $\mathcal{B}(X)$. Now for every $f \in \mathcal{C}(X)$,

$$\int_X f \, d\eta^+ - \int_X f \, d\eta^- = \int_X f \, d\eta = 0. \tag{6.72}$$

Define $F : \mathcal{C}(X) \to \mathbb{R}$ by $F(f) = \int_X f \, d\eta^+ = \int_X f \, d\eta^-$. By 6.72, F is well-defined. By linearity of the integral, F is a positive linear functional. By uniqueness part of Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$, $\eta^+ = \eta^-$. Then $\mu - \nu = \eta = \eta^+ - \eta^- = 0$ which implies $\mu = \nu$.

■

6.7 Fourier Analysis

In this section, we will introduce the Fourier Transform and state its main properties. Fourier Transform played an essential role in proving that certain activation functions were discriminatory. In previous sections, we have been working with real-valued functions. In this section, we will assume that $\mathcal{L}^1(\mathbb{R}^n)$ consists of complex-valued functions. In other words, we write $\mathcal{L}^p(\mathbb{R}^n), 1 \leq p < \infty$ for the normed linear space of complex-valued, Borel-measurable functions on $\mathbb{R}^n$ such that

$$\|f\|_p = \left(\int_{\mathbb{R}^n} |f|^p \, d\lambda \right)^{\frac{1}{p}} < \infty,$$

where two functions are identified if they are λ-almost everywhere equivalent. Here, λ is the n-dimensional Lebesgue measure. For the sake of simplicity, we will simply refer to representative functions of those equivalence classes instead of equivalence classes themselves. Also, it is worth noting that we presented integration theorems only in terms of real-valued functions. However, by definition of the integral of a complex-valued function, all these results generalize to complex-valued functions and we may use them in this section.

6.7.1 Fourier Transform on $\mathcal{L}^1(\mathbb{R}^n)$

We begin by introducing the Fourier Transform on $\mathcal{L}^1(\mathbb{R}^n)$.

Definition 74 (Fourier Transform on $\mathcal{L}^1(\mathbb{R}^n)$)**.** Let $f \in \mathcal{L}^1(\mathbb{R}^n)$. We define the Fourier Transform of f, denoted $\widehat{f} : \mathbb{R}^n \to \mathbb{C}$, by

$$\widehat{f}(\mathbf{u}) = \int_{\mathbb{R}^n} e^{i\langle \mathbf{u}, \mathbf{x} \rangle} f(\mathbf{x}) \, d\lambda(\mathbf{x}).$$

Remark 35. Some authors put a negative sign before the inner product $\langle \mathbf{u}, \mathbf{x} \rangle$ or introduce a scaling factor of $(2\pi)^{-1}$ or $(2\pi)^{-\frac{1}{2}}$ before the integral. I decided to follow the convention in [Bas14] for the sake of consistency.

Remark 36. Because in this section we work only with the Lebesgue measure, we will write $\int_{\mathbb{R}^n} e^{i\langle \mathbf{u}, \mathbf{x} \rangle} f(\mathbf{x}) \, d\mathbf{x}$ instead of $\int_{\mathbb{R}^n} e^{i\langle \mathbf{u}, \mathbf{x} \rangle} f(\mathbf{x}) \, d\lambda(\mathbf{x})$.

The Fourier Transform satisfies various convenient algebraic properties.

Proposition 23 (Properties of the Fourier Transform on $\mathcal{L}^1(\mathbb{R}^n)$)**.** *Suppose f and g are in $\mathcal{L}^1(\mathbb{R}^n)$. Then*

1. *$\widehat{f}$ is bounded and continuous;*
2. *$\widehat{(f+g)}(\mathbf{u}) = \widehat{f}(\mathbf{u}) + \widehat{g}(\mathbf{u})$;*
3. *for every $a \in \mathbb{C}$, $\widehat{(af)}(\mathbf{u}) = a\widehat{f}(\mathbf{u})$;*
4. *if $\mathbf{a} \in \mathbb{R}^n$ and $f_{\mathbf{a}}(\mathbf{x}) = f(\mathbf{x} + \mathbf{a})$ then $\widehat{f_{\mathbf{a}}}(\mathbf{u}) = e^{-i\langle \mathbf{u}, \mathbf{a} \rangle} \widehat{f}(\mathbf{u})$;*
5. *if $\mathbf{a} \in \mathbb{R}^n$ and $g_{\mathbf{a}}(\mathbf{x}) = e^{i\langle \mathbf{a}, \mathbf{x} \rangle} g(\mathbf{x})$ then $\widehat{g_{\mathbf{a}}}(\mathbf{u}) = \widehat{g}(\mathbf{u} + \mathbf{a})$;*
6. *if $a \in \mathbb{R}, a \neq 0$ and $h_a(x) = f(a\mathbf{x})$ then $\widehat{h_a}(\mathbf{u}) = a^{-n} \widehat{f}(\frac{\mathbf{u}}{a})$.*

Proof. Continuity follows from Dominated Convergence Theorem. All other results are elementary calculations justified by the linearity of the integral or the change of variables. See Proposition 16.1 in [Bas14] for the details. ■

6.7.2 Convolution on $\mathcal{L}^1(\mathbb{R}^n)$

An important operation related to integral transforms is a convolution.

Definition 75 (Convolution on $\mathcal{L}^1(\mathbb{R}^n)$)**.** Let $f, g \in \mathcal{L}^1(\mathbb{R}^n)$. The convolution of f with g is a function $(f * g) : \mathbb{R}^n \to \mathbb{C}$ given by

$$(f * g)(\mathbf{x}) = \int_{\mathbb{R}^n} f(\mathbf{x} - \mathbf{y}) g(\mathbf{y})\, d\mathbf{y}.$$

Lemma 27 (Convolution Norm Lemma)**.** *Let $f, g \in \mathcal{L}^1(\mathbb{R}^n)$. Then $(f * g)$ is integrable and*

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1 .$$

Proof. We proceed with a direct calculation,

$$\begin{aligned}
\|(f * g)\|_1 &= \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} f(\mathbf{x} - \mathbf{y}) g(\mathbf{y})\, d\mathbf{y} \right| d\mathbf{x} \\
&\leq \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} |f(\mathbf{x} - \mathbf{y}) g(\mathbf{y})|\, d\mathbf{y} \right) d\mathbf{x} \\
&\leq \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} |f(\mathbf{x} - \mathbf{y})| |g(\mathbf{y})|\, d\mathbf{x} \right) d\mathbf{y} \text{ by Fubini-Tonelli Theorem} \\
&\leq \int_{\mathbb{R}^n} |g(\mathbf{y})| \left(\int_{\mathbb{R}^n} |f(\mathbf{x} - \mathbf{y})|\, d\mathbf{x} \right) d\mathbf{y} \\
&\leq \int_{\mathbb{R}^n} |g(\mathbf{y})| \left(\int_{\mathbb{R}^n} |f(\mathbf{t})|\, d\mathbf{t} \right) d\mathbf{y} \\
&\leq \left(\int_{\mathbb{R}^n} |g(\mathbf{y})|\, d\mathbf{y} \right) \left(\int_{\mathbb{R}^n} |f(\mathbf{t})|\, d\mathbf{t} \right) = \|f\|_1 \cdot \|g\|_1 .
\end{aligned}$$

The application of Fubini-Tonelli Theorem was permissible because the integrand is non-negative. Since $f, g \in \mathcal{L}^1(\mathbb{R}^n)$, so is $(f * g)$ by the inequality above. ■

The following result describes a deep connection between the Fourier Transform and the convolution operation.

Theorem 43 (Fourier Transform of a convolution is a multiplication)**.** *If $f, g \in \mathcal{L}^1(\mathbb{R}^n)$, then $\widehat{(f * g)}(\mathbf{u}) = \widehat{f}(\mathbf{u}) \cdot \widehat{g}(\mathbf{u})$.*

Proof. Proof is by direct calculation and it is very similar to the argument above. See Proposition 16.4 in [Bas14]. ■

Although its proof is elementary, the result above is very important in applications of the Fourier transform. For instance, in signal processing, an efficient implementation of signal filters relies on the multiplicative property of convolution guaranteed by the result above.

6.7.3 Approximate identites

We begin by introducing the idea of an approximate identity with respect to the convolution. That is a family of functions that act as an identity with respect to the convolution, under limiting conditions. We will use the approximate identities to establish the Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$. The following lemma justifies the name.

Lemma 28 (Approximate Identity Lemma)**.** *Let $\varphi \in \mathcal{L}^1(\mathbb{R}^n)$ and suppose that $\int_{\mathbb{R}^n} \varphi(\mathbf{x})d\mathbf{x} = 1$. For $\delta > 0$, define $\varphi_\delta : \mathbb{R}^n \to \mathbb{C}$ by $\varphi_\delta(\mathbf{x}) = \frac{1}{\delta^n}\varphi(\frac{\mathbf{x}}{\delta})$.*

1. *If g is continuous with compact support, then $g*\varphi_\delta \to g$ pointwise as $\delta \to 0$.*
2. *If g is continuous with compact support, then $g * \varphi_\delta \to g$ in $\|\cdot\|_1$ as $\delta \to 0$.*
3. *If $f \in \mathcal{L}^1(\mathbb{R}^n)$ then $f * \varphi_\delta \to f$ in $\|\cdot\|_1$ as $\delta \to 0$.*

Proof. By change of variables $\mathbf{x} = \frac{1}{\delta}\mathbf{y}$,

$$\int_{\mathbb{R}^n} \varphi_\delta(\mathbf{y})\, d\mathbf{y} = \int_{\mathbb{R}^n} \frac{1}{\delta^n}\varphi(\mathbf{y})\, d\mathbf{y} = \int_{\mathbb{R}^n} \varphi(\mathbf{x})\, d\mathbf{x} = 1,$$
$$\|\varphi_\delta\|_1 = \int_{\mathbb{R}^n} |\varphi_\delta(\mathbf{y})|\, d\mathbf{y} = \int_{\mathbb{R}^n} |\varphi(\mathbf{t})|\, d\mathbf{t} = \|\varphi\|_1 .$$

Step 1 (Proof of 1). Let $\mathbf{x} \in \mathbb{R}^n$. Then

$$\begin{aligned} |(g * \varphi_\delta)(\mathbf{x}) - g(\mathbf{x})| &= \left| \int_{\mathbb{R}^n} (g(\mathbf{x}-\mathbf{y}) - g(\mathbf{x}))\varphi_\delta(\mathbf{y})\, d\mathbf{y} \right| && \text{since } \int_{\mathbb{R}^n} \varphi_\delta(\mathbf{y}) = 1 \\ &\le \left| \int_{\mathbb{R}^n} (g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x}))\varphi(\mathbf{t})\, d\mathbf{t} \right| && \text{change of variables } \mathbf{t} = \frac{\mathbf{y}}{\delta} \\ &\le \int_{\mathbb{R}^n} |g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x})|\, |\varphi(\mathbf{t})|\, d\mathbf{t}. \end{aligned}$$

Since g is continuous and compactly supported, $|g|$ is bounded above by $\|g\|_\infty$ which is finite. Thus $|g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x})|\,|\varphi| \le 2\|g\|_\infty |\varphi|$. Since $\varphi \in \mathcal{L}^1(\mathbb{R}^n)$, $2\|g\|_\infty |\varphi| \in \mathcal{L}^1(\mathbb{R}^n)$. By Dominated Convergence Theorem and continuity of g,

$$0 \le \lim_{\delta\to 0} |(g * \varphi_\delta)(\mathbf{x}) - g(\mathbf{x})| \le \int_{\mathbb{R}^n} \lim_{\delta\to 0} |g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x})|\, |\varphi(\mathbf{t})|\, d\mathbf{t} = 0.$$

Step 2 (Proof of 2). We proceed by direct calculation,

$$\begin{aligned} \int_{\mathbb{R}^n} |(g * \varphi_\delta)(\mathbf{x}) - g(\mathbf{x})|\, d\mathbf{x} &= \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} (g(\mathbf{x}-\mathbf{y}) - g(\mathbf{x}))\varphi_\delta(\mathbf{y})\, d\mathbf{y} \right| d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} (g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x}))\varphi(\mathbf{t})\, d\mathbf{t} \right| d\mathbf{x} \\ &\le \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} |g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x})||\varphi(\mathbf{t})|\, d\mathbf{x}\, d\mathbf{t} \\ &\le \int_{\mathbb{R}^n} |\varphi(\mathbf{t})| \left(\int_{\mathbb{R}^n} |g(\mathbf{x}-\delta\mathbf{t}) - g(\mathbf{x})|\, d\mathbf{x} \right) d\mathbf{t}. \end{aligned}$$

Consider $G_\delta(\mathbf{t}) = \int_{\mathbb{R}^n} |g(\mathbf{x} - \delta\mathbf{t}) - g(\mathbf{x})|\, d\mathbf{x}$. Now, $|g(\mathbf{x} - \delta\mathbf{t}) - g(\mathbf{x})| \leq |g(\mathbf{x} - \delta\mathbf{t})| + |g(\mathbf{x})|$. Now $\int_{\mathbb{R}^n}(|g(\mathbf{x} - \delta\mathbf{t})| + |g(\mathbf{x})|)d\mathbf{x} = 2\|g\|_1$. Since g is compactly supported and λ is regular, $\|g\|_1 < \infty$. By Dominated Convergence Theorem, $\lim_{\delta\to 0} G_\delta(\mathbf{t}) = \int_{\mathbb{R}^n} \lim_{\delta\to 0} |g(\mathbf{x} - \delta\mathbf{t}) - g(\mathbf{x})|\, d\mathbf{x} = 0$. Since for every $\mathbf{t} \in \mathbb{R}^n, |G_\delta(\mathbf{t})||\varphi(\mathbf{t})| \leq 2\|g\|_1 |\varphi(\mathbf{t})|$ and $\int_{\mathbb{R}^n} \varphi(\mathbf{x})d\mathbf{x} = 1$, by Dominated Convergence Theorem,

$$0 \leq \lim_{\delta\to 0} \|(g * \varphi_\delta) - g\|_1 \leq \int_{\mathbb{R}^n} |\varphi(\mathbf{t})| \lim_{\delta\to 0} G_\delta(\mathbf{t})\, d\mathbf{t} = 0.$$

Thus, $\lim_{\delta\to 0} \|(g * \varphi_\delta) - g\|_1 = 0$.

Step 3 (Proof of 3). Let $\epsilon > 0$. By, Density of compactly supported functions in $\mathcal{L}^p$, there exists compactly supported $g \in \mathcal{L}^1(\mathbb{R}^n)$ such that $\|f - g\|_1 < \epsilon$. Set $h = f - g$. Then $\|h\|_1 < \epsilon$. By Minkowski Inequality,

$$\|(f * \varphi_\delta) - f\|_1 = \|(h + g) * \varphi_\delta - (h + g)\|_1 \leq \|(g * \varphi_\delta) - g\|_1 + \|(h * \varphi_\delta) - h\|_1 . \tag{6.73}$$

By Convolution Norm Lemma, $\|h * \varphi_\delta\|_1 \leq \|h\|_1 \|\varphi_\delta\|_1$ and $\|\varphi_\delta\|_1 = \|\varphi\|_1$ so

$$\|(h * \varphi_\delta) - h\|_1 \leq \|h * \varphi_\delta\|_1 + \|h\|_1 \leq \|h\|_1 \|\varphi_\delta\|_1 + \|h\|_1 < \epsilon(1 + \|\varphi\|_1) \tag{6.74}$$

By Step 2, $\limsup_{\delta\to 0} \|(g * \varphi_\delta) - g\|_1 = 0$. Combining 6.73 and 6.74 gives

$$\limsup_{\delta\to 0} \|(f * \varphi_\delta) - f\|_1 \leq \epsilon(1 + \|\varphi\|_1).$$

Since ϵ was arbitrary, $\limsup_{\delta\to 0} \|(f * \varphi_\delta) - f\|_1 = 0$. Thus, $f * \varphi_\delta \to f$ in $\|\cdot\|_1$.

■

6.7.4 Gaussians and their Fourier Transforms

An example of a family of functions satisfying conditions of the Approximate Identity Lemma is the family of Gaussians, parameterized by the variance. In this section, we will explore their Fourier Transforms. We will use Gaussians to prove Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$. We start with the simplest Gaussians.

Proposition 24 (Fourier Transforms of Gaussians)**.** *Suppose* $f_1 : \mathbb{R} \to \mathbb{R}$ *and* $f_n : \mathbb{R}^n \to \mathbb{R}$ *are given by*

$$f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad f_n(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{\|\mathbf{x}\|^2}{2}}.$$

Then $\widehat{f}_1(u) = e^{-\frac{u^2}{2}}$ *and* $\widehat{f}_n(\mathbf{u}) = e^{-\frac{\|\mathbf{u}\|^2}{2}}$.

Proof Idea. There are several ways to prove this proposition. A common approach is to use contour integration with a Residue Theorem. However, it is possible to give a more elementary argument, such as the following based on page 107 in [JP04] .

Proof. We begin by proving the claim for f_1.

Step 1 (Proof for f_1). We begin with a direct calculation,

$$\begin{aligned}\widehat{f_1}(u) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{iux}e^{-\frac{x^2}{2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \cos(ux)e^{-\frac{x^2}{2}}\,dx + i\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \sin(ux)e^{-\frac{x^2}{2}}\,dx\right) \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \cos(ux)e^{-\frac{x^2}{2}}\,dx.\end{aligned}$$

The imaginary part vanishes because $x \to \sin(ux)e^{-\frac{x^2}{2}}$ is odd. Since $|\cos(ux)e^{-\frac{x^2}{2}}| \leq e^{-\frac{x^2}{2}}$ and $\int_{\mathbb{R}} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}$, by Dominated Convergence Theorem,

$$\frac{\partial \widehat{f_1}}{\partial u}(u) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \frac{\partial}{\partial u}\left(\cos(ux)\right)e^{-\frac{x^2}{2}}\,dx = -\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \sin(ux)xe^{-\frac{x^2}{2}}\,dx. \tag{6.75}$$

Integration by parts gives

$$\int_{\mathbb{R}} \sin(ux)xe^{-\frac{x^2}{2}}\,dx = -u\sin(ux)xe^{-\frac{x^2}{2}}\Big|_{x=-\infty}^{x=\infty} + u\int_{\mathbb{R}} \cos(ux)e^{-\frac{x^2}{2}}\,dx. \tag{6.76}$$

Since $|-u\sin(ux)xe^{-\frac{x^2}{2}}| \leq |u||x|e^{-\frac{x^2}{2}}$ and $\lim_{x\to\pm\infty}|u||x|e^{-\frac{x^2}{2}} = 0$, $-u\sin(ux)xe^{-\frac{x^2}{2}}\Big|_{x=-\infty}^{x=\infty}$ vanishes. Applying this observation to 6.76 and substituting the result in 6.75 gives the initial value problem

$$\left\{\frac{\partial \widehat{f_1}}{\partial u}(u) = -u\widehat{f_1}(u) \text{ subject to } \widehat{f_1}(0) = 1.\right.$$

Observe that the differential equation implies $\frac{\partial}{\partial u}\ln|\widehat{f_1}(u)| = -u$. By Fundamental Theorem of Calculus, $\ln|\widehat{f_1}(u)| = -\frac{1}{2}u^2 + C$ for $C \in \mathbb{R}$. Exponentiating both sides gives $\widehat{f_1}(u) = e^C \cdot e^{-\frac{1}{2}u^2}$. Setting the initial value condition results in $C = 0$. Hence $\widehat{f_1}(u) = e^{-\frac{1}{2}u^2}$.

Step 2 (Proof for f_n). We proceed with a direct calculation,

$$\begin{aligned}\widehat{f_n}(\mathbf{u}) &= \frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\mathbb{R}^n} e^{i\langle\mathbf{u},\mathbf{x}\rangle}e^{-\frac{\|\mathbf{x}\|^2}{2}}\,d\mathbf{x} = \frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\mathbb{R}^n} e^{i\left(\sum_{k=1}^n u_k x_k\right)}e^{-\frac{\sum_{k=1}^n x_k^2}{2}}\,d\mathbf{x} \\ &= \int_{\mathbb{R}^n}\prod_{k=1}^n \frac{1}{\sqrt{2\pi}}e^{iu_kx_k}e^{-\frac{1}{2}x_k^2}d\mathbf{x} \\ &= \int_{\mathbb{R}}\cdots\int_{\mathbb{R}}\prod_{k=1}^n \frac{1}{\sqrt{2\pi}}e^{iu_kx_k}e^{-\frac{1}{2}x_k^2}\,dx_1\ldots dx_n \text{ by Fubini-Tonelli Theorem} \\ &= \prod_{k=1}^n \widehat{f_1}(u_k) = \prod_{k=1}^n e^{-\frac{1}{2}u_k^2} = e^{-\frac{1}{2}\|\mathbf{u}\|^2}.\end{aligned}$$

■

Using Properties of the Fourier Transform on $\mathcal{L}^1(\mathbb{R}^n)$ and Fourier Transforms of Gaussians, we can calculate the Fourier Transform of Gaussians which will play essential role in the proof of Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$.

Corollary 9. *Let $a \neq 0$ and suppose that $h_a : \mathbb{R}^n \to \mathbb{R}$ is given by*

$$h_a(\mathbf{x}) = \frac{1}{(2\pi)^n} e^{-\frac{1}{2a^2}\|x\|^2}.$$

Then $\widehat{h_a}(\mathbf{u}) = (2\pi)^{-\frac{n}{2}} a^n e^{-\frac{a^2\|\mathbf{u}\|^2}{2}}$.

Proof. Write $h_a(\mathbf{x}) = (2\pi)^{-\frac{n}{2}}(2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2a^2}\|x\|^2}$. Then $h_a(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} f_n(\frac{1}{a}\mathbf{x})$. By Proposition 23, $\widehat{h_a}(\mathbf{u}) = (2\pi)^{-\frac{n}{2}} a^n \widehat{f_n}(a\mathbf{u})$. By Proposition 24, $\widehat{f_n}(a\mathbf{u}) = e^{-\frac{a^2\|\mathbf{u}\|^2}{2}}$ so $\widehat{h_a}(\mathbf{u}) = (2\pi)^{-\frac{n}{2}} a^n e^{-\frac{a^2\|\mathbf{u}\|^2}{2}}$. ■

6.7.5 Fourier inversion theorem on $\mathcal{L}^1(\mathbb{R}^n)$

We are ready to state and prove the Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$.

Theorem 44 (Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$)**.** *Suppose that f and $\widehat{f}$ are both in $\mathcal{L}^1(\mathbb{R}^n)$. Then*

$$f(\mathbf{y}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle} \widehat{f}(\mathbf{u})\, d\mathbf{u}, \text{ almost everywhere.}$$

Proof. For $a \neq 0$ define $h_a : \mathbb{R}^n \to \mathbb{R}$ by

$$h_a(\mathbf{x}) = \frac{1}{(2\pi)^n} e^{-\frac{1}{2a^2}\|x\|^2}.$$

By Corollary 9, $\widehat{h_a}(\mathbf{u}) = (2\pi)^{-\frac{n}{2}} a^n e^{-\frac{a^2\|\mathbf{u}\|^2}{2}}$. For every $\mathbf{y} \in \mathbb{R}^n$ and $a \neq 0$,

$$\begin{aligned}
\int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle} \widehat{f}(\mathbf{u}) h_a(\mathbf{u})\, d\mathbf{u} &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} e^{i\langle \mathbf{u},\mathbf{x}-\mathbf{y}\rangle} f(\mathbf{x}) h_a(\mathbf{u})\, d\mathbf{x} \right) d\mathbf{u} \\
&= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^n} e^{i\langle \mathbf{u},\mathbf{x}-\mathbf{y}\rangle} f(\mathbf{x}) h_a(\mathbf{u})\, d\mathbf{u} \right) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} f(\mathbf{x}) \left(\int_{\mathbb{R}^n} e^{i\langle \mathbf{u},\mathbf{x}-\mathbf{y}\rangle} h_a(\mathbf{u})\, d\mathbf{u} \right) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} f(\mathbf{x}) \left(\int_{\mathbb{R}^n} e^{i\langle \mathbf{x}-\mathbf{y},\mathbf{u}\rangle} h_a(\mathbf{u})\, d\mathbf{u} \right) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} f(\mathbf{x}) \widehat{h_a}(\mathbf{x}-\mathbf{y})\, d\mathbf{x} \\
&= \int_{\mathbb{R}^n} f(\mathbf{y}-\mathbf{t}) \widehat{h_a}(-\mathbf{t})\, d\mathbf{t} \\
&= \int_{\mathbb{R}^n} f(\mathbf{y}-\mathbf{t}) \widehat{h_a}(\mathbf{t})\, d\mathbf{t} = (f * h_a)(\mathbf{y}).
\end{aligned}$$

The interchange of integrals is justified by Fubini-Tonelli Theorem, since

$$\int_{\mathbb{R}^n}\int_{\mathbb{R}^n} |e^{i\langle \mathbf{u},\mathbf{x}-\mathbf{y}\rangle}||f(\mathbf{x})||h_a(\mathbf{u})|\,d\mathbf{x}\,d\mathbf{u} \leq \|f\|_1 \|h_a\|_1 < \infty.$$

We performed the change of variables with $\mathbf{t} = \mathbf{y} - \mathbf{x}$ and applied the symmetry of $\langle\cdot,\cdot\rangle$ and $\widehat{h_a}$. Observe that $|e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})h_a(\mathbf{u})| \leq |\widehat{f}(u)||h_a(\mathbf{u})| \leq \frac{1}{(2\pi)^n}|\widehat{f}(u)|$. Since $\widehat{f}$ in $\mathcal{L}^1(\mathbb{R}^n)$, $\frac{1}{(2\pi)^n}|\widehat{f}| \in \mathcal{L}^1(\mathbb{R}^n)$. Since for $\mathbf{u} \in \mathbb{R}^n$, $\lim_{a\to\infty} h_a(\mathbf{u}) = 1$, by Dominated Convergence Theorem,

$$\lim_{a\to\infty}\int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})h_a(\mathbf{u})\,d\mathbf{u} = \int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})\,d\mathbf{u}. \tag{6.77}$$

Applying Approximate Identity Lemma with $\delta = \frac{1}{\alpha}$ gives $(f * h_a) \to f$ in $\mathcal{L}^1(\mathbb{R}^n)$ as $a \to \infty$. Let $\{a_n\}_{n=1}^{\infty}$ be any sequence of strictly increasing real numbers. Then $(f*h_{a_n}) \to f$ in $\mathcal{L}^1(\mathbb{R}^n)$ as $n \to \infty$. By Proposition 3.1.5 in [Coh13], $(f*h_{a_n}) \to f$ in measure. By Proposition 3.1.3 in [Coh13], there exists a subsequence $\{(f * h_{a_{n_k}})\}_{k=1}^{\infty}$ such that $(f * h_{a_{n_k}}) \to f$ pointwise almost everywhere as $k \to \infty$. By 6.77, $(f * h_{a_{n_k}})(\mathbf{y}) = \int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})h_{a_{n_k}}(\mathbf{u})\,d\mathbf{u} \to \int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})\,d\mathbf{u}$ as $k \to \infty$. By uniqueness of the limit, $f(\mathbf{y}) = \int_{\mathbb{R}^n} e^{-i\langle \mathbf{u},\mathbf{y}\rangle}\widehat{f}(\mathbf{u})\,d\mathbf{u}$ almost everywhere, as claimed. ■

6.7.6 Fourier Transform of a measure

Definition 76. Let μ be a finite signed measure on $\mathbb{R}^n$. We define the Fourier Transform of μ by

$$\widehat{\mu}(\mathbf{u}) = \int_{\mathbb{R}^n} e^{i\langle \mathbf{u},\mathbf{x}\rangle}\,d\mu(\mathbf{x}).$$

Remark 37. In probability theory, $\widehat{\mu}$ is known as a characteristic function. The following result, which is a nontrivial consequence of Fourier Transforms of Gaussians and Stone-Weierstrass Theorem, justifies that name.

Theorem 45. *Let μ and ν be finite signed regular measures on $\mathcal{B}(\mathbb{R}^n)$. If for every $\mathbf{u} \in \mathbb{R}^n$, $\widehat{\mu}(\mathbf{u}) = \widehat{\nu}(\mathbf{u})$, then $\mu = \nu$.*

Proof. Omitted. See Exercise 16.6 in [Bas14]. The proof for probability measures is Theorem 14.1 in [JP04]. ■

Corollary 10. *Let μ be a finite signed regular measure on $\mathcal{B}(\mathbb{R}^n)$. If for every $\mathbf{u} \in \mathbb{R}^n$, $\widehat{\mu}(\mathbf{u}) = 0$, then $\mu = 0$.*

Proof. Define $\nu(A) = 0$ for every $A \in \mathcal{B}(\mathbb{R}^n)$. Clearly, $\widehat{\nu} = 0$. Since $\widehat{\mu} = \widehat{\nu} = 0$, by Theorem 45, $\mu = \nu = 0$. ■

6.8 Statement of originality

In this section, the relevant resources and bibliography are acknowledged. Acknowledgments are grouped by chapters. Each subsection in this section corresponds to a chapter in this thesis.

6.8.1 Introduction

Machine Learning

Definitions in this section are based on Section 2.1 in [SB14].

Deep Learning

Definitions in this section and descriptions of loss functions and algorithms are based on [HH18]. Neural network graph figures are based on figures displayed on websites [Neu] and [Stu20]. The proof of Backpropagation equations is based on the proof of Lemma 1 in [HH18] and the discussion from Section 6.2.3 in [Cal20].

6.8.2 Universality

Universal approximation of continuous functions via Cybenko's method

The proof of Lemma 2 is based on the proof of Lemma 1 in [Cyb89] and the proof of Lemma 2.2.3 in [Cal20]. The proof of Lemma 2.2.3 in [Cal20] contains the following possibly incorrect argument. [Cal20] establishes that F vanishes on indicators of intervals. Then the following conclusion is made.

> "...it follows from linearity that F vanishes on any indicator function. Applying the linearity again, we obtain that F vanishes on simple functions,
> $$F\left(\sum_{i=1}^{n} \alpha_i \chi_{J_i}\right) = \sum_{i=1}^{n} \alpha_i F(\chi_{J_i}) = 0,$$
> for any $\alpha_j \in \mathbb{R}$ and J_i intervals. Since simple functions are dense in $\mathcal{L}^\infty(\mathbb{R})$, it follows that $F = 0$." (*Deep Learning Architectures* [Cal20], Lemma 2.2.3 (p.34))

Firstly, it is worth noting the important difference between step functions and simple functions. It is not the case that step functions and simple functions are the same families. In our context, simple functions are linear combinations of indicators of Borel sets in $\mathbb{R}$, while step functions are linear combinations of intervals on $\mathbb{R}$. It is the case that intervals are contained in $\mathcal{B}(\mathbb{R})$, but not every Borel set is an interval. Hence proving that F vanishes on indicators of intervals does not directly prove that F vanishes on indicators of all Borel sets. However, the quoted argument presented in [Cal20] can be justified with a reference to the following result addressing density of step functions in $\mathcal{L}^p$.

Proposition 25 (Density of step functions in $\mathcal{L}^p$, Proposition 3.4.3 in [Coh13])**.** *Suppose that $[a,b]$ is a closed bounded interval and that p satisfies $1 \leq p < \infty$. Then the subspace of $\mathcal{L}^p([a,b],\mathcal{B}([a,b]),\lambda)$ determined by the step functions on $[a,b]$ is dense in $\mathcal{L}^p([a,b],\mathcal{B}([a,b]),\lambda)$.*

Remark 38. According to the following discussion in [Coh13], Proposition 3.4.3 holds for $\mathcal{L}^p(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda)$.

> "Let us call a function on $\mathbb{R}$ a step function if for each interval $[a,b]$ its restriction to $[a,b]$ is a step function. Analogue of Proposition 3.4.3 holds for $\mathcal{L}^p(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda)$ if we replace the set of step functions on $[a,b]$ with the set of step functions on $\mathbb{R}$ which vanish outside some bounded interval." (*Measure theory* [Coh13], p.102)

The proof of Proposition 3.4.3 relies on the regularity of Lebesgue measure and it is nontrivial. For instance, its extension to $\mathcal{L}^p(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda)$ was not presented in the textbook. However, it seems that this result has been taken for granted in [Cal20] without a reference. However, [Cyb89] provided a reference for the similar density result.

The original part of the proof in this thesis is the use of Dynkin's $\lambda-\pi$ theorem. Dynkin's $\lambda-\pi$ theorem was not used in the proof of Lemma 1 in [Cyb89] nor in the proof of Lemma 2.2.3 in [Cal20]. The proof of Lemma 1 in [Cyb89] proves that F vanishes on indicators of Borel sets in the following way. [Cyb89] establishes that F vanishes on indicators of intervals and appeals to the result similar to Proposition 3.4.3 in [Coh13] to deduce that F vanishes on indicators of Borel sets in $\mathbb{R}$. The proof in this thesis uses Dynkin's $\lambda-\pi$ theorem to directly prove that F vanishes on Borel sets, assuming it vanishes on indicators of intervals. The use of Dynkin's $\lambda-\pi$ theorem is a standard method in measure theory. For instance, it was discussed in *Essentials in Analysis and Probability*.
The proof of Proposition 3 is based on the proof of Proposition 2.2.4 in [Cal20]. However, the proof of Proposition 2.2.4 is addressing continuous sigmoidal function and the proof in this thesis is addressing bounded measurable sigmoidal functions. The proof of Proposition 2.2.4 invokes Dominated Convergence Theorem, implicitly assuming the theorem for signed measures. However, this theorem is often stated only for measures. This technical detail is discussed in detail in the proof of Proposition 3.
The proof of Separation functional lemma is a combination of proofs of Lemma 9.3.1 in [Cal20] and Lemma 9.3.2 in [Cal20]. The original part of the proof of Separation functional lemma is a discussion of a few omitted technical details in [Cal20]. For instance, it is proved that the representation of elements in subspace $\mathcal{T}$ is unique. This is not discussed in the proof of Lemma 9.3.1 in [Cal20]. This detail is important because it guarantees that the definition of the functional L is well-defined. The proof of well-definedness is missing in [Cal20]. Similarly, [Cal20] does not discuss a case when $\lambda = 0$. This detail is important because we argue that $\frac{1}{\lambda}\mathbf{u} \in \mathcal{T}$. When $\lambda = 0$, the vector $\frac{1}{\lambda}\mathbf{u}$ is not well-defined. The proof in [Cal20] also does not prove that $\mathcal{T}$ is a linear subspace.
The proof of The Universal Approximation Theorem for continuous functions is the proof of Proposition 9.3.5 in [Cal20].

Universal approximation of square-integrable functions

This section is based on Section 9.3.2 in [Cal20].
Definition of $\mathcal{L}^2$-discriminatory activation function is based on Definition 9.3.10 in [Cal20]. In this thesis, we use the implication $g = 0$ almost everywhere instead of g identically zero, as in [Cal20]. This difference can be attributed to a possible typo in [Cal20]. It could also be the case that [Cal20] treats functions in $\mathcal{L}^2$ as equivalence classes under λ almost everywhere equality. This abuse of notation is often used in this thesis.
The proof of Lemma 5 is based on the proof of Lemma 9.3.12 in [Cal20]. Similarly to the proof of Lemma 2, the argument in [Cal20] possibly contains a logical flaw related to simple and step functions. On page 264, [Cal20] establishes that F vanishes on indicator functions. Then the following conclusion is made.

> "...By linearity, F vanishes on combinations of indicator functions, such as χ_A, with A **interval**, and then vanishes on finite sums of these types of functions, i.e, on **simple** functions." (*Deep Learning Architectures* [Cal20], Lemma 9.3.12 (p.264))

However, as discussed in 6.8.2, simple functions and step functions are not the same family. This problem can be resolved in the same way as in the case of Lemma 2 - by appealing to Proposition 3.4.3 in [Coh13]. The original part of the proof in this thesis is the use of Dynkin's $\lambda - \pi$ theorem instead of Proposition 3.4.3 in [Coh13].
The proof of Lemma 6 is based on Example 9.3.14 in [Cal20].
The proof of Lemma 7 is based on Example 9.3.15 in [Cal20].
The proof of The Universal Approximation Theorem for square-integrable functions is based on the proof of Proposition 9.3.11 in [Cal20].

Universal approximation of integrable functions

This section is based on Universal approximation of square-integrable functions.

Universal approximation of measurable functions on compact sets

The proof of Cybenko's Universal Approximation Theorem for measurable functions, [Cyb89] is the proof of Theorem 3 in [Cyb89].

Universal approximation of measurable functions in probabilistic sense

This section is based on Section 9.3.4 in [Cal20].
Proposition 7 is Lemma 2.1 in [HSW89]. However, its proof is different. Proposition 8 is Proposition 9.3.21 in [Cal20]. Theorem 16 is Theorem 9.3.22 in [Cal20]. Lemma 8 is my own. Lemma 9 is my own.
Proposition 9 is Proposition 9.3.23 in [Cal20]. Its proof is based on the proof of Proposition 9.3.23 in [Cal20]. However, the explicit construction of g_i is replaced with an application of Proposition 9.
The proof of The Probabilistic Universal Approximation Theorem is proof of Theorem 9.3.24 in [Cal20].

6.8.3 Appendix

Stone-Weierstrass Theorem

Definitions and proofs in this section are based on Chapter 27 of [RL15] and Chapter 10 of [Wad14]. Proof of Bernstein Approximation Theorem is a proof of Theorem 27.4 in [RL15]. Weierstrass Approximation Theorem is Theorem 27.5 in [RL15] whose proof was omitted in the textbook and provided in this thesis. Closure under min and max is Lemma 10.67 in [Wad14]. The proof of Closure under min and max is an adaption of the proof of Lemma 10.67 in [Wad14]. Instead of approximating $|t|$ with the binomial series as in [Wad14], the approximation of $|t|$ was proved using Weierstrass Approximation Theorem. The proof of Stone-Weierstrass Theorem is based on the proof of Theorem 10.69 from [Wad14].

Measure Theory and Integration

Definitions and proofs in this section are based on various chapters from [Coh13] and [Bas14]. The proof of Negative set lemma is the proof of Lemma 4.1.4 in [Coh13]. The proof of Hahn Decomposition Theorem is an adaption of the proof of Theorem 4.1.5 in [Coh13]. The proof of Hahn-Jordan decomposition is the proof of Corollary 4.1.6 in [Coh13]. The proof of Radon-Nikodym Theorem for measures is based on the proof of Theorem 4.2.2 in [Coh13]. The reduction to σ-finite case is outlined in [Coh13] and completed in the proof of Radon-Nikodym Theorem for measures. The statement of Radon-Nikodym Theorem for signed measures is based on the statement of Theorem 4.2.4 in [Coh13].

Functional Analysis

Definitions are based on Chapter 4 from [RY08] and the proof of the Hahn-Banach Theorem is a combination of Theorem 5.6 in [Fol99] and the proof of Theorem 5.13 from [RY08]. However, significant parts of the proof are not discussed in detail in those textbooks. For instance, the fact $\prec$ is a partial order is not proved, the construction of the maximal element on Ω and verification that it is maximal is omitted. All such issues are clarified in the proof of Hahn-Banach Theorem presented in this thesis. The proof of Hahn-Jordan decomposition for bounded linear functionals on $\mathcal{C}(X)$ is an adaption of the proof of Lemma 8.13 in [Bar95]. The result in this thesis addresses bounded linear functionals on $\mathcal{C}(X)$ while Lemma 8.13 addresses $\mathcal{L}^p$. Another similar argument is the proof of Proposition 17.7 in [Bas14].

$\mathcal{L}^p$ spaces

Definitions are based on Chapter 15 from [Bas14] and Section 3.3 from [Coh13]. The proof of Lemma 21 is a combination of proofs of Theorem 15.9 from [Bas14], Corollary 15.10 in [Bas14] and Proposition 15.11 from [Bas14]. The proof of Riesz Representation Theorem for the Dual of $\mathcal{L}^p$ is based on the proof of Theorem 15.12 from [Bas14]. However, many subtle parts of the argument are different.

For instance, verification that ν is a signed measure is proved in a different way. The generalization from simple functions to $\mathcal{L}^p$ is performed using a different argument based on Lemma 21.

Linear functionals on $\mathcal{C}(X)$

Definitions in this section are based on [Bas14] and [Mun14]. The proof of Lemma 24 is an adaptation of the proof of Proposition 17.2 in [Bas14]. The proof of Lemma 25 is the proof of Proposition 17.6 in [Bas14]. However, many details have been added. The proof of Corollary 8 is a straightforward application of Lemma 25. The proof of Riesz Representation Theorem for positive linear functionals on $\mathcal{C}(X)$ is based on the proof of Theorem 17.3 in [Bas14]. However, the argument is considerably more detailed. For instance, Theorem 17.3 in [Bas14] does not address uniqueness and regularity of the resulting measure. Both uniqueness and regularity are proved in this thesis. The proof of Riesz Representation Theorem for bounded linear functionals on $\mathcal{C}(X)$ is based on the proof of Theorem 17.8 in [Bas14]. However, Theorem 17.8 in [Bas14] does not discuss uniqueness of the finite signed measure μ. In this thesis, the uniqueness is also proved.

Fourier Analysis

This section is based on Chapters 15 and 16 from [Bas14]. The proof of Convolution Norm Lemma is based on the proof of Proposition 15.7 in [Bas14]. The proof of Approximate Identity Lemma is the proof of Proposition 16.6 in [Bas14]. The proof of Fourier Transforms of Gaussians is based on the discussion in Example 5 (p.107) in [JP04]. The proof of Fourier inversion theorem for $\mathcal{L}^1(\mathbb{R}^n)$ is based on the proof of Theorem 16.7 in [Bas14]. The proof in [Bas14] ends with the fact that $f * \widehat{h}_a$ converges to f in $\|\cdot\|_1$. However, the claim is about the convergence pointwise almost everywhere. The proof in this thesis justifies that the convergence of $f * \widehat{h}_a$ to f in $\|\cdot\|_1$ implies the convergence of $f * \widehat{h}_a$ to f almost everywhere under the conditions relevant to the theorem.

6.9 Code for Experiments

6.9.1 Code for Model

```
from collections import namedtuple
import torch.nn as nn

image_width = 28
image_height = 28

LayerConfig = namedtuple("LayerConfig", ["width", "activation"])


class FashionNNMultiHiddens(nn.Module):

    def __init__(self, layers_config):
        super(FashionNNMultiHiddens, self).__init__()

        previous_width = image_width * image_height
        hidden_layers = []
        for config in layers_config:
            hidden_layers.append(
                nn.Linear(in_features=previous_width,
                          out_features=config.width))
            hidden_layers.append(config.activation)
            previous_width = config.width

        self.nn = nn.Sequential(*hidden_layers)

    def forward(self, x):
        return self.nn(x)
```

6.9.2 Code for Methodology

```
import torch
from torch.autograd import Variable
from torch.optim.adam import Adam
from torch.utils.data.dataloader import DataLoader
from torchvision.datasets import FashionMNIST
from torchvision import transforms
import torch.nn as nn
import matplotlib.pyplot as plt
import sklearn.metrics as metrics

from nn_multi_hidden_layers import FashionNNMultiHiddens
from utils import get_label_name

torch.manual_seed(42)
device = 'cpu'


def evaluate_accuracy(model: any, loader: any):
    total = 0
    correct = 0
    loss = 0
    loss_fn = nn.CrossEntropyLoss()

    with torch.no_grad():
        for images, labels in loader:
            images, labels = images.to(device), labels.to(device)
            x_images = Variable(images.view(images.size(0), -1))
            outputs = model(x_images)
            loss += loss_fn(outputs, labels)
            predictions = torch.max(outputs, 1)[1].to(device)
            correct += (predictions == labels).sum()
            total += len(labels)
    return (loss, (correct / total) * 100)
```

```
def display_loss_plot(iterations: list, train_losses: list,
                      validation_losses: list):
    plt.plot(iterations, train_losses, label='train loss')
    plt.plot(iterations, validation_losses, label='val loss')
    plt.xlabel('Iteration')
    plt.ylabel('Loss')
    plt.legend(loc='lower right')
    plt.title('Iterations vs Loss')
    plt.show()


def display_accuracy_plot(iterations: list, train_acc: list,
                          validation_acc: list):
    plt.plot(iterations, train_acc, label='train acc')
    plt.plot(iterations, validation_acc, label='val acc')
    plt.xlabel('Iteration')
    plt.ylabel('Accuracy')
    plt.legend(loc='lower right')
    plt.title('Iterations vs Accuracy')
    plt.show()


def print_confusion_matrix(model: any, loader: any):

    labels_list = []
    predictions_list = []

    with torch.no_grad():
        for images, labels in loader:
            images, labels = images.to(device), labels.to(device)
            x_images = Variable(images.view(images.size(0), -1))
            y_outs = model(x_images)
            y_pred = torch.max(y_outs, 1)[1].to(device)
            predictions_list.extend(y_pred)
            labels_list.extend(labels)

    metrics.confusion_matrix(labels_list, predictions_list)

    print('classification report for NN :\n%s\n' %
          (metrics.classification_report(
              labels_list,
              predictions_list,
              target_names=[get_label_name(c) for c in range(10)])))


def display_per_class_accuracy(model, loader):

    class_correct = [0. for _ in range(10)]
    total_correct = [0. for _ in range(10)]

    with torch.no_grad():
        for images, labels in loader:
            images, y_true = images.to(device), labels.to(device)
            x_images = Variable(images.view(images.size(0), -1))
            y_outs = model(x_images)
            y_pred = torch.max(y_outs, 1)[1].to(device)
            c = (y_pred == y_true).squeeze()

            for i in range(images.size(0)):
                label = labels[i]
                class_correct[label] += c[i].item()
                total_correct[label] += 1

    for c in range(10):
        print('val accuracy for {}: {:.2f}%'.format(
            get_label_name(c), class_correct[c] * 100 / total_correct[c]))
```

```
def train(model: any,
          train_loader: any,
          val_loader: any,
          epochs: int,
          model_name: str,
          lr=1e-3):
    print('--------------training--------------')
    print('model:')
    print(model)

    model.to(device)
    optimizer = Adam(model.parameters(), lr=lr)
    train_loss_fn = nn.CrossEntropyLoss()

    iterations_list = []
    train_losses = []
    val_losses = []
    train_accs = []
    val_accs = []
    best_val_acc = None
    best_val_loss = None

    iteration = 0

    for epoch in range(epochs):
        for images, labels in train_loader:
            # Transfering images and labels to GPU if available
            images, labels = images.to(device), labels.to(device)
            x_train = Variable(images.view(images.size(0), -1))
            labels = Variable(labels)

            # Forward pass
            outputs = model(x_train)
            # Compute train loss
            loss = train_loss_fn(outputs, labels)
            # Initializing a gradient as 0 so there is no mixing of gradient among the batches
            optimizer.zero_grad()
            # Compute gradients
            loss.backward()
            # Apply optimisation step
            optimizer.step()

            iteration += 1

            if not (iteration % 100):
                # Evaluate train and validation statistics
                (train_loss,
                 train_accuracy) = evaluate_accuracy(model, train_loader)
                (val_loss, val_accuracy) = evaluate_accuracy(model, val_loader)
                # Maintain the best model
                if (not best_val_acc) or (val_accuracy > best_val_acc):
                    best_val_acc = val_accuracy
                    best_val_loss = val_loss
                    torch.save(model.state_dict(), f'{model_name}.pth')
                # Register calculated statistics
                train_losses.append(train_loss)
                train_accs.append(train_accuracy)
                val_losses.append(val_loss)
                val_accs.append(val_accuracy)
                iterations_list.append(iteration)
                # Display evaluated statistics
                if not (iteration % 500):
                    print(
                        f'epoch:{epoch}, iteration:{iteration}, train_loss:{train_loss}, train_accuracy:{train_acc
                    )

    # restore the best model
    model.load_state_dict(torch.load(f'{model_name}.pth'))
    model.eval()
```

```
    return {
        'iterations_list': iterations_list,
        'train_losses': train_losses,
        'val_losses': val_losses,
        'train_accs': train_accs,
        'val_accs': val_accs,
        'best_val_acc': best_val_acc,
        'best_val_loss': best_val_loss
    }


def run_experiment(experiment_config: dict):
    epochs = experiment_config['epochs']
    batch_size = experiment_config['batch_size']
    layers_config = experiment_config['layers_config']
    model_name = experiment_config['name']

    torch.manual_seed(42)
    nn = FashionNNMultiHiddens(layers_config)

    train_set = FashionMNIST('./data',
                             download=True,
                             train=True,
                             transform=transforms.Compose(
                                 [transforms.ToTensor()]))
    val_set = FashionMNIST('./data',
                           download=True,
                           train=False,
                           transform=transforms.Compose(
                               [transforms.ToTensor()]))

    train_loader = DataLoader(train_set, batch_size=batch_size)
    val_loader = DataLoader(val_set, batch_size=batch_size)

    history = train(nn, train_loader, val_loader, epochs, model_name)
    best_val_acc = history['best_val_acc']
    best_val_loss = history['best_val_loss']

    with torch.no_grad():
        display_loss_plot(history['iterations_list'], history['train_losses'],
                          history['val_losses'])
        display_accuracy_plot(history['iterations_list'],
                              history['train_accs'], history['val_accs'])
        print('--------------best model--------------')
        print('validation accuracy: {:.2f}%'.format(best_val_acc))
        print(f'validation loss: {best_val_loss}')
        display_per_class_accuracy(nn, val_loader)
        print_confusion_matrix(nn, val_loader)
```

Bibliography

[Bar95] Robert G Bartle. "Decomposition of Measures". In: *The Elements of Integration and Lebesgue Measure*. John Wiley & Sons, Ltd, 1995. Chap. 8, pp. 80–95. ISBN: 9781118164471. DOI: https://doi.org/10.1002/9781118164471.ch8. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118164471.ch8. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118164471.ch8.

[Bas14] R.F. Bass. *Real Analysis for Graduate Students: Measure and Integration Theory (Version 4.2)*. Bass, R.F., 2014. ISBN: 9781502514455. URL: https://bass.math.uconn.edu/real.html.

[Bis98] Christopher M Bishop. *Neural networks and machine learning*. Springer, 1998.

[Bra+18] James Bradbury et al. *JAX: composable transformations of Python+NumPy programs*. Version 0.2.5. 2018. URL: http://github.com/google/jax.

[Bri] John Bridle. *Training Stochastic Model Recognition Algorithms 211 Training Stochastic Model Recognition Algorithms as Networks can lead to Maximum Mutual Information Estimation of Parameters*. URL: https://proceedings.neurips.cc/paper/1989/file/0336dcbab05b9d5ad24f4333c7658a0e-Paper.pdf.

[Bro+20] Tom Brown et al. *Language Models are Few-Shot Learners*. 2020. URL: https://arxiv.org/pdf/2005.14165.pdf.

[BWL20] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. Apr. 2020. URL: https://arxiv.org/pdf/2004.10934.pdf.

[Cal20] Ovidiu Calin. *Deep Learning Architectures*. Springer International Publishing, 2020. DOI: 10.1007/978-3-030-36721-3. (Visited on 03/09/2022).

[Coh13] Donald L Cohn. *Measure theory*. Birkhäuser, 2013.

[CUH15] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. arXiv.org, 2015. URL: https://arxiv.org/abs/1511.07289.

[Cyb89] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2 (Dec. 1989), pp. 303–314. DOI: 10.1007/bf02551274. URL: https://link.springer.com/article/10.1007%2FBF02551274.

[Den12] Li Deng. "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.

[Fol99] Gerald B Folland. *Real analysis : modern techniques and their applications*. John Wiley and Sons, 1999.

[GB] Xavier Glorot and Yoshua Bengio. *Understanding the difficulty of training deep feedforward neural networks*. URL: https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf.

[He+15a] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 2015. URL: https://arxiv.org/pdf/1512.03385.pdf.

[He+15b] Kaiming He et al. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. URL: https://arxiv.org/pdf/1502.01852.pdf.

[HH18] Catherine F. Higham and Desmond J. Higham. "Deep Learning: An Introduction for Applied Mathematicians". In: *arXiv:1801.05894 [cs, math, stat]* (Jan. 2018). URL: https://arxiv.org/abs/1801.05894 (visited on 03/09/2022).

[HHS18] Elad Hoffer, Itay Hubara, and Daniel Soudry. *Train longer, generalize better: closing the generalization gap in large batch training of neural networks*. Jan. 2018. URL: https://arxiv.org/pdf/1705.08741.pdf.

[Hor91] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4 (1991), pp. 251–257. DOI: 10.1016/0893-6080(91)90009-t. (Visited on 02/28/2020).

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[JP04] Jean Jacod and Philip E Protter. *Probability essentials*. Springer, 2004.

[KL20] Patrick Kidger and Terry Lyons. "Universal Approximation with Deep Narrow Networks". In: *Proceedings of Machine Learning Research* TBD (2020), pp. 1–22. URL: https://arxiv.org/pdf/1905.08539.pdf (visited on 02/22/2022).

[Kla+17] G Klambauer et al. *Self-Normalizing Neural Networks*. 2017. URL: https://arxiv.org/pdf/1706.02515.pdf (visited on 08/27/2020).

[Les+93] Moshe Leshno et al. "Multilayer feedforward networks with a non-polynomial activation function can approximate any function". In: *Neural Networks* 6 (Jan. 1993), pp. 861–867. DOI: 10.1016/s0893-6080(05)80131-5. (Visited on 02/28/2020).

[Li+18] Hao Li et al. "Visualizing the Loss Landscape of Neural Nets". In: *arXiv:1712.09913 [cs, stat]* (Nov. 2018). URL: https://arxiv.org/abs/1712.09913 (visited on 02/26/2022).

[Li+20] Mingzhen Li et al. *The Deep Learning Compiler: A Comprehensive Survey*. Aug. 2020. URL: https://arxiv.org/pdf/2002.03794.pdf (visited on 02/25/2022).

[Lu+] Zhou Lu et al. *The Expressive Power of Neural Networks: A View from the Width*. URL: https://proceedings.neurips.cc/paper/2017/file/32cbf687880eb1674a07bf717761dd3a-Paper.pdf (visited on 02/22/2022).

[Mar+15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[Mun14] James Raymond Munkres. *Topology*. Pearson, 2014.

[Mur22] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL: probml.ai.

[Neu] Izaak Neutelings. *Neural networks*. URL: https://tikz.net/neural_networks/ (visited on 03/10/2022).

[Par+20] Sejun Park et al. "Minimum Width for Universal Approximation". In: *arxiv.org* (June 2020). URL: https://arxiv.org/abs/2006.08859.

[Pas+19] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv.org, 2019. URL: https://arxiv.org/abs/1912.01703.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (Oct. 1986), pp. 533–536. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0?error=cookies_not_supported&code=2926f83e-9c3a-46a7-9b28-b3d19d46768a.

[RL15] Kenneth A Ross and Jorge M López. *Elementary analysis : the theory of calculus*. Springer, 2015.

[Rud17] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. June 2017. URL: https://arxiv.org/pdf/1609.04747.pdf.

[RY08] Bryan P Rynne and Martin A Youngson. *Linear functional analysis*. Springer, 2008.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. URL: https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6 (visited on 01/04/2022).

[Sen+20] Andrew W. Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577 (Jan. 2020), pp. 706–710. DOI: 10.1038/s41586-019-1923-7. URL: https://www.nature.com/articles/s41586-019-1923-7.

[Shi+17] Nitish Shirish Keskar et al. *ON LARGE-BATCH TRAINING FOR DEEP LEARNING: GENERALIZATION GAP AND SHARP MINIMA*. Feb. 2017. URL: https://arxiv.org/pdf/1609.04836.pdf.

[Sil+17] David Silver et al. *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*. arXiv.org, 2017. URL: https://arxiv.org/abs/1712.01815.

[Stu20] David Stutz. *Illustrating (Convolutional) Neural Networks in LaTeX with TikZ • David Stutz*. David Stutz, June 2020. URL: https://davidstutz.de/illustrating-convolutional-neural-networks-in-latex-with-tikz/ (visited on 03/10/2022).

[Sze+14] Christian Szegedy et al. *Going Deeper with Convolutions*. arXiv.org, 2014. URL: https://arxiv.org/abs/1409.4842.

[TGC18] Ludovic Trottier, Philippe Giguère, and Brahim Chaib-draa. "Parametric Exponential Linear Unit for Deep Convolutional Neural Networks". In: *arXiv:1605.09332 [cs]* (Jan. 2018). URL: https://arxiv.org/abs/1605.09332 (visited on 02/27/2022).

[Vas+17] Ashish Vaswani et al. *Attention Is All You Need*. arXiv.org, 2017. URL: https://arxiv.org/abs/1706.03762.

[Wad14] William R Wade. *Introduction to Analysis: Pearson New International Edition*. Pearson, 2014.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: cs.LG/1708.07747 [cs.LG].

[Zal20] Research Zalando. *Fashion MNIST Github*. GitHub, Nov. 2020. URL: https://github.com/zalandoresearch/fashion-mnist.